

THÈSE

AMIRAL, UNE PLATEFORME GÉNÉRIQUE POUR LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR — DE L'AUTHENTIFICATION À L'INDEXATION

Présentée et soutenue publiquement le 18 novembre 2004
pour obtenir le grade de Docteur en Sciences
de l'Université d'Avignon et des Pays de Vaucluse

SPÉCIALITÉ : INFORMATIQUE

par

Teva MERLIN

Composition du jury :

Mme Régine ANDRÉ-OBRECHT	PR, IRIT, Toulouse	Présidente du jury
MM. Frédéric BIMBOT Paul DELÉGLISE	CR-HDR, IRISA, Rennes PR, LIUM, Le Mans	Rapporteur Rapporteur
MM. Marc ACHEROY Laurent BESACIER	PR, ERM, Bruxelles MC, CLIPS/IMAG, Grenoble	Examinateur Examinateur
MM. Jean-François BONASTRE Patrick VERLINDE	MC-HDR, LIA, Avignon PR, ERM, Bruxelles	Directeur de thèse Co-Directeur de thèse

0

Remerciements

Je tiens tout d'abord à remercier tous les membres de mon jury pour leur présence et leur participation à la soutenance de cette thèse : madame Régine André-Obrecht, qui a accepté de présider ce jury, ainsi que messieurs Marc Acheroy et Laurent Besacier ; tous sont venus de loin juste pour le plaisir de participer à cette journée. Un grand merci également à messieurs Frédéric Bimbot et Paul Deléglise qui ont de plus accepté la charge d'être rapporteurs de ce travail.

Je voudrais également remercier Patrick Verlinde pour avoir été co-directeur de cette thèse mais aussi et surtout pour sa gentillesse à toute épreuve, qui fait que c'est toujours un plaisir de travailler avec lui.

Enfin, bien sûr, je dois exprimer mes plus grands remerciements à Jean-François Bonastre, pour avoir été un directeur de thèse formidable et m'avoir supporté (à tous les sens du terme) à fond durant toutes ces années. Pour cela, il a toute ma gratitude. Si c'était à refaire, je resignerais avec lui !

Au-delà de son chef, je remercie également le reste de l'équipe des RALEurs du LIA pour le bon esprit qui y régnait et pour tous les bons moments passés ensemble, au travail et en dehors. Honneur à la plus ancienne, équipe des RALEurs à elle toute seule au moment de mon arrivée et qui a eu le discutable honneur de m'accueillir dans son bureau jusqu'alors paisible : Corinne Fredouille. Merci Corinne. Merci aussi à Sylvain Meignier, venu très vite compléter le trio en s'installant sur un coin de table du même bureau. Merci enfin à Nicolas Scheffer qui l'a remplacé ensuite. Ce fut un plaisir de travailler avec tous.

Ce fut un plaisir aussi de travailler avec les autres membres du LIA et de l'IUP d'Avignon, grâce à cette bonne ambiance très particulière qui contribue à en faire presque une deuxième maison (voire une première pour certains) plus qu'un lieu de travail. Ainsi, merci aux "anciens" : le Spriet, Philou, Fred, Patrice, Driss, Pierrot pour ses leçons sur la révolution, Georges, El Sté, Christian, Pascal, Thierry Valet, et j'en oublie sûrement. Merci aux doctorants : Domi, Christophe, Olivier, Jens, Christian, Benoit, Sophie, Mi-reille, Loïc, Laurianne, Hichem, et avant ça Yannick, David, Loïc (l'autre), Serigne, et j'en oublie ici aussi.

Je souhaite adresser un remerciement particulier au directeur actuel du LIA, Renato De Mori, pour sa gentillesse mais aussi pour n'avoir cessé de m'assurer son soutien tout au long de ma thèse, en me poussant à continuer dans les moments difficiles. Merci Renato.

En dehors du LIA, cette thèse s'est également faite au sein du laboratoire Signal and Image Center de l'École Royale Militaire de Bruxelles, dirigé par Marc Acheroy. Je

voudrais remercier les membres du SIC pour l'accueil chaleureux qu'ils m'ont réservé, propre à faire oublier la couleur du ciel bruxellois. Merci en particulier à Patrick Verlinde, Xavier Neyt et Idrissa. Et merci encore à Marc qui est pour beaucoup, par sa façon de diriger l'équipe, dans la bonne ambiance qui règne là-bas.

Remerciement également aux amis que j'ai pu rencontrer au sein du consortium ÉLISA, au premier rangs desquels Raphaël Blouet, Ivan Magrin-Chagnolleau et Frédéric Bimbot.

Enfin, un remerciement spécial à Yannick et Sylvain pour m'avoir aidé à apporter la touche finale à cette thèse, même si le prix à payer a été de les rejoindre dans la grisaille sarthoise. Sans ce coup de pouce, je n'y serais pas arrivé.

J'adresse un autre grand merci à Yannick pour son amitié qui dure depuis bien avant que nous ayons eu cette idée étrange d'aller à Avignon pour y faire une thèse. Je ne lui en veux pas trop de ne pas m'avoir attendu pour sa soutenance ce qui nous aurait permis de terminer ensemble ce que nous avions commencé ensemble.

Au chapitre de l'amitié, je voudrais également mentionner de nouveau Jef, car bien plus qu'un directeur de thèse il est devenu pour moi un très bon ami.

Et bien sûr, pour finir, un grand merci à My Linh, supportrice de ce travail au point d'accepter de quitter avec moi le soleil du Sud dans le seul but de me voir terminer ce travail. Merci.

Sommaire

1 Introduction	13
I Contexte de travail	17
2 La reconnaissance automatique du locuteur	19
2.1 Généralités	19
2.2 Applications et tâches	20
2.2.1 Identification automatique du locuteur	21
2.2.2 Vérification automatique du locuteur	23
2.2.3 Détection de locuteur dans un flux multi-locuteurs	24
2.2.4 Suivi de locuteur	24
2.2.5 Segmentation en locuteurs	24
2.2.6 Apprentissage à partir de documents multi-locuteurs	25
2.3 Principales difficultés rencontrées	26
2.3.1 Variabilité intra-locuteur	26
2.3.2 Contenu linguistique	27
2.3.3 Difficultés liées à l'environnement	27
2.3.4 Difficultés liées à l'acquisition et la transmission du signal de parole	28
2.3.5 Niveau de coopération des locuteurs	29
2.3.6 Problèmes induits par le mode opératoire	29
2.4 Niveau de dépendance au texte	30
2.5 Clients, imposteurs et pseudo-imposteurs	31
2.6 Chaîne de traitement	31

2.6.1 Paramétrisation du signal de parole	32
2.6.2 Traitement post-paramétrisation	36
2.6.3 Modélisation du locuteur	38
2.6.4 Normalisation des scores	41
3 Le contexte de travail	49
3.1 Les campagnes d'évaluation NIST	50
3.1.1 Le consortium ELISA	50
3.1.2 La tâche initiale : vérification automatique du locuteur	51
3.1.3 Évolution : les tâches multi-locuteurs	52
3.1.4 Méthodes d'évaluation	53
3.2 Autres applications	57
3.2.1 Le projet MTM	57
3.2.2 Le projet Certivox	59
3.2.3 La convention LIARMA	59
3.2.4 Le projet RAVOL	59
II Travail réalisé	61
4 “Cœur” du système AMIRAL	63
4.1 Architecture multi-reconnasseurs segmentale	64
4.2 Paramétrisation	68
4.3 Traitements post-paramétrisation	69
4.3.1 Suppression des trames de basse énergie	69
4.3.2 Normalisation des vecteurs de paramètres acoustiques	71
4.4 Normalisation des scores	74
4.4.1 Rapport de vraisemblances	74
4.4.2 Normalisation WMAP	75
4.4.3 Autres techniques de normalisation des scores	80
4.5 Modélisation des locuteurs	80
4.5.1 Structure des modèles	81

4.5.2 Estimation des modèles de locuteurs	82
4.6 Modèle du monde	86
4.6.1 Triple intervention	86
4.6.2 Données d'apprentissage	86
4.6.3 Estimation	90
5 Évaluation dans le cadre de la tâche NIST “One-Speaker Detection”	91
5.1 NIST 98	91
5.2 NIST 99	92
5.3 NIST 2000	94
5.4 NIST 2001	94
5.5 NIST 2002	97
5.6 NIST 2003	98
5.7 Bilan	98
6 Les tâches multi-locuteurs lors des campagnes NIST	103
6.1 Suivi de locuteur et segmentation en locuteurs	103
6.1.1 Utilisation d'une méthode de détection de ruptures	104
6.1.2 HMM évolutif	109
6.2 Détection de locuteur dans des documents bi-locuteurs	113
6.2.1 Apprentissage sur données mono-locuteur	113
6.2.2 Apprentissage à partir de données multi-locuteurs (NIST 2002)	116
7 Adaptation du système à diverses applications	121
7.1 Le projet MTM	122
7.1.1 La plateforme MTM	122
7.1.2 Problématique de l'intégration des technologies vocales	124
7.1.3 Mise en œuvre	125
7.2 La collaboration LIA-RMA	133
7.2.1 Problématique	134
7.2.2 La modalité parole	134
7.2.3 La modalité visage	135

7.2.4 Normalisation par modèle du monde pour la modalité visage	136
7.2.5 Fusion	136
7.2.6 Démonstrateur	138
7.3 Bilan	138
III Conclusion et perspectives	141
8 Conclusion et perspectives	143
8.1 AMIRAL, outil pour la recherche	144
8.2 AMIRAL, outil de transfert de technologie	145
8.3 Perspectives	146
IV Annexes	149
A Résultats divers	151
B Structure logicielle d'AMIRAL	155
C Bibliographie personnelle	157

Liste des figures

2.1	<i>Principe de base de l'identification du locuteur</i>	22
2.2	<i>Principe de base de la vérification du locuteur</i>	23
2.3	<i>Principe de base du suivi de locuteur</i>	24
2.4	<i>Principe de base de la segmentation en locuteurs</i>	25
2.5	<i>Principe de base de l'apprentissage à partir de documents multi-locuteurs</i>	26
2.6	<i>Modules de base d'un système de reconnaissance du locuteur</i>	32
2.7	<i>Calcul d'une représentation cepstrale d'un signal de parole à partir d'une analyse en banc de filtres.</i>	34
2.8	<i>Illustration du principe de Znorm — Un jeu de paramètres $(\mu_{imp}^x, \sigma_{imp}^x)$ est calculé pour le locuteur client X après l'apprentissage de son modèle \mathcal{X} en le confrontant à un ensemble de signaux imposteurs ; ces paramètres sont ensuite utilisés pour appliquer la normalisation lors du test d'un signal Y (d'identité inconnue) par rapport à \mathcal{X}.</i>	44
2.9	<i>Illustration du principe de Tnorm — Un jeu de paramètres $(\mu_{imp}^y, \sigma_{imp}^y)$ est calculé, lors de la phase de test, pour chaque signal y à comparer au modèle \mathcal{X} ; ces paramètres sont obtenus en comparant le signal y à un ensemble de modèles d'imposteurs (Z_1, \dots, Z_n).</i>	45
3.1	<i>Exemple d'une courbe DET</i>	55
3.2	<i>Illustration sur un exemple fictif du calcul des deux taux d'erreurs (mauvaise détection parole/non parole et mauvaise affectation de segments aux locuteurs) pour la tâche de segmentation en locuteurs.</i>	58
4.1	<i>Illustration de l'architecture multi-reconnaisseurs segmentale du système AMIRAL.</i>	65
4.2	<i>Définition de la structure des trames — Illustration par un exemple de l'effet des trois paramètres (nombre de vecteurs couverts par une trame, indices des coefficients de ces vecteurs présents dans la trame et décalage entre deux trames successives).</i>	68

4.3 Illustration de la suppression des trames de basse énergie après paramétrisation, sur un fichier issu du corpus NIST 2001 — La distribution du log-énergie des trames est estimée par une bi-gaussienne ; seule la gaussienne de moyenne la plus élevée (moyenne μ_P , variance σ_P^2) est utilisée pour déterminer le seuil d'énergie ($\mu_P - 2\sigma_P$) en-dessous duquel les trames seront supprimées ; pour ce fichier, 26,2 % des trames sont supprimées.	70
4.4 Intérêt de la suppression des trames de basse énergie pour la vérification du locuteur — Les deux courbes ci-dessus présentent les résultats obtenus pour la tâche “One Speaker” de l’évaluation NIST 2001, avec et sans suppression de ces trames, toutes choses égales par ailleurs (application de la normalisation des trames restantes après la suppression (cf. 4.3.2), modèles à 128 composantes, normalisation des scores par rapport de vraisemblances).	72
4.5 Intérêt de la normalisation des vecteurs de paramètres pour la vérification du locuteur — Les trois courbes ci-dessus présentent les résultats obtenus pour la tâche “One Speaker” de l’évaluation NIST 2001, sans normalisation des vecteurs de paramètres, avec normalisation des coefficients cepstraux statiques uniquement et avec normalisation des coefficients statiques et dynamiques, toutes choses égales par ailleurs (suppression des trames de basse énergie avant la normalisation (cf. 4.3.1), modèles à 128 composantes, normalisation des scores par rapport de vraisemblances).	73
4.6 Normalisation WMAP – Distributions des rapports de vraisemblances pour les tests de types “clients” et “imposteurs” avant normalisation.	76
4.7 Normalisation WMAP – Fonction de normalisation des rapports de vraisemblances. Cette fonction a été obtenue à partir des distributions de rapports de vraisemblances présentées par la figure 4.6 et des probabilités a priori suivantes : $p(X = Y) = 0,1$, $p(X \neq Y) = 0,9$	77
4.8 Normalisation WMAP – Distributions des scores pour les tests de types “clients” et “imposteurs” après normalisation.	78
4.9 Normalisation WMAP – Récapitulation du principe.	79
4.10 Choix des paramètres α et β pour l’apprentissage des modèles de locuteurs par adaptation du modèle du monde — Illustration de l’influence de ces paramètres sur les résultats.	85
4.11 Illustration de la triple intervention du modèle du monde dans le processus de RAL – Ce modèle intervient deux fois lors de l’apprentissage d’un modèle de locuteur, puis une fois de plus lors de la normalisation des scores par rapport de vraisemblances.	87
4.12 Influence de la qualité du modèle du monde sur les performances du reconnaisseur – Illustration sur les résultats de la tâche “One Speaker” de l’évaluation NIST 2001, en faisant varier la quantité de données utilisée pour l’apprentissage du modèle du monde, de 5 minutes à 1 heure de parole.	89
5.1 Résultats du système primaire pour la tâche “One-Speaker Detection” de l’évaluation NIST 99.	93
5.2 Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 2000.	95

5.3 Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 2001 pour les corpus filaire et cellulaire.	96
5.4 Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 02 pour le corpus cellulaire.	97
5.5 Évolution des résultats pour la tâche “One-Speaker Detection” des campagnes d’évaluation NIST de 1999 à 2002.	100
6.1 Détection des changements de locuteur – Calcul de distance par fenêtres glissantes.	106
6.2 Segmentation en locuteurs – Illustration du processus d’agrégation des segments utilisé lors de la seconde phase de l’approche basée sur la détection de ruptures.	108
6.3 Suivi de locuteur – Résultats obtenus lors de la campagne d’évaluation NIST 99 par l’approche basée sur une détection de ruptures suivie d’une phase de vérification du locuteur.	109
6.4 Algorithme de segmentation par HMM évolutif.	111
6.5 Suivi de locuteur — Technique à base de HMM évolutif.	114
6.6 Détection de locuteur dans un document bi-locuteurs — La segmentation en 2 locuteurs est réalisée par la technique du HMM évolutif.	115
6.7 Apprentissage d’un modèle de locuteur à partir d’un ensemble d’enregistrements bi-locuteurs.	118
6.8 Détection de locuteur dans des documents bi-locuteurs — Apprentissage du locuteur cible à partir de documents multi-locuteurs — Résultats obtenus lors de la campagne d’évaluation NIST 2002.	119
7.1 Projet MTM — Prototype du terminal.	123
7.2 Choix de la taille des modèles de locuteur pour le projet MTM — Taux d’égal erreur (EER) obtenu pour diverses valeurs du nombre de composantes des GMM, confronté à l’évolution correspondante du temps de traitement.	127
7.3 Décimation de trames pour le projet MTM — Taux d’égal erreur (EER) obtenus avec des modèles à 64 composantes en fonction de la quantité de trames utilisées lors du test (de toutes les trames à 1 trame sur 12).	129
7.4 Collaboration LIARMA — Illustration de la motivation du choix des yeux et du nez comme caractéristiques biométriques exploitées par le module de reconnaissance du visage : les 4 portraits présentés ici ne diffèrent que par les yeux et le nez ; la forme du visage, les cheveux, la bouche, sont strictement identiques.	136
7.5 Collaboration LIARMA — Illustration du principe de base du module de reconnaissance de visage.	137

7.6 Collaboration LIARMA —Application de la technique de normalisation par rapport de vraisemblances à la reconnaissance de visage : visage moyen utilisé comme modèle du monde.	137
A.1 Influence de l'ordre d'application des traitements post-paramétrisation – Comparaison des résultats obtenus sur le corpus NIST 2001 en effectuant d'abord la suppression des trames de basse énergie, puis la normalisation des vecteurs acoustiques, avec les résultats obtenus en appliquant d'abord la normalisation.	152
A.2 Choix du nombre de composantes pour les modèles – Comparaison des résultats obtenus sur le corpus NIST 2001 pour des modèles matrices diagonales à 32, 64, 128, 256 et 512 gaussiennes appris par adaptation d'un modèle du monde dépendant du type de combiné.	153
A.3 Choix des coefficients α et β pour l'apprentissage des modèles de locuteurs par adaptation – Résultats obtenus sur le corpus NIST 2001 en faisant varier α/β de 0,1/0,9 à 0,9/ 0,1 par incrément de 0,1.	154
B.1 Structure en couche des divers blocs logiciels composant le système AMIRAL – Chaque bloc est conçu en tirant parti des fonctions offertes par le(s) bloc(s) de niveau directement inférieur.	155

Liste des acronymes

- CMS** *Cepstral Mean Subtraction*, soustraction de la moyenne cepstrale – cf. chapitre 2, page 36.
- DET** *Detection Error Tradeoff* – cf. chapitre 3, page 53.
- DTW** *Dynamic Time Warping*, alignement temporel dynamique – cf. chapitre 2, page 38.
- EER** *Equal Error Rate*, taux d'égale erreur – cf. chapitre 3, page 53.
- ELISA** Consortium regroupant plusieurs laboratoires francophones en vue de participer aux campagnes d'évaluation des systèmes de reconnaissance du locuteur organisées par l'institut américain NIST – cf. chapitre 3, page 50.
- EM** *Expectation-Maximization* – cf. chapitre 4, page 82.
- FA** Fausse acceptation : en détection de locuteur, test produisant à tort une réponse positive – cf. chapitre 3, page 53.
- FFT** *Fast Fourier Transform* – Transformée de Fourier.
- FR** Faux rejet : en détection de locuteur, test produisant à tort une réponse négative – cf. chapitre 3, page 53.
- GMM** *Gaussian Mixture Model*, modèle à mixture de gaussiennes – cf. chapitre 2, page 38.
- HMM** *Hidden Markov Model*, modèle de Markov caché – cf. chapitre 2, page 38.
- IAL** Identification Automatique du Locuteur – cf. chapitre 2, page 21.
- LFCC** *Linear Frequency Cepstral Coefficients*, coefficients cepstraux issus d'une analyse en banc de filtres à échelle linéaire – cf. chapitre 2, page 32.
- MAP** *Maximum A Posteriori*.
- MFCC** *Mel Frequency Cepstral Coefficients*, coefficients cepstraux issus d'une analyse en banc de filtres à échelle Mel – cf. chapitre 2, page 32.
- MTM** *Multimedia Terminal Mobile* – cf. chapitre 3, page 57.
- NIST** *National Institute of Standards and Technology*, agence fédérale américaine pour la promotion des technologies, dépendant du ministère du commerce – cf. chapitre 3, page 50.
- RAL** Reconnaissance Automatique du Locuteur.
- RMA** *Royal Military Academy*, École Royale Militaire de Bruxelles – cf. chapitre 3, page 59.
- SVM** *Support Vector Machine*, machine à support vectoriel – cf. chapitre 2, page 40.
- UBM** *Universal Background Model*, modèle du monde universel (par opposition aux modèles dépendants du genre et/ou du type de combiné) – cf. chapitre 4, page 90.

VAL Vérification Automatique du Locuteur – *cf.* chapitre 2, page 23.

WMAP *World+MAP*, technique de normalisation des scores combinant une normalisation par rapport de vraisemblances et un calcul de probabilités *a posteriori* – *cf.* chapitre 4, page 75.

Chapitre 1

Introduction

Ce travail de thèse s'inscrit dans le cadre de la reconnaissance automatique du locuteur (RAL). Le domaine de la RAL est abordé dans ce document sous deux angles correspondant à des problématiques au premier abord différentes mais qui présentent une large interconnexion. Le premier point concerne le développement, la mise en œuvre et l'évolution d'une plateforme de reconnaissance du locuteur destinée à servir de support à différents travaux de recherche dans le domaine de la RAL. Pour valider les résultats de ces travaux de recherche, il est nécessaire d'évaluer les performances de l'outil de base et de les comparer avec l'état de l'art, par exemple à travers la participation à des campagnes d'évaluation des systèmes de RAL. Le premier objectif de cette thèse est donc double : il s'agit d'une part de mettre au point la plateforme logicielle adaptée et d'autre part de démontrer le niveau de performance de celle-ci à travers la participation à des campagnes d'évaluation. La seconde vision du domaine de la RAL proposée dans cette thèse est liée aux applications pratiques dans ce domaine. De manière pragmatique, ce travail de thèse a été financé par l'intermédiaire de deux projets à visée applicative, le projet LIARMA né d'une collaboration entre le LIA et l'École Royale Militaire de Bruxelles concernant le développement d'un prototype de vérificateur d'identité biométrique et bimodal (voix et visage) et le projet européen IST/MTM concernant l'intégration d'un module de RAL à un assistant numérique personnel (PDA). La participation à ces deux projets permet d'aborder la problématique posée par le transfert de technologies du domaine de la recherche vers le monde applicatif.

Le domaine de la reconnaissance automatique du locuteur s'est longtemps limité à la détection ou la vérification de l'identité d'une personne à partir d'un échantillon de sa voix (à travers les tâches connues dans la littérature sous les noms d'Identification du Locuteur et de Vérification du Locuteur). Mais depuis quelques années le champ d'application des techniques de reconnaissance automatique du locuteur s'est considérablement élargi, suite au progrès à la fois des algorithmes utilisés et de la puissance de traitement disponible. Les nouvelles directions de recherches qui ont vu le jour ont notamment porté sur le traitement de documents audio plus longs et impliquant plusieurs locuteurs. Les tâches réalisées sur ce type de documents sont diverses, tant par leur nature que par leur complexité, allant de la détection de la présence d'un locuteur connu dans une conversation donnée, à l'analyse complète d'un enregistrement en termes de locuteurs — incluant la découverte automatique du nombre de participants ainsi que des instants (et de la durée) de leurs interventions dans la conversation.

Chacune de ces tâches implique une chaîne de traitement des documents audio bien

spécifique. Cependant, un noyau commun de méthodes et de techniques (concernant par exemple l'extraction de paramètres acoustiques ou la modélisation statistique) peut assez facilement être identifié. Dégager ce noyau pour en faire une base commune aux différents systèmes et applications permet de factoriser les efforts avec en corollaire un gain en termes de maintenance.

Le domaine de la RAL se montre très actif depuis une dizaine d'années, avec une progression constante des performances. Pour servir de base à des propositions innovantes, en termes de travaux de recherche, une plateforme logicielle de RAL doit montrer des performances proches de l'état de l'art. L'évaluation des performances (en termes de qualité de la reconnaissance) représente une étape cruciale (néanmoins parfois négligée) de tout travail de développement. La volonté de suivre les avancées du domaine amène un besoin de mise à jour régulière de la plateforme logicielle avec en corollaire une nécessité d'évaluer ou de valider régulièrement les performances de l'outil. L'évaluation des performances peut représenter une part importante de l'effort total de développement. En effet, une telle évaluation implique tout d'abord la définition (et le respect) d'un protocole strict, limitant les biais et offrant une facilité de comparaison et de reproduction des résultats. Mais pour être significative, l'évaluation des performances doit également reposer sur des tests variés et en nombre important, permettant d'établir des statistiques concluantes.

Les campagnes d'évaluation organisées annuellement par le *National Institute of Standards and Technology* (NIST) américain offrent un cadre intéressant pour évaluer les performances dans le cadre de la recherche. Elles reposent sur des bases d'enregistrements de très grande taille et sur un jeu réduit de protocoles simples et bien définis. Ces campagnes sont devenues le standard de fait à l'heure actuelle et un passage obligé pour la validation de tout travail de recherche en RAL. De plus, les connaissances et l'expérience acquises à travers une participation régulière à ces campagnes d'évaluation peuvent ensuite être transposées dans le domaine applicatif. Pour des applications pratiques, la nécessité d'évaluer les performances est au moins aussi importante que pour le monde de la recherche académique. Il est cependant généralement difficile, sinon impossible, d'établir une base de données de taille significative intégrant les conditions d'utilisation auxquelles sera confronté le système lors de son déploiement, et ce avant même que ce déploiement soit effectif. Une solution permettant de minimiser les efforts de développement, de mise à jour et autorisant une évaluation des performances dans le cadre de systèmes applicatifs consiste à construire un système de base, un noyau technologique, comme évoqué dans le début de cette partie, à évaluer ce système dans le cadre "de la recherche" et à dériver l'ensemble des systèmes destinés aux applications depuis cette base.

C'est ce principe qui a été mis en œuvre au cours du travail présenté ici : le développement d'un système de reconnaissance du locuteur, complet et modulaire, servant à la fois de plateforme de recherche et de base pour le développement de projets applicatifs dans le domaine de la RAL. Le qualificatif de "complet" correspond ici à la capacité de ce système à traiter l'ensemble des tâches définies dans le domaine de la reconnaissance du locuteur indépendante du texte : de l'identification et la vérification du locuteur, à la segmentation automatique de documents multi-locuteurs. Celui de "modulaire" fait référence à la structure adoptée pour permettre l'adaptation du système à chaque nouvelle tâche.

Organisation du document

La première partie est consacrée à la présentation du domaine de la reconnaissance automatique du locuteur ainsi que du contexte de travail. Le chapitre 2 offre une description de la RAL, suivie d'un bref exposé de l'état de l'art des techniques utilisées dans cette discipline. Le chapitre 3 quant à lui présente les besoins définis au LIA dans le domaine de la RAL. L'effort de recherche y est présenté notamment à travers la participation régulière aux campagnes NIST d'évaluation des systèmes de RAL; les diverses tâches composant ces évaluations, correspondant à autant de champs de recherche, y sont décrites. La présentation du projet MTM et de la collaboration avec l'École Royale Militaire, ainsi que de quelques autres applications, complète ce chapitre.

La plateforme développée pour répondre à ces besoins fait l'objet de la deuxième partie de ce document. Le chapitre 4 décrit l'ensemble des outils communs aux diverses tâches de RAL, formant le "cœur" du système, et les méthodes statistiques sur lesquelles ils reposent. Au chapitre 5 sont présentés les résultats obtenus par cette base (et leur évolution) lors des campagnes d'évaluation NIST, pour la tâche fondamentale de la RAL, la vérification du locuteur. Les développements effectués pour répondre aux autres tâches proposées lors de ces évaluations sont décrits dans le chapitre 6. Puis le chapitre 7 termine cette deuxième partie par la présentation des développements spécifiques aux deux projets à l'origine du financement de ce travail, le projet MTM et la collaboration avec l'ERM de Bruxelles.

Enfin, un ensemble de conclusions et de perspectives clôturent ce travail de thèse.

Première partie

Contexte de travail

Chapitre 2

La reconnaissance automatique du locuteur

Sommaire

2.1 Généralités	19
2.2 Applications et tâches	20
2.2.1 Identification automatique du locuteur	21
2.2.2 Vérification automatique du locuteur	23
2.2.3 Détection de locuteur dans un flux multi-locuteurs	24
2.2.4 Suivi de locuteur	24
2.2.5 Segmentation en locuteurs	24
2.2.6 Apprentissage à partir de documents multi-locuteurs	25
2.3 Principales difficultés rencontrées	26
2.3.1 Variabilité intra-locuteur	26
2.3.2 Contenu linguistique	27
2.3.3 Difficultés liées à l'environnement	27
2.3.4 Difficultés liées à l'acquisition et la transmission du signal de parole	28
2.3.5 Niveau de coopération des locuteurs	29
2.3.6 Problèmes induits par le mode opératoire	29
2.4 Niveau de dépendance au texte	30
2.5 Clients, imposteurs et pseudo-imposteurs	31
2.6 Chaîne de traitement	31
2.6.1 Paramétrisation du signal de parole	32
2.6.2 Traitement post-paramétrisation	36
2.6.3 Modélisation du locuteur	38
2.6.4 Normalisation des scores	41

2.1 Généralités

La Reconnaissance Automatique du Locuteur (RAL) consiste à reconnaître l'identité d'une personne par l'analyse de sa voix.

Objet d'un intérêt accru depuis quelque temps au même titre que l'ensemble des méthodes d'authentification dites biométriques, elle ne figure pas parmi les plus fiables de ces techniques, au premier rang desquelles on retrouve l'analyse des empreintes digitales et génétiques. Cependant la RAL présente un certain nombre de qualités qui la distingue de ces dernières notamment en termes de facilité de déploiement. Tout d'abord, le mode opératoire, un simple enregistrement audio, permet une acceptation plus aisée de la part des utilisateurs par rapport à d'autres techniques d'identification plus intrusives (notamment du fait que la reconnaissance du locuteur ne requiert aucun contact physique). De même le coût du matériel impliqué est des plus réduits. Enfin, la RAL offre l'unique avantage d'être utilisable à distance, sans nécessiter d'autre terminal qu'un simple téléphone.

Les caractéristiques de la reconnaissance du locuteur lui ouvrent d'autres champs applicatifs que la simple authentification d'utilisateur (voir 2.2).

Cependant, le principe même de la RAL induit un certain nombre de difficultés auxquelles il faut faire face lors de la mise en œuvre d'un système de reconnaissance du locuteur. En effet la capacité à identifier les locuteurs repose sur les différences entre les voix de divers locuteurs. Mais cette variabilité inter-locuteurs se retrouve en concurrence avec la variabilité intra-locuteur (changement de la voix d'un même locuteur entre deux enregistrements, volontaire (dans le cas d'une tentative d'imposture) ou non), la variabilité de l'environnement d'opération (bruit, niveau d'enregistrement) et du canal de transmission du signal de parole (par exemple lors d'une transmission par téléphone), etc.

2.2 Applications et tâches

Au même titre que pour les autres techniques d'identification biométriques, le type d'application qui apparaît de prime abord comme le plus évident pour la reconnaissance automatique du locuteur est l'authentification de l'utilisateur au sein d'un système de sécurité (dans le but de contrôler l'accès à un bâtiment, un réseau ou toute autre ressource sensible), concept déjà assimilé par le grand public notamment grâce à sa présence récurrente dans les récits de science-fiction. À cela s'ajoutent des applications policières telle l'automatisation d'écoutes téléphoniques.

Cependant, comme évoqué dans la section précédente (et détaillé en 2.3), la mise en œuvre d'un système de RAL se heurte à un certain nombre de difficultés dont la fiabilité des systèmes se ressent. Les connaissances actuelles dans le domaine de la reconnaissance automatique du locuteur ne permettent pas de considérer cette technique comme un moyen d'authentification biométrique fiable comme peuvent l'être l'analyse des empreintes digitales ou génétiques. La communauté scientifique, s'exprimant notamment par la voix de l'Association Francophone de la Communication Parlée¹, combat l'utilisation du terme abusif d'"empreinte vocale", produit d'une analogie trop poussée entre ces techniques d'identification biométriques et la RAL. En particulier, une opposition formelle à l'utilisation, hors de toute évaluation préalable, de la reconnaissance automatique du locuteur dans le domaine judiciaire est exprimée régulièrement ([Boë 1999], [Boë 2001], [Bonastre 2003a]).

Cependant, au delà de son utilisation au sein d'applications nécessitant une authen-

¹<http://www.afcp-parole.org>

tification sûre, la reconnaissance automatique du locuteur trouve des débouchés dans nombre d'autre domaines moins exigeants quant à la fiabilité de la reconnaissance et où une erreur d'identification n'aura pas de conséquences dramatiques. La vitesse de traitement et les besoins matériels des systèmes actuels permettent par exemple d'envisager l'intégration de la RAL comme une option de confort dans des produits grand public (pour une sélection automatique de profil d'utilisateur). D'autres applications plus ambitieuses commencent également à voir le jour, notamment dans le domaine de l'archivage et de l'indexation automatique de documents audiovisuels : dans un futur proche, il sera envisageable de proposer l'identité des locuteurs comme critère de recherche au sein d'une base de données de documents multimédia.

Toutes ces applications se distinguent par les contraintes de nature diverse qu'elles imposent au système de RAL. Dans certains cas, la performance en termes de taux d'erreurs est le critère à privilégier, dans d'autres cas la vitesse de traitement prime. Cette dernière peut se trouver contrainte par des ressources limitées en termes de puissance de calcul. La robustesse au bruit est souvent considérée comme essentielle lorsque les conditions d'utilisation du système ne peuvent être maîtrisées (par exemple pour la reconnaissance du locuteur intégrée à un assistant personnel numérique (PDA)) ; mais ces conditions peuvent au contraire être imposées (comme dans un cadre de reconnaissance du locuteur par téléphone). Enfin, bien sûr, le mode opératoire varie d'une application à l'autre, notamment en fonction de soucis d'ergonomie et d'éventuelle transparence du système pour l'utilisateur.

Cependant, toutes les applications de la RAL reposent sur des principes communs et peuvent être définies comme variantes de quelques tâches de base, dont les principales sont présentées ci-après.

2.2.1 Identification automatique du locuteur

L'identification automatique du locuteur (IAL) fut historiquement l'une des premières applications de la RAL ([Atal 1976], [Doddington 1985], [O'Shaughnessy 1986]) avec la vérification automatique du locuteur (voir ci-dessous). Le principe de l'IAL, illustré par la figure 2.1, consiste, en présence d'un signal de parole, à retrouver l'identité du locuteur associé parmi un ensemble de locuteurs connus.

Deux phases se distinguent dans le fonctionnement d'un système d'IAL. Dans un premier temps, il est nécessaire pour tous les locuteurs de se faire connaître auprès du système, qui apprendra leurs caractéristiques à partir d'un ou plusieurs enregistrements de leur voix. Dans un second temps, lors de la phase de test, un individu se présente devant le système et l'enregistrement de sa voix effectué alors est comparé à la voix de chacun des locuteurs connus à la recherche de la voix la plus proche.

Deux modes de décision sont alors envisageables en fonction de l'application visée. L'identification en ensemble fermé suppose que le locuteur à identifier est forcément un des locuteurs connus du système ; la réponse du système est ici l'identité du locuteur dont la voix est la plus proche de celle testée. L'identification en ensemble ouvert, en revanche, ne fait aucun *a priori* concernant l'appartenance du locuteur de test à l'ensemble des locuteurs connus ; elle impose une étape supplémentaire dans le processus de décision afin d'accepter ou de rejeter le locuteur testé (cette étape correspondant en fait à la tâche de vérification du locuteur décrite ci-dessous).

Il va de soi que la difficulté de l'identification automatique du locuteur s'accroît avec

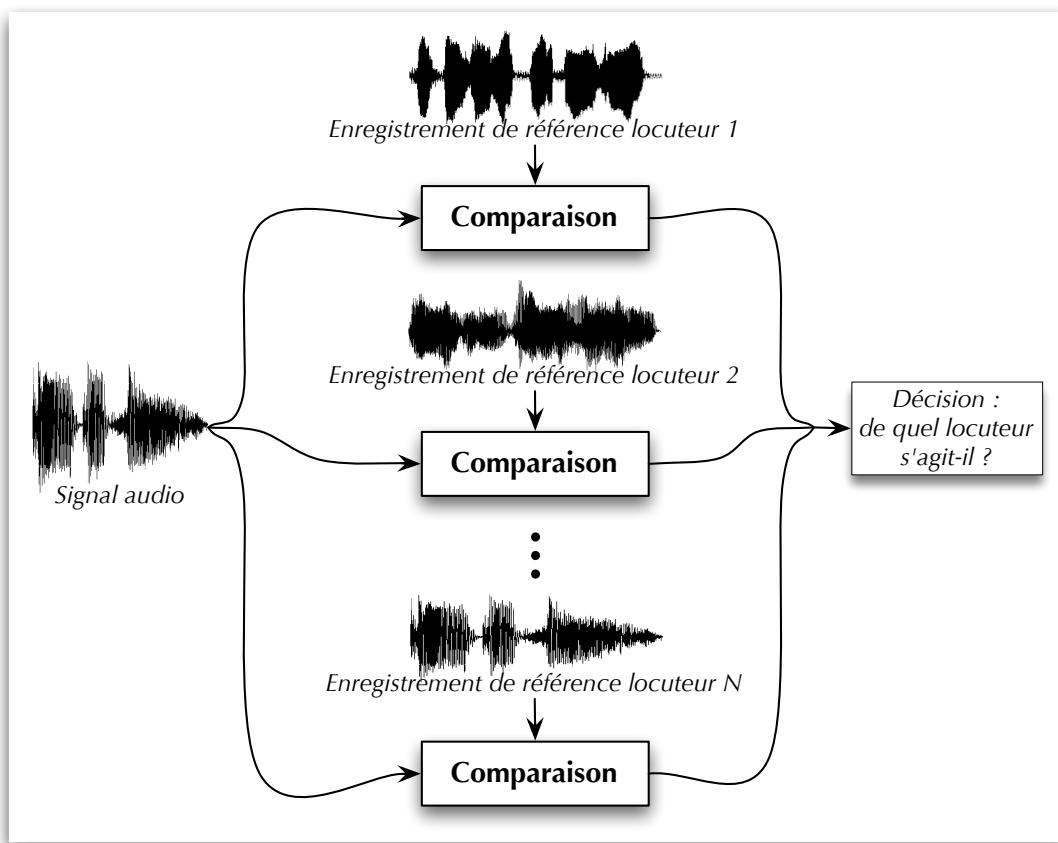


FIG. 2.1 – Principe de base de l’identification du locuteur

le nombre de locuteurs connus du système, tant en termes de vitesse d'exécution qu'en termes de taux d'identification correcte. Cependant concernant ce dernier critère, les performances obtenues par les systèmes actuels d'IAL en ensemble fermé sont excellentes, y compris sur des centaines de locuteurs. En revanche la tâche d'un système d'IAL en ensemble ouvert, de part l'étape de décision supplémentaire, est d'une difficulté plus élevée et les performances obtenues dépendent des progrès réalisés dans le cadre de la vérification automatique du locuteur.

D'un point de vue applicatif, si les systèmes de sécurité par authentification vocale se doivent évidemment de reposer sur le principe de l'identification en ensemble ouvert, d'autres applications se satisfont du cadre plus restrictif de l'identification en ensemble fermé : la sélection automatique de profil d'utilisateur évoquée précédemment entre dans ce cadre, ainsi que l'intégration de l'IAL au sein d'un système de reconnaissance de la parole pour assurer une adaptation automatique à l'utilisateur.

2.2.2 Vérification automatique du locuteur

Un locuteur est connu du système grâce à un ou plusieurs échantillon(s) de sa voix enregistré(s) au préalable. La vérification automatique du locuteur (VAL) consiste à comparer ensuite cet enregistrement de référence à un autre échantillon de parole afin de déterminer s'il s'agit bien du même locuteur dans les deux cas ([Rosenberg 1976], [Atal 1976], [Doddington 1985], [O'Shaughnessy 1986], [Furui 1981a], [Naik 1994], [Furui 1994]). Ce principe est résumé par la figure 2.2.

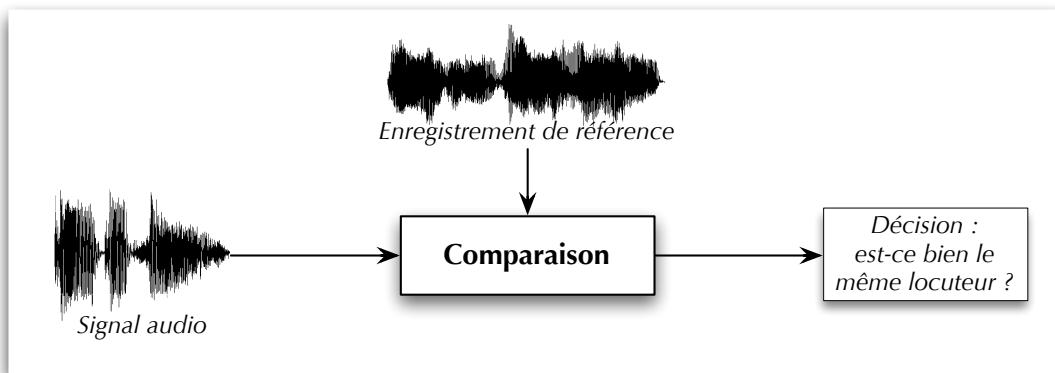


FIG. 2.2 – Principe de base de la vérification du locuteur

Ce principe simple trouve son utilité dans un grand nombre d'applications. De plus, les techniques utilisées dans le cadre de la vérification du locuteur forment la base de la plupart des autres tâches de la RAL (leur utilisation dans le cadre de l'identification du locuteur en ensemble ouvert a déjà été évoquée). Leur maîtrise est donc indispensable à la réalisation d'un système de RAL, quelle que soit la tâche visée.

Cependant, la diversité des conditions d'application du principe de la VAL fait varier considérablement la difficulté de la tâche, qui dépend principalement des variations de la voix pouvant intervenir entre l'enregistrement de référence et l'enregistrement de test. La section 2.3 (page 26) offre un récapitulatif des diverses causes de variation possibles.

2.2.3 Détection de locuteur dans un flux multi-locuteurs

Il s'agit d'une extension de la VAL à un test en environnement multi-locuteurs. Le principe est, toujours à partir de l'enregistrement de référence d'un locuteur, de déterminer si ce locuteur est présent au sein d'un enregistrement multi-locuteurs, par exemple une conversation ([Rosenberg 1998], [Przybocki 1999]).

2.2.4 Suivi de locuteur

Extension naturelle de la tâche précédente, le suivi de locuteur consiste à trouver les frontières des interventions du locuteur recherché au sein du document multi-locuteurs. Il s'agit donc de déterminer si ce locuteur intervient et si oui, quand ([Rosenberg 1998], [Sonmez 1999], [Bonastre 2000b], [Bonastre 2000a]).

Une illustration de ce principe est fournie par la figure 2.3.

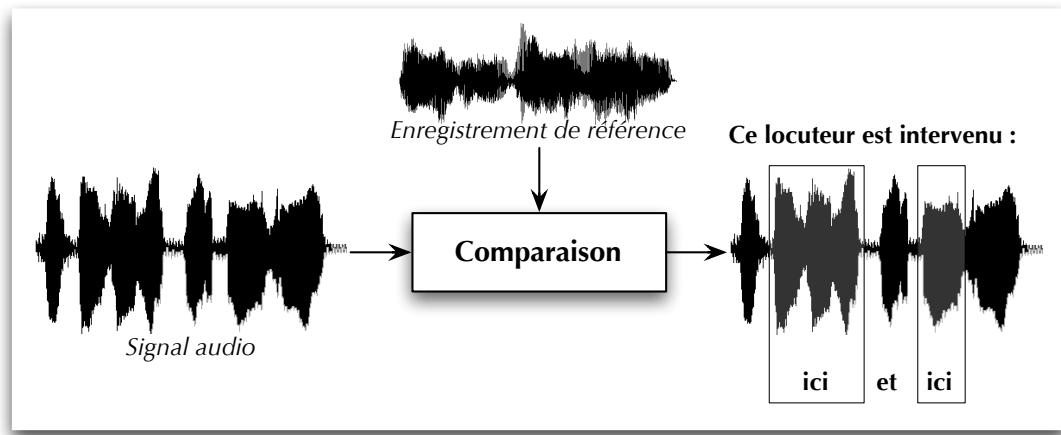


FIG. 2.3 – Principe de base du suivi de locuteur

2.2.5 Segmentation en locuteurs

La segmentation en locuteurs consiste à déterminer le nombre de locuteurs présents dans un document audio tout en délimitant leurs interventions (*cf.* figure 2.4).

La difficulté de cette tâche provient du traitement de documents pour lesquels peu ou pas d'informations sont connues *a priori*. Notamment, aucune information n'est disponible au préalable concernant les locuteurs intervenant dans le document : ni leur nombre, ni leur identité, ni aucun échantillon de leur voix permettant d'avoir une référence. Toutes ces informations doivent être extraites du document étudié.

Les travaux fondamentaux définissant la segmentation en locuteurs ont été réalisés par la société BBN sous la direction de H. Gish ([Siu 1991], [Siu 1992]) et concernaient

la segmentation automatique d'échanges radio entre pilotes et contrôleurs aériens. Depuis, le champ d'application de la segmentation en locuteurs s'est étendu et cette tâche se retrouve intégrée dans le cadre plus vaste de l'indexation en locuteurs de bases de données de documents multimédia ([Meignier 2002b], [Meignier 2002a]). Le spectre des types de documents traités s'en trouve élargi : conversations téléphoniques, enregistrement de journaux télévisés ou radiophoniques, films, enregistrements de réunions, etc. La variété de conditions rencontrées (parole plus ou moins spontanée, conditions d'enregistrement variables, nombre d'intervenants...) contribue à faire de la segmentation en locuteurs une tâche très complexe.

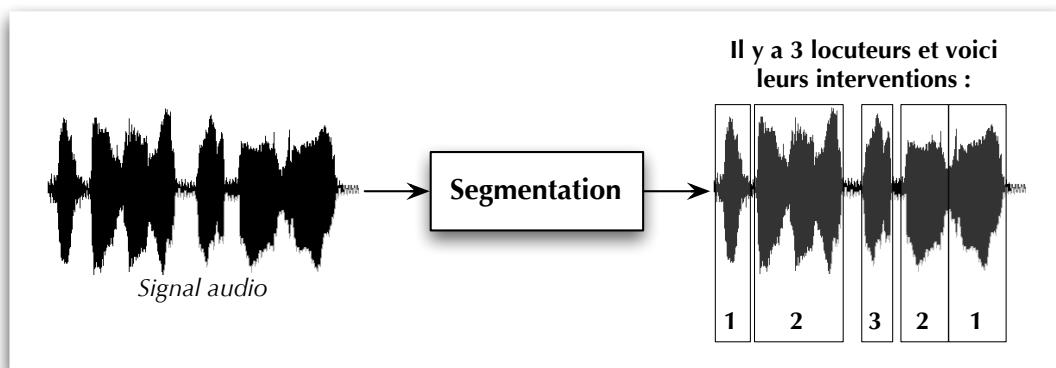


FIG. 2.4 – *Principe de base de la segmentation en locuteurs*

2.2.6 Apprentissage à partir de documents multi-locuteurs

Dans le cadre des applications nécessitant l'établissement d'une référence de la voix des locuteurs (applications basées sur l'IAL, la VAL, la détection dans un flux multi-locuteurs ou le suivi de locuteur), cette référence peut, selon les cas, être créée sans réelle coopération du locuteur visé, à partir d'un ou de plusieurs enregistrements de sa voix disponibles par ailleurs.

L'apprentissage à partir de documents multi-locuteurs permet d'introduire une plus grande souplesse lors de ce type d'apprentissage. Dans ce cas, en lieu et place d'un échantillon de parole mono-locuteur à l'identité bien déterminée, le système dispose, pour établir une référence, de plusieurs documents multi-locuteurs au sein desquels intervient systématiquement le locuteur visé. La tâche du système est alors de détecter et extraire de ces documents les interventions de ce locuteur afin de les utiliser ensuite comme référence dans le cadre d'une application de RAL (*cf. figure 2.5*).

Cette tâche, d'apparition très récente, a été rendue possible par le développement d'outils pour la segmentation en locuteurs ([Bonastre 2003b]).

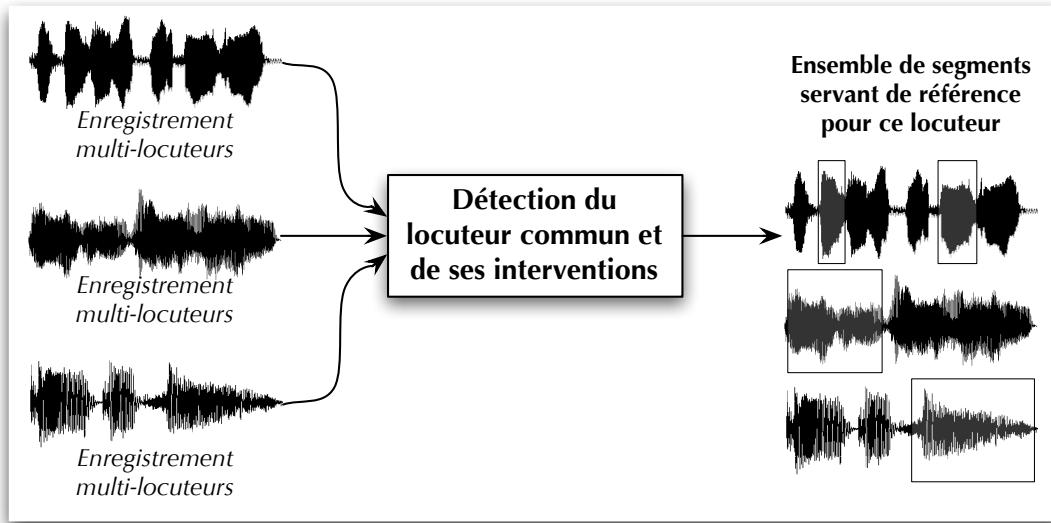


FIG. 2.5 – Principe de base de l'apprentissage à partir de documents multi-locuteurs

2.3 Principales difficultés rencontrées

La plupart des problèmes rencontrés par les systèmes de reconnaissance automatique du locuteur proviennent de la nature même des données manipulées : le signal de parole. Ce signal n'a en effet pas pour vocation première de véhiculer des informations relatives à l'identité du locuteur. Les informations linguistiques y sont bien sûr également présentes, mais aussi nombre d'autres informations, relatives par exemple à l'état de stress du locuteur, à son état de santé, etc. Enfin, le signal de parole se voit également modifié lors de sa transmission par les caractéristiques des médias utilisés. Ces différentes composantes interfèrent, rendant du même coup plus difficile l'extraction des seules informations caractéristiques du locuteur.

De plus, à cette difficulté intrinsèque de la RAL viennent s'ajouter des problèmes spécifiques à ses diverses applications et aux modes opératoires qui leur correspondent.

2.3.1 Variabilité intra-locuteur

Quelle que soit l'application envisagée, la reconnaissance du locuteur implique la comparaison de deux signaux de parole censés provenir du même locuteur. Selon l'intervalle de temps qui sépare l'enregistrement de ces deux signaux, des variations plus ou moins importantes peuvent intervenir sur la voix du locuteur considéré, compliquant la tâche de reconnaissance.

Divers facteurs sont à l'origine de cette évolution de la voix :

- Des variations peuvent être induites par l'état pathologique (fatigue, rhume, etc.) ou émotionnel (stress) du locuteur ([Homayounpour 1995], [Karlsson 1998], [Scherer 1998], [Banziger 2000]). Plus généralement, un être humain est incapable

de reproduire à l'identique le même signal de parole deux fois de suite, des variations plus ou moins légères étant toujours observées.

- Dans le cas d'une interaction volontaire et consciente avec un système de reconnaissance du locuteur (par exemple dans le cadre d'un accès sécurisé), le comportement des utilisateurs face au système évolue au cours du temps, notamment pour des raisons d'assurance accrue ou de lassitude. Leur voix (au moment de l'enregistrement) s'en trouve modifiée.
- Enfin, à plus long terme, le vieillissement modifie la voix d'un individu.

L'influence de ces divers facteurs varie selon l'application visée. Des travaux ont montré l'importance des variations à long terme dans un cadre d'IAL et de VAL ([Rosenberg 1976], [Furui 1977], [Setlur 1994]), où les performances se dégradent lorsqu'augmente l'intervalle de temps séparant l'enregistrement de référence et l'enregistrement de test. Cependant, dans le cas de la segmentation en locuteurs, où une comparaison s'effectue entre deux extraits d'un même document, les variations à long terme n'ont pas de réelle influence. En revanche, les facteurs responsables de variations à court terme de la voix sont présents quelle que soit la tâche considérée.

2.3.2 Contenu linguistique

Parmi les diverses informations caractéristiques du locuteur présentes dans le signal de parole, certaines sont directement liées au contenu linguistique, telles le choix des mots ou des structures de phrases, etc. Divers travaux récents révèlent le potentiel offert par l'utilisation de ces informations ([Andrews 2001], [Doddington 2001], [Peskin 2003], [Reynolds 2003b]). Mais leur nature même, qui impose de disposer d'enregistrements variés et de longue durée pour apprendre les caractéristiques d'un locuteur ou pour le reconnaître, rend ces informations difficilement exploitables dans le cadre de la plupart des applications de la RAL. De plus, l'indépendance de ces informations par rapport au contexte conversationnel n'a pas été démontrée.

Cependant, le contenu linguistique, par la modulation qu'il applique au signal de parole, exerce également une influence sur d'autres caractéristiques du locuteur qui ne lui sont pas directement liées. Notamment, certains phonèmes sont reconnus pour discriminer entre les locuteurs ([Eatoeck 1994], [Olsen 1997]). Dès lors, l'absence de ces phonèmes dans les enregistrements de référence et/ou de test peut être préjudiciable à qualité de la reconnaissance. De même, la présence de certains phénomènes de co-articulation porteurs d'informations caractéristiques du locuteur dépend du contenu linguistique du signal.

De manière plus générale, une importante différence de contenu linguistique entre les enregistrements de référence et de test rend plus difficile la reconnaissance. Ce cas se présente en particulier dans le cadre d'applications multilingues.

2.3.3 Difficultés liées à l'environnement

Les variations de l'environnement sonore entre l'enregistrement de référence et l'enregistrement du signal à tester constituent une source de difficulté importante pour la reconnaissance du locuteur.

Ce problème se pose pour de nombreuses applications pour lesquelles le contrôle de

l'environnement sonore (notamment la nature et le niveau du bruit) lors de l'enregistrement est impossible : par exemple lors de la segmentation en locuteurs de documents sonores d'origines diverses, ou dans le cadre de la reconnaissance du locuteur intégrée à un téléphone portable.

2.3.4 Difficultés liées à l'acquisition et la transmission du signal de parole

La transmission du signal de parole du producteur (le locuteur) au système de RAL chargé de l'analyser nécessite plusieurs étapes et emprunte divers types de supports. À chacune de ces étapes, le média utilisé pour transporter ce signal y imprime sa marque, le déforme voire le dégrade.

Ces dégradations, mais aussi et surtout les variations de la chaîne de transport entre les enregistrements de référence et de tests influent sur le processus de reconnaissance.

La première phase, la transmission du son par l'air ambiant, peut présenter de telles variations du fait de l'environnement acoustique et notamment des effets de réverbération du son qui peuvent survenir.

L'étape suivante, l'acquisition du signal sonore par un microphone, a fait l'objet de nombreuses études. Des différences notables existent entre les divers types de microphones, ne serait-ce qu'en termes de qualité, et de nombreux travaux expérimentaux ont montré d'importantes dégradations de performances en vérification du locuteur en cas de variation de matériel d'acquisition entre les enregistrements de référence et de test, comme par exemple l'utilisation de combinés téléphoniques de types différents ([van Vuuren 1996], [Reynolds 1996], [Auckenthaler 2000]).

Les mêmes observations peuvent être formulées à propos du ou des canaux par lesquels le signal de parole sera transmis jusqu'au système de RAL, dans le cas d'une reconnaissance à distance (par exemple, par téléphone). Outre la distorsion qu'induit le canal de transmission sur le signal de parole (notamment en termes de limitation de bande passante et d'ajout de bruit), le changement de canal d'un enregistrement à l'autre est source de dégradation de performance dans de nombreuses applications.

De très nombreux travaux ont été consacrés à tenter de compenser la contribution du canal de transmission afin de s'affranchir de ces problèmes de variabilité ([Furui 1981a], [Hermansky 1994], [Koolwaaij 2000], [Heck 2000], [Pelecanos 2001], [Reynolds 2003a]). Néanmoins ce problème voit peu à peu son importance s'amoindrir dans de nombreuses applications où la transmission de la parole, autrefois analogique, se fait de plus en plus sous forme numérique (par exemple dans les cas de la téléphonie mobile ou de la transmission de la voix sur IP). La question de la contribution du canal se voit dans ce cas remplacée par celle des dégradations induites par le codage utilisé. De plus en plus d'études portent donc sur la RAL à partir des standards actuels de codage de la parole, s'intéressant plus particulièrement aux conséquences de la compression et des pertes de paquets lors de la transmission ([Besacier 2000c], [Quatieri 2000], [Dunn 2001], [Gazit 2001], [Besacier 2004]).

Enfin, il convient de noter que l'importance des facteurs présentés ici varie selon le type d'application visée. En effet, comme évoqué plus haut, l'influence de la chaîne de transmission sur la qualité de la reconnaissance provient principalement des variations qu'elle ajoute entre l'enregistrement de référence d'un locuteur et l'enregistrement de

comparaison. Dans le cadre d'une application où la chaîne de transmission du signal est maîtrisée et stable d'un enregistrement à l'autre, ce problème n'existe pas ; dans ce cas, toute tentative de compensation de la contribution du canal de transmission est inutile, voire néfaste car génératrice de dégradations supplémentaires du signal. La segmentation en locuteurs, telle que définie en 2.2.5 (page 24), constitue quant à elle un cas particulier où l'influence du canal de transmission du signal peut se révéler bénéfique ; en effet, si par exemple les intervenants d'un document à segmenter ont été enregistrés avec différents matériel, la variaibilité inter-locuteurs s'en trouve augmentée et la variaibilité du matériel d'enregistrement devient par là même une aide à la segmentation.

2.3.5 Niveau de coopération des locuteurs

Le degré de coopération des intervenants distingue également les diverses applications de la reconnaissance automatique du locuteur et en conditionne en partie la difficulté. Ce niveau de coopération est sensible aussi bien lors de l'établissement de la référence d'un locuteur que lors d'une reconnaissance. Il se traduit par la réponse de l'utilisateur aux attentes du système (en termes d'effort d'articulation, de durée de parole, etc.). Un locuteur coopératif fera son possible pour répondre aux demandes du système dans le but d'être reconnu.

Nombre d'applications impliquent une participation volontaire des utilisateurs. Il s'agit par exemple d'applications de sécurité ou de confort, reposant sur le principe de l'identification ou de la vérification du locuteur. Dans ce cas, un utilisateur du système a tout intérêt à se montrer coopératif et à faire un effort tant au moment de s'enregistrer pour établir sa référence qu'au moment de se faire reconnaître.

Une toute autre classe d'applications de la reconnaissance automatique du locuteur repose sur une participation involontaire des locuteurs étudiés. L'exemple type en est la segmentation en locuteurs de documents divers tels que des conversations publiques ou des journaux télévisés ou radiophoniques. Dans ce cas, les locuteurs ne sont généralement pas conscients lors de l'enregistrement qu'un tel traitement sera effectué, voire même pas conscients d'être enregistrés. Il est donc impossible d'attendre d'eux une quelconque tentative de coopération en vue de faciliter la reconnaissance du locuteur et celle-ci s'en trouve d'autant plus difficile, car effectuée sur de la parole spontanée, sur des durées pouvant être très courtes, etc.

Enfin, dans certains cas la reconnaissance peut être effectuée contre le gré du locuteur. Ce cas se présente notamment dans le cadre judiciaire, où la plupart des locuteurs peuvent être tentés de transformer leur voix, dans un cadre où ils n'auraient pas intérêt à être reconnus, qu'ils soient bien le locuteur recherché ou qu'ils ne le soient pas.

2.3.6 Problèmes induits par le mode opératoire

Certaines difficultés supplémentaires plus spécifiques peuvent être introduites par l'application visée, notamment de par son mode d'interaction avec l'utilisateur. Ce mode peut influer en particulier sur :

- la quantité de données disponibles, que ce soit pour enregistrer une référence ou effectuer un test ; ce paramètre peut varier considérablement, même entre deux applications comparables (par exemple deux applications de vérification du locuteur) en fonction du public visé, du confort attendu ;

- la distance et position du microphone lors de l'enregistrement ; les applications de reconnaissance du locuteur par téléphone, en particulier, doivent être nettement plus robustes face à la variation de ce paramètre que d'autres applications à microphone fixe ;
- la qualité du matériel d'enregistrement utilisé ; l'intégration de la RAL aux sites Web, par exemple, souffre de la qualité souvent médiocre des microphones et convertisseurs analogique/numérique fournis avec les ordinateurs ([Boves 1998]) ;

2.4 Niveau de dépendance au texte

Diverses applications reposant sur une même tâche (telle que définie en 2.2) peuvent se différencier entre autres par leur degré de dépendance au texte.

Les systèmes de RAL dits indépendants du texte ne tiennent aucun compte du contenu linguistique du signal de parole. À l'opposé, les systèmes dits dépendants du texte utilisent la connaissance de tout ou partie de ce contenu linguistique pour affiner la reconnaissance du locuteur. Parmi ces derniers, plusieurs niveaux de dépendance au texte peuvent être distingués, correspondant à des besoins et des contraintes différents pour les applications :

- Systèmes à message fixé : la reconnaissance s'effectue sur un message fixé au préalable par l'application ou par l'utilisateur ([Jacob 2000]). Cette contrainte forte limite les applications envisageables à celles basées sur le principe d'identification ou de vérification du locuteur. En contrepartie ce mode de fonctionnement autorise l'emploi de techniques suffisamment "légères" pour être utilisées sur des systèmes à faible puissance de calcul (de type PDA ou téléphone portable par exemple).
- Systèmes à message prompté : ici, le message à prononcer par l'utilisateur est fixé par le système au moment du test, de manière à être différent lors de chaque test ([Higgins 1991], [Matsui 1994], [Lindberg 1997]). Les applications visées sont également de type identification ou vérification du locuteur. Par rapport à un système à message fixé, la sécurité se trouve ici nettement renforcée par l'universalité du "mot de passe", qui réduit la possibilité pour un imposteur de se faire reconnaître grâce à l'enregistrement de la voix d'une personne autorisée.

Les techniques de reconnaissance du locuteur dépendantes du texte tirent profit de la connaissance *a priori* de tout ou partie du texte prononcé par l'utilisateur, éliminant ainsi une source de variabilité du signal de parole (cf. 2.3.2, page 27). Cette caractéristique leur permet d'afficher de meilleures performances que les techniques indépendantes du texte, en particulier dans le cas de tests de très courte durée.

Ce niveau de performances se paye cependant par une souplesse moindre, particulièrement dans le cas des systèmes à messages fixes ou promptés, que leur principe de fonctionnement confine à des applications de type identification ou vérification du locuteur. À l'opposé, les techniques de reconnaissance du locuteur indépendantes du texte peuvent être appliquées dans un contexte où le contenu linguistique est connu et y montrer dans certains cas des performances équivalentes à celles qu'obtiendrait une méthode exploitant cette connaissance.

2.5 Clients, imposteurs et pseudo-imposteurs

À ce niveau du document, il est utile de préciser le sens de termes qui seront très utilisés par la suite : “client” et “imposteur”.

Le terme de “client” est utilisé pour désigner tout locuteur connu du système (“connu” dans le sens où un modèle de ce locuteur a été calculé à partir d’un ou plusieurs échantillon(s) de sa voix). Par extension, il est aussi utilisé, lors de la comparaison d’un signal de parole avec un modèle de locuteur, pour désigner ce test comme un test de type “client” (ou plus simplement un test client) s’il s’agit d’un test intra-locuteur, en d’autres termes si le signal testé a bien été émis par le locuteur correspondant à ce modèle. De la même façon, les scores issus de tests clients sont souvent qualifiés de scores clients.

À l’inverse, un test inter-locuteurs (impliquant un signal émis par un autre locuteur que celui à qui correspond le modèle) est désigné comme un test de type “imposteur” (ou simplement “test imposteur”). Les scores correspondant à des tests imposteurs sont bien sûr appelés des scores imposteurs.

Cependant, il convient de différencier les concepts de “vrai” imposteur et de “pseudo” imposteur. Un vrai imposteur est un individu qui correspond en fait au sens commun du mot : une personne qui, dans le cas de la RAL, tente de se faire accepter d’un système en imitant la voix d’un locuteur client. C’est ce genre de tentative que se doivent de déjouer les systèmes de RAL. Le concept de pseudo-imposteur a été introduit pour désigner le cas d’un test imposteur où le signal testé ne correspond pas à une réelle tentative d’imposture : un locuteur a été enregistré et sa voix est comparée à celle d’un autre locuteur, mais sans qu’il n’y ait une quelconque volonté de se faire passer pour cet autre locuteur. Ce cas se présente notamment dans le cadre de la segmentation en locuteurs, lorsqu’une portion du signal à segmenter est comparée à une autre, prononcée par un autre locuteur. Mais les tests pseudo-imposteurs ne se limitent pas à ce cas. Il s’en trouve également beaucoup au cours du développement d’un système de RAL et de l’évaluation de ses performances. En effet, l’évaluation d’un système de vérification du locuteur, par exemple, devrait logiquement reposer sur un ensemble de tests clients et un ensemble de tests imposteurs. Il est malheureusement impossible de trouver une base de donnée importante de tests imposteurs. À la place, le développement et l’évaluation des systèmes se font donc en ayant recours à un ensemble de tests pseudo-imposteurs, bien que cela ne soit pas strictement la même chose.

Dans la suite de ce document, le terme “pseudo-imposteur” est utilisé lorsqu’il paraît nécessaire de préciser (ou de rappeler) qu’il ne s’agit pas d’un vrai test imposteur. Le reste du temps, le terme “imposteur” doit être compris soit dans le sens de “non-client” (signifiant vrai ou pseudo-imposteur mais dans un contexte où la distinction n’a pas d’importance), soit dans le sens de “pseudo-imposteur” si le contexte rend ce sens évident.

2.6 Chaîne de traitement

Malgré les différences entre les tâches présentées précédemment, leur mise en œuvre implique nécessairement les mêmes outils de base permettant de réaliser une phase d’apprentissage (estimation du modèle d’un locuteur) ainsi qu’une phase de test durant

laquelle un enregistrement est comparé à un modèle de locuteur. En amont de ces deux modules, un traitement est appliqué au signal afin d'en extraire les informations pertinentes.

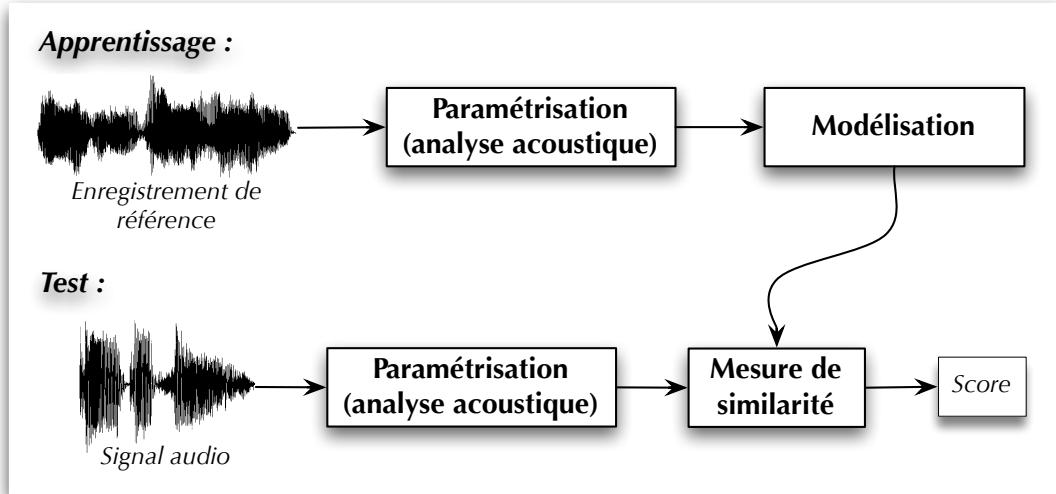


FIG. 2.6 – Modules de base d'un système de reconnaissance du locuteur

2.6.1 Paramétrisation du signal de parole

L'objectif du processus de paramétrisation est d'extraire du signal de parole les informations caractéristiques du locuteur, afin d'obtenir une nouvelle représentation du signal, plus compacte car se focalisant sur les seules informations pertinentes en vue de la reconnaissance du locuteur.

Cependant, cet objectif reste un idéal incompli. En effet, bien que posées depuis longtemps, les questions de la nature même de ces informations ([Voiers 1964]) ainsi que des techniques à utiliser pour les extraire du signal ([Wolf 1972], [Goldstein 1975], [Sambur 1975], [Atal 1976], [Cheung 1978]) n'ont toujours pas trouvé de réponse définitive.

Un consensus s'est toutefois dégagé : une grande partie de l'information se trouve dans les caractéristiques spectrales à court terme ([Furui 1981a]) et leurs variations ([Furui 1981b]).

Analyse à court terme, bas niveau, du signal de parole — Paramètres cepstraux

Avant son analyse, le signal de parole subit une phase (optionnelle) de pré-accentuation, dont l'objectif est d'amplifier les hautes fréquences du spectre, qui sont assourdis par le processus de production de la parole. Le filtre appliqué est le suivant :

$$x_p(t) = x(t) - a \cdot x(t-1) \quad (2.1)$$

où $x(t)$ représente l'échantillon de signal à l'instant t . La valeur de a est généralement comprise dans l'intervalle $[0, 95; 0, 98]$. Notons que ce pré-traitement n'est pas systématiquement appliqué, la question de sa pertinence n'ayant pas encore trouvé de réponse définitive ; le choix de l'appliquer ou non est uniquement empirique.

Analyse par bancs de filtres L'analyse du signal de parole est effectuée localement par application d'une fenêtre de courte durée, déplacée le long du signal et dont chaque application produit un vecteur spectral. Le type de fenêtre utilisé (les fenêtres de Hamming et de Hanning sont les plus utilisées en RAL), ainsi que sa largeur et le décalage entre deux applications consécutives influent sur le processus. Les deux largeurs de fenêtre revenant le plus fréquemment sont 20 milli-secondes et 30 milli-secondes, correspondant à la durée moyenne qui permet de vérifier l'hypothèse de stationnarité. Le décalage, quant à lui, est choisi de façon à obtenir un recouvrement entre deux fenêtres consécutives ; la valeur de 10 milli-secondes est couramment utilisée.

Après application de la fenêtre sur le signal, sa transformée de Fourier (FFT — *Fast Fourier Transform*) est calculée. Le nombre de points pour le calcul de la FFT, généralement une puissance de 2 supérieure au nombre d'échantillons présents dans la fenêtre, est le plus souvent 512 pour les largeurs de fenêtres classiques données ci-dessus. Enfin, le module de la FFT est extrait et un spectre est obtenu (échantillonné sur 512 points mais symétrique, donc seuls 256 points en sont conservés).

Le spectre est ensuite lissé afin de récupérer l'enveloppe spectrale débarrassée des variations les plus fines. Ce lissage, qui a pour autre conséquence la réduction de la taille des vecteurs spectraux, est réalisé par la multiplication du spectre par un banc de filtres. Le banc de filtres est défini par la forme (triangulaire ou autre) des filtres qui le composent ainsi que par leur position dans la gamme des fréquences. En particulier, l'échelle Mel, similaire à l'échelle de fréquences de l'oreille humaine, peut être utilisée pour positionner les filtres.

Enfin, le logarithme de chaque coefficient de l'enveloppe spectrale est extrait et multiplié par 20 afin d'obtenir l'enveloppe spectrale exprimée en dB. Une dernière transformation est appliquée aux vecteurs spectraux ainsi obtenus pour produire des vecteurs cepstraux ([Oppenheim 1989], [Bogert 1963], [Oppenheim 1968]). Il s'agit d'une transformée en cosinus discrète, exprimée sous la forme :

$$c_n = \sum_{k=1}^K S_k \cdot \cos\left(\frac{n\pi}{K}(k - \frac{1}{2})\right), \quad n = 1, 2, \dots, N \quad (2.2)$$

où K est le nombre de coefficients des vecteurs spectraux, S_k sont ces coefficients et N est le nombre de coefficients cepstraux désirés (avec la contrainte $N \leq K$). Le résultat obtenu à la fin du processus est un vecteur cepstral à N coefficients pour chaque fenêtre d'analyse du signal.

La figure 2.7 récapitule le processus décrit ici.

Analyse LPC Une autre approche très utilisée dans la littérature pour le calcul de coefficients cepstraux est le recours à l'analyse LPC (*Linear Predictive Coding*). Cette analyse se base sur un modèle linéaire de la production de parole, généralement un modèle auto-régressif. Le principe de l'analyse LPC est d'estimer les paramètres de ce modèle pour une portion de signal, sélectionnée par une fenêtre glissant le long du

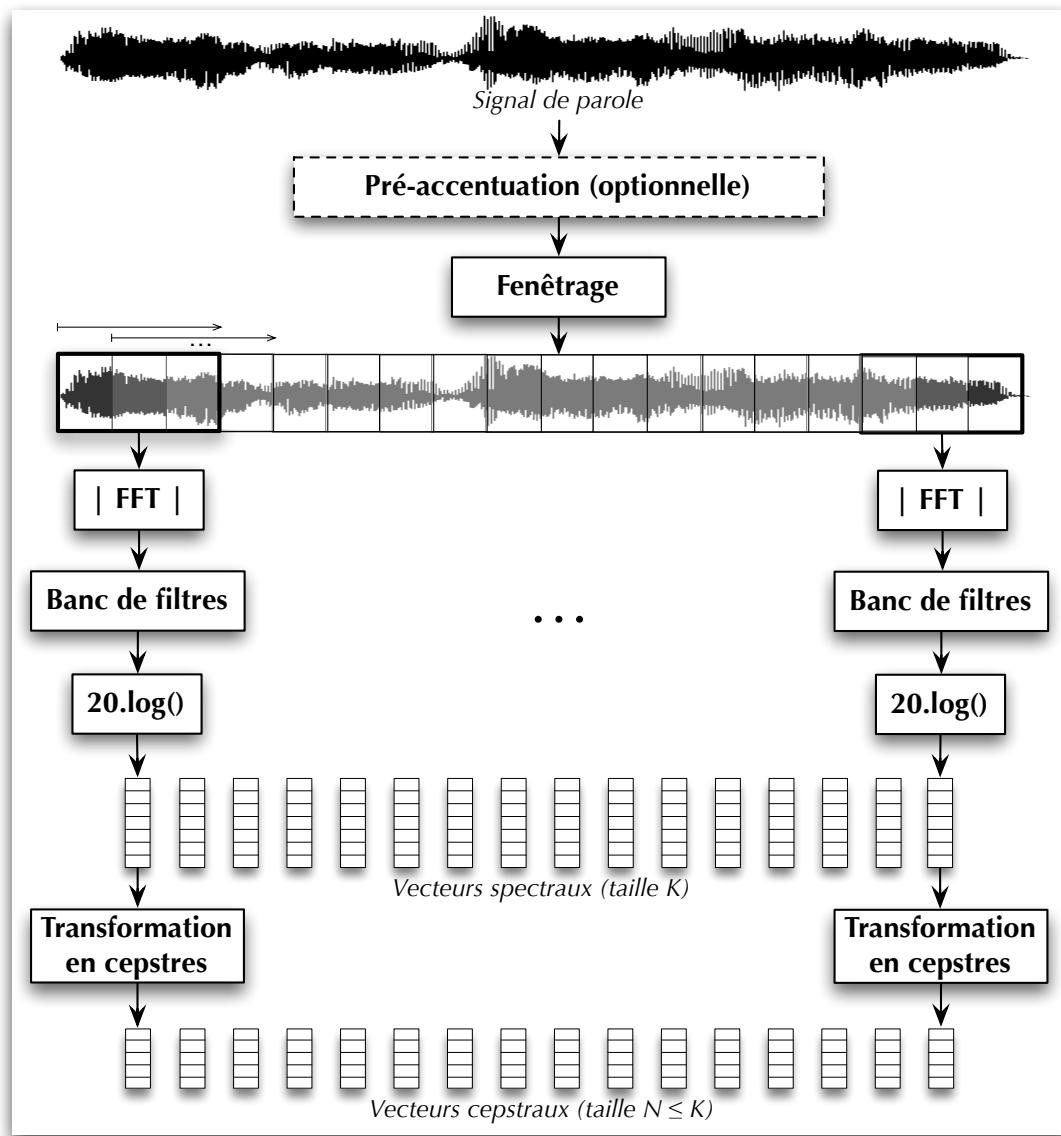


FIG. 2.7 – Calcul d'une représentation cepstrale d'un signal de parole à partir d'une analyse en banc de filtres.

signal (le lecteur se reportera à [Oppenheim 1989] ou [Petrovska-Delacrétaz 2000] pour une présentation de divers algorithmes d'estimation des coefficients issus de l'analyse LPC). De ces coefficients LPC (également appelés coefficients prédictifs), il est possible de tirer directement un ensemble de coefficients cepstraux ([Petrovska-Delacrétaz 2000]).

Analyse dynamique à court terme

Les variations à court terme du signal de parole sont porteuses d'informations caractéristiques du locuteur. L'extraction de ces informations au cours de la paramétrisation passe par l'analyse de la dynamique des vecteurs de paramètres.

L'approche la plus utilisée à l'heure actuelle, proposée par [Furui 1981b], repose sur l'utilisation des dérivées première et seconde des coefficients cepstraux. Ces dérivées sont généralement estimées à partir de fonctions polynomiales comme suit :

$$\frac{\partial c_n(t)}{\partial t} \approx \Delta c_n(t) = \frac{\sum_{k=-K}^K k \cdot c_n(t+k)}{\sum_{k=-K}^K |k|} \quad (2.3)$$

$$\frac{\partial^2 c_n(t)}{\partial t^2} \approx \Delta\Delta c_n(t) = \frac{\sum_{k=-K}^K k^2 \cdot c_n(t+k)}{\sum_{k=-K}^K k^2} \quad (2.4)$$

où $c_n(t)$ représente le $n^{ième}$ coefficient cepstral au temps t , $\Delta c_n(t)$ et $\Delta\Delta c_n(t)$ (communément appelés coefficients Delta et Delta-Delta respectivement) les estimations de ses dérivées première et seconde et K détermine la largeur de la fenêtre utilisée. La valeur de K , objet de plusieurs études destinées à déterminer son optimum, est classiquement fixée à 2 (soit une fenêtre de 5 trames) pour des trames espacées de 10 milli-secondes ([Reynolds 1994]).

Après leur estimation, les vecteurs de coefficients Delta et/ou Delta-Delta sont concaténés à la fin du vecteur de paramètres correspondant (dont la taille passe donc à $2N$ ou $3N$ coefficients).

Analyse à long terme du signal de parole — Extraction d'informations de "haut niveau"

Au delà des informations révélées par l'analyse spectrale à court terme, d'autres informations caractéristiques du locuteur sont présentes dans le signal de parole à un niveau plus élevé, portées par des modulations à long terme du signal, telles que la prosodie ou le contenu linguistique (à travers le choix des mots ou des structures de phrases, par exemple).

Ces informations ont été longtemps ignorées en reconnaissance du locuteur au profit de l'analyse spectrale à court terme. D'une part, leur nature (les rattachant plus à l'acquis qu'à l'inné) les rend intuitivement plus sujettes à contrefaçon, à l'opposé des informations spectrales à court terme, plus dépendantes des caractéristiques physiques des individus. D'autre part, l'extraction de ces informations est un problème moins maîtrisé que l'analyse spectrale. L'exploitation de ce type d'informations ne peut s'envisager que pour le traitement d'enregistrements de longue durée, aussi bien pour la phase d'apprentissage que pour la phase de test. En particulier, une quantité importante de données est nécessaire durant la phase d'apprentissage d'un modèle de locuteur. Enfin, des outils efficaces de reconnaissance de la parole sont souvent requis pour

décoder le contenu linguistique, cette contrainte s'accompagnant d'un besoin de forte puissance de calcul.

Cependant, ces obstacles tendent à disparaître peu à peu, notamment grâce aux progrès accomplis dans le domaine des outils de reconnaissance de la parole et dans celui de la puissance de calcul disponible. De plus, l'apparition de nouvelles tâches de RAL traitant des documents audio de plus longue durée, accompagnée de la création de bases de données correspondantes, rend possible le recours aux informations de haut niveau. Celles-ci sont perçues comme un moyen de dépasser les limites des informations spectrales à court terme, notamment grâce à leur moindre sensibilité aux dégradations acoustiques.

Des travaux récents se sont orientés dans cette direction. Les informations caractéristiques du locuteur portées par la prosodie sont étudiées dans [Adami 2003] à travers l'observation des trajectoires de l'énergie et de la fréquence fondamentale du signal. D'autres travaux ([Andrews 2001], [Doddington 2001], [Peskin 2003], [Reynolds 2003b], [Campbell 2003a]) s'intéressent à divers types d'informations de haut niveau et à leur utilisation conjointe avec les informations issues de l'analyse spectrale à court terme pour renforcer la robustesse des systèmes traditionnels. Les résultats obtenus par la fusion des sorties des deux types de systèmes sont encourageants (il est à noter cependant que les systèmes à base d'analyse spectrale à court terme combinée à des modèles GMM (*cf.* section 2.6.3) conservent de meilleures performances dans le cadre des campagnes d'évaluation NIST).

2.6.2 Traitement post-paramétrisation

Dans le cas d'une paramétrisation par analyse spectrale à court terme, une fois les vecteurs de paramètres calculés, deux types de traitements leur sont généralement appliqués. Il s'agit d'une part de compenser autant que possible les distorsions du signal provoquées par le canal de transmission et d'autre part de sélectionner les vecteurs porteurs d'information utile pour la suite du processus. L'ordre d'application de ces deux traitements fait l'objet d'une discussion au chapitre 4, page 71.

Compensation de la contribution du canal

Plusieurs techniques sont proposées dans la littérature pour tenter de compenser les distorsions du signal induites par le canal de transmission (*cf.* section 2.3.4, p. 28).

Le retrait de la moyenne cepstrale (CMS — *Cepstral Mean Subtraction*) repose sur l'hypothèse que ces distorsions sont suffisamment stables sur un intervalle de temps long pour que la moyenne des coefficients cepstraux en soit une estimation raisonnable (ce qui rend cette technique peu appropriée pour les très courtes durées). La CMS peut être appliquée de manière globale, calculant la moyenne sur toute la durée de l'enregistrement ([Atal 1974], [Furui 1981a]). Elle peut aussi se faire de manière locale, en retirant la moyenne sur une fenêtre glissant le long du signal ([Rosenberg 1994]), afin de prendre en compte les variations lentes des distorsions du signal. Le retrait de moyenne par fenêtre glissante est également envisageable dans le cas d'un enregistrement composé de segments issus de canaux de transmission différents, en faisant l'hypothèse que la largeur choisie pour la fenêtre est inférieure à la durée d'un segment (il subsiste tout de même une difficulté au niveau des frontières de segments, où l'effet du retrait

de la moyenne locale n'est pas garanti).

La normalisation des vecteurs cepstraux complète la CMS en y ajoutant la division par l'écart-type de la distribution des vecteurs (cf. chapitre 4, page 71).

Le filtrage RASTA (*Relative Spectral Processing*), proposé dans [Hermansky 1994], atténue les composantes du signal de parole portées par les basses fréquences ainsi que les hautes fréquences, éliminant ainsi une grande part des distorsions liées au canal de transmission. Il a rencontré un certain succès dans le domaine de la reconnaissance automatique de la parole pour cette raison. Il a cependant été montré qu'une partie des basses fréquences de modulation filtrées par RASTA étaient porteuses d'information utile pour la vérification du locuteur ([van Vuuren 1999], [Besacier 2000b]).

Le *Feature Warping*, proposé dans [Pelecanos 2001], qui consiste à déformer la distribution locale (sur une fenêtre glissante) des vecteurs de paramètres pour la rapprocher d'une distribution gaussienne, montre une grande efficacité dans le cadre des campagnes d'évaluation NIST.

Le lecteur se reportera à [Pelecanos 2001] pour une évaluation comparative, dans le cadre de la vérification du locuteur, des différents traitements présentés ici.

Plus récemment, la technique de *Feature Mapping* a été proposée par [Reynolds 2003a]. Elle repose sur la modélisation de chaque canal de transmission par adaptation d'un modèle générique indépendant du canal (étant appris sur un ensemble de données regroupant tous les types de canaux connus). Après la paramétrisation d'un signal, celui-ci est comparé à l'ensemble de ces modèles, le plus probable déterminant le type de canal utilisé pour ce signal. Une transformation est alors appliquée à chaque vecteur de paramètres, basée sur la différence entre la plus proche composante du modèle de canal et la composante correspondante du modèle indépendant du canal.

Sélection de l'information utile

Un enregistrement comprend des zones non porteuses d'information relative au locuteur, en particulier les zones de silence mais aussi les zones de bruit ou de musique. Il convient de détecter ces zones et de les supprimer afin de n'utiliser dans la suite du processus que des vecteurs de paramètres porteurs d'information utile.

Cette détection se limite généralement au traitement des zones de silence, reposant sur l'étude de la distribution des vecteurs de paramètres ou de la distribution de l'énergie du signal. Un exemple d'une telle technique de séparation silence/parole est détaillé au chapitre 4, page 69.

La détection des zones de bruit et de musique est un problème plus complexe et implique un traitement beaucoup plus lourd, proche par le principe de la segmentation en locuteurs, nécessitant une modélisation des classes de son (silence, parole, bruit, musique) à distinguer. Cependant ce traitement ne concerne que certaines catégories d'enregistrements (généralement de longue durée), la détection des seules zones de silence restant satisfaisante pour la majorité des applications.

2.6.3 Modélisation du locuteur

Les techniques les plus utilisées pour modéliser la voix d'un locuteur sont exposées ici.

Alignment temporel dynamique

La technique de l'alignment temporel dynamique (DTW — *Dynamic Time Warping*) consiste à comparer deux enregistrements, vus comme deux séquences de vecteurs de paramètres, en réalisant un alignment temporel de la première séquence de vecteurs avec la seconde. Une distance moyenne entre les deux séquences alignées est ensuite calculée.

Dans ce cadre, un modèle de locuteur est simplement la séquence de vecteurs de paramètres correspondant à ses données d'apprentissage.

De par son principe, de l'alignment temporel dynamique dans le domaine de la RAL se limite au mode dépendant du texte ([Furui 1981a], [Booth 1993],[Yu 1995]). Les bonnes performances obtenues par l'algorithme de DTW dans le cas d'enregistrements courts, alliées à sa complexité calculatoire relativement faible, sont à l'origine de son utilisation très répandue pour les applications correspondantes (telle la reconnaissance du locuteur à partir de mot de passe).

Quantification vectorielle

La quantification vectorielle (VQ — *Vector Quantization*) repose sur un partitionnement de l'espace acoustique en sous-espaces, chacun représenté par un vecteur centroïde. Un modèle de locuteur consiste alors en un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification (*codebook*). Lors de la phase de test, la comparaison d'un vecteur de paramètres à ce modèle commence par le calcul d'une distance entre le vecteur et chacun des centroïdes ; la distance minimale est conservée. La distance d'une séquence de vecteurs au modèle est la moyenne des distances minimales obtenues pour les vecteurs la composant.

La quantification vectorielle trouve son application en reconnaissance du locuteur en mode dépendant du texte comme en mode indépendant du texte ([Mason 1989], [Soong 1992], [Matsui 1992]). Sa rapidité et ses performances dépendent directement de la taille du dictionnaire. Cependant, elles varient de façon inverse : un dictionnaire plus grand augmentera les performances de reconnaissance au prix d'une perte de vitesse d'exécution.

Méthodes statistiques

L'approche statistique repose sur la modélisation de la distribution des vecteurs de paramètres correspondant à un locuteur. Ce principe relativement simple offre néanmoins de très bonnes performances, notamment en mode indépendant du texte avec l'apparition des GMM.

Spectre/cepstre moyen Les premiers travaux adoptant une approche statistique de la modélisation des locuteurs reposent sur l'estimation du spectre moyen à long terme ([Pruzansky 1963]). La comparaison d'un signal de parole à ce modèle consiste alors à calculer une distance entre le spectre moyen estimé sur le signal de test et celui estimé sur les données d'apprentissage.

Méthodes statistiques d'ordre 2 L'approche statistique est une évolution des travaux préliminaires présentés au paragraphe précédent. Elle consiste à employer des statistiques d'ordre 2, prenant en compte la variation des paramètres acoustiques par le calcul d'une matrice de covariance ([Bimbot 1995]).

Modèles à mixture de gaussiennes Un modèle à mixture de gaussiennes (GMM — *Gaussian Mixture Model*) représente une estimation d'une distribution multidimensionnelle quelconque par une somme pondérée de distributions gaussiennes (chacune caractérisée par un vecteur moyen et une matrice de covariance). Un GMM est défini par l'ensemble des poids affectés à ses composantes et par l'ensemble des vecteurs moyens et des matrices de covariances définissant les composantes elles-mêmes.

Dans le cas de la reconnaissance du locuteur, la distribution estimée est celle des vecteurs de paramètres des données d'apprentissage d'un locuteur. Seule la distribution des vecteurs est modélisée, un GMM n'intégrant aucune notion de temporalité.

L'estimation d'un GMM à partir d'un signal d'apprentissage peut être réalisée soit par l'algorithme EM (*Expectation-Maximization* — [Dempster 1977], [Reynolds 1992], [Reynolds 1995]) soit en appliquant une méthode d'adaptation bayésienne (MAP — *Maximum A Posteriori* — [Gauvain 1994]) à un modèle de locuteur générique ([Reynolds 1997], [Reynolds 2000]). Cette dernière solution permet de bonnes performances même pour un modèle appris à partir d'une faible quantité de données.

Les GMM sont rapidement devenus les modèles les plus utilisés dans le cadre de la reconnaissance automatique du locuteur en mode indépendant du texte depuis leur introduction dans ce domaine ([Reynolds 1992], [Reynolds 1995]). Ce succès est dû à de bonnes performances générales, inégalées par les techniques concurrentes, particulièrement dans le cas d'apprentissage à partir des signaux de courte durée, où l'estimation de GMM par adaptation bayésienne permet de conserver un bon niveau de performances.

Une présentation plus détaillée du principe des GMM et des techniques d'estimation associées est faite au chapitre 4, page 81.

Modèles de Markov cachés Un modèle de Markov caché (HMM — *Hidden Markov Model*) est un automate probabiliste à états finis, qui change d'état à chaque unité de temps. A chaque état est associée une fonction de densité de probabilité d'émission (généralement, un GMM) et chaque transition entre états porte une probabilité. Un HMM est complètement défini par ce triplet : l'ensemble des états, l'ensemble des probabilités de transition entre états et l'ensemble des fonctions de densité de probabilité d'émission associées aux états. Cette structure permet au modèle d'intégrer des informations temporelles, absentes par exemple des GMM. Les modèles de Markov cachés sont abondamment documentés dans la littérature. Le lecteur se reporterà entre autres

à [Rabiner 1989] pour une explication détaillée de leur structure et de leur fonctionnement.

Dans le cadre de la reconnaissance automatique du locuteur, l'emploi de HMM se retrouve en mode dépendant du texte comme en mode indépendant du texte.

En mode dépendant du texte, la structure des modèles de Markov cachés est bien adaptée à la prise en compte des informations temporelles issues de la connaissance *a priori* du texte prononcé.

L'extraction d'informations temporelles en mode indépendant du texte est également envisageable ([Peskin 1993]), mais à ce jour l'utilisation de HMM dans ce cadre n'a pas montré de réel avantage par rapport à l'utilisation de GMM ([Weber 2000]).

Machines à support vectoriel

Les machines à support vectoriel (SVM — *Support Vector Machines*) sont des classificateurs adaptés à la partition en deux classes de régions complexes de l'espace par la définition d'une frontière optimale et non linéaire ([Vapnik 1995]). Pour cette raison, l'utilisation des SVM en reconnaissance automatique du locuteur dans le cadre d'un apprentissage discriminant fait l'objet d'études depuis quelques années ([Schmidt 1996], [Gu 2001], [Campbell 2002], [Campbell 2003b], [Wan 2003]).

Plus récemment, une voie de recherche a été explorée, envisageant les SVM en RAL comme solution complémentaire aux GMM, diverses approches étant étudiées ([Kharroubi 2001], [Fine 2001], [Dong 2002]). En particulier, la fusion des sorties d'un reconnaissendeur à base de SVM avec celles d'un système traditionnel à base de GMM apporte un gain de performances intéressant ([Campbell 2004]).

Modèles connexionnistes

Les modèles dits connexionnistes sont un ensemble de techniques utilisant un réseau de neurones en vue de faire de la discrimination entre locuteurs. Le réseau est entraîné avec l'ensemble des signaux d'apprentissage des locuteurs connus afin qu'il apprenne à les discriminer les uns des autres (un tel apprentissage correspondant en fait à une tâche de classification). Le test d'un enregistrement consiste à calculer la vraisemblance que la séquence de vecteurs correspondante soit produite par le réseau.

Différents types de réseaux ont été proposés dans la littérature au fil du temps : MLP (*Multi-Layer Perceptron* — [Oglesby 1990]), RBF (*Radial Basis Functions* — [Oglesby 1991], [Frederickson 1994]), LVQ (*Learning Vector Quantisation* — [Bennani 1990]).

L'orientation vers la discrimination entre des locuteurs connus fait de l'identification du locuteur le champ d'application logique de ces techniques — mais non exclusif, des modèles connexionnistes ayant été proposés également dans le cadre de la vérification du locuteur. L'utilisation de ces techniques en RAL reste cependant limitée, notamment du fait de la complexité de l'apprentissage du réseau et de sa mise à jour lors de l'ajout d'un nouveau locuteur.

Modèles prédictifs

Partant du principe que l'observation d'une série de trames du signal de parole peut permettre de prédire la trame suivante, les modèles prédictifs reposent sur la modélisation d'un locuteur par une fonction de prédiction estimée sur ses données d'apprentissage. L'estimation de cette fonction se fait dans la littérature par un modèle auto-régressif vectoriel (ARV — [Grenier 1980], [Bimbot 1992], [Montacié 1992], [Griffin 1994], [Magrin-Chagnolleau 1996]) ou un réseau de neurones ([Hattori 1992], [Artières 1993], [Bennani 1994], [Paoloni 1996], [Chetouani 2004]).

Deux approches sont possibles pour exploiter la fonction de prédiction lors de la phase de test. Dans la première approche ([Grenier 1980]), la fonction de prédiction au modèle d'un locuteur donné est appliquée, pour chaque trame du signal de test, sur les trames qui la précédent. La différence entre le résultat obtenu (la trame prédictive) et la trame réellement observée est évaluée. Cette erreur de prédiction est moyennée sur toutes les trames pour obtenir le score de la séquence.

La seconde approche consiste à estimer une fonction de prédiction sur le signal de test et à la comparer (par un calcul de distance) à celle du locuteur ([Bimbot 1992], [Montacié 1992], [Griffin 1994], [Magrin-Chagnolleau 1996]). Cette approche n'est cependant envisageable que pour des durées de test suffisamment longues pour autoriser l'estimation fiable d'une fonction de prédiction.

Les modèles prédictifs montrent de bonnes performances en identification du locuteur, mais ne parviennent pas au niveau de l'état de l'art dans le cadre de la vérification du locuteur où ils sont par conséquent peu utilisés à l'heure actuelle. De plus, le principe fondamental des modèles prédictifs, l'exploitation des informations dynamiques du signal de parole pour apprendre une fonction de prédiction, a été remis en question dans le cas des modèles ARV par [Magrin-Chagnolleau 1996].

2.6.4 Normalisation des scores

Les difficultés auxquelles doit faire face un système de reconnaissance du locuteur, présentées à la section 2.3 (page 26), sont la source d'une grande variabilité des scores obtenus lors du test d'un signal de parole par rapport à un modèle de locuteur. Cette variabilité est un obstacle au choix du seuil permettant de décider, d'après le score obtenu, si le signal testé correspond bien au modèle considéré. La difficulté est encore accrue si ce seuil doit avoir la même valeur pour tous les locuteurs, plutôt qu'être déterminé locuteur par locuteur, puisqu'à la variabilité intra-locuteur (causée par les variations de la voix mais aussi par des facteurs extérieurs tels que le bruit environnant) vient alors s'ajouter une part de la variabilité inter-locuteurs. Pourtant, un tel seuil indépendant du locuteur est souhaitable dans de nombreuses applications.

Pour minimiser ce problème, les scores obtenus après comparaison du signal de parole à un modèle de locuteur passent par une phase de normalisation dont le but est évidemment de réduire la variabilité constatée avant de comparer les scores au seuil de décision.

Normalisation par rapport de vraisemblances

La normalisation par rapport de vraisemblances représente un cas particulier. Elle repose sur un test d'hypothèse bayésien, entre, d'une part, l'hypothèse que le signal de parole testé ait été émis par le locuteur considéré, et d'autre part l'hypothèse inverse (que ce signal n'ait pas été émis par ce locuteur). Cependant, les effets de ce type de technique sur les distributions de scores sont très similaires à ceux obtenus par normalisation.

Ce type de technique fut proposé pour la première fois par Higgins et al. en 1991 ([Higgins 1991]), suivis par Matsui et Furui en 1993 ([Matsui 1993]), pour qui le test d'hypothèse prend la forme d'un rapport de vraisemblances :

$$LR_{\mathcal{X}}(y) = \frac{L_{\mathcal{X}}(y)}{L_{\bar{\mathcal{X}}}(y)} \quad (2.5)$$

où $L_{\mathcal{X}}(y)$ est la vraisemblance du signal de test y par rapport au modèle de locuteur \mathcal{X} et $L_{\bar{\mathcal{X}}}(y)$ la vraisemblance de l'hypothèse inverse, correspondant au cas où y a été prononcé par un locuteur autre que X (le lecteur se reportera au chapitre 4, page 74, pour une présentation plus détaillée de ce principe).

Dans les deux approches, la vraisemblance $L_{\bar{\mathcal{X}}}(y)$ est estimée à partir d'une cohorte de modèles de locuteurs. Dans [Higgins 1991], la cohorte de locuteurs (également appelée cohorte d'imposteurs) est choisie de façon à être proche du locuteur X tout en excluant celui-ci. Au contraire, dans [Matsui 1993], la cohorte retenue inclut X . Malgré cette différence, ces deux techniques améliorent les performances de manière identique en vérification du locuteur.

Dans le but de réduire la quantité de calcul nécessaire, la cohorte de modèles imposteurs a été remplacée par la suite par un unique modèle appris sur les mêmes données. Cette idée constitue la base de la normalisation par modèle du monde, introduite par [Carey 1991]. Plusieurs études ont depuis montré l'intérêt d'une telle technique de normalisation ([Reynolds 1997], [Heck 1997], [Gravier 1998]).

Les autres techniques de normalisation présentées ci-dessous (à l'exception de Tnorm et de ses dérivés) sont généralement appliquées sur des scores issus d'une normalisation par modèle du monde (appelés communément rapports de vraisemblances).

Normalisation par distribution imposteur centrée/réduite

Cette famille de techniques de normalisation est la plus utilisée. Le score normalisé $\tilde{L}_{\mathcal{X}}(y)$ du signal y testé par rapport au modèle de locuteur \mathcal{X} est donné par² :

$$\tilde{L}_{\mathcal{X}}(y) = \frac{L_{\mathcal{X}}(y) - \mu_{imp}}{\sigma_{imp}} \quad (2.6)$$

²En pratique, pour des raisons de précision des calculs, la normalisation est généralement effectuée sur le logarithme du score plutôt que sur le score lui-même ; il convient alors de remplacer, dans l'équation 2.6, $L_{\mathcal{X}}(y)$ par $\log(L_{\mathcal{X}}(y))$ (ou par $\log(LR_{\mathcal{X}}(y))$ dans les cas où une pré-normalisation par rapport de vraisemblances est appliquée).

où la moyenne μ_{imp} et l'écart-type σ_{imp} sont estimés sur une distribution de scores (pseudo) imposteurs.

Diverses possibilités existent pour calculer cette distribution de scores imposteurs. Elles définissent autant de techniques de normalisation, dont les principales sont présentées ci-dessous.

Znorm La normalisation Znorm (*Zero Normalisation*) est directement dérivée du travail effectué dans [Li 1988]. Elle a été très utilisée en vérification du locuteur au milieu des années 90.

Un modèle de locuteur est testé par rapport à un ensemble de signaux de parole produits par des imposteurs, fournissant une distribution de scores imposteurs. La moyenne μ_{imp}^x et l'écart-type σ_{imp}^x (qui sont les deux paramètres de cette normalisation) sont estimées sur cette distribution et appliquées (cf. [Rosenberg 1996]) sur les scores produits lors de l'utilisation du système de vérification du locuteur, tel qu'illustré par la figure 2.8.

L'un des avantages de Znorm est la possibilité d'estimer les paramètres de normalisation lors de l'apprentissage du modèle de locuteur, plutôt que lors de la phase de test.

Hnorm Dans le cas de la parole téléphonique, la plupart des modèles de locuteurs répondent différemment selon le type de combiné utilisé lors de la phase de test (qui n'est pas forcément le même que lors de l'apprentissage du modèle). Une variante de la technique Znorm, nommée Hnorm (pour *handset normalization*), a été introduite dans [Reynolds 1996] dans le but de minimiser les effets de la différence de combinés entre apprentissage et test.

Les paramètres de cette normalisation sont estimés en confrontant chaque modèle de locuteur à des signaux de parole produits par des imposteurs en utilisant divers types de combinés. Un jeu de paramètres de normalisation est ainsi calculé pour chaque modèle et chaque type de combiné.

Lors de la phase de test, c'est le type de combiné utilisé pour le signal testé qui détermine le jeu de paramètres à utiliser pour la normalisation.

Tnorm Tnorm (*test normalization*), proposée dans [Auckenthaler 2000], représente une fusion entre une normalisation de type Znorm — reposant sur l'estimation d'une moyenne et d'une variance pour normaliser la distribution des scores imposteurs — et une normalisation par cohorte de locuteurs.

À l'opposé de Znorm dont les paramètres de normalisation ne dépendent que du modèle du locuteur considéré, Tnorm repose sur des paramètres dépendant du signal de test. Ces paramètres sont ici calculés par l'utilisation de modèles d'imposteurs au lieu de signaux de test imposteurs.

Durant le test, le signal de parole à tester est confronté au modèle du locuteur considéré, mais aussi à une cohorte de modèles imposteurs afin d'estimer une distribution de scores imposteurs et d'en tirer les paramètres de la normalisation (μ_{imp}^y et σ_{imp}^y), comme illustré par la figure 2.9.

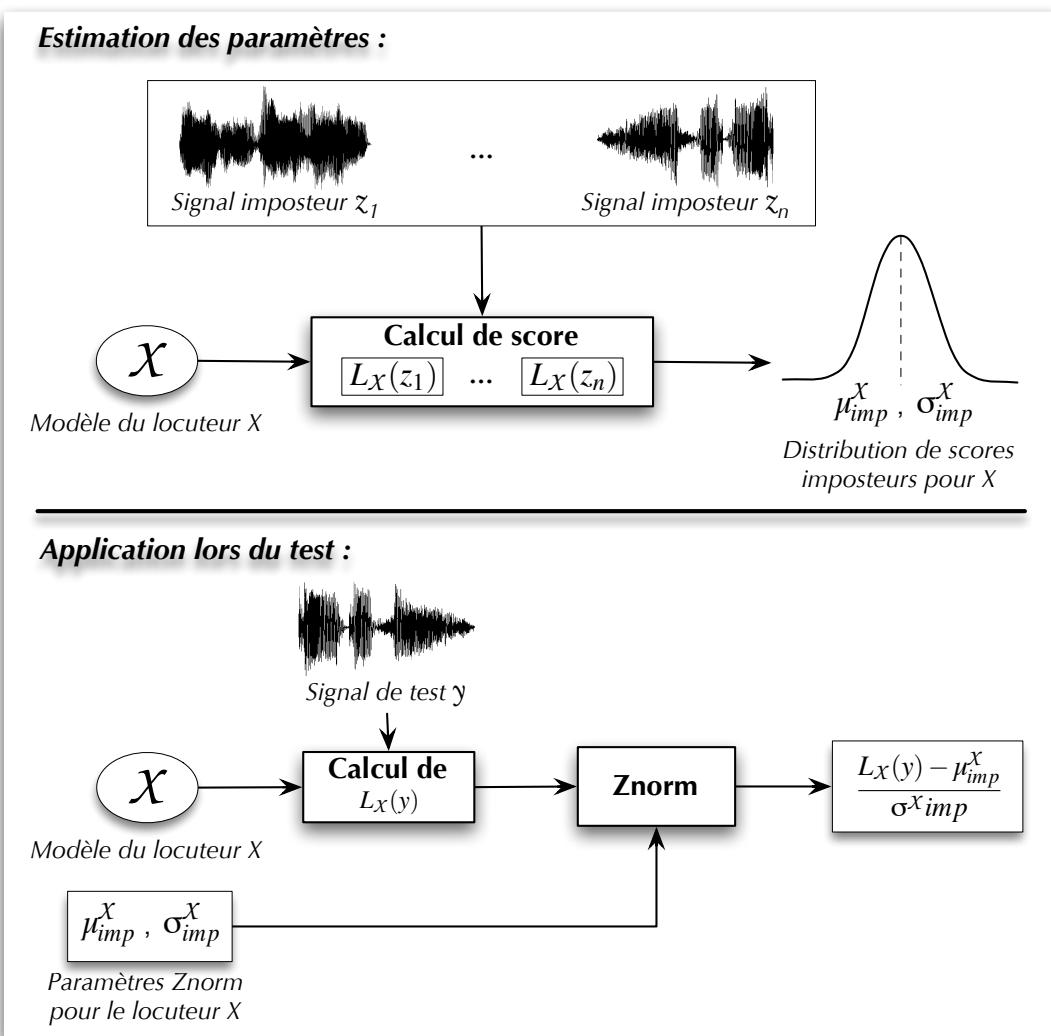


FIG. 2.8 – Illustration du principe de Znorm — Un jeu de paramètres $(\mu_{imp}^X, \sigma_{imp}^X)$ est calculé pour le locuteur client X après l'apprentissage de son modèle X en le confrontant à un ensemble de signaux imposteurs ; ces paramètres sont ensuite utilisés pour appliquer la normalisation lors du test d'un signal Y (d'identité inconnue) par rapport à X .

Comme le même signal de parole est utilisé à la fois pour le test et pour l'estimation des paramètres de normalisation, Tnorm évite un problème potentiel de Znorm posé par un écart trop important entre les signaux de test et de normalisation, reprenant en cela un avantage de la normalisation par cohorte. Contrairement à Znorm, une pré-normalisation par rapport de vraisemblances est inutile dans le cas de Tnorm qui en intègre les effets. En revanche, Tnorm hérite aussi de la normalisation par cohorte son principal inconvénient, à savoir que les paramètres de normalisation ne peuvent être estimés que durant la phase de test, pour chaque signal à tester.

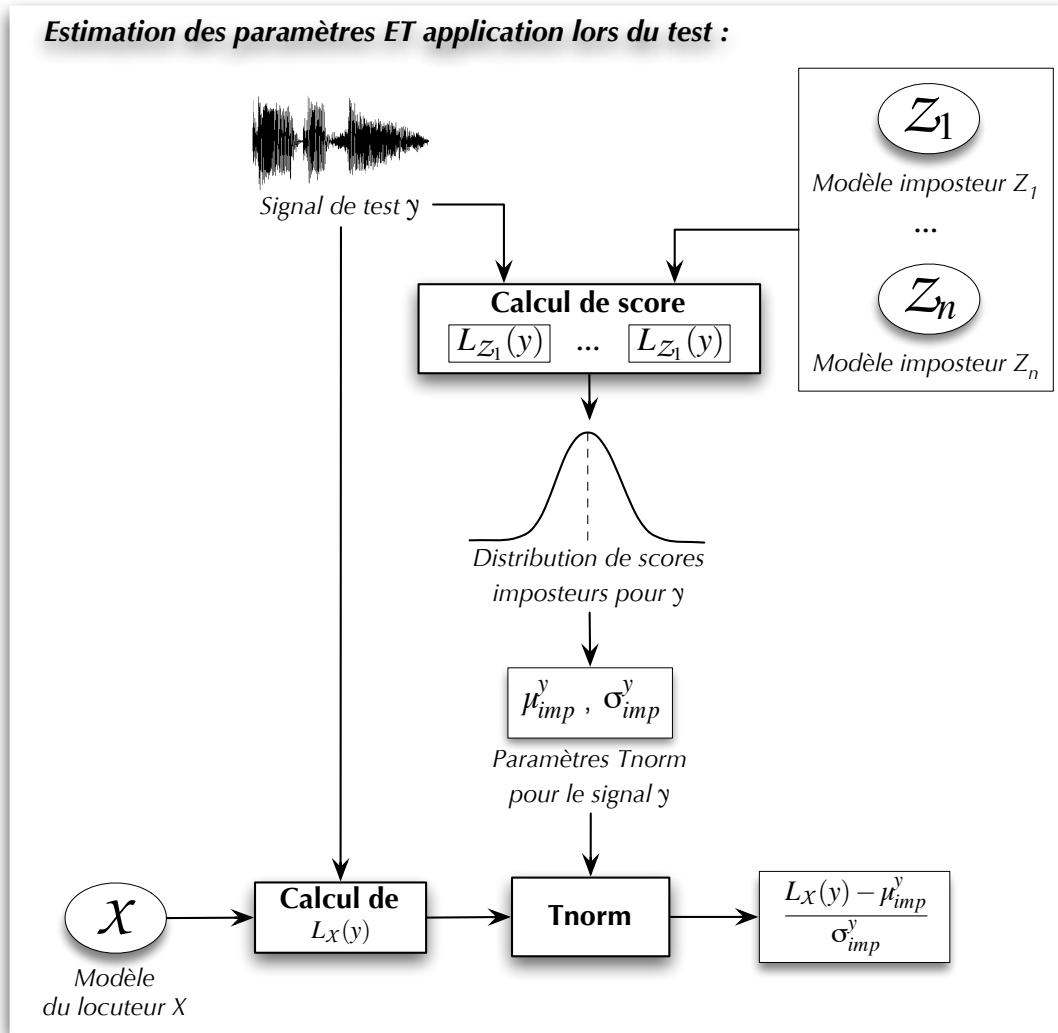


FIG. 2.9 – Illustration du principe de Tnorm — Un jeu de paramètres $(\mu_{imp}^y, \sigma_{imp}^y)$ est calculé, lors de la phase de test, pour chaque signal y à comparer au modèle X ; ces paramètres sont obtenus en comparant le signal y à un ensemble de modèles d'imposteurs (Z_1, \dots, Z_n) .

HTnorm Reposant sur le même principe que Hnorm, une variante de Tnorm a été proposée pour prendre en compte les effets de la différence de type de combiné lors de traitement de la parole téléphonique.

Dans le cas de HTnorm, les paramètres de normalisation (dépendants du type de combiné) sont estimés en confrontant chaque signal de test à un ensemble de modèles imposteurs correspondant à des signaux enregistrés par le même type de combiné que pour le modèle du locuteur testé.

C(T)norm Cnorm a été introduite pour étendre le principe de Hnorm à des enregistrements pour lesquels le type de combiné utilisé n'est pas spécifié (à l'origine, ce problème a été soulevé lors de l'introduction lors de la campagne d'évaluation NIST 2002 d'enregistrements provenant de téléphones cellulaires). Une classification en aveugle des données de normalisation est effectuée, chaque classe étant ensuite considérée comme correspondant à un type de combiné et utilisée de manière classique pour une normalisation Hnorm.

Bien sûr, une variante CTnorm est possible, utilisant le résultat de la classification pour calculer une normalisation Tnorm.

ZTnorm La combinaison de Tnorm et de Znorm est appelée ZTnorm. Effectuée en appliquant une normalisation Znorm à des scores préalablement normalisés par Tnorm, elle intègre des paramètres de normalisation dépendants du signal de test et des paramètres dépendants du modèle de locuteur.

Dnorm

Les techniques de normalisation de type Znorm, Tnorm et dérivés ont été développées pour une utilisation dans le cadre de la vérification du locuteur en mode indépendant du texte et elles y sont particulièrement bien adaptées. En effet, dans ce cadre, il est relativement aisé de trouver des données en quantité suffisante pour générer la distribution des scores pseudo-imposteurs, notamment à partir des données d'apprentissage des différents locuteurs. Cependant, dans d'autres cas, par exemple pour un système de vérification du locuteur par mot de passe (donc en mode dépendant du texte), les données disponibles sont en quantité largement inférieure (généralement l'ensemble des données disponibles est constitué d'un petit nombre de répétitions du mot de passe par chaque locuteur lors de l'apprentissage). L'estimation d'une distribution de scores imposteurs est de ce fait plus difficile.

La normalisation Dnorm, introduite dans [Ben 2002], s'attaque à ce problème de manque de données "imposteur" en générant des données à partir des modèles de locuteurs et du modèle du monde. Une méthode de Monte Carlo est utilisée pour créer un jeu de données "client" et un jeu de données "imposteur", respectivement à partir des modèles du locuteur considéré et du monde. Le score normalisé est donné par :

$$\tilde{L}_{\mathcal{X}}(y) = \frac{L_{\mathcal{X}}(y)}{KL2(\mathcal{X}, \bar{\mathcal{X}})} \quad (2.7)$$

où $KL2(\mathcal{X}, \bar{\mathcal{X}})$ est une estimation de la distance de Kullback-Leibler symétrisée entre

le modèle du locuteur et celui du monde. Cette estimation est faite à partir des données générées.

À l'instar de Znorm et Hnorm, Dnorm est appliquée en pratique sur des logarithmes de scores préalablement normalisés par rapport de vraisemblances, conduisant à une réécriture de l'équation 2.7 en :

$$\tilde{L}_{\mathcal{X}}(y) = \frac{\log(LR_{\mathcal{X}}(y))}{KL2(\mathcal{X}, \bar{\mathcal{X}})} \quad (2.8)$$

WMAP

Les techniques de normalisation décrites ci-dessus produisent des scores dans un espace non borné. De ce fait, l'interprétation (et par conséquent, le choix) de la valeur du seuil de décision est malaisée. La normalisation WMAP (ou "World+MAP"), présentée dans [Fredouille 1999], vise à apporter une solution à ce problème en combinant les avantages d'une normalisation par rapport de vraisemblances et de l'approche bayésienne pour projeter les scores dans un espace probabiliste. Une présentation plus détaillée du principe de cette normalisation est faite au chapitre 4, page 75.

Discussion

Des techniques de normalisation qui viennent d'être présentées, deux se détachent nettement dans l'état de l'art actuel : la normalisation par modèle du monde et Tnorm. La première est considérée comme le minimum acceptable, la normalisation à appliquer lorsque les ressources (en termes de quantité de données ou de ressources de calcul) nécessaires à d'autres techniques sont indisponibles ; le gain qu'elle apporte à la précision de la reconnaissance, par rapport à l'utilisation de scores non normalisés, est très important et pour cette raison elle sert de première étape à d'autres techniques de normalisation comme Znorm et ses dérivés. Tnorm, elle, est citée au titre de son efficacité. Dans le cadre des campagnes d'évaluation NIST en vérification du locuteur en mode indépendant du texte, Tnorm est la technique de normalisation apportant le meilleur gain de performances selon la métrique définie par NIST (sa variante ZTnorm apportant dans certains cas un gain supplémentaire).

Cependant, au delà de la réduction constatée du taux d'erreurs, le choix d'une technique de normalisation pour une application réelle de reconnaissance du locuteur est guidé (sinon dicté) en premier lieu par les ressources disponibles pour cette application. Les ressources nécessaires sont de deux types : tout d'abord les données utilisables pour la normalisation ; le problème posé par leur quantité a été présenté dans la section consacrée à Dnorm. Le second type de ressource à considérer est la puissance de calcul (ou le temps, ce qui revient au même). Pour une application d'identification ou de détection de locuteur, ce problème du temps de calcul est généralement considéré comme critique au cours de la phase de test, bien plus que lors de l'apprentissage d'un modèle de locuteur. Pour cette raison, des techniques comme la normalisation par modèle du monde et Znorm, qui placent la plus grosse partie des calculs (estimation du modèle du monde ou de la distribution des scores imposteurs) au niveau de la phase d'apprentissage, se révèlent bien plus avantageuses de ce point de vue que Tnorm, pour laquelle des calculs lourds doivent être effectués au moment du test. Pour cette raison, Tnorm n'est à l'heure actuelle que peu envisageable dans le cadre d'applications nécessitant une réponse en temps réel sur du matériel à faible puissance de calcul.

Notons enfin une particularité de la plupart des techniques de normalisation par distribution imposteur centrée/réduite présentées ici, de Znorm à Tnorm : l'accent est mis sur la normalisation des scores imposteurs, en utilisant une distribution de scores imposteurs et en ignorant la distribution des scores clients. Cette particularité peut être vue comme un effet pervers d'un développement guidé ces dernières années principalement par le cadre des campagnes d'évaluation NIST en vérification du locuteur. La stratégie d'évaluation retenue lors de ces campagnes se focalise sur les erreurs de fausse acceptation et confère une grande importance à la détection de tous les imposteurs. Une normalisation des scores favorisant cette détection est donc plus intéressant dans ce cadre, au risque de négliger l'erreur de faux rejet qui est pourtant un point important à considérer pour de nombreuses applications commerciales (le lecteur se reportera au chapitre suivant pour une présentation des campagnes d'évaluation NIST et plus particulièrement à la section 3.1.4 (page 53) pour une explication des termes "fausse acceptation" et "faux rejet").

Chapitre 3

Le contexte de travail

Sommaire

3.1 Les campagnes d'évaluation NIST	50
3.1.1 Le consortium ELISA	50
3.1.2 La tâche initiale : vérification automatique du locuteur	51
3.1.3 Évolution : les tâches multi-locuteurs	52
3.1.4 Méthodes d'évaluation	53
3.2 Autres applications	57
3.2.1 Le projet MTM	57
3.2.2 Le projet Certivox	59
3.2.3 La convention LIARMA	59
3.2.4 Le projet RAVOL	59

Le travail effectué dans le cadre de cette thèse a conduit au développement d'un système complet de reconnaissance du locuteur reposant sur l'état de l'art dans ce domaine. Le présent chapitre s'attache à exposer le contexte dans lequel ce travail a été réalisé, à travers la présentation des activités du LIA dans le domaine de la reconnaissance du locuteur.

Le système de RAL cité ci-dessus occupe une place centrale dans ces activités. Ce système sert de base pour le test et l'intégration de nouvelles techniques à différents niveaux du processus de reconnaissance.

La validation de ces techniques se fait à travers la participation annuelle à des campagnes internationales d'évaluation des systèmes de reconnaissances du locuteur. Cette participation est également l'occasion de développements en vue de l'adaptation du système aux nouvelles tâches de la RAL au fur et à mesure de leur intégration au sein de ces campagnes d'évaluation. C'est ainsi que l'effort de recherche du LIA, focalisé initialement sur les tâches d'identification et de vérification du locuteur, s'est étendu peu à peu pour couvrir l'ensemble des tâches multi-locuteurs présentées dans le chapitre précédent.

Enfin, en dehors du cadre strict de la recherche et de son évaluation par la participation à ces campagnes, d'autres développements prennent place dans le cadre de projets visant à intégrer les technologies de reconnaissance du locuteur dans des applications réelles. Chacun de ces projets applicatifs correspond, à travers l'ensemble de contraintes spécifiques qu'il définit, à une nouvelle problématique scientifique.

3.1 Les campagnes d'évaluation NIST

Depuis 1996, l'institut américain NIST (*National Institute of Standards and Technology*) organise chaque année une campagne internationale d'évaluation des systèmes de reconnaissance du locuteur ([Przybocki 1998], [Przybocki 1999], [Martin 2000], [Przybocki 2004]). La participation à cette campagne est ouverte à tous les laboratoires travaillant dans ce domaine, publics et privés.

L'objectif de cette évaluation est de tester et comparer les divers systèmes de reconnaissance du locuteur en les confrontant à des données et des problèmes aussi proches que possible des conditions réelles d'utilisation de la RAL. Ainsi les corpus de données utilisés sont composés d'enregistrements de parole spontanée en conditions réelles et les diverses tâches proposées reprennent la liste présentée à la section 2.2 (p. 20).

La définition exacte des tâches proposées à l'évaluation est précisée chaque année par NIST quelques semaines avant le début de la campagne. Cette définition et les corpus de données utilisés évoluent d'année en année, au gré de l'évolution des techniques et des pôles d'intérêt de la RAL, de l'apparition de nouveaux problèmes, des données disponibles, etc. Les premières années ont été consacrées intégralement au problème de la vérification du locuteur sur données téléphoniques; puis les tâches multi-locuteurs sont apparues, sous forme de détection multi-locuteurs et de suivi de locuteur d'abord, de segmentation en locuteurs enfin; dans le même temps le type de données utilisé est passé d'enregistrements téléphoniques standards à des enregistrements issus de téléphones cellulaires. Il faut cependant noter que l'orientation des tâches proposées lors de ces campagnes d'évaluation est évidemment influencé par les centres d'intérêt de leur source de financement, l'état américain. De plus, malgré la volonté très nette de NIST de rendre les conditions d'évaluation aussi proche que possible des conditions d'utilisation en applications réelles, il est impossible de recréer totalement la variété de ces applications et les développements actuels en matière du locuteur, conditionnés par les règles d'évaluation des campagnes NIST, sont susceptibles d'ignorer certaines directions de recherche n'apportant pas de gain immédiat dans ce cadre mais qui pourraient se révéler judicieuses dans d'autres conditions. Malgré tout, ces campagnes d'évaluation sont devenues un standard de fait, incontournable pour qui veut développer et évaluer les performances d'un système de reconnaissance du locuteur, et théâtre de la plupart des avancées dans le domaine de la reconnaissance du locuteur au cours de la dernière décennie.

3.1.1 Le consortium ELISA

Le LIA participe chaque année depuis 1998 aux campagnes d'évaluation NIST. Les phases successives de développement du système de reconnaissance du locuteur du LIA (AMIRAL) suivent en fait l'historique des campagnes NIST, notamment en ce qui concerne l'intégration des nouvelles tâches portant sur des documents multi-locuteurs.

Aux côtés de l'ENST (Paris) et l'IRISA (Rennes) le LIA a fondé en 1997 le consortium ELISA ([Gravier 1999], [ELISA 2000], [Magrin-Chagnolleau 2001]), auto-financé 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'évaluation internationales telles que les campagnes NIST. Sa composition est à géométrie variable, d'autres laboratoires s'étant joints à ELISA au cours des années, pour une durée plus ou moins longue. Le consortium organise des réunions

régulières, maintient une plateforme logicielle et fournit aide technique et scientifique à ses membres pour se présenter aux campagnes d'évaluation. La majorité des laboratoires du consortium participent conjointement aux évaluations NIST et présentent certains de leurs travaux dans des publications communes.

C'est dans le cadre d'ELISA que se sont inscrites toutes les participations du LIA aux campagnes d'évaluation NIST.

3.1.2 La tâche initiale : vérification automatique du locuteur

Initialement, les campagnes d'évaluation NIST proposaient uniquement une tâche de vérification du locuteur portant sur segments mono-locuteur (*one-speaker detection* ou *1sp* dans la terminologie NIST), telle que définie au chapitre 2, page 23. Le traitement de cette tâche, la seule disponible lors de la première participation du LIA, a orienté la définition de l'architecture de base du système de reconnaissance du locuteur du LIA, le système AMIRAL (cf. chapitre 4, page 63).

Cette tâche est restée au cours des années la tâche de référence des campagnes NIST, du fait que les méthodes développées pour la vérification du locuteur forment la base des méthodes utilisées pour les autres tâches. Sa définition a peu évolué, seul le corpus de données utilisé a changé. Des variantes de cette tâche ont cependant été ajoutées peu à peu, travaillant sur d'autres données et/ou étendant la définition de la tâche.

Tâche de référence Chaque signal de test est confronté à un ensemble de modèles de locuteurs, dont un seul est le locuteur du signal de test, les autres étant des pseudo-imposteurs (en ce sens que ni l'enregistrement de test ni les enregistrements d'apprentissage des modèles ne correspondent à de réelles tentatives d'impostures, mais simplement à des locuteurs différents — cf. chapitre 2, p. 31). Toutefois, la règle imposée lors de la participation à cette tâche est la stricte indépendance entre les tests. Chaque comparaison du signal de test à un modèle donné doit être réalisée sans intégrer aucune connaissance des autres locuteurs clients.

Le type de données utilisé correspond à des conversations téléphoniques à sujet non imposé (donc de la parole spontanée). Les deux parties des conversations sont séparées pour obtenir des segments de parole mono-locuteur, dont le silence est ensuite supprimé. Les mêmes types d'enregistrements sont utilisés pour les phases d'apprentissage et de test, seules les durées diffèrent. Les modèles des locuteurs clients sont appris sur environ 2 minutes de parole, issues de deux conversations différentes les premières années, puis d'une seule conversation à partir de 2000. Les signaux utilisés pour le test varient de quelques secondes à une minute. Les tests ne mixent pas les genres : tous les locuteurs auxquels est comparé un signal de test sont du même sexe que le locuteur ayant émis ce signal.

Le corpus de données utilisé a été Switchboard Phase I en 1996 et 1997, puis Switchboard Phase II (conversations sur téléphone filaire) jusqu'en 2001, avant d'être remplacé par Switchboard Cellular (conversations sur téléphone cellulaire) en 2002 (un petit corpus de données cellulaires ayant déjà été introduit lors de la campagne 2001).

Afin d'illustrer l'ampleur de la tâche, voici quelques chiffres de la campagne d'évaluation de 1999 : il y avait 539 locuteurs clients, répartis en 230 hommes et 309

femmes ; le nombre d'enregistrements à tester était de 3420, pour un total de 37620 tests dont 3157 étaient des tests clients et 34463 des tests pseudo-imposteurs. Malgré les changements de jeux de données d'une année à l'autre, les chiffres sont restés du même ordre de grandeur.

Variantes et extensions En 2000 et 2001, une tâche de vérification du locuteur utilisant des données du corpus AHUMADA (en Espagnol) a été proposée. Hormis le langage utilisé, la définition de cette tâche était similaire à celle de la tâche de référence.

Depuis 2001, une version étendue de la tâche de référence est proposée aux participants. Ayant pour objectif d'explorer les possibilités offertes par l'exploitation de caractéristiques du signal de parole plus complexes et de plus haut niveau que les paramètres cepstraux standards, cette tâche repose sur une quantité de données d'apprentissage beaucoup plus importante — jusqu'à une heure par locuteur, accompagnée de transcriptions automatiques et manuelles du contenu des enregistrements.

Enfin, en 2002, une tâche de vérification du locuteur simulant un environnement judiciaire a été introduite, reposant sur une base d'enregistrements du FBI. La composition du corpus permettait d'expérimenter la vérification du locuteur dans le cas où les données d'apprentissage sont enregistrées dans des environnements et avec des matériels très différents des données de test.

3.1.3 Évolution : les tâches multi-locuteurs

Au fil des années, les campagnes d'évaluation NIST se sont adaptées aux nouveaux axes de recherche se développant en reconnaissance du locuteur, notamment en complétant la tâche initiale de VAL par d'autres, portant sur des documents multi-locuteurs.

Détection de locuteur dans des documents bi-locuteurs La première tâche multi-locuteurs a été proposée en 1999 avec la détection de locuteur dans des documents bi-locuteurs (cf. chapitre 2, page 24). Cette tâche (appelée *two-speaker detection* par NIST, abrégée en 2sp) est très similaire dans sa définition à la vérification du locuteur et reprend les mêmes données, à ceci près que dans les enregistrements de tests les deux parties de la conversation ne sont plus séparées. Le système doit déterminer si un des deux participants (ou les deux) correspond à l'un (ou deux) des clients proposés. Cette tâche a été proposée sous cette forme lors des campagnes de 1999, 2000 et 2001.

Suivi de locuteur La seconde tâche multi-locuteurs apparue lors de la campagne 1999 est le suivi de locuteur, ou *speaker tracking* (cf. chapitre 2, page 24). Utilisant un sous-ensemble du jeu de données de la tâche *two-speaker detection*, cette tâche consiste à déterminer, pour un fichier de test donné (correspondant à une conversation à deux locuteurs), quels locuteurs clients parmi ceux proposés, participent à la conversation et délimiter leurs interventions. La tâche de suivi de locuteur a été proposée jusqu'en 2001.

Segmentation en locuteurs Depuis l'édition 2000, la segmentation en locuteurs (cf. chapitre 2, page 24) figure parmi les tâches proposées aux participants. Il s'agit d'une segmentation en aveugle, sans aucune connaissance préalable sur les locuteurs participant à la conversation. Deux sous-tâches ont été proposées jusqu'en 2002, différencierées par les données utilisées. La première partie du jeu de données utilisé est composée de conversations téléphoniques entre deux locuteurs, d'une durée d'une minute, issues du corpus utilisé pour la tâche *two-speaker detection*. L'autre partie du jeu de données regroupe des documents issus d'un autre corpus, comportant 1 à 10 locuteurs, pouvant durer jusqu'à 10 minutes et susceptibles d'être dans une autre langue que l'Anglais. Depuis 2003, la tâche de segmentation en locuteurs a été transférée de la campagne d'évaluation des systèmes de reconnaissance du locuteur (dite NIST SRE *Speaker Recognition Evaluation*) à la campagne NIST RT¹ (*Rich Transcription*) sous le nom de *speaker diarization*. Si le principe de segmentation en aveugle reste le même, le corpus de données sur lequel repose la tâche a évolué considérablement, incluant des enregistrements radio-diffusés comportant notamment des sections musicales.

Apprentissage à partir de documents bi-locuteurs Le dernier ajout dans la catégorie des tâches multi-locuteurs est l'apprentissage à partir de documents bi-locuteurs (textitcf. chapitre 2, page 25), apparue lors de la campagne d'évaluation de 2002. Il s'agit en fait d'une redéfinition de la tâche *two-speaker detection*. La phase de test reste définie telle qu'avant (détection du locuteur dans une conversation téléphonique à deux locuteurs d'une durée d'une minute). Mais l'apprentissage des modèles de locuteurs clients a été redéfini. Les données d'apprentissage pour un client donné se composent de 3 enregistrements de conversations complètes (1 minute, 2 locuteurs). Aucune connaissance n'est disponible sur les locuteurs présents dans ces enregistrements (ni leur identité, ni les moments et durées de leurs interventions dans la conversation). La seule information connue est que le client considéré se trouve parmi ces locuteurs, et qu'il est le seul locuteur à être présent dans les 3 conversations. La phase d'apprentissage de son modèle doit donc passer par une segmentation des 3 enregistrements, puis la détection du locuteur commun et le regroupement de ses interventions, dont les données sont finalement utilisées pour l'estimation du modèle.

3.1.4 Méthodes d'évaluation

Détection de locuteur

L'évaluation des systèmes de détection de locuteur repose sur l'analyse des deux types d'erreurs qu'ils peuvent commettre :

- l'erreur de fausse acceptation, qui se produit lorsqu'un signal testé est déclaré correspondre au modèle de locuteur considéré alors qu'il a été émis par un autre locuteur;
- l'erreur de faux rejet, qui se produit lorsqu'un test est déclaré négatif alors même que le signal testé correspond bien au locuteur considéré.

Il est à noter que le taux de fausse acceptation est également appelé taux de fausse alarme et que le taux de faux rejet est également appelé taux de non détection, ces appellations renvoyant à la théorie de la détection.

¹<http://www.nist.gov/speech/tests/rt/rt2003/index.htm>

Les deux taux d'erreurs correspondants sont notés FA et FR respectivement. Ils dépendent directement du choix du seuil de décision. Un système avec un seuil de décision bas aura tendance à accepter à tort de nombreux signaux de test ne correspondant pas réellement aux modèles auxquels ils sont comparés, générant ainsi un fort taux d'erreurs de fausse acceptation, mais un faible taux de faux rejet. À l'inverse, un seuil de décision élevé entraînera l'échec de nombreux tests, avec un taux de faux rejet élevé mais un faible taux de fausse acceptation. Le choix du seuil de décision revient à trouver un compromis entre les deux taux d'erreurs. Les valeurs de FA et de FR alors obtenues définissent le *point de fonctionnement* du système.

Les taux FA et FR étant tous deux fonctions du seuil de décision, il est possible d'exprimer l'un en fonction de l'autre (fonction qui est alors monotone et décroissante). La courbe correspondante est généralement tracée en utilisant une échelle garantissant une courbe linéaire de pente -1 si les distributions de scores clients et imposteurs sont toutes deux gaussiennes et de même variance, et appelée courbe DET (*Detection Error Tradeoff* — [Martin 1997]). Une telle courbe permet d'avoir un aperçu des performances d'un système de détection dans diverses conditions d'utilisation. Les axes représentant les taux d'erreurs, un meilleur système obtient une courbe plus proche de l'origine. Ces caractéristiques font de la courbe DET l'outil privilégié d'évaluation, et surtout de comparaison, des performances des systèmes de détection de locuteur. Des courbes DET sont systématiquement utilisées pour présenter les résultats de tels systèmes lors des campagnes d'évaluation NIST. La figure 3.1 montre un exemple de courbe DET.

Cependant, au delà du fonctionnement global d'un système que présente la courbe DET, il est également intéressant d'avoir une mesure des performances de ce système à son point de fonctionnement (une fois le seuil de décision fixé), pour une application précise. Cette mesure est réalisée par l'utilisation d'une fonction de coût (DCF — *Detection Cost Function*) qui prend en compte les deux taux d'erreurs et le coût qui leur est associé dans le cadre de l'application visée (les applications à visée sécuritaire donnant par exemple un coût très élevé à l'erreur de fausse acceptation) :

$$C = C_{FA} \times FA + C_{FR} \times FR \quad (3.1)$$

L'utilisation de cette fonction permet de représenter par un simple nombre les performances du système pour application donnée. Le minimum de la fonction de coût est atteint pour un seuil de décision réglé correctement pour l'application visée. La comparaison de ce minimum à la valeur obtenue au point de fonctionnement réel permet dès lors d'évaluer la qualité du choix du seuil de décision.

Le classement des systèmes de détection du locuteur lors des campagnes d'évaluation NIST se fait par rapport à la valeur obtenue au point de fonctionnement pour une fonction de coût qui définit C_{FA} et C_{FR} de manière un peu plus fine. Chacune de ces deux valeurs est vue comme le produit du coût de l'erreur considérée et de la probabilité *a priori* d'être en présence d'un test susceptible de générer ce type d'erreur :

$$\begin{aligned} C_{FA} &= C_{FA}^{NIST} \times P_{imposteur}, \\ C_{FR} &= C_{FR}^{NIST} \times P_{client} \end{aligned} \quad (3.2)$$

où $P_{imposteur}$ est la probabilité *a priori* d'être un présence d'un test de type imposteur et P_{client} celle d'être en présence d'un test de type client. Les valeurs typiquement utilisées au cours des années sont $C_{FA}^{NIST} = 1$, $C_{FR}^{NIST} = 10$, $P_{imposteur} = 0,99$ et $P_{client} = 0,01$, soit des valeurs pour C_{FA} et C_{FR} de 0,99 et 1 respectivement.

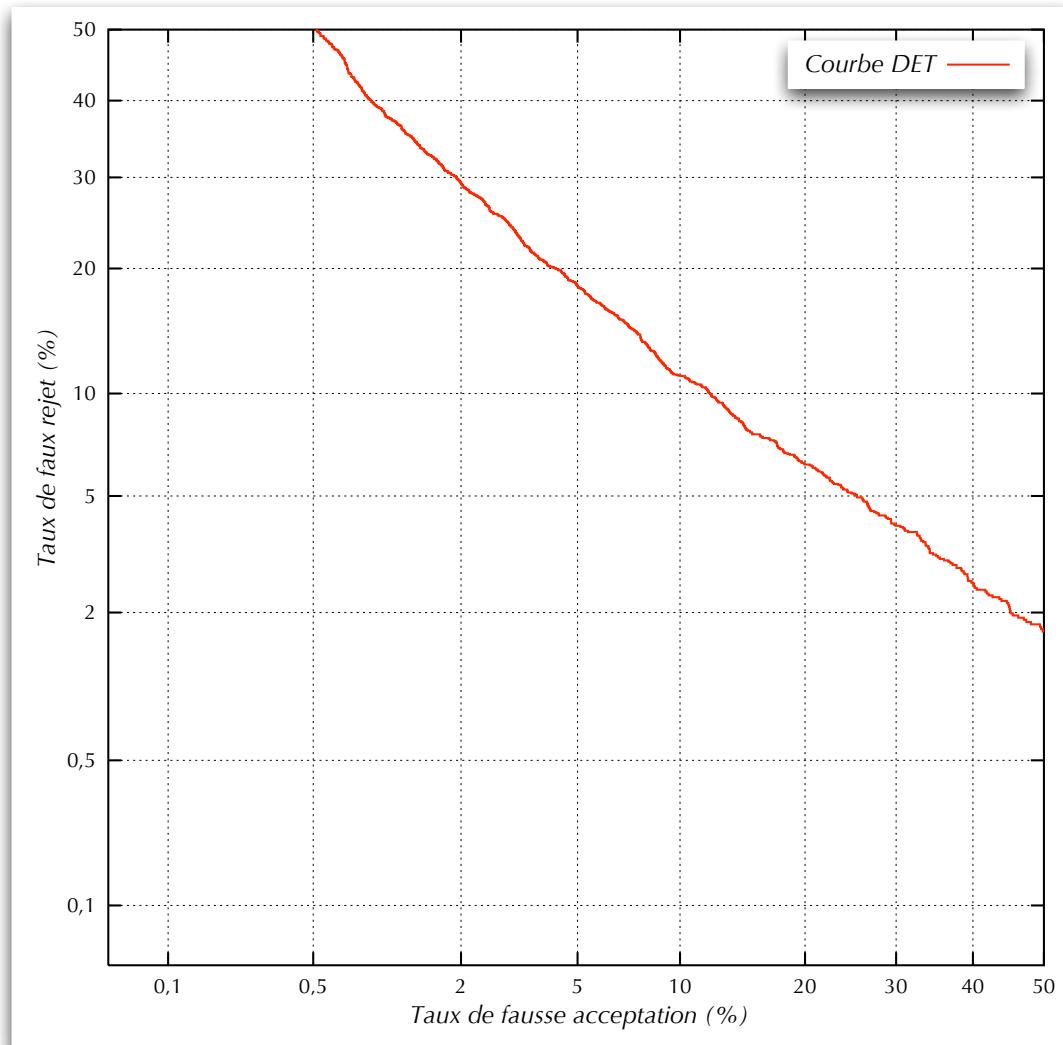


FIG. 3.1 – Exemple d'une courbe DET

Une fonction de coût particulière régulièrement utilisée est la demi erreur totale (HTER — *Half Total Error Rate*), qui est la moyenne arithmétique de FA et FR (soit $C_{FA} = C_{FR} = 0,5$).

Enfin, une autre mesure fréquemment utilisée pour évaluer les performances d'un système est le taux d'égale erreur (EER — *Equal Error Rate*), qui correspond au point pour lequel $FA = FR$. Il peut être trouvé très facilement sur une courbe DET par l'intersection avec la première diagonale. L'EER permet d'exprimer facilement la capacité d'un système de détection à séparer les clients des imposteurs.

Suivi de locuteur et segmentation en locuteur

L'évaluation des performances des systèmes de segmentation en locuteur (ou de suivi de locuteur, qui est à considérer comme un problème de segmentation simplifié) est une tâche plus compliquée que dans le cas des systèmes de détection de locuteur. Après avoir utilisé de manière peu convaincante en 1999 des outils basés sur les méthodes d'évaluation de la vérification du locuteur pour évaluer les résultats en suivi de locuteur, NIST propose depuis 2000 (année d'apparition de la segmentation en locuteur dans les campagnes d'évaluation) une méthode d'évaluation des résultats adaptée aux systèmes de suivi et de segmentation ([NIST 2000], [NIST 2001]).

La première étape consiste à produire, pour chaque enregistrement à segmenter, une segmentation de référence à laquelle comparer les soumissions des participants. Cette segmentation de référence est établie par une méthode semi-automatique ou manuellement, selon les types d'enregistrements disponibles. Lors de la production de la référence, il est difficile de fixer avec précision les frontières entre les segments. Le positionnement de ces frontières a fait l'objet de nombreuses discussions dans la communauté et il est généralement admis de tolérer un décalage de plus ou moins 0,25 seconde (soit une demie seconde au total) des frontières lors du calcul de l'erreur de classification.

L'évaluation de la segmentation produite par un système repose sur la détection de deux types d'erreurs : les erreurs de détection des zones de parole et de non-parole, et les erreurs d'affectation des zones de parole aux divers locuteurs composant le document. Il est à noter que la détection du premier type d'erreurs n'a été prise en compte pour les campagnes d'évaluation NIST qu'à partir de 2001.

Le calcul du taux d'erreurs de détection parole/non-parole est simple dans la mesure où les deux classes considérées sont exclusives : il n'y a jamais de recouvrement. La durée des portions du document mal étiquetées vis-à-vis de ces classes (les portions de parole étiquetées non-parole ou l'inverse) est calculée puis divisée par la durée totale du document pour fournir le taux d'erreurs.

L'évaluation du second taux d'erreurs (la mauvaise affectation des segments aux locuteurs) est plus complexe. Tout d'abord, dans le cadre de la tâche de segmentation, les soumissions sont envoyées en étiquetant les locuteurs détectés avec des numéros anonymes. Il est donc nécessaire de commencer par faire correspondre ces numéros aux étiquetages des locuteurs dans la segmentation de référence (étape qui n'est bien entendu pas réalisé dans le cas du suivi de locuteur). Cette association est réalisée en recherchant les paires de locuteurs (un locuteur défini dans la segmentation de référence et un locuteur défini dans la segmentation générée par le système) qui, parmi toutes les paires de locuteurs possibles, minimisent l'erreur d'affectation. Une fois cette cor-

respondance établie, le taux d'erreurs d'affectation est calculé là aussi en additionnant les durées des zones mal étiquetées et en divisant ce résultat par la durée totale des zones de parole du document. Cependant, cette durée totale peut se révéler supérieure à la durée effective du document, car les zones où plusieurs locuteurs s'expriment simultanément sont comptés autant de fois qu'il y a de locuteurs.

Il convient de noter que cette évaluation du taux d'erreurs d'affectation en locuteurs intègre indirectement l'évaluation du nombre de locuteurs et des frontières : les erreurs sur le nombre de locuteurs ou sur les ruptures génèrent des erreurs d'affectation aux locuteurs.

3.2 Autres applications

En marge de la participation régulière aux campagnes d'évaluation NIST, le moteur de reconnaissance du locuteur du LIA est mis à contribution pour des projets intégrant la mise en œuvre d'applications réelles de reconnaissance du locuteur. Parmi ces projets, les deux principaux ont été le projet MTM et la collaboration LIARMA, sources de financement de ce travail de thèse.

Ces diverses applications présentent des contraintes différentes de celles rencontrées lors des évaluations NIST. La souplesse gagnée en s'affranchissant des règles strictes des campagnes d'évaluation est compensée par les contraintes spécifiques définies pour chacune par le cadre d'utilisation prévu pour la reconnaissance du locuteur. Cet ensemble de contraintes contribue à définir pour chaque application une problématique scientifique propre, différente de celle abordée à travers les travaux de recherche.

3.2.1 Le projet MTM

Le projet européen MTM (*Multimedia Terminal Mobile*) du programme IST², réalisé sur une période de 2 ans, consistait à définir un nouveau terminal multimedia de type PDA. Ce projet comprenait notamment l'intégration de la reconnaissance du locuteur à l'interface utilisateur.

Le travail à réaliser par le LIA pour la partie RAL comportait deux aspects : offrir aux développeurs d'applications des fonctions d'accès de haut niveau au moteur de reconnaissance du locuteur et adapter ce dernier au fonctionnement sur un système embarqué. Le choix de l'utilisation effective de cette technologie restait à la discréption des développeurs d'applications.

L'utilisation de la reconnaissance du locuteur définie pour ce projet correspondait plus à un souci de confort que de sécurité : sélection de profil d'utilisateur, reconnaissance simultanée de la parole et du locuteur, etc.

Une présentation détaillée du projet MTM et du travail réalisé pour le mener à bien est faite au chapitre 7, page 122.

²Projet IST-1999-11100 — cf. http://dbs.cordis.lu/fep-cgi/srchidadb?ACTION=D&CALLER=PROJ_IST&QF_EP_RPG=IST-1999-11100

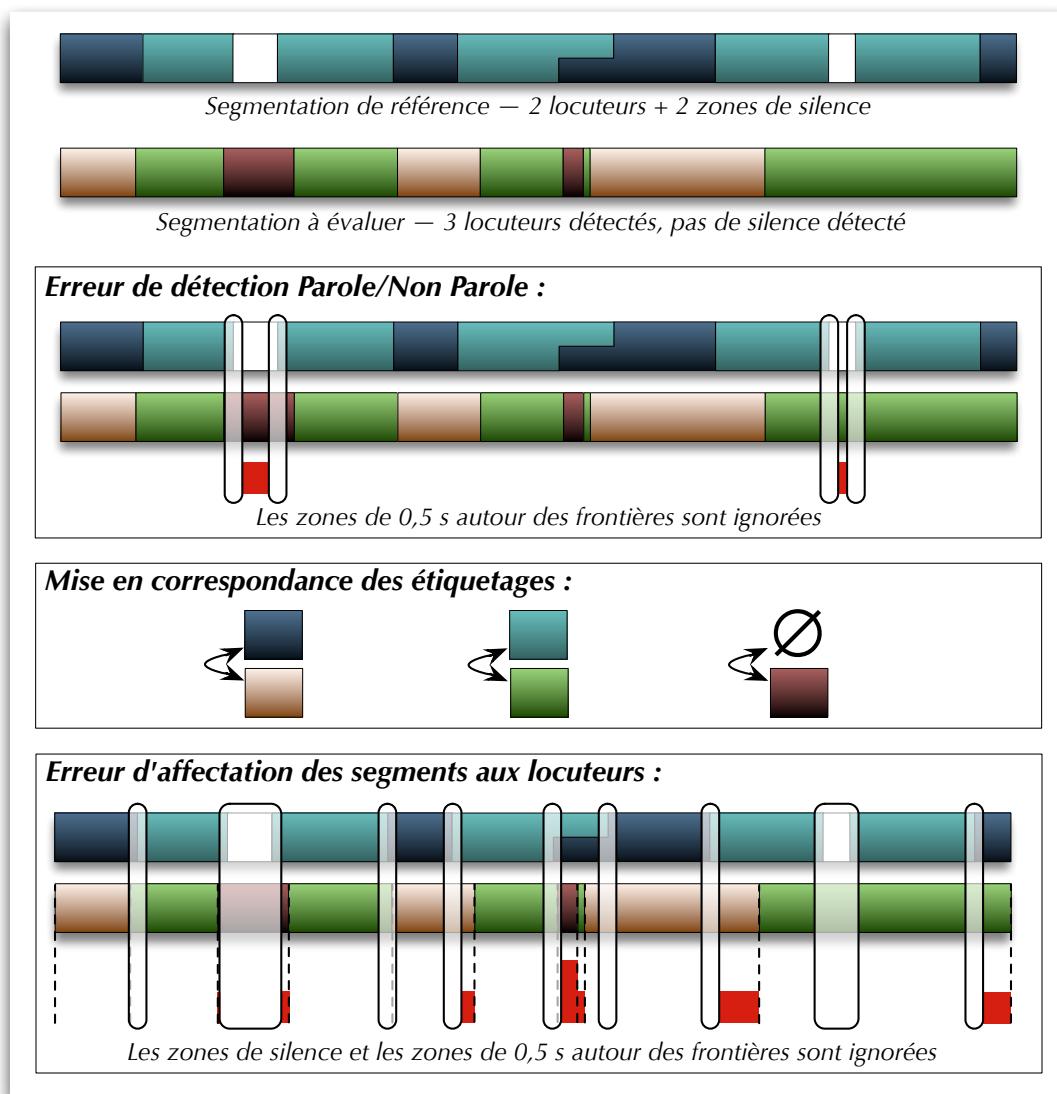


FIG. 3.2 – Illustration sur un exemple fictif du calcul des deux taux d'erreurs (mauvaise détection parole/non parole et mauvaise affectation de segments aux locuteurs) pour la tâche de segmentation en locuteurs.

3.2.2 Le projet Certivox

Le projet Certivox, au contraire du projet MTM, définissait une utilisation de la reconnaissance du locuteur orientée vers la sécurité. Ce projet proposait en effet de réaliser une authentification de l'utilisateur par téléphone dans le cadre de paiements électroniques à distance.

Le niveau de performance requis est élevé (en privilégiant la qualité du rejet des imposteurs) malgré les courtes durées de test et le manque de données disponibles, notamment pour créer un modèle du monde adapté.

3.2.3 La convention LIARMA

Le projet LIARMA s'inscrit dans le cadre d'une convention passée entre l'École Royale Militaire de Bruxelles et le LIA. Ce projet consistait à développer un prototype de reconnaiseur multimodal combinant la reconnaissance du visage et la reconnaissance du locuteur destiné à des applications de sécurisation d'accès par authentification biométrique.

Une particularité de l'étude portait sur la définition de la plateforme matérielle, à base d'ordinateur personnel standard et de composants peu coûteux pour l'acquisition des données biométriques. Un démonstrateur réalisé sur une plateforme de ce type a été livré à la fin du projet.

Plus de détails concernant les motivations de ce projet et les réalisations associées sont donnés au chapitre 7, page 133.

3.2.4 Le projet RAVOL

Le projet de Reconnaissance de l'Appelant au Vol (RAVOL) était un projet liant la région PACA, la société DigiFrance et le LIA. Il visait à mettre au point un dispositif assurant toutes les tâches de reconnaissance du locuteur (identification/vérification et indexation) dans le cadre d'applications "en ligne". Plus particulièrement, le système devait permettre de retrouver ou confirmer l'identité d'un appelant à partir de sa voix et cela indépendamment du lieu et du média d'appel.

Les travaux effectués au sein du LIA, en collaboration avec l'entreprise DigiFrance, ont porté sur la mise en place d'un démonstrateur permettant l'authentification de l'appelant, dans le cadre du couplage téléphonie/informatique (CTI). La mise en place d'un prototype a permis la première évaluation du système de reconnaissance automatique du locuteur du LIA hors du contexte des campagnes d'évaluations NIST.

Ce projet a soulevé des problèmes liés à la quantité et à la qualité d'informations disponibles. En particulier, l'enrôlement des locuteurs à partir d'enregistrements de courtes durées et les problèmes de normalisation des résultats (acceptation ou rejet des locuteurs) ont été abordés.

Deuxième partie

Travail réalisé

Chapitre 4

“Cœur” du système AMIRAL

Sommaire

4.1 Architecture multi-reconnaisseurs segmentale	64
4.2 Paramétrisation	68
4.3 Traitements post-paramétrisation	69
4.3.1 Suppression des trames de basse énergie	69
4.3.2 Normalisation des vecteurs de paramètres acoustiques	71
4.4 Normalisation des scores	74
4.4.1 Rapport de vraisemblances	74
4.4.2 Normalisation WMAP	75
4.4.3 Autres techniques de normalisation des scores	80
4.5 Modélisation des locuteurs	80
4.5.1 Structure des modèles	81
4.5.2 Estimation des modèles de locuteurs	82
4.6 Modèle du monde	86
4.6.1 Triple intervention	86
4.6.2 Données d'apprentissage	86
4.6.3 Estimation	90

Le système AMIRAL tire son nom de l'acronyme de : “Architecture Multi-reconnaisseurs pour l'Indexation et la Reconnaissance Automatique du Locuteur”. Ce nom reflète la nature d'AMIRAL : un système de reconnaissance automatique du locuteur conçu pour traiter les tâches mono et multi-locuteurs du domaine (de l'identification et la vérification du locuteur à la segmentation en locuteurs). Il fait apparaître également une caractéristique majeure du système : sa structure est conçue pour permettre l'utilisation de plusieurs reconnaiseurs travaillant en parallèle sur la même tâche.

La section 4.1 s'attache à décrire l'architecture générale du système, qui découle en grande partie de cet aspect multi-reconnaisseurs. Les sections suivantes détaillent chaque composant et les choix techniques qui ont été faits.

4.1 Architecture multi-reconnaisseurs segmentale

L'architecture multi-reconnaisseurs du système AMIRAL trouve son origine dans l'étude menée par Laurent Besacier sur l'information utile pour la reconnaissance du locuteur portée par diverses sous-bandes de fréquences ([Besacier 2000a], [Besacier 1998]). Ce travail définissait le besoin de plusieurs moteurs de reconnaissance, différent par la sous-bande de fréquences traitée par chacun, appliqués en parallèle sur les mêmes données, et dont les résultats étaient fusionnés par la suite.

Ce principe a été conservé et généralisé pour définir la structure du système AMIRAL, présentée dans [Fredouille 2000b] et illustrée par la figure 4.1. Comme évoqué au chapitre 2, diverses sources d'informations caractéristiques du locuteur peuvent être distinguées dans le signal de parole, telles que le spectre à court ou à long terme, la prosodie, les phénomènes coarticulatoires, le contenu linguistique, etc. Les méthodes d'extraction et de traitement de l'information diffèrent d'une catégorie d'informations à l'autre. Afin d'être en mesure de prendre en compte simultanément plusieurs types d'informations pour la reconnaissance du locuteur, l'architecture proposée ici permet de définir autant de reconnaiseurs que nécessaire, traitant en parallèle le même flux de données.

Chaque reconnaisseur a la possibilité d'intégrer sa propre technique de paramétrisation ou peut partager la paramétrisation utilisée par d'autres reconnaiseurs, sans redondance des traitements ni des résultats de la paramétrisation. Ainsi plusieurs reconnaiseurs peuvent traiter les mêmes vecteurs de paramètres, ne différant que par la suite du processus : modélisation, calcul et normalisation de scores. Chacun de ces modules peut également être spécifique à un reconnaisseur ou être partagé par plusieurs.

La seule contrainte imposée aux reconnaiseurs porte sur la fréquence de production de scores. En effet, la difficulté majeure d'une architecture multi-reconnaisseurs est la fusion des scores en sortie des reconnaiseurs, du fait de la possible hétérogénéité de ces scores ainsi que de la nature des informations traitées par les divers reconnaiseurs. De par cette nature, les reconnaiseurs peuvent nécessiter plus ou moins de données, situées à différents moments du signal de parole, pour produire des scores significatifs. Dans ce contexte, un système forçant une production de scores trame par trame pour chaque reconnaisseur peut introduire une redondance importante dans les sorties de certains reconnaiseurs tout en obligeant d'autres reconnaiseurs à produire des scores non significatifs. Une solution à ce problème passe par une approche segmentale, dans laquelle chaque reconnaisseur génère un score uniquement à la fin d'un segment temporel comportant de l'information utile pour lui. La redondance des scores se trouve ainsi éliminée et la robustesse renforcée par la suppression des zones non porteuses d'information significative. De plus, la quantité de valeurs à fusionner s'en trouve réduite.

Cependant une telle approche implique d'être capable de détecter les zones informatives pour chaque reconnaisseur et complexifie la fusion des scores inter-reconnaisseurs en introduisant une asynchronie entre leurs sorties respectives. Pour cette raison, une approche plus simple est proposée ici, constituant un compromis entre performance et complexité. Le signal de parole est divisé en blocs temporels de longueur fixe (d'où le nom retenu d'"approche bloc-segmentale"), identiques pour tous les reconnaiseurs. Une telle segmentation arbitraire règle le problème de la synchronisation inter-reconnaisseurs, tout en apportant tout de même une réduction du nombre

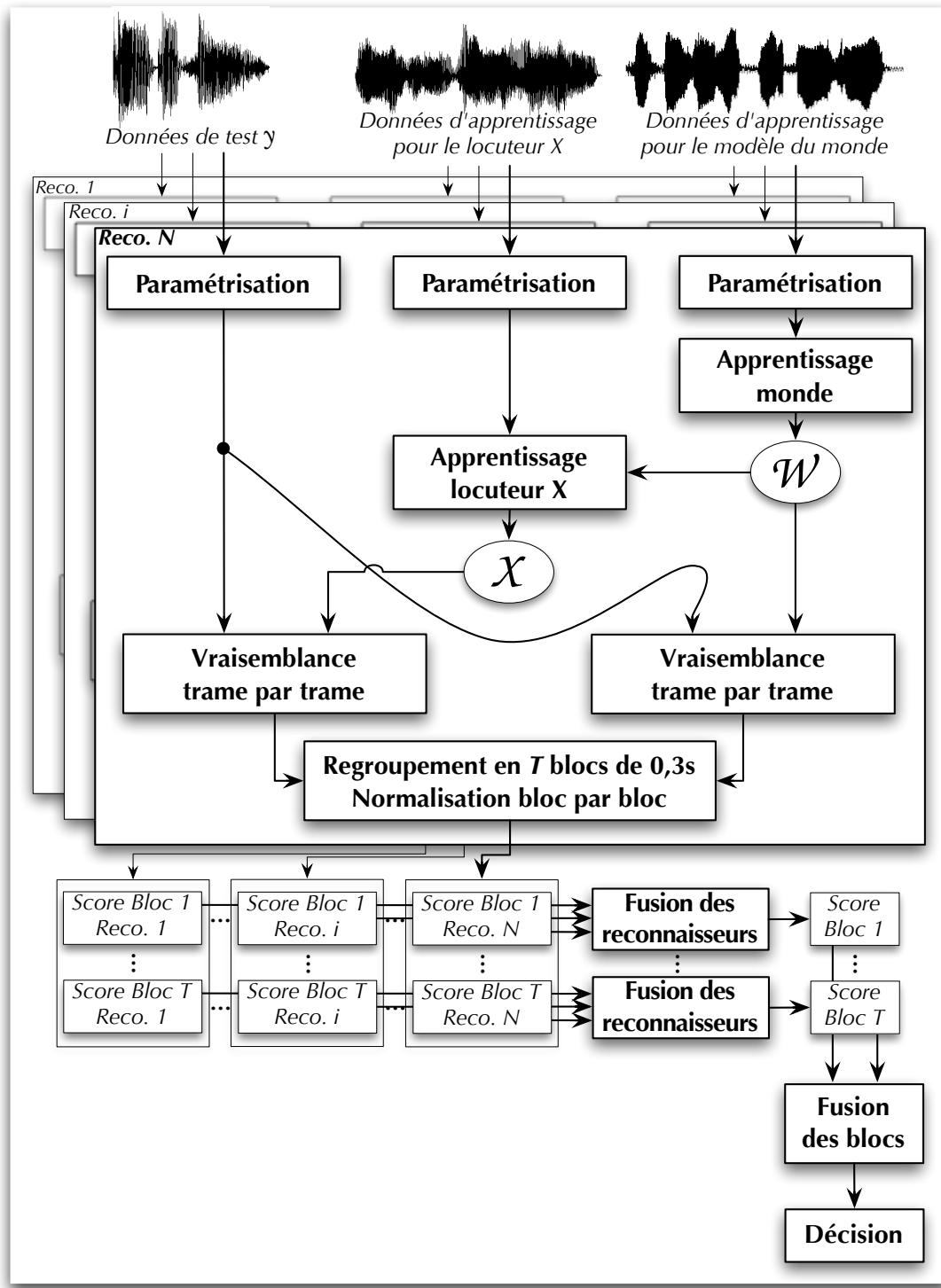


FIG. 4.1 – Illustration de l'architecture multi-reconnaisseurs segmentale du système AMIRAL.

de scores à fusionner et en préservant la possibilité de supprimer des zones non informatives. De plus, la longueur fixe des blocs simplifie la normalisation des scores d'un bloc à l'autre pour un reconnaisseur donné.

Un bloc B_i tel que décrit dans cette approche est défini comme l'ensemble des T trames y_t du signal y pour t tel que $i \times T < t \leq (i+1) \times T$. Le score $S_n(B_i|\mathcal{X}_n)$ produit pour ce bloc par le reconnaisseur n ($n \in \{1, \dots, N\}$) en le comparant au modèle de locuteur \mathcal{X}_n est donné par :

$$S_n(B_i|\mathcal{X}_n) = f_n(y_{i \times T+1}, \dots, y_{i \times (T+1)}, \mathcal{X}_n) \quad (4.1)$$

où la fonction f_n , produisant le score normalisé d'un bloc, est dépendante du reconnaisseur n .

Le choix de la longueur T des blocs est évidemment dépendant de la nature des fonctions f_n définies par les divers reconnaiseurs. Cependant d'autres facteurs doivent aussi être pris en compte. Des blocs plus larges signifient moins de scores à fusionner sur l'ensemble du signal et surtout des scores de blocs plus fiables car estimés sur plus de données. À l'inverse, des blocs plus courts permettent une détection plus précise des zones non informatives et une précision plus élevée lors de la détection des changements de locuteur dans le cadre des tâches multi-locuteurs.

Lors de la mise en œuvre de cette approche, outre la longueur des blocs, un autre paramètre à fixer est la stratégie à appliquer pour la fusion des scores des blocs. Cette étape peut reposer sur toute technique classique de fusion, telle une moyenne arithmétique ou géométrique, le choix des N meilleurs ou des N pires scores ou encore un vote majoritaire. L'ordre d'application de la fusion doit également être fixé. L'architecture présentée autorise le choix de n'importe quel ordre, allant de la fusion des scores inter-reconnaiseurs pour chaque bloc suivie de la fusion des scores de tous les blocs, à la solution inverse (fusion temporelle pour chaque reconnaisseur, puis fusion des scores inter-reconnaiseurs), ou toute combinaison intermédiaire.

Mise en œuvre En pratique, l'architecture multi-reconnaiseurs bloc-segmentale décrite ici n'a été jusqu'à présent mise en œuvre qu'avec des reconnaiseurs reposant sur l'utilisation de GMM¹ pour la modélisation des locuteurs, la normalisation des scores étant assurée par l'une des techniques présentées au chapitre 2 (page 41). C'est un reconnaisseur de ce type, utilisant une normalisation des scores par modèle du monde, qui est illustré dans le bloc central de la figure 4.1 (la description des divers modules d'AMIRAL dans les sections suivantes portera sur ce type de reconnaisseur).

Dans ce cadre, la fonction f_n retenue pour chaque reconnaisseur est une moyenne géométrique des scores obtenus trame par trame, sur laquelle est appliquée la normalisation :

$$f_n(y_{i \times T+1}, \dots, y_{i \times (T+1)}, \mathcal{X}_n) = \text{norm} \left(\prod_{j=1}^T L_{\mathcal{X}_n}(y_{i \times T+j}) \right)^{1/T} \quad (4.2)$$

où norm est la fonction de normalisation retenue. La valeur de T est fixée à 30 trames.

Dans le cas de l'utilisation de la normalisation WMAP (présentée en section 4.4.2,

¹Les reconnaiseurs utilisés reposaient selon les cas sur des GMM à matrices de covariance diagonales ou à matrices pleines, ou encore un mélange des deux types grâce à l'architecture décrite ici.

p. 75), cette fonction s'écrit :

$$f_n(y_{i \times T+1}, \dots, y_{i \times (T+1)}, \mathcal{X}_n) = P_n \left(\left(\prod_{j=1}^T \frac{L_{\mathcal{X}_n}(y_{i \times T+j})}{L_{W_n}(y_{i \times T+j})} \right)^{1/T} \middle| \mathcal{X}_n \right) \quad (4.3)$$

où W_n est le modèle du monde correspondant au reconnaiseur n et f_n est la fonction de normalisation WMAP propre à ce reconnaiseur, donnant la probabilité *a posteriori* que le score obtenu par le bloc B_i par rapport au modèle \mathcal{X}_n soit un score de type client. Cette sortie sous forme de probabilité, *vía* une fonction de normalisation prenant en compte les caractéristiques du reconnaiseur considéré, est une particularité qui fait de WMAP une normalisation parfaitement adaptée à l'architecture décrite ici. En effet, la fusion des scores entre reconnaiseurs et le long de l'axe temporel devient triviale, une simple moyenne étant suffisante.

Description de la structure des trames Les systèmes multi-reconnaiseurs qui ont été mis en œuvre grâce à l'architecture présentée ici reposent sur deux types de variations entre les divers reconnaiseurs : d'une part, des reconnaiseurs traitant différentes sous-bandes de fréquences, d'autre part des reconnaiseurs intégrant (ou non) la dynamique à court terme du signal par concaténation des coefficients cepstraux statiques ([Fredouille 2000b]).

La réalisation de ces différents reconnaiseurs est grandement facilitée par une particularité technique d'AMIRAL : la possibilité de réutiliser pour toutes ces variantes les mêmes vecteurs de paramètres, en décrivant la structure des trames en termes de coefficients à utiliser et d'intervalle entre deux trames successives. La structure et la composition des trames utilisées par les divers reconnaiseurs deviennent dès lors indépendantes de celles des vecteurs générés lors de la paramétrisation du signal.

La définition de la structure des trames repose sur trois paramètres : le nombre de vecteurs de paramètres que recouvre une trame, les indices des coefficients de ces vecteurs qui sont présents dans la trame, ainsi que le décalage (en nombre de vecteurs) entre deux trames successives. La figure 4.2 illustre par un exemple d'utilisation l'effet de ces paramètres.

Cette approche permet de nombreuses combinaisons qui autorisent, à partir du même jeu de vecteurs de paramètres calculé une seule fois, de tester plusieurs variantes. La première possibilité offerte est la sélection d'un sous-ensemble de coefficients à utiliser : coefficients cepstraux et/ou leur(s) dérivée(s), énergie du signal, ou toute combinaison. Il est également possible, tel qu'évoqué précédemment, d'opérer ainsi une concaténation de vecteurs pour réaliser des trames complexes intégrant la dynamique du signal. De plus, cette définition de la structure des trames rend triviale l'application de la technique de la décimation de trames, qui consiste à n'utiliser, lors de la phase de test, qu'une trame sur N afin d'accélérer les calculs ; d'après [McLaughlin 1999], la dégradation des performances est minime pour N valant 5 ou 10, dans le cadre de la tâche "one-speaker detection" des campagnes d'évaluation NIST. Enfin, toute combinaison des solutions citées ci-dessus est envisageable.

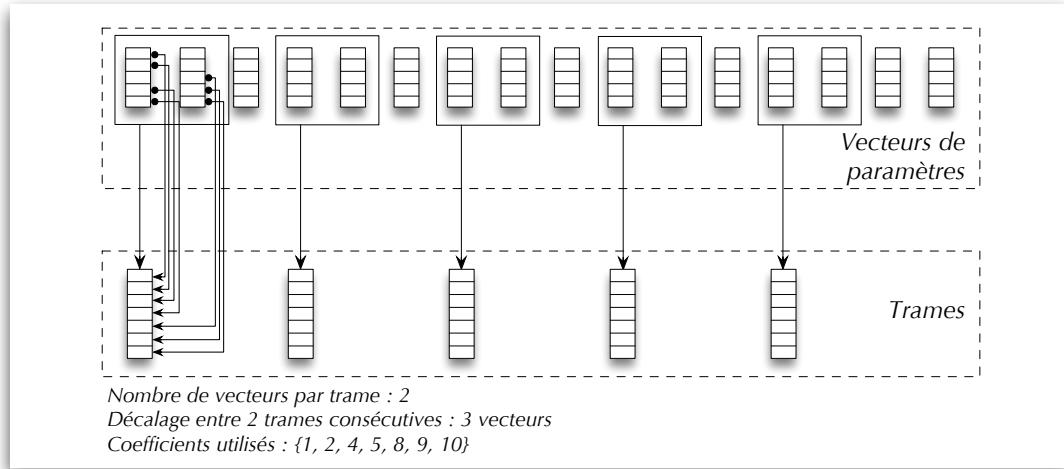


FIG. 4.2 – *Définition de la structure des trames — Illustration par un exemple de l’effet des trois paramètres (nombre de vecteurs couverts par une trame, indices des coefficients de ces vecteurs présents dans la trame et décalage entre deux trames successives).*

4.2 Paramétrisation

AMIRAL n’intègre pas directement de module de paramétrisation acoustique et a recours à un outil externe pour cela. La structure interne du système permet un fonctionnement indépendant du choix de l’outil de paramétrisation. Dans le cadre des campagnes d’évaluation NIST et de la collaboration au sein du consortium ELISA, il s’agit du module SPRO, développé par Guillaume Gravier (IRISA) et inclus dans la plate-forme commune de développement du consortium (cf. chapitre 3, page 50).

La paramétrisation retenue pour AMIRAL repose sur des coefficients de type LFCC (coefficients cepstraux issus d’une analyse en banc de filtres à échelle linéaire — cf. chapitre 2, p. 32). Hormis l’utilisation d’une échelle linéaire plutôt qu’une échelle Mel, ce choix correspond au type de paramètres le plus fréquemment rencontré dans l’état de l’art actuel. Le recours à une échelle linéaire trouve son origine dans l’application majoritaire d’AMIRAL : le traitement d’enregistrements téléphoniques dans le cadre des campagnes d’évaluation NIST. En effet, dans les limites de la bande passante réduite du téléphone, le positionnement des filtres sur une échelle Mel ou sur une échelle linéaire présente très peu de différences.

Ces coefficients cepstraux, au nombre de 16, sont calculés toutes les 10 ms sur une fenêtre de 20 ms. Lorsque le signal de parole traité provient du réseau téléphonique, il est au préalable filtré pour ne conserver que les fréquences de 300 à 3400 Hz.

Les vecteurs cepstraux sont complétés de leur dérivée première afin de prendre en compte la dynamique à court terme du signal. Ni leur dérivée seconde, ni l’énergie du signal ne sont utilisés pour la reconnaissance, leur présence n’ayant montré aucun apport en termes de performances. Ici encore, ce choix rejoint l’état de l’art actuel.

Il convient cependant de noter que l’énergie est estimée malgré tout, pour être utilisée

lors des traitements post-paramétrisation (voir section suivante). La description de la structure des trames exposée à la section précédente est mise à profit pour utiliser l'énergie lors de ces traitements et l'ignorer dans le reste du processus.

4.3 Traitements post-paramétrisation

Après calcul des vecteurs cepstraux et de leur dérivée, deux types de traitements optionnels peuvent leur être appliqués : une suppression des zones de silence, basée sur le niveau d'énergie du signal, et une normalisation des vecteurs de paramètres.

Ces deux traitements, détaillés ci-dessous, apportent une amélioration significative de performance pour la tâche de vérification du locuteur. Ils peuvent en revanche avoir un effet inverse dans certains cadres multi-locuteurs, notamment en segmentation (cf. [Meignier 2002a], p. 60), et ne sont donc pas appliqués dans ce cas.

4.3.1 Suppression des trames de basse énergie

Du fait du principe même de la reconnaissance du locuteur, les silences présents naturellement au sein du signal de parole constituent des zones non informatives quant à l'identité du locuteur et gagnent donc à être ignorés.

L'algorithme retenu au sein du système AMIRAL pour détecter ces zones de silence se base sur l'étude de l'énergie des vecteurs (ou trames) du signal de parole afin de déterminer un niveau d'énergie minimal pour les trames considérées comme informatives. Les trames dont l'énergie est inférieure à ce seuil sont ignorées lors du processus de reconnaissance.

Le seuil d'énergie est calculé localement, pour chaque enregistrement considéré, et non globalement, une fois pour toutes. Il est issu de l'observation de la distribution de l'énergie des trames composant l'enregistrement. Une estimation en est faite par une distribution bi-gaussienne, dont la gaussienne de moyenne la plus élevée devrait correspondre à la distribution des trames de parole, porteuses d'information relative au locuteur. Cette gaussienne, caractérisée par sa moyenne μ_P et sa variance σ_P^2 , sera utilisée pour fixer le seuil d'énergie.

La gaussienne de plus faible moyenne ($\mu_{\bar{P}}$, $\sigma_{\bar{P}}^2$), en revanche, recouvre des trames correspondant à des événements de nature plus variée (silence et parole de niveau trop faible pour être réellement informative, mais aussi bruit, etc.), lui offrant une variance plus large. Pour cette raison, cette gaussienne n'est pas prise en considération lors du calcul du seuil. Celui-ci est obtenu par la formule suivante :

$$\text{seuil} = \mu_P - 2 \times \sigma_P \quad (4.4)$$

L'algorithme d'estimation du seuil est récapitulé par la figure 4.3, qui montre son application sur un enregistrement issu du corpus de la campagne d'évaluation NIST 2001.

Le gain observé sur ce même corpus pour la tâche de vérification du locuteur (“One Speaker” selon la dénomination NIST) est illustré par la figure 4.4. Les systèmes ayant

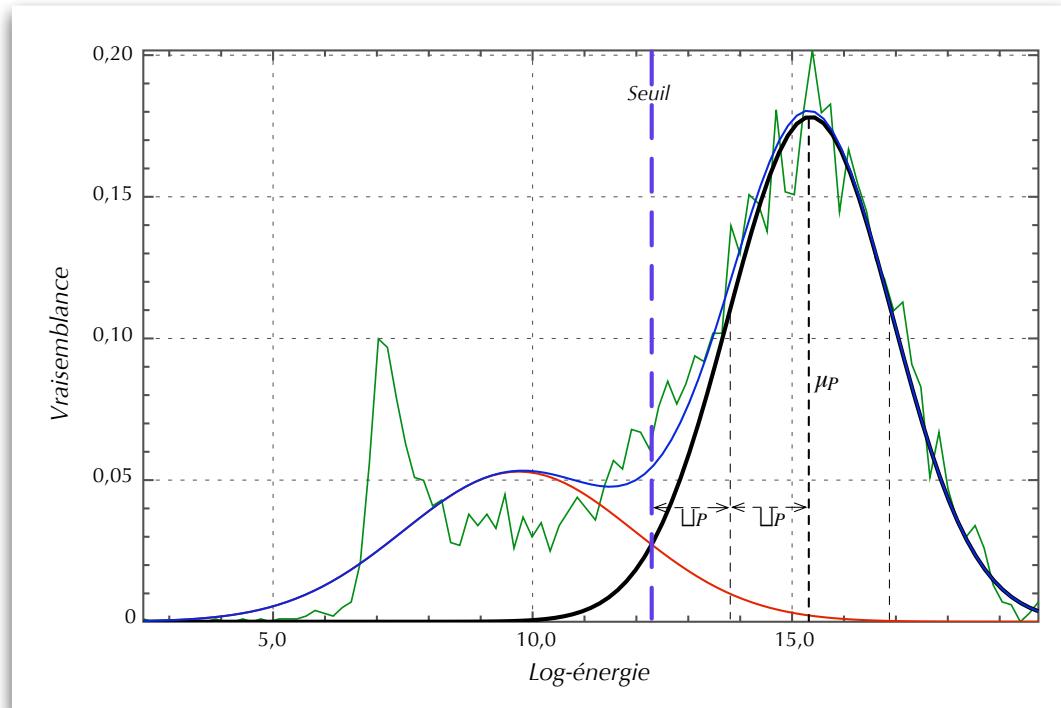


FIG. 4.3 – Illustration de la suppression des trames de basse énergie après paramétrisation, sur un fichier issu du corpus NIST 2001 — La distribution du log-énergie des trames est estimée par une bi-gaussienne ; seule la gaussienne de moyenne la plus élevée (moyenne μ_P , variance σ_P^2) est utilisée pour déterminer le seuil d'énergie ($\mu_P - 2\sigma_P$) en-dessous duquel les trames seront supprimées ; pour ce fichier, 26,2 % des trames sont supprimées.

servi à générer les deux courbes DET qui y sont présentées ne diffèrent que par l'application ou non de la suppression de trames de basse énergie ; les autres caractéristiques sont identiques (à savoir, application de la normalisation des vecteurs de paramètres (voir ci-dessous), modèles GMM à 128 composantes et normalisation des scores par rapport de vraisemblances).

4.3.2 Normalisation des vecteurs de paramètres acoustiques

Les techniques de normalisation des vecteurs de paramètres acoustiques ont pour but de réduire les distorsions présentes dans le signal de parole, dues notamment aux canaux de transmissions utilisés. L'objectif recherché est la minimisation des variations entre les signaux d'apprentissage et de test d'un même locuteur, par l'application à chacun d'eux de cette normalisation.

L'approche retenue dans AMIRAL pour la normalisation des vecteurs de paramètres acoustiques repose sur le centrage de la distribution de ces paramètres et la réduction de la variance de cette même distribution. Après normalisation, la distribution de chacun des coefficients est ramenée à une moyenne nulle et une variance de 1.

L'estimation de la distribution des vecteurs de paramètres acoustiques est effectuée sur la totalité d'un enregistrement (qu'il s'agisse d'un signal de test ou destiné à l'apprentissage d'un modèle de locuteur). La moyenne des coefficients cepstraux étant considérée comme une estimation des distorsions cepstrales engendrées par le canal de transmission ([Furui 1981a], [Rosenberg 1994]), cette normalisation ne présente par conséquent de réel intérêt que si l'enregistrement considéré est mono-canal. Dans le cas contraire, le centrage de la distribution des coefficients ne représenterait plus une compensation des distorsions dues aux canaux de transmissions. Le champ d'application de cette normalisation se restreint donc essentiellement aux tâches traitant des enregistrements mono-locuteur, soit l'identification ou la vérification du locuteur. En effet, les tâches multi-locuteurs, sauf rares exceptions, impliquent potentiellement plusieurs canaux de transmission ; de plus, dans la plupart des cas, les différences de canaux de transmissions constituent, en tant qu'information supplémentaire de discrimination entre les locuteurs, une aide bien plus qu'un obstacle.

La figure 4.5 illustre le gain qu'elle apporte pour cette tâche en confrontant les résultats obtenus sans normalisation, avec normalisation des seuls coefficients statiques (les 16 coefficients cepstraux) et avec normalisation de la totalité des composantes des vecteurs de paramètres acoustiques (les 16 coefficients cepstraux et leur dérivée première). Le dernier cas (normalisation des vecteurs complets) produit les meilleurs résultats.

De plus, une autre série d'expériences révèle que le plus gros gain de performance est observé lorsque cette normalisation des paramètres acoustiques est appliquée après la suppression des trames de basse énergie présentée à la section précédente, plutôt qu'avant (le lecteur se reporterà à l'annexe A pour le détail de ces résultats).

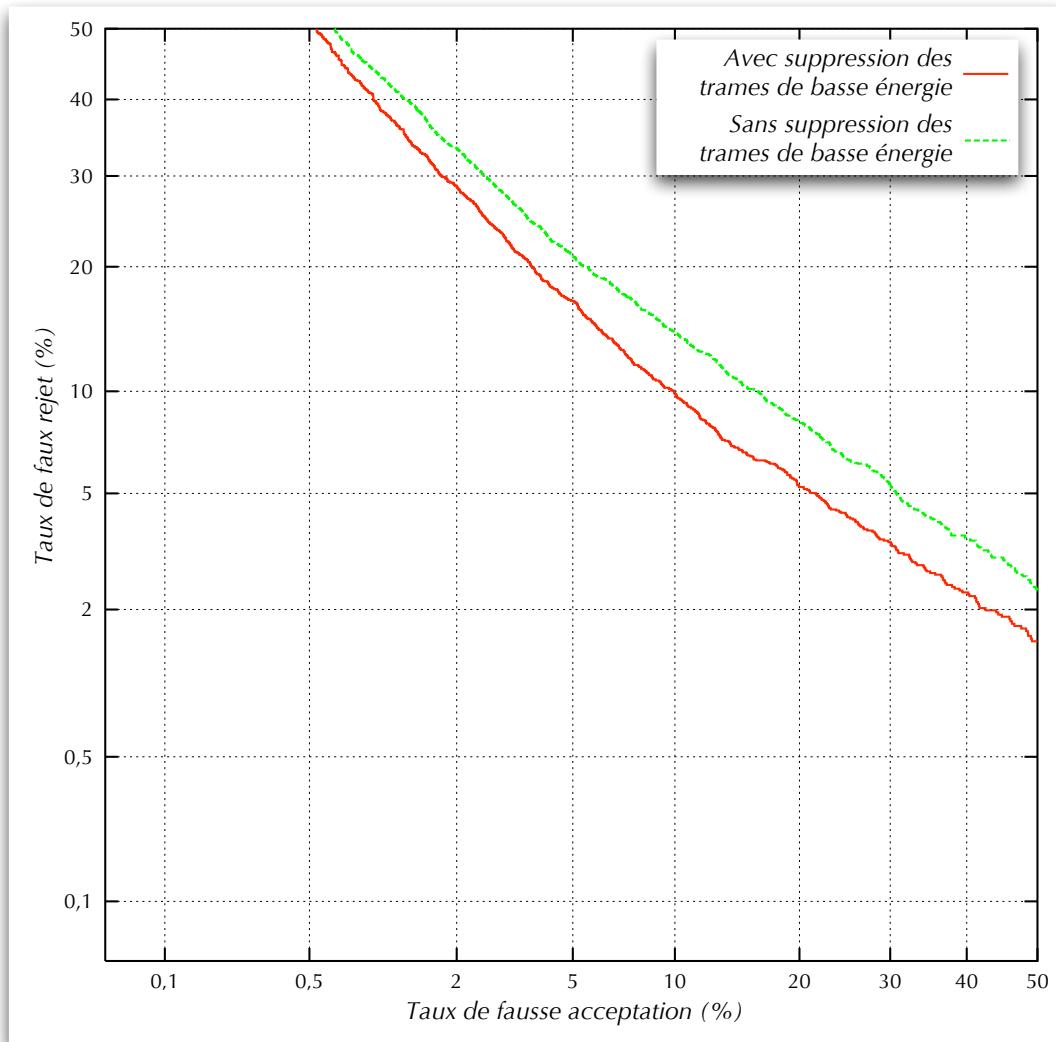


FIG. 4.4 – Intérêt de la suppression des trames de basse énergie pour la vérification du locuteur — Les deux courbes ci-dessus présentent les résultats obtenus pour la tâche “One Speaker” de l’évaluation NIST 2001, avec et sans suppression de ces trames, toutes choses égales par ailleurs (application de la normalisation des trames restantes après la suppression (cf. 4.3.2), modèles à 128 composantes, normalisation des scores par rapport de vraisemblances).

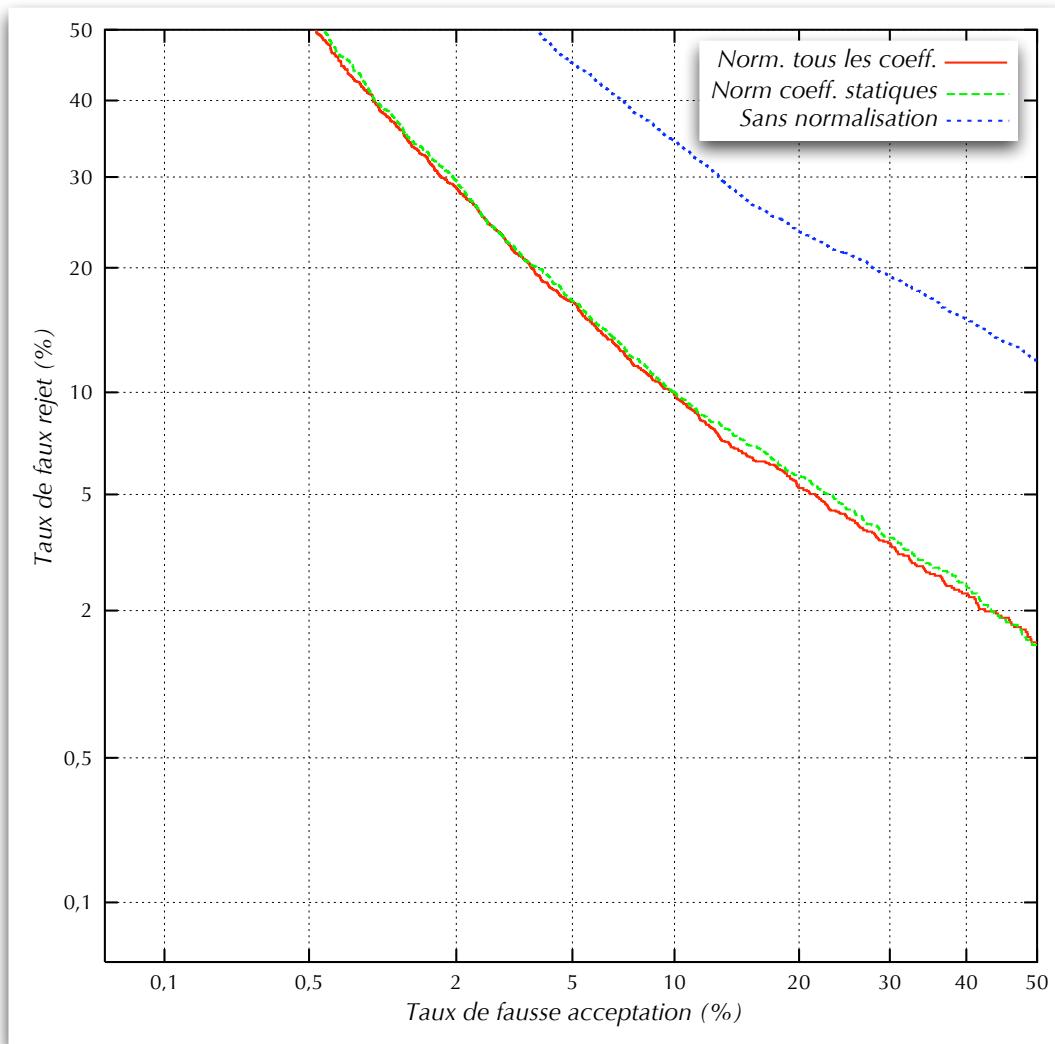


FIG. 4.5 – Intérêt de la normalisation des vecteurs de paramètres pour la vérification du locuteur — Les trois courbes ci-dessus présentent les résultats obtenus pour la tâche “One Speaker” de l’évaluation NIST 2001, sans normalisation des vecteurs de paramètres, avec normalisation des coefficients cepstraux statiques uniquement et avec normalisation des coefficients statiques et dynamiques, toutes choses égales par ailleurs (suppression des trames de basse énergie avant la normalisation (cf. 4.3.1), modèles à 128 composantes, normalisation des scores par rapport de vraisemblances).

4.4 Normalisation des scores

L'une des principales difficultés rencontrées lors du développement d'un système de reconnaissance automatique du locuteur, quelle que soit la tâche visée, concerne le choix du seuil de décision. Ce seuil devant évidemment être fixé *a priori*, avant la mise en exploitation du système, il se doit d'être global (indépendant du locuteur). Il doit également prendre en compte les caractéristiques de l'application visée pour permettre au système de répondre aux contraintes de cette application en termes de point de fonctionnement.

Cependant, les mesures de similarités (vraisemblances) entre un signal de test et un modèle de locuteur présentent une forte variabilité, notamment en présence de variations de conditions d'enregistrement. Cette variabilité constitue un obstacle majeur à l'existence d'un seuil possédant les caractéristiques définies ci-dessus. Pour cette raison, le développement de techniques de normalisation des vraisemblances, visant à réduire cette variabilité, représente une part importante des travaux menés dans le domaine de la RAL (cf. chapitre 2).

Le système AMIRAL intègre deux méthodes de normalisation des scores, décrites ci-après, tout en offrant la possibilité d'en ajouter de nouvelles.

4.4.1 Rapport de vraisemblances

La première étape de la normalisation des scores correspond à un test d'hypothèse bayésien, sous la forme d'un classique rapport de vraisemblances entre l'hypothèse de la production du signal de test par le modèle considéré (hypothèse H_0) et l'hypothèse inverse H_1 (cf. chapitre 2, page 42).

Le principe en est rappelé ici :

- soit $L_{\mathcal{X}}(y)$ la vraisemblance pour que le signal de parole à tester, représenté par un ensemble de vecteurs acoustiques y , ait été produit par le locuteur X , représenté par modèle \mathcal{X} (vraisemblance de l'hypothèse H_0) ;
- soit $L_{\bar{\mathcal{X}}}(y)$ la vraisemblance pour que ce signal ait été produit par un locuteur autre que le locuteur X (vraisemblance de l'hypothèse H_1) ;
- le résultat du rapport de vraisemblances est obtenu par :

$$LR_{\mathcal{X}}(y) = \frac{L_{\mathcal{X}}(y)}{L_{\bar{\mathcal{X}}}(y)} \quad (4.5)$$

Dans le cadre d'AMIRAL, la vraisemblance de l'hypothèse inverse H_1 est estimée par la vraisemblance du signal de test par rapport à un modèle générique de locuteur, ou "modèle du monde", reprenant ainsi l'état de l'art actuel en matière de reconnaissance automatique du locuteur en mode indépendant du texte.

Soit \mathcal{W} ce modèle du monde et $L_{\mathcal{W}}(y)$ la vraisemblance pour que le signal y soit produit par \mathcal{W} . Le rapport de vraisemblances se réécrit alors :

$$LR_{\mathcal{X}}(y) = \frac{L_{\mathcal{X}}(y)}{L_{\mathcal{W}}(y)} \quad (4.6)$$

Le modèle du monde utilisé ici pour la normalisation des vraisemblances correspond au modèle utilisé lors de l'apprentissage des modèles de locuteurs (cf. section 4.5). Les détails de l'estimation de ce modèle ainsi que les enjeux qui y sont liés sont présentés en section 4.6.

4.4.2 Normalisation WMAP

L'objectif principal de la normalisation WMAP (ou “World+MAP”), outre la réduction de la variabilité des vraisemblances, est d'apporter une solution au problème d'interprétation du seuil de décision. En effet, la plupart des techniques de normalisation proposées dans la littérature (cf. section 2.6.4, page 47) génèrent des scores dans un espace non borné, dans lequel l'interprétation d'une valeur de seuil donnée est malaisée.

Afin de pallier cette difficulté, la normalisation WMAP concilie les avantages d'une normalisation par rapport de vraisemblances et de l'approche bayésienne pour projeter les vraisemblances dans un espace probabiliste.

Le principe de cette normalisation, présenté dans [Fredouille 1999] et [Fredouille 2001], est exposé ci-dessous. Le lecteur se reportera à [Fredouille 2000a] pour une présentation plus complète des motivations et des aspects théoriques de la normalisation WMAP.

Principe

Le principe de la normalisation WMAP est basé sur la combinaison de deux concepts :

- la normalisation des vraisemblances par rapport de vraisemblances en utilisant un modèle du monde ;
- l'estimation de probabilités *a posteriori*.

L'utilisation de ces deux concepts pour réaliser la normalisation WMAP est détaillée ci-dessous.

La première phase consiste, étant donnés le signal de parole y , le locuteur X , représenté par son modèle \mathcal{X} , et le modèle du monde \mathcal{W} , à calculer le rapport de vraisemblances $LR_{\mathcal{X}}(y)$, tel que présenté par l'équation 4.6.

La deuxième phase de la normalisation (MAP) s'appuie sur la théorie bayésienne. Elle consiste à remplacer le rapport de vraisemblances $LR_{\mathcal{X}}(y)$ par la probabilité *a posteriori* que le locuteur X ait prononcé le signal y connaissant le rapport de vraisemblances $LR_{\mathcal{X}}(y)$. En d'autres termes, il s'agit de la probabilité d'être en présence d'un test de type “client”, connaissant $LR_{\mathcal{X}}(y)$.

Cette probabilité, notée $p(X = Y | LR_{\mathcal{X}}(y))$, où Y est le locuteur ayant prononcé le signal y , s'exprime sous la forme :

$$p(X = Y | LR_{\mathcal{X}}(y)) = \frac{p(LR_{\mathcal{X}}(y) | X = Y) \cdot p(X = Y)}{p(LR_{\mathcal{X}}(y) | X = Y) \cdot p(X = Y) + p(LR_{\mathcal{X}}(y) | X \neq Y) \cdot p(X \neq Y)} \quad (4.7)$$

où

- $p(LR_{\mathcal{X}}(y) | X = Y)$ est la probabilité du rapport de vraisemblances $LR_{\mathcal{X}}(y)$ sachant que le test est de type “client” et $p(X = Y)$ est la probabilité *a priori* d'être en présence d'un test de type “client” ;

- $p(LR_{\mathcal{X}}(y)|X \neq Y)$ est la probabilité du rapport de vraisemblances $LR_{\mathcal{X}}(y)$ sachant que le test est de type "imposteur" et $p(X \neq Y)$ est la probabilité *a priori* d'être en présence d'un test de type "imposteur".

La mise en place de la normalisation WMAP consiste à estimer la fonction :

$$f_{WMAP}(LR_{\mathcal{X}}(y)) = p(X = Y|LR_{\mathcal{X}}(y))$$

basée sur la formule présentée ci-dessus.

Les probabilités $p(LR_{\mathcal{X}}(y)|X = Y)$ et $p(LR_{\mathcal{X}}(y)|X \neq Y)$ sont estimées à partir de distributions de rapports de vraisemblances correspondant à des tests de types "client" et "imposteur", calculés sur un jeu de données de développement. Un exemple de telles distributions est donné en figure 4.6.

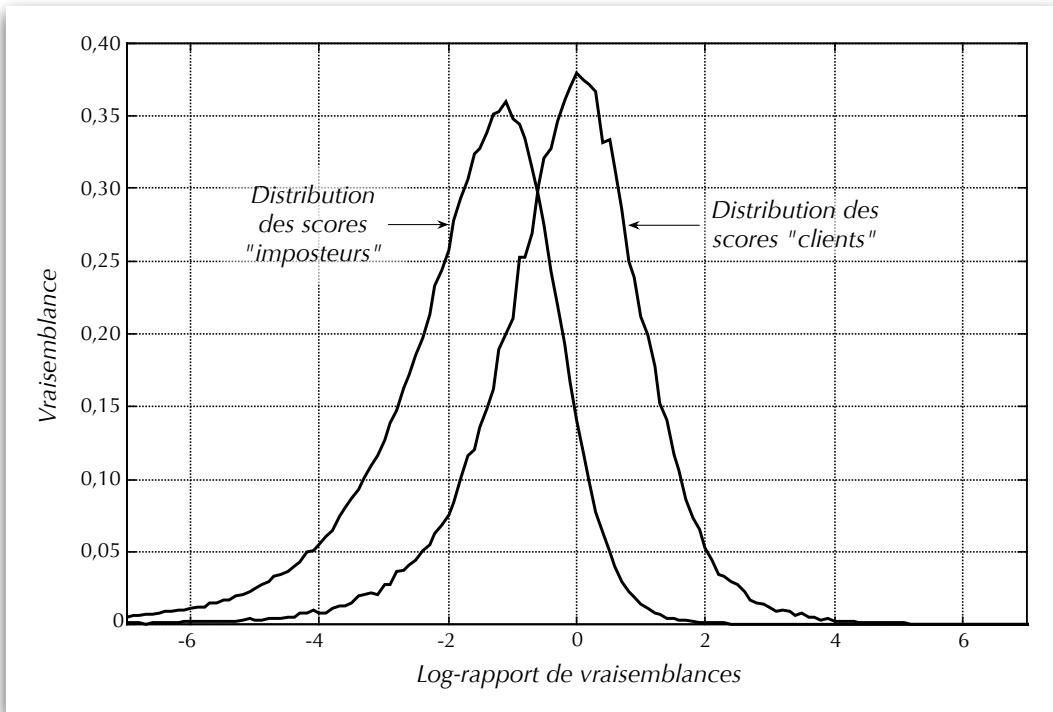


FIG. 4.6 – Normalisation WMAP – Distributions des rapports de vraisemblances pour les tests de types "clients" et "imposteurs" avant normalisation.

Les probabilités $p(X = Y)$ et $p(X \neq Y)$, quant à elles, sont dépendantes de l'application visée. Elles définissent la connaissance *a priori* des conditions d'utilisation du système de RAL (concernant les proportions de tests de types "client" et "imposteur") dans le cadre de cette application.

La figure 4.7 montre la fonction de normalisation f_{WMAP} obtenue en combinant, selon l'équation 4.7, les distributions de probabilités présentées par la figure 4.6 et les probabilités *a priori* $p(X = Y) = 0,1$ et $p(X \neq Y) = 0,9$ (ces valeurs sont celles fixées pour la tâche "One Speaker" des campagnes d'évaluation NIST (cf. chapitre 3, page 53)). Dans

le cadre de l'architecture bloc-segmentale présentée en section 4.1, cette fonction est en fait calculée sur des moyennes de rapports de vraisemblances obtenues sur des blocs de taille fixe.

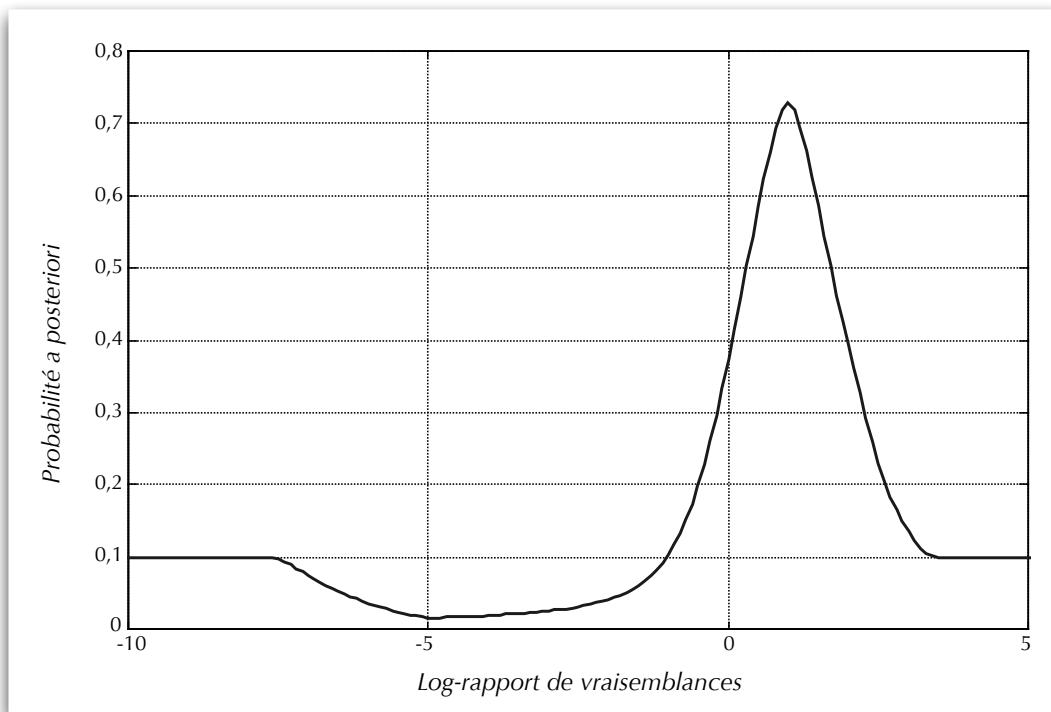


FIG. 4.7 – Normalisation WMAP – Fonction de normalisation des rapports de vraisemblances. Cette fonction a été obtenue à partir des distributions de rapports de vraisemblances présentées par la figure 4.6 et des probabilités *a priori* suivantes : $p(X = Y) = 0,1$, $p(X \neq Y) = 0,9$.

L'application de la normalisation WMAP lors de la phase de test se déroule en deux étapes : dans un premier temps, calcul du rapport de vraisemblances moyen sur un bloc, puis application à ce rapport de vraisemblances de la fonction de normalisation f_{WMAP} . Les scores ainsi obtenus présentent une distribution telle que celle illustrée par la figure 4.8.

La figure 4.9 offre un récapitulatif de l'ensemble du processus de la normalisation WMAP, de l'estimation de la fonction de normalisation à son application.

Intérêt

L'intérêt le plus évident de la normalisation WMAP tient à la nature des scores de sortie, qui correspondent à des probabilités. La signification de cette probabilité (à savoir, la probabilité que l'enregistrement testé ait bien été prononcé par le locuteur considéré, connaissant les proportions *a priori* de tests de types “client” et “imposteur”) permet une interprétation plus aisée des résultats. Par là même, cela facilite le choix du

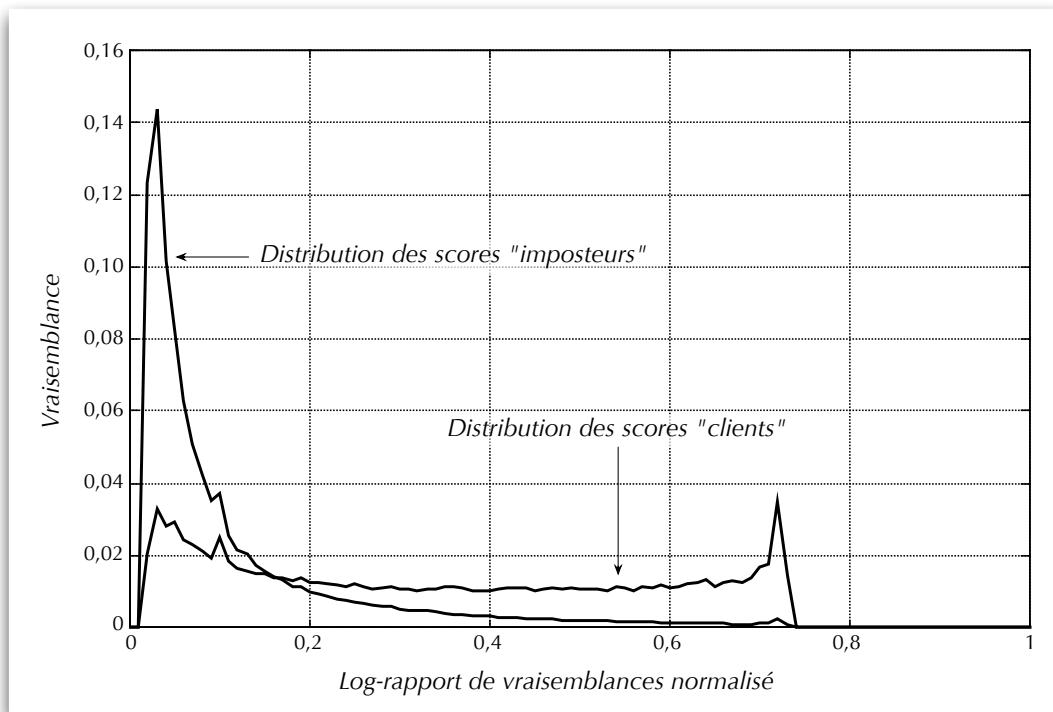


FIG. 4.8 – Normalisation WMAP – Distributions des scores pour les tests de types “clients” et “imposteurs” après normalisation.

seuil de décision, qui représente d'ordinaire une des étapes les plus délicates de la mise en place d'un système de reconnaissance automatique du locuteur. Par exemple, dans le contexte de WMAP, multiplier par deux le seuil de décision revient à diviser par deux le risque de fausse acceptation.

L'intégration de deux types de connaissances (*a priori* et *a posteriori*) lors de l'estimation de la fonction de normalisation présente également des avantages. Le premier type d'information utilisé est la connaissance *a posteriori* du fonctionnement du reconnaiseur sur l'ensemble de données de réglage. Cette connaissance est représentée par la distribution des scores obtenus sur ces données. Elle permet la prise en compte par la fonction de normalisation de la qualité intrinsèque du reconnaiseur. Cette prise en compte rend notamment triviale la fusion des scores entre reconnaiseurs dans un système multi-reconnaiseurs (le lecteur se référera à [Fredouille 2000b] pour un exemple d'application de ce principe).

La seconde source d'information utilisée par la fonction de normalisation est la connaissance *a priori* des conditions d'utilisation du système, en termes de proportion de tests de types “client” et “imposteur”. La façon dont cette connaissance est intégrée lors du calcul de la fonction de normalisation autorise un recalcul rapide de cette fonction pour prendre en compte un éventuel changement des conditions d'utilisation du système. En effet, la partie la plus lourde, d'un point de vue calculatoire, du calcul de la fonction de normalisation, est l'estimation de la distribution des scores obtenus

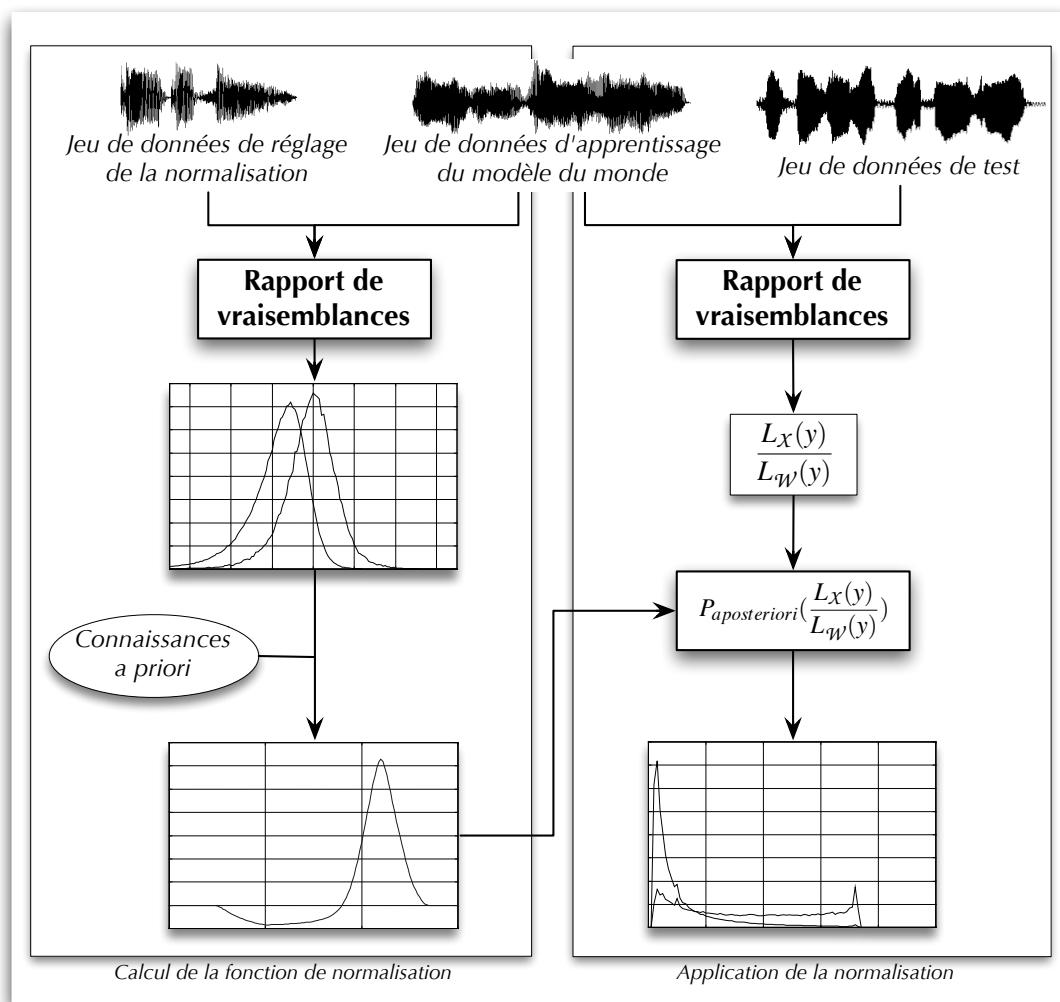


FIG. 4.9 – Normalisation WMAP – Récapitulation du principe.

sur les données de réglages. Une fois ce résultat obtenu et conservé, l'intégration de nouvelles valeurs pour les probabilités *a priori* $p(X = Y)$ et $p(X \neq Y)$ est très rapide. Ce point se révèle important lors du déploiement d'un système de RAL, au même titre que l'interprétation aisée du seuil de décision, car il autorise l'adaptation du système à ses conditions d'utilisation (et à leurs variations) par le biais de réglages facilement compréhensibles par un administrateur non spécialiste de la reconnaissance automatique du locuteur.

Limites

Il convient de noter la limite majeure de la normalisation WMAP, soulignée dans [Fredouille 2000a] : dans le cas où la fonction de normalisation est apprise sur un jeu de données différent des conditions d'exploitation, l'efficacité de cette normalisation est dépendante du degré de similitude des deux jeux de données considérés.

Cette limite n'en est pas forcément un dans un cadre applicatif lorsque rien n'interdit de récupérer les données d'exploitation pour compléter l'estimation de la fonction de normalisation ; l'obstacle de la différence entre les jeux de données disparaît alors. En revanche, ce point se révèle un handicap dans un cadre tel que les campagnes d'évaluation NIST, où est imposée une séparation stricte entre données de développement et données de test.

4.4.3 Autres techniques de normalisation des scores

L'architecture d'AMIRAL rend l'intégration d'autres méthodes de normalisation des mesures de similarité relativement aisée. Cette possibilité offerte aux autres membres du consortium ELISA a été exploitée notamment pour la technique de normalisation Dnorm, mise au point par l'IRISA ([Ben 2002]). Les techniques de normalisation classiques Znorm, Hnorm et Tnorm, constituant l'état de l'art en la matière, ont également été intégrées.

4.5 Modélisation des locuteurs

La modélisation de la voix d'un locuteur repose, dans le cadre du système AMIRAL, sur l'utilisation d'un modèle à mélange de gaussiennes (GMM) pour estimer la distribution des vecteurs acoustiques correspondant aux données d'apprentissage du locuteur.

Comme indiqué au chapitre 2 (page 38), ce type de modèle est le plus fréquemment utilisé aujourd'hui pour la reconnaissance du locuteur en mode indépendant du texte. Depuis leur introduction dans le domaine ([Reynolds 1992]), les GMM se sont en effet montrés particulièrement bien adaptés à cette tâche, et les meilleures performances à l'heure actuelle sont obtenues par des systèmes basés sur des modèles à mélange de gaussiennes. Ce choix s'est donc imposé naturellement pour la modélisation des locuteurs dans le cadre d'AMIRAL.

4.5.1 Structure des modèles

Un modèle GMM consiste en une combinaison linéaire de M distributions gaussiennes, multi-dimensionnelles dans le cas présent, chacune caractérisée par un vecteur moyen et une matrice de covariance.

La vraisemblance d'un vecteur acoustique \vec{y} de dimension D par rapport au GMM \mathcal{X} modélisant le locuteur X est exprimée de la façon suivante :

$$L_{\mathcal{X}}(\vec{y}) = p(\vec{y}|\mathcal{X}) = \sum_{i=1}^M w_i p_i(\vec{y}) \quad (4.8)$$

où M est le nombre de composantes du modèle \mathcal{X} , $p_i(\vec{y})$ représente la vraisemblance de \vec{y} par rapport à la $i^{\text{ème}}$ composante et w_i est le poids de cette composante au sein de la mixture.

$p_i(\vec{y})$ s'exprime sous la forme :

$$p_i(\vec{y}) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{\frac{-1}{2} (\vec{y} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{y} - \vec{\mu}_i)} \quad (4.9)$$

où $\vec{\mu}_i$ est la moyenne de la $i^{\text{ème}}$ distribution gaussienne et Σ_i sa matrice de covariance.

Enfin, le poids total des composantes du modèle est égal à 1 :

$$\sum_{i=1}^M w_i = 1 \quad (4.10)$$

Le modèle \mathcal{X} est alors représenté par ces trois ensembles : poids, moyennes et matrices de covariances.

$$\mathcal{X} = \{w_i, \vec{\mu}_i, \Sigma_i\}, \quad i = 1, \dots, M \quad (4.11)$$

Les matrices de covariance des gaussiennes peuvent être pleines, utilisant tous leurs éléments pour représenter les corrélations entre dimensions, ou simplement diagonales. Typiquement, seules des matrices diagonales sont utilisées dans la littérature. En effet, une distribution quelconque peut théoriquement être modélisée par un GMM à M matrices de covariance pleines comme par un GMM avec un nombre plus élevé de composantes à matrices de covariance diagonales. Mais les GMM à matrices diagonales ont l'avantage de se montrer bien plus efficaces d'un point de vue calculatoire lors de l'apprentissage d'un modèle, qui requiert de nombreuses inversions de matrices de covariance ; l'inversion d'une matrice diagonale est triviale, au contraire de celle d'une matrice pleine. Enfin, il a été constaté empiriquement que les GMM à base de matrices diagonales obtiennent de meilleures performances que les GMM à matrices de covariance pleine.

Le système AMIRAL permet le recours à des matrices de covariances pleines ou diagonales pour les composantes des GMM, les deux approches ayant été évaluées (cf. chapitre 5, p. 91).

Le nombre M de composantes est le second paramètre déterminant la structure d'un modèle GMM. Sa valeur est fonction bien sûr du choix précédent (matrices pleines ou diagonales) mais d'autres critères sont également à prendre en considération. Une valeur plus élevée, autorisant une modélisation plus fine de la distribution des vecteurs de paramètres, conduit théoriquement à une reconnaissance plus efficace. Cependant cette augmentation de performances n'est pas linéaire et tend à atteindre un plafond au delà duquel l'accroissement de la taille des GMM n'apporte plus de gain significatif. La valeur de M correspondant à ce plafond n'est pas absolue et dépend, entre autres choses, de la quantité de données disponible pour l'apprentissage des modèles de locuteurs ainsi que de l'algorithme retenu pour leur estimation (voir la section suivante à ce sujet), mais aussi de l'application visée, à travers la présence ou non des sources de variabilité décrites au chapitre 2 (page 26). Enfin, le choix de la taille des GMM peut également se voir dicté par un souci de vitesse d'exécution, inversement proportionnelle à la taille retenue.

L'annexe A illustre (p. 153) les résultats guidant le choix de la taille des GMM dans le cadre de la tâche de vérification du locuteur des campagnes d'évaluation NIST. Les valeurs retenues lors des diverses participations à ces campagnes sont données aux chapitres 5 et 6.

4.5.2 Estimation des modèles de locuteurs

Deux approches sont disponibles dans le système AMIRAL pour l'apprentissage d'un modèle de locuteur.

Estimation par l'algorithme EM associé au critère du maximum de vraisemblance La première approche repose sur l'algorithme EM (*Expectation - Maximization* — [Dempster 1977]) associé à un critère de maximum de vraisemblance (ML — *Maximum Likelihood*). Étant donnée un ensemble de vecteurs d'apprentissage, les paramètres du GMM (w_i , μ_i , Σ_i) sont estimés de façon itérative en maximisant la vraisemblance des vecteurs par rapport à ce modèle. Les paramètres sont affinés à chaque itération, accroissant la vraisemblance de la séquence de vecteurs par rapport au modèle. Chaque itération se déroule en deux étapes :

- la première étape consiste à calculer la vraisemblance de chaque vecteur par rapport à chacune des composantes du modèle ;
- cette vraisemblance est ensuite utilisée dans la seconde étape comme pondération lors de l'intégration de ce vecteur dans l'estimation des statistiques nécessaires au calcul des nouveaux paramètres du modèle.

Selon la distribution des vecteurs considérés, une dizaine d'itérations est généralement suffisante pour observer une convergence des paramètres. Cette convergence correspond à un maximum local de la vraisemblance, l'algorithme EM ne garantissant pas d'atteindre le maximum absolu. Les paramètres du GMM utilisés à l'initialisation de l'algorithme jouent de ce fait un rôle important pour démarrer l'évolution au plus près de la configuration correspondant au maximum absolu. Pour cette raison, un modèle de locuteur est initialisé, au début de son apprentissage, avec les composantes du modèle du monde (le modèle du monde aux caractéristiques les plus proches des données d'apprentissage est choisi dans le cas où plusieurs modèles, dépendants du genre ainsi que du type de combiné, sont disponibles).

Apprentissage par adaptation bayésienne La seconde approche pour l'apprentissage d'un modèle de locuteur consiste en une forme d'adaptation bayésienne, ou MAP (*Maximum A Posteriori* — [Gauvain 1994]). Le principe mis en œuvre ici, par opposition à l'approche précédente, n'est pas de modéliser précisément la distribution des données d'apprentissage du locuteur, mais de dériver le GMM à partir d'un modèle de locuteur générique en adaptant ce dernier pour le rapprocher des données d'apprentissage. La différence fondamentale entre les deux approches tient dans le fait que le modèle obtenu par adaptation ne modélise pas uniquement les données d'apprentissage du locuteur ; il reprend également une partie des informations portées par le modèle d'origine. De ce fait, l'intérêt de cette approche est d'éviter l'écueil que représentent des données d'apprentissage en quantité insuffisante. En effet, en utilisant comme point de départ un modèle "bien" appris, sur un large ensemble de données, couvrant l'ensemble des classes de paramètres acoustiques susceptibles de se trouver dans un signal de parole, cette approche permet au nouveau modèle d'intégrer également ces classes mêmes si elles ne sont que peu, voire pas du tout, représentées dans les données d'apprentissage du locuteur. Par conséquent, les performances de ce modèle lors de la phase de test se trouvent moins affectées en cas de confrontation à des événements acoustiques non vus dans les données d'apprentissage mais qui sont tout de même du domaine de la parole.

Les résultats des divers systèmes ayant participé aux campagnes d'évaluation NIST ces dernières années confirment les conclusions de l'étude décrite dans [Reynolds 1997] indiquant que l'apprentissage des modèles de locuteurs par adaptation conduit à de bien meilleures performances que l'approche reposant sur l'estimation directe par EM.

Le modèle utilisé dans la littérature comme point de départ de l'adaptation est typiquement le modèle du monde (cf. section 4.6), qui remplit la condition exprimée plus haut d'apprentissage sur un large ensemble de données indépendantes du locuteur. Les diverses équations proposées dans la littérature pour définir l'adaptation diffèrent légèrement, en particulier concernant la forme du coefficient d'adaptation (voir ci-dessous), mais montrent des performances similaires. Un point fait l'unanimité : le meilleur niveau de performance est atteint en n'adaptant que les moyennes des composantes du GMM (les matrices de covariance ainsi que les poids de la mixture restant alors, dans le modèle final, identiques à ceux du modèle initial). Ce résultat a été retrouvé lors des expériences menées au cours du développement de la technique d'adaptation des modèles dans AMIRAL. Pour cette raison, seule l'adaptation des moyennes est abordée ici.

L'adaptation d'un modèle — représenté par ses paramètres (w_i, μ_i, Σ_i) — est réalisée par AMIRAL en deux temps. Les paramètres (w'_i) et (μ'_i) d'un modèle intermédiaire sont tout d'abord estimés sur les données d'apprentissage en appliquant les deux étapes de l'algorithme EM décrites pour l'approche précédente : calcul de la vraisemblance de chaque vecteur par rapport au modèle initial, puis prise en compte de cette vraisemblance comme pondération lors de l'estimation de statistiques sur les données. Les paramètres $(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)$ du modèle final sont ensuite obtenus par :

$$\begin{aligned} \hat{w}_i &= w_i, \\ \hat{\mu}_i &= \alpha_i \times \mu_i + \beta_i \times \mu'_i, & i = 1, \dots, M \\ \hat{\Sigma}_i &= \Sigma_i \end{aligned} \tag{4.12}$$

Les coefficients α_i et β_i définissant l'adaptation de la composante i sont dépendants des données d'apprentissage. Ils sont définis de manière à ce que les composantes du

modèle auxquelles correspond une quantité importante de données d'apprentissage subissent une adaptation plus forte, les faisant reposer plus sur les statistiques correspondant à ces données. Au contraire, les composantes peu représentées dans les données d'apprentissage sont plus faiblement adaptées et restent par conséquent plus proches du modèle initial, évitant de reposer sur des statistiques potentiellement mal estimées. Cet effet est obtenu par l'intégration des poids du modèle initial (w_i) et du modèle estimé (w'_i) dans le calcul des coefficients α_i et β_i de la façon suivante :

$$\alpha_i = \frac{\alpha \times w_i}{\alpha \times w_i + \beta \times w'_i} \quad (4.13)$$

$$\beta_i = \frac{\beta \times w'_i}{\alpha \times w_i + \beta \times w'_i} \quad (4.14)$$

Les coefficients α et β sont des paramètres permettant, indépendamment de l'effet propre à chaque composante obtenu par le calcul ci-dessus, de régler globalement l'importance que doivent avoir les données d'apprentissage par rapport aux connaissances représentées par le modèle initial. L'influence de ces paramètres sur les performances du système est illustré par la figure 4.10 à travers les résultats obtenus sur le corpus de développement (NIST 2001) pour trois jeux de valeurs.

Une valeur élevée de α par rapport à β , ayant pour conséquence des modèles de locuteurs plus proches du modèle du monde, tend logiquement, associée à la normalisation par rapport de vraisemblances, à déséquilibrer le fonctionnement du système par un taux de rejet plus élevé, le cas inverse ($\beta > \alpha$) accroissant le taux d'acceptation. Malheureusement ces effets sont produits par une dégradation des performances des fonctions d'acceptation et rejet, plutôt que par une amélioration de l'une de ces fonctions, comme le montrent les taux d'erreurs systématiquement plus élevés que pour le cas $\alpha = \beta$. Une étude plus complète de couples de valeurs possibles pour α et β , dont les résultats sont présentés en annexe A, page 154, a conduit au choix des valeurs suivantes :

$$\begin{aligned} \alpha &= 0,25 \\ \beta &= 0,75 \end{aligned} \quad (4.15)$$

dans le cadre des campagnes d'évaluation NIST, ces valeurs montrant sur le corpus de développement des performances légèrement supérieures autour du point de fonctionnement évalué pour la tâche "one-speaker detection".

Enfin, en complément de la technique d'adaptation de modèle présentée ici, une variante du même principe est également intégrée au système AMIRAL. Il s'agit de la méthode d'adaptation décrite dans [Reynolds 2000], intégrée ici du fait de son utilisation dans un système représentant l'état de l'art. La différence avec la technique présentée précédemment tient dans le mode de calcul des coefficients α_i et β_i :

$$\begin{aligned} \beta_i &= \frac{T \cdot w'_i}{T \cdot w'_i + r} \\ \alpha_i &= 1 - \beta_i \end{aligned} \quad (4.16)$$

où T est le nombre de trames des données d'apprentissage et r , appelé *relevance factor*, un paramètre jouant le même rôle que les paramètres α et β de l'approche précédente. La comparaison des deux techniques fait apparaître des performances sensiblement équivalentes.

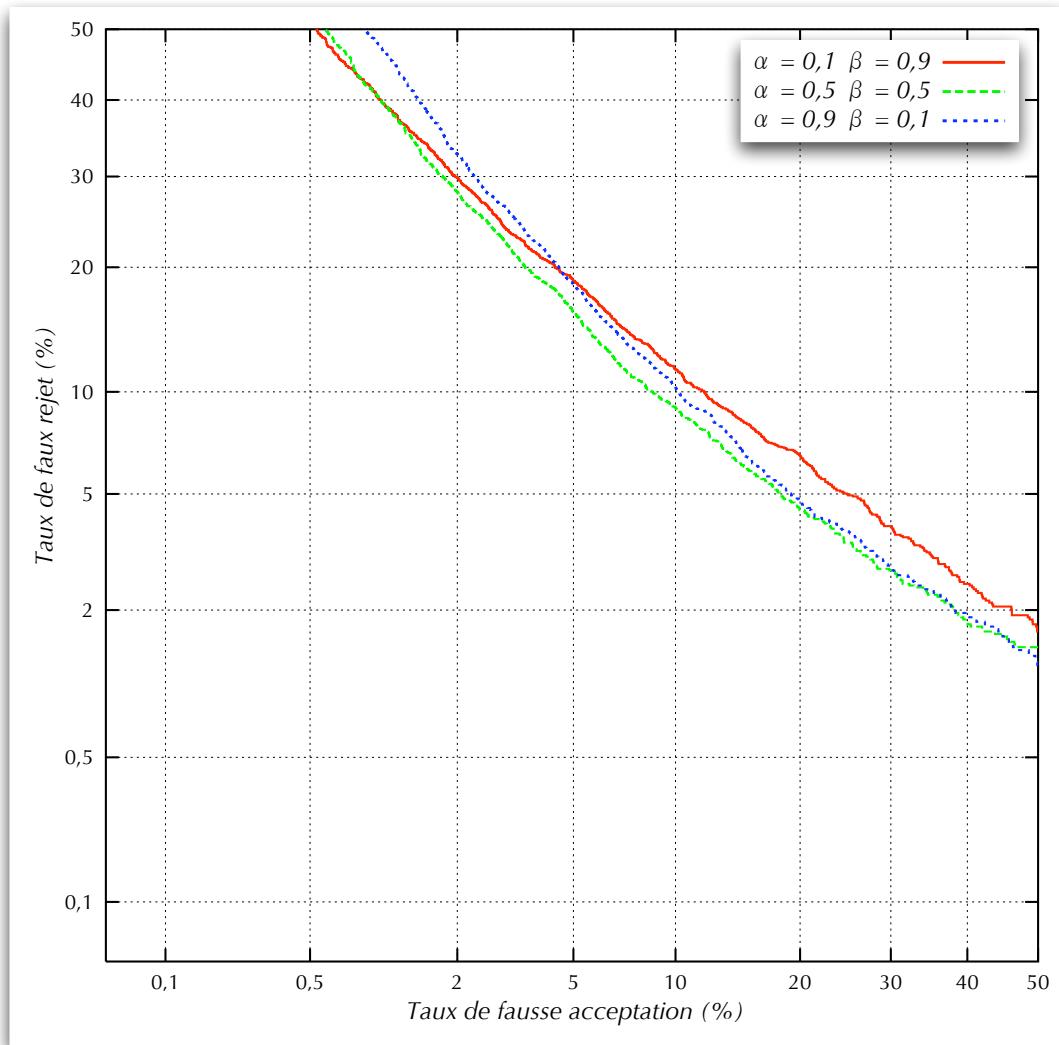


FIG. 4.10 – Choix des paramètres α et β pour l'apprentissage des modèles de locuteurs par adaptation du modèle du monde — Illustration de l'influence de ces paramètres sur les résultats.

4.6 Modèle du monde

Dans un système de RAL basé sur une approche de type GMM, le modèle du monde a un rôle central, autant pour la modélisation que pour la normalisation et la décision. Ce modèle est appris à l'aide de l'algorithme EM, en général en maximisant le critère du maximum de vraisemblance. Le choix de la structure du modèle et des données d'apprentissage sont des éléments très importants et il est nécessaire de comprendre le rôle exact du modèle du monde pour effectuer les bons choix.

4.6.1 Triple intervention

Des sections 4.5 et 4.4, il ressort que le modèle du monde joue un rôle considérable dans le processus de reconnaissance. En effet, ce modèle intervient en trois points, comme illustré par la figure 4.11.

Au cours de l'apprentissage d'un modèle de locuteur, le rôle du modèle du monde est double. Il est utilisé une première fois comme initialisation de l'algorithme EM pour l'estimation d'un modèle sur les données d'apprentissage du locuteur. Sa seconde intervention se situe lors de la phase d'adaptation proprement dite, où il représente l'information *a priori* qui sera combinée avec les statistiques issues de l'ensemble des données d'apprentissage d'un locuteur pour produire le modèle de ce locuteur.

Enfin, le modèle du monde est impliqué une troisième fois, lors de la phase de normalisation des scores par rapport de vraisemblances.

Cette multiple intervention dans la chaîne de traitement est la raison de l'importance déterminante, pour les performances d'un reconnaisseur, de la qualité du modèle du monde utilisé.

4.6.2 Données d'apprentissage

La qualité d'un modèle du monde dépend de deux facteurs : d'une part, l'efficacité de l'algorithme utilisé pour estimer ce modèle (ou plus précisément les paramètres utilisés pour contrôler cet algorithme, qui est en général l'algorithme EM) ; d'autre part, l'adéquation des données d'apprentissage au cadre d'utilisation du modèle. Dans le cas du modèle du monde, ce second point signifie l'adéquation des données sur lesquelles il est estimé au triple rôle présenté ci-dessus.

La double fonction remplie par le modèle du monde au cours de l'apprentissage des modèles de locuteurs (initialisation de l'estimation et information *a priori* lors de l'adaptation) fait apparaître la nécessité d'un modèle du monde aussi proche que possible (en termes notamment de conditions d'enregistrement et de transmission du signal) des données d'apprentissage des locuteurs.

De même, l'objectif de la normalisation par rapport de vraisemblances étant de réduire la variabilité des vraisemblances en compensant les variations observées entre les signaux de test et ceux ayant servi à l'apprentissage des modèles de locuteurs, le modèle du monde se doit là aussi d'être proche de ces derniers. Cette similitude lui permettra de présenter une réponse similaire face aux signaux de test.

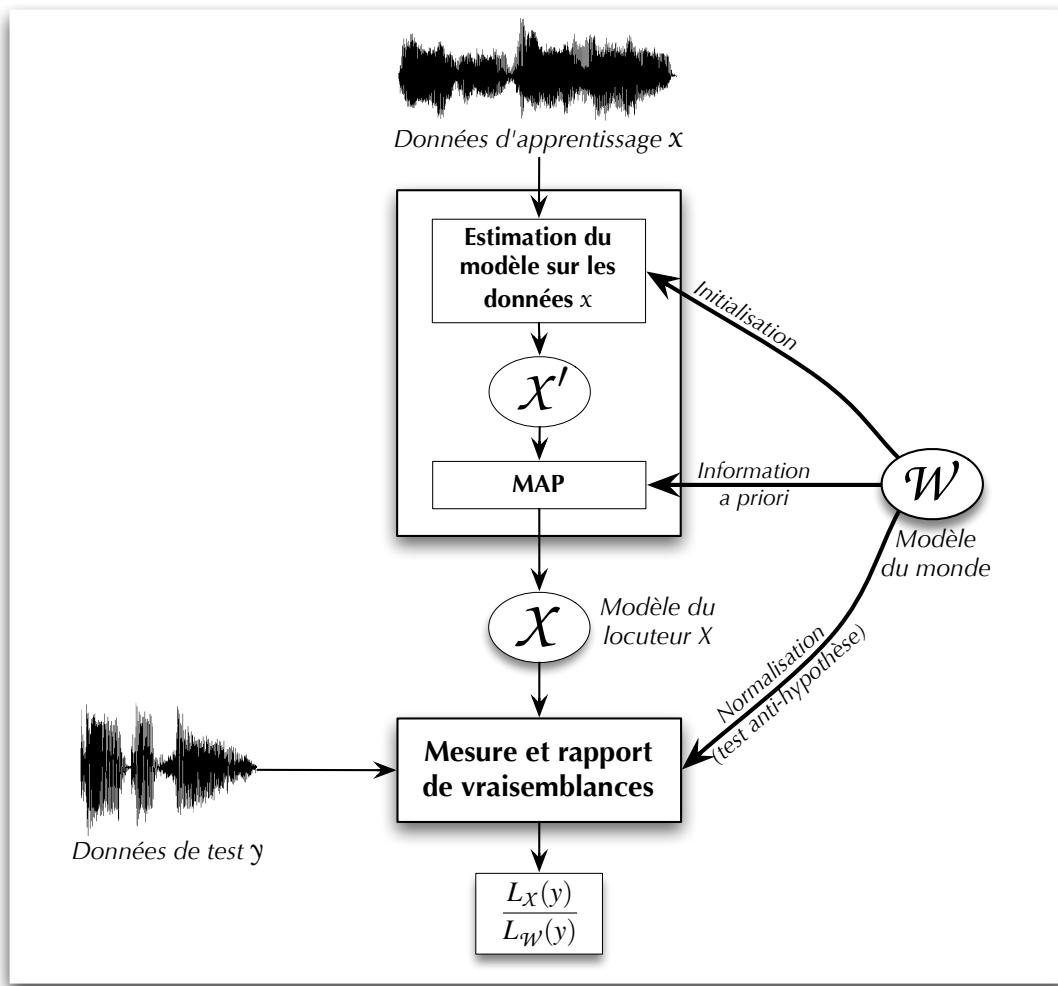


FIG. 4.11 – Illustration de la triple intervention du modèle du monde dans le processus de RAL – Ce modèle intervient deux fois lors de l'apprentissage d'un modèle de locuteur, puis une fois de plus lors de la normalisation des scores par rapport de vraisemblances.

Une des clés d'un modèle du monde efficace est donc la disponibilité de données reflétant suffisamment bien les conditions dans lesquelles se fera l'apprentissage des modèles de locuteurs. Naturellement, ces données devront être issues d'un nombre de locuteurs assez élevé pour prétendre estimer un modèle générique, indépendant du locuteur.

Cependant, l'existence même de telles données suppose au minimum une connaissance préalable des conditions d'utilisation du système. Cette hypothèse est facilement vérifiée dans le cadre bien défini des campagnes d'évaluation NIST². Mais elle ne peut, au mieux, qu'être partiellement vérifiée dans de nombreux cadres applicatifs, notamment lorsqu'un reconnaisseur est appelé à fonctionner sur divers types de matériels d'enregistrement et dans un environnement inconnu au préalable (et potentiellement changeant). Le chapitre 7 (page 121) est consacré à deux projets applicatifs présentant cette caractéristique.

De plus, au-delà de cette nécessaire connaissance des conditions d'utilisation du système, se pose le problème de la disponibilité de données correspondantes en quantité suffisante pour permettre une bonne estimation de ces conditions. En dehors des divers cadres applicatifs présentés dans ce document (*cf. chapitre 7*), cette difficulté s'est présentée lors de la campagne d'évaluation NIST 2001, à l'occasion de l'introduction de la tâche de vérification du locuteur sur des données issues de téléphones cellulaires, pour laquelle l'ensemble de données fourni pour le développement était de taille très réduite (voir à ce sujet le chapitre 5).

La figure 4.12 est présentée à titre d'illustration de l'influence sur les performances du reconnaisseur de la quantité de données pertinentes disponibles pour l'estimation du modèle du monde. Elle montre les résultats obtenus sur le corpus de la campagne d'évaluation NIST 2001 en limitant la quantité de données d'apprentissage du modèle du monde. Il s'agit donc ici d'un cadre où les conditions d'exploitation sont connues lors du développement du système, mais où les données correspondantes sont disponibles uniquement en quantité limitée. Ces résultats font apparaître une très nette dégradation des performances dès lors que le modèle du monde est estimé sur moins de 30 minutes de données. Cette valeur n'est bien entendu à prendre qu'à titre d'exemple, la quantité de données nécessaire variant selon la complexité des conditions que le modèle du monde doit représenter (le cas présenté ici est à considérer comme "facile" de ce point de vue, l'ensemble de données traité étant limité à un genre et un type de combiné).

Il peut être judicieux, lorsque l'application le permet, de récupérer une partie des données d'exploitation (signaux d'apprentissage, ou de test, ou les deux) pour compléter le jeu de données d'estimation du modèle du monde, si celui-ci est de taille insuffisante. Un système de segmentation en locuteurs travaillant sur une base de documents préalablement enregistrés, par exemple, pourra utiliser tout ou partie de la base pour obtenir un modèle du monde le plus proche possible des conditions d'utilisation.

Cependant, dans de nombreux cas cette solution n'est pas applicable (particulièrement lorsque les données d'exploitation ne sont pas pré-enregistrées). Il est alors envisageable de compléter les données d'apprentissage du modèle du monde par d'autres correspondant moins au cadre d'application. Un mode d'action possible consiste à réaliser une estimation d'un modèle du monde correspondant aux peu de données d'exploitation disponibles grâce à l'adaptation bayésienne d'un modèle estimé sur ce nouveau jeu de données de taille plus importante. Le recours à cette solution

²Dans le cadre de la campagne d'évaluation NIST 2004, cette hypothèse n'était plus vérifiée ; les taux d'erreurs obtenus par les différents systèmes ont augmenté de façon significative par rapport à l'année précédente.

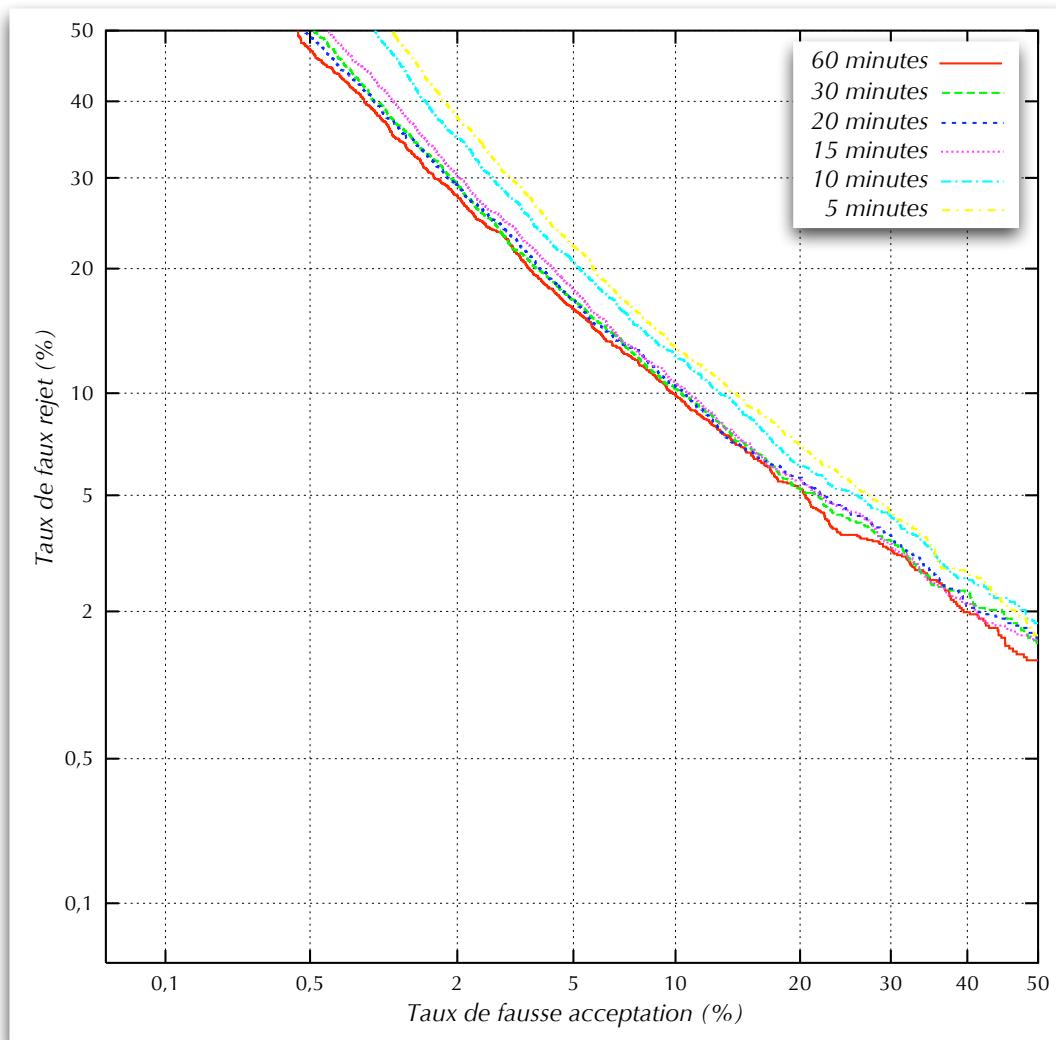


FIG. 4.12 – Influence de la qualité du modèle du monde sur les performances du reconnaisseur – Illustration sur les résultats de la tâche “One Speaker” de l’évaluation NIST 2001, en faisant varier la quantité de données utilisée pour l’apprentissage du modèle du monde, de 5 minutes à 1 heure de parole.

dans le cadre de la tâche de vérification du locuteur sur données téléphoniques cellulaires lors de la campagne d'évaluation NIST 2001 est présentée au chapitre 5.

Enfin, lorsqu'un modèle du monde est estimé sur des données intégrant l'ensemble des diverses conditions d'exploitation possibles (par opposition à un modèle du monde dépendant des conditions d'utilisation), il est qualifié de modèle du monde universel ou UBM (*Universal Background Model*— [Reynolds 1997]). Deux approches sont envisageables pour l'estimation d'un tel modèle. La première consiste à regrouper toutes les données correspondant aux diverses conditions et à estimer le modèle directement sur cet ensemble. Cependant, en cas de déséquilibre (en termes de taille) des divers ensembles de données, la surreprésentation d'une condition par rapport aux autres introduit un biais dans le modèle. La seconde solution contourne cet obstacle par l'apprentissage d'un modèle sur chaque ensemble de données, les composantes des divers modèles étant ensuite regroupées pour former le modèle UBM (par exemple, deux modèles à 128 composantes permettent de générer un UBM à 256 composantes).

4.6.3 Estimation

L'estimation du modèle du monde se fait grâce à l'algorithme EM. Le lecteur se reportera à la section 4.5.2 (p. 82) pour la présentation de cet algorithme, seuls les points spécifiques à son utilisation dans le cadre de l'estimation d'un modèle du monde étant détaillés ci-après.

En dehors des contraintes techniques imposées par la quantité de données à gérer, la principale différence par rapport à l'utilisation de l'algorithme EM dans le cadre de l'apprentissage d'un modèle de locuteur tient dans l'initialisation de l'apprentissage. En effet, le modèle du monde étant estimé avant tout autre modèle, l'initialisation dans ce cas ne peut reposer sur un modèle existant, au contraire de l'apprentissage d'un modèle de locuteur qui est généralement initialisé en utilisant le modèle du monde. L'approche retenue dans le cadre d'AMIRAL pour effectuer l'initialisation de l'apprentissage du modèle du monde repose sur l'estimation de moyennes locales et de la covariance globale. Pour chaque composante du modèle, une trame des données d'apprentissage est sélectionnée aléatoirement ; une estimation de la moyenne est réalisée sur une fenêtre de 10 trames autour de cette trame et sert de valeur initiale pour la moyenne de cette composante. La matrice de covariance de chaque composante est quant à elle initialisée à l'estimation de la covariance sur l'ensemble des données. Enfin, les poids des M composantes sont tous initialisés à la même valeur, à savoir $1/M$.

Une autre particularité de l'apprentissage d'un modèle du monde est la contrainte imposée à l'évolution de la variance des diverses composantes au cours des itérations de EM. Un minimum est en effet fixé pour les coefficients de la matrice de covariance de chaque composante. L'objectif d'un tel seuillage est de préserver le caractère générique du modèle du monde en évitant une trop grande spécialisation de certaines de ses composantes sur une partie des données d'apprentissage. Le gain de performances observé en vérification du locuteur après application de ce seuillage est significatif, l'étude de diverses valeurs sur le corpus de développement conduisant au choix de 0,5 fois la matrice de covariance globale comme minimum.

Chapitre 5

Évaluation dans le cadre de la tâche NIST “*One-Speaker Detection*”

Sommaire

5.1 NIST 98	91
5.2 NIST 99	92
5.3 NIST 2000	94
5.4 NIST 2001	94
5.5 NIST 2002	97
5.6 NIST 2003	98
5.7 Bilan	98

Le système AMIRAL tel qu'il est présenté dans le chapitre précédent est le fruit de plusieurs années de développement visant à obtenir un système intégrant l'état de l'art en matière de reconnaissance automatique du locuteur, les diverses techniques utilisées ayant été adoptées progressivement au cours des ans et non d'un bloc lors de la création d'AMIRAL. Comme il en a déjà été fait mention, de bonnes performances en vérification du locuteur sont cruciales pour espérer traiter correctement les autres tâches de la RAL, qui reposent toutes sur l'utilisation des techniques de base de la VAL. La tâche de vérification du locuteur (*one-speaker detection*) des campagnes d'évaluation NIST a servi de référence depuis 1999 pour l'évaluation des performances du système AMIRAL, permettant d'évaluer le gain obtenu à chaque étape de son développement, et ce grâce à une participation annuelle aux campagnes d'évaluation NIST. Le présent chapitre est consacré à un historique de cette participation, qui permet de découvrir les diverses étapes de l'évolution de l'architecture d'AMIRAL.

5.1 NIST 98

La première participation du LIA à une campagne d'évaluation NIST s'est faite en 1998. Elle n'entre pas dans le cadre de ce travail de thèse, mais l'architecture du

système utilisé alors est tout de même présentée (très succinctement) pour référence, comme point de départ de l'évolution d'AMIRAL.

Il s'agissait déjà d'un système de reconnaissance reposant sur l'utilisation de statistiques du second ordre ([Bimbot 1995]). Les locuteurs étaient modélisés par une gaussienne à matrice de covariance pleine. Le score obtenu par un signal lors de la phase de test était la moyenne des vraisemblances des trames du signal par rapport au modèle, normalisé ensuite par rapport de vraisemblances reposant sur un modèle du monde estimé sur un sous-ensemble des données de la campagne de 1996.

5.2 NIST 99

La campagne NIST 1999 a vu la première participation du système AMIRAL basé sur l'architecture présentée au chapitre 4.

L'introduction de nombreuses nouveautés par rapport à l'année précédente a été permise par, entre autres, l'utilisation des données de l'évaluation 1998 pour le développement du système et l'expérience acquise lors de la participation à la campagne de 1998. Les nouveautés sont notamment (en dehors de la réécriture complète du système) l'utilisation de GMM, l'architecture bloc-segmentale multi-reconnaisseurs et la normalisation WMAP. Une première partie de ces données (un ensemble de signaux de 30 secondes, produits par 100 hommes et 100 femmes et sélectionnés pour être également répartis suivant le type de combiné ayant servi à l'enregistrement (“electret” ou “carbon”)) a servi à l'estimation de 4 modèles du monde, dépendants du genre et du type de combiné. Un autre sous-ensemble des données de 1998, composé de signaux de 100 locuteurs, sert de jeu de développement et de réglage. Ce jeu de données est utilisé pour estimer la fonction de normalisation WMAP appliquée lors de la campagne d'évaluation, en utilisant les probabilités *a priori* de 0,1 pour les tests clients et 0,9 pour les tests imposteurs (ces probabilités sont issues des proportions annoncées de tests clients et imposteur au cours de la campagne d'évaluation).

La paramétrisation est classique, reposant sur des vecteurs cepstraux à échelle linéaire issus d'une analyse par banc de filtres. Mais seuls les 16 coefficients statiques sont utilisés, ni l'énergie ni les dérivées première ou seconde ne sont présents. Un retrait de la moyenne cepstrale est appliqué aux vecteurs de paramètres.

La modélisation des locuteurs repose sur des GMM à 16 composantes à matrices pleines pour le système soumis comme système primaire (dont les résultats sont présentés par la figure 5.1). Ces modèles ne sont pas issus d'une adaptation du modèle du monde, mais estimés à l'aide de l'algorithme EM (l'algorithme est le même que celui utilisé pour le calcul du modèle du monde, n'intègrant pas encore de seuillage de la variance) ; l'initialisation de l'apprentissage se fait en utilisant le modèle du monde correspondant au genre et au combiné utilisé pour enregistrer les données d'apprentissage du locuteur considéré.

Les scores en sortie du calcul de vraisemblance sont normalisés par WMAP (en utilisant le modèle du monde correspondant au genre du locuteur cible et au type de combiné utilisé pour l'enregistrement de ses données d'apprentissage — le type de combiné du signal de test n'est pas pris en compte), ce qui constitue la principale originalité du système par rapport aux autres participants.

Enfin, l'architecture multi-reconnaisseurs et le principe de la description de la structure des trames sont mis à profit par une étude sur l'exploitation des informations dynamiques du signal par concaténation des coefficients statiques, dans le cadre du travail de thèse de Corinne Fredouille. Les résultats de cette étude ne sont pas reproduits ici. Le lecteur se reportera à [Fredouille 2000a] pour une présentation détaillée de cette étude.

Les performances affichées par le système primaire montrent l'intérêt de WMAP comme technique de normalisation. Mais l'écart avec les meilleurs systèmes présentés à l'évaluation fait apparaître combien le reste du système est en retrait par rapport à l'état de l'art sur certains points, notamment l'utilisation de modèles de locuteurs à gaussiennes pleines et appris par EM, ainsi que la non utilisation des coefficients delta dans la paramétrisation.

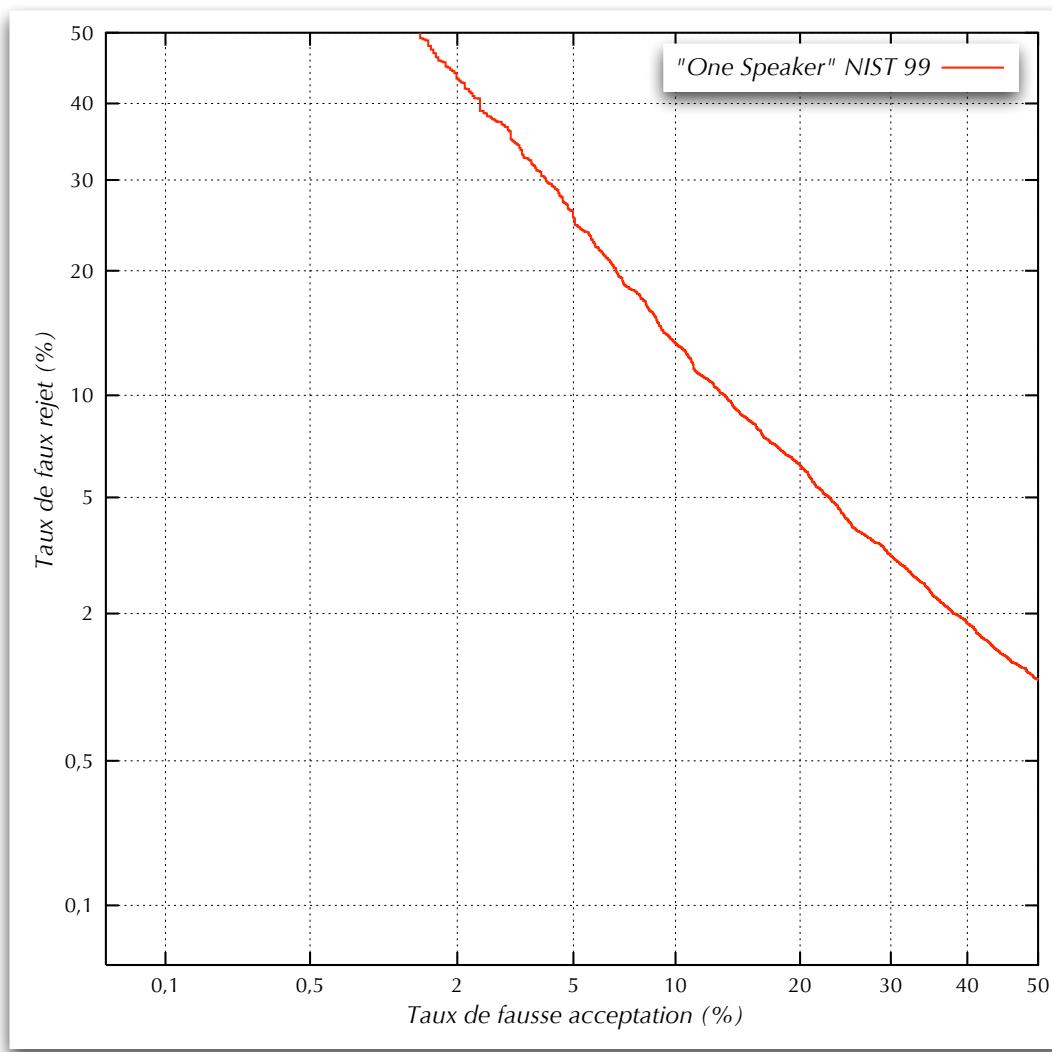


FIG. 5.1 – Résultats du système primaire pour la tâche “One-Speaker Detection” de l’évaluation NIST 99.

5.3 NIST 2000

La campagne d'évaluation NIST 2000 reprend les données de l'édition 1998 (mais organisées différemment), la fin du développement du corpus Switchboard II ne permettant pas d'obtenir de nouvelles données. Une différence importante avec les éditions précédentes rend la tâche de vérification plus difficile : les données d'apprentissage des locuteurs, jusque là issues de deux sessions d'enregistrement différentes pour chaque locuteur, ne correspondent plus qu'à une seule session. Les modèles de locuteurs intègrent de ce fait moins de variabilité intra-locuteur.

Le développement du système AMIRAL en vue de la participation à la campagne 2000 s'est fait en utilisant le jeu de données de la campagne 1999. Le système utilise toujours 4 modèles du monde (dépendants du genre et du combiné utilisé pour les données d'apprentissage des locuteurs), calculés chacun sur une centaine de minutes de parole issues des données 1999. Une autre partie de ces données sert à l'estimation des fonctions de normalisation WMAP. Ces fonctions sont dépendantes du genre, du combiné utilisé pour les données d'apprentissage, ainsi que du combiné utilisé pour le signal de test. Huit fonctions de normalisation sont ainsi calculées.

La paramétrisation utilisée pour le système primaire en 2000 est toujours basée sur des vecteurs cepstraux, mais intègre cette fois la dérivée première de ces vecteurs. Une soustraction de la moyenne cepstrale est appliquée sur les coefficients statiques des vecteurs de paramètres.

L'autre nouveauté marquante d'AMIRAL en 2000 concerne l'apprentissage des modèles de locuteurs, qui se fait maintenant par adaptation du modèle du monde, selon la technique décrite au chapitre 4 (page 82). La structure utilisée pour les modèles du monde comme les modèles de locuteurs est maintenant fixée à 128 gaussiennes à matrice diagonale.

Les résultats obtenus par ce système (présentés par la figure 5.2) le classent dans la moyenne des systèmes présentés. Ces résultats restent toutefois en retrait par rapport aux meilleurs systèmes.

5.4 NIST 2001

L'édition 2001 de la campagne d'évaluation NIST reprend, pour la tâche de vérification du locuteur, les mêmes données que l'année précédente, faute de nouvelles données. Malgré le changement de nom de tous les fichiers pour éviter toute tentation de tricherie, cette réutilisation — annoncée tardivement — introduit un biais, du fait que de nombreux participants ont utilisé pendant l'année les données de la campagne 2000 pour développer leur système.

Cependant, une variante de la tâche est proposée, offrant de travailler sur un nouveau corpus, enregistré à partir de téléphones cellulaires. Malheureusement la quantité de données de développement fournies par NIST pour travailler sur cette tâche est extrêmement réduite.

Les caractéristiques du système AMIRAL présenté cette année-là sont similaires à celles de 2000, à l'exception d'une nouveauté au niveau du traitement post-paramétrisation : l'adoption de la suppression des trames de basse énergie telle que

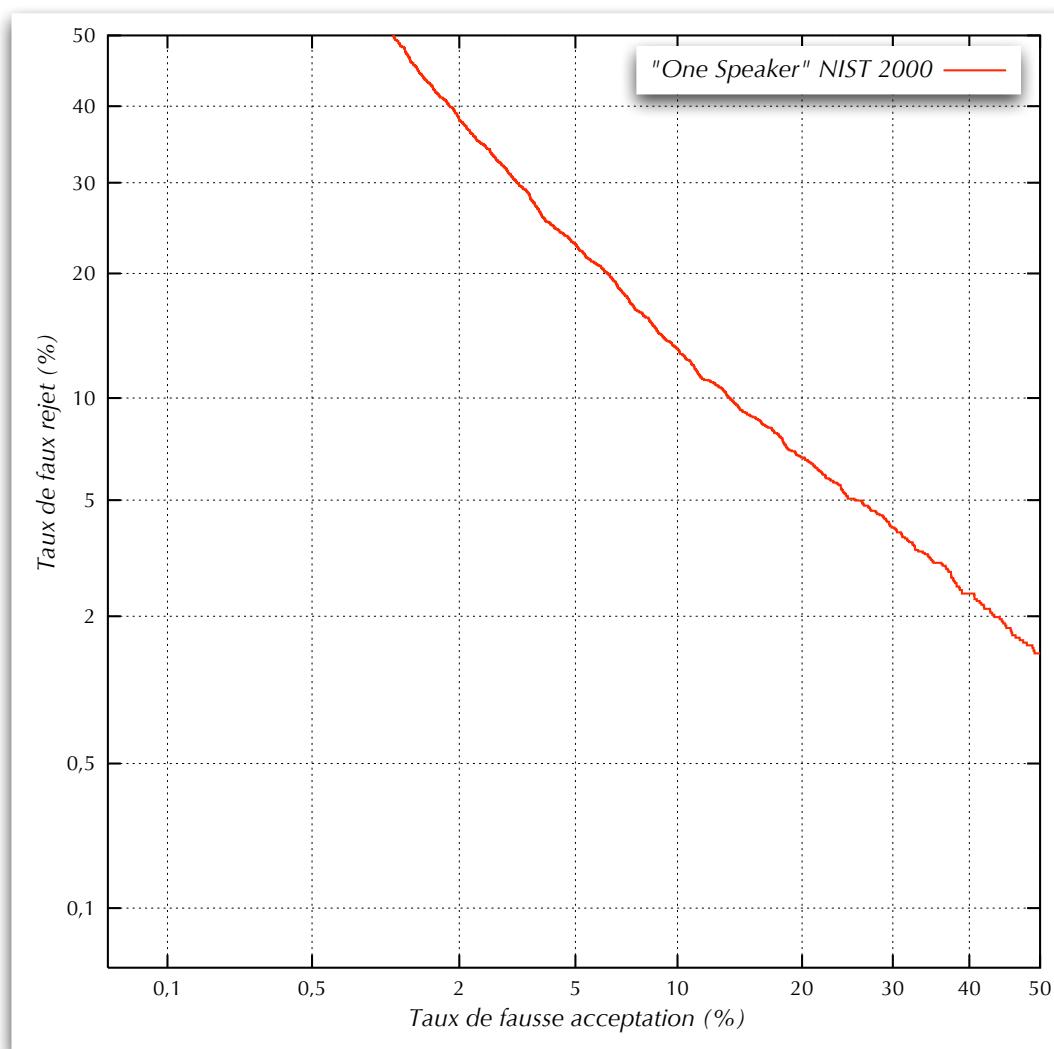


FIG. 5.2 – Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 2000.

décrite au chapitre 4, page 69.

Pour la participation à la tâche de vérification sur données cellulaires, deux modèles du monde (dépendants du genre) sont estimés par adaptation de modèles du monde (dépendants du genre également) calculés sur le corpus téléphonique filaire, la quantité de données de développement cellulaires ne permettant pas le calcul direct de modèles du monde.

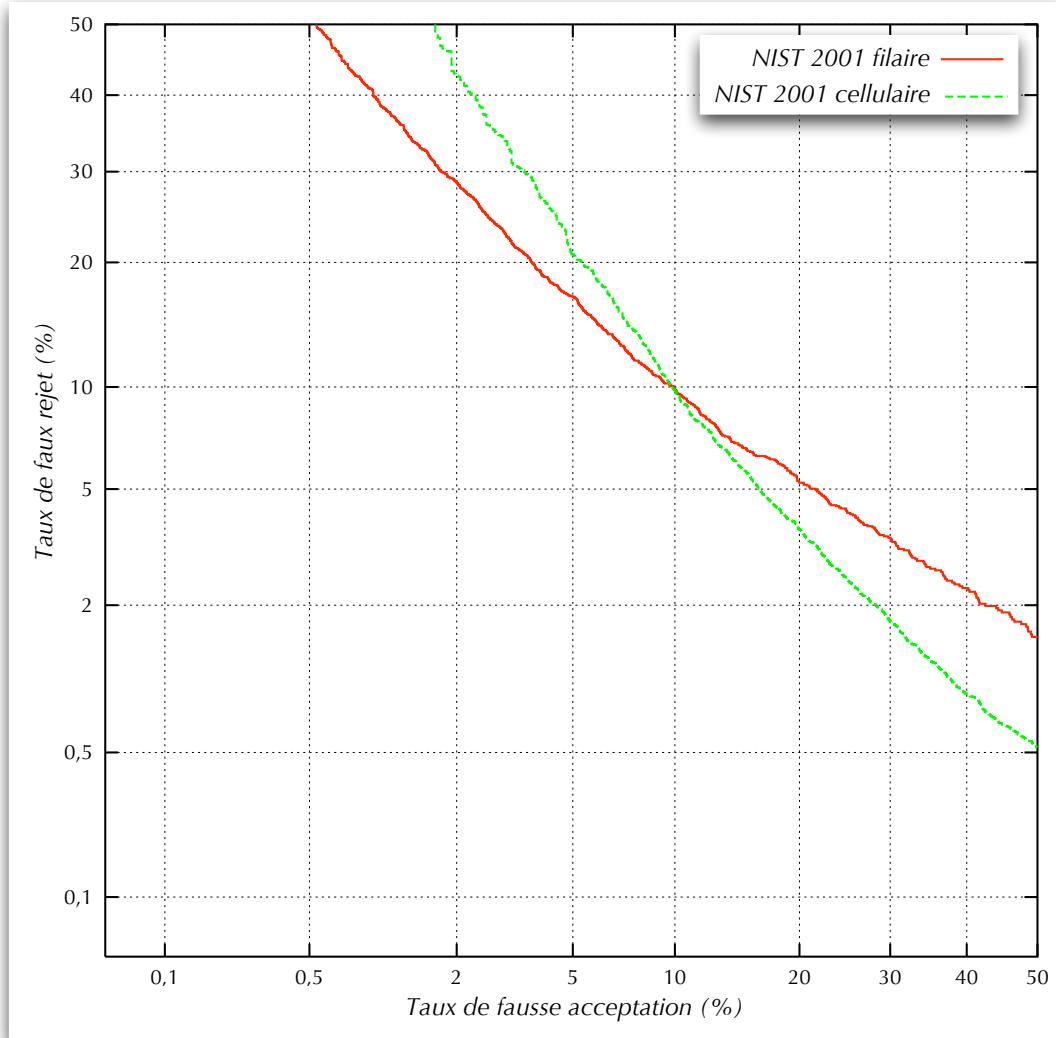


FIG. 5.3 – Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 2001 pour les corpus filaire et cellulaire.

Les résultats sur le corpus filaire (cf. figure 5.3), en nette progression par rapport à l'édition 2000, font apparaître l'intérêt de la suppression des trames de basse énergie. Les résultats obtenus sur le corpus cellulaire, en revanche, montrent la difficulté de la tâche (due notamment à la faible quantité de données de développement) associée au manque de préparation à cette tâche (aucun développement spécifique n'avait pu être

réalisé, toujours par manque de données).

5.5 NIST 2002

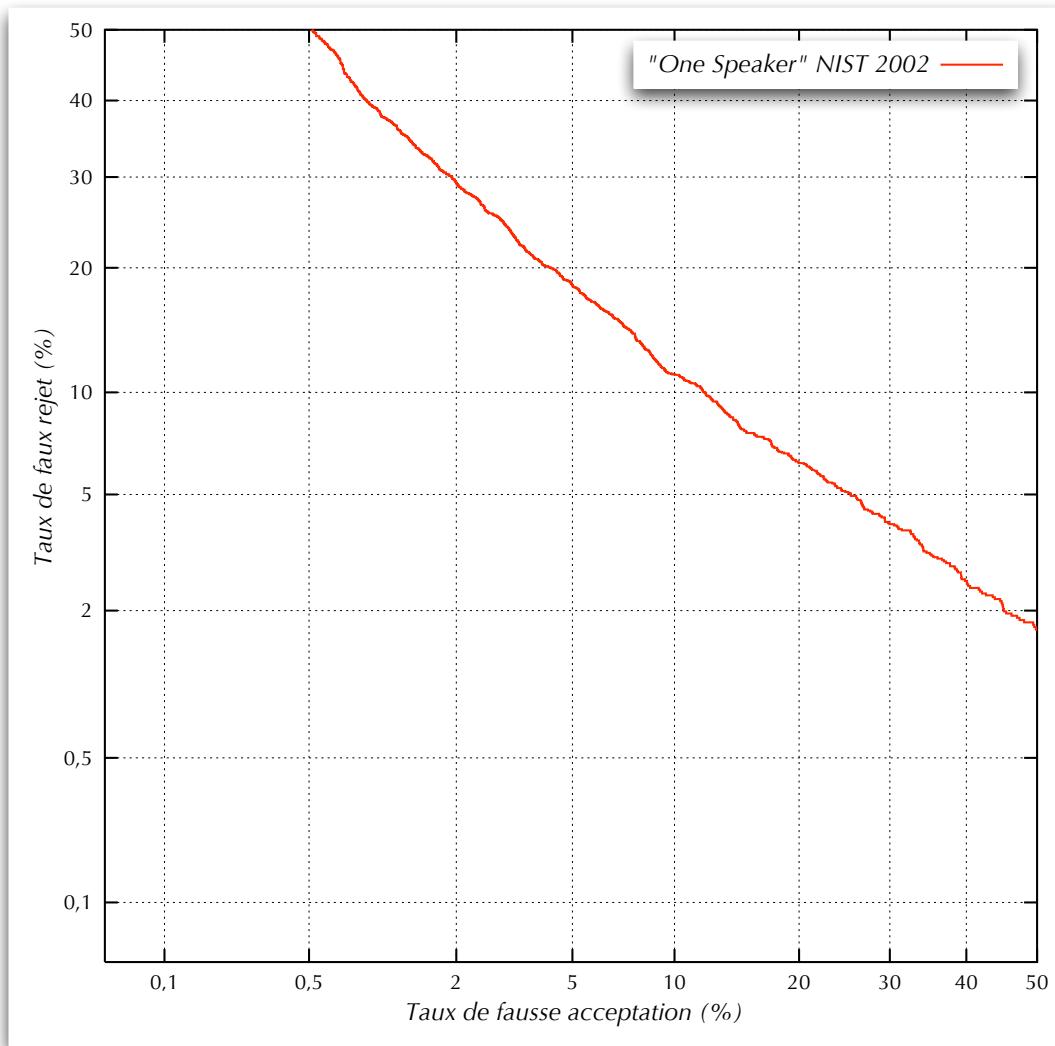


FIG. 5.4 – Résultats pour la tâche “One-Speaker Detection” de l’évaluation NIST 02 pour le corpus cellulaire.

La campagne d'évaluation NIST 2002 poursuit le mouvement amorcé lors de l'édition 2001 en adoptant pour la tâche de vérification du locuteur un jeu de données intégralement issues de téléphones cellulaires, extrait du corpus Switchboard Cellular Part 2, et composé de près de 400 locuteurs et 3500 enregistrements de test (soit une taille similaire aux corpus issus de téléphones filaires les années précédentes). Hormis l'absence d'information sur les types de combinés utilisés lors des enregistrements, les

conditions et règles de l'évaluation sont les mêmes que les années précédentes (apprentissage sur 2 minutes de parole enregistrées en une seule session, tests de quelques secondes à une minute, 11 identités vérifiées pour chaque enregistrement de test, etc.). Les données de développement sont les données cellulaires du corpus d'évaluation de la campagne 2001.

Plusieurs nouvelles techniques sont intégrées au système AMIRAL pour cette campagne. Le traitement post-paramétrisation s'enrichit d'une phase de normalisation des vecteurs cepstraux (telle que présentée page 71) en remplacement de la soustraction de moyenne cepstrale. Les expériences montrent que les meilleures performances sont obtenus en effectuant cette normalisation après la suppression des trames de basse énergie et en l'appliquant à tous les coefficients, statiques comme dynamiques.

La modélisation adopte le seuillage de la variance (cf. chapitre 4, page 90) lors de l'estimation d'un modèle du monde. Le modèle du monde utilisé est de type UBM (*Universal Background Model*) à 256 composantes, obtenu en combinant deux modèles du monde à 128 composantes dépendants du genre (la dépendance au combiné n'existant plus) appris, dans le cas du système primaire, à partir de données téléphoniques filaires et non cellulaires.

Enfin, les scores de vraisemblance ne sont plus normalisés par WMAP mais par T_{norm} .

La figure 5.4 montre les résultats obtenus par ce système.

5.6 NIST 2003

Le LIA n'a présenté aucun système de vérification du locuteur lors de la campagne d'évaluation NIST 2003, les efforts de développement cette année-là s'étant concentrés sur la participation à la campagne d'évaluation des tâches multi-locuteurs RT'03 ([Meignier 2004]) :

5.7 Bilan

Le tableau 5.7 récapitule l'historique de l'intégration de diverses techniques au système AMIRAL au cours des campagnes d'évaluation NIST de 1999 à 2002.

Parallèlement, la figure 5.5 présente sur un même graphe les courbes des résultats obtenus par AMIRAL lors de ces campagnes d'évaluation. L'évolution visible, d'année en année, vers de meilleures performances, suit assez logiquement le rythme d'introduction de nouvelles techniques au sein du système.

Il convient cependant de rappeler que, dans le même temps, la difficulté de la tâche considérée s'est accrue. Tout d'abord, le type de données utilisées pour l'apprentissage des modèles locuteurs est passé d'enregistrements multi-sessions à des enregistrements mono-session. Puis le corpus utilisé est passé de données téléphoniques filaires à des données téléphoniques cellulaires. Ces dernières sont plus difficiles pour les systèmes de RAL du fait de la variabilité accrue qu'elles présentent (provenant de la variabilité des types de combinés ainsi que de la distance au micro) et des dégradations induites par les

	00	01-F	01-C	02
Modèle du monde	4 (dép. genre & combiné)	4 (dép. genre & combiné)	4 (dép. genre & combiné)	2 (dép. genre)
Taille des modèles	16	128	128	256
Type de matrices	Pleines	Diagonales	Diagonales	Diagonales
Apprentissage locuteurs	EM	Adaptation	Adaptation	Adaptation
Normalisation des scores	WMAP	WMAP	WMAP	Tnorm
Coefficients Δ	✗	✓	✓	✓
Sup. des trames de basse énergie	✗	✗	✓	✓
Compensation de canal	CMS	CMS	CMS	Centrage/ réduction
Seuillage variance pour app. monde	✗	✗	✗	✓

TAB. 5.1 – Évolution d'AMIRAL pour le traitement de la tâche de vérification du locuteur — Historique de l'apparition des diverses techniques dans AMIRAL : le type de modèles du monde, le nombre de composantes des modèles de locuteurs et le type de matrices associées, l'algorithme d'apprentissage des modèles de locuteurs, la technique de normalisation des scores utilisée, l'utilisation de la dérivée première des paramètres, la suppression des trames de basse énergie, le traitement appliqué pour la compensation de canal, le recours au seuillage lors de l'estimation des modèles du monde.

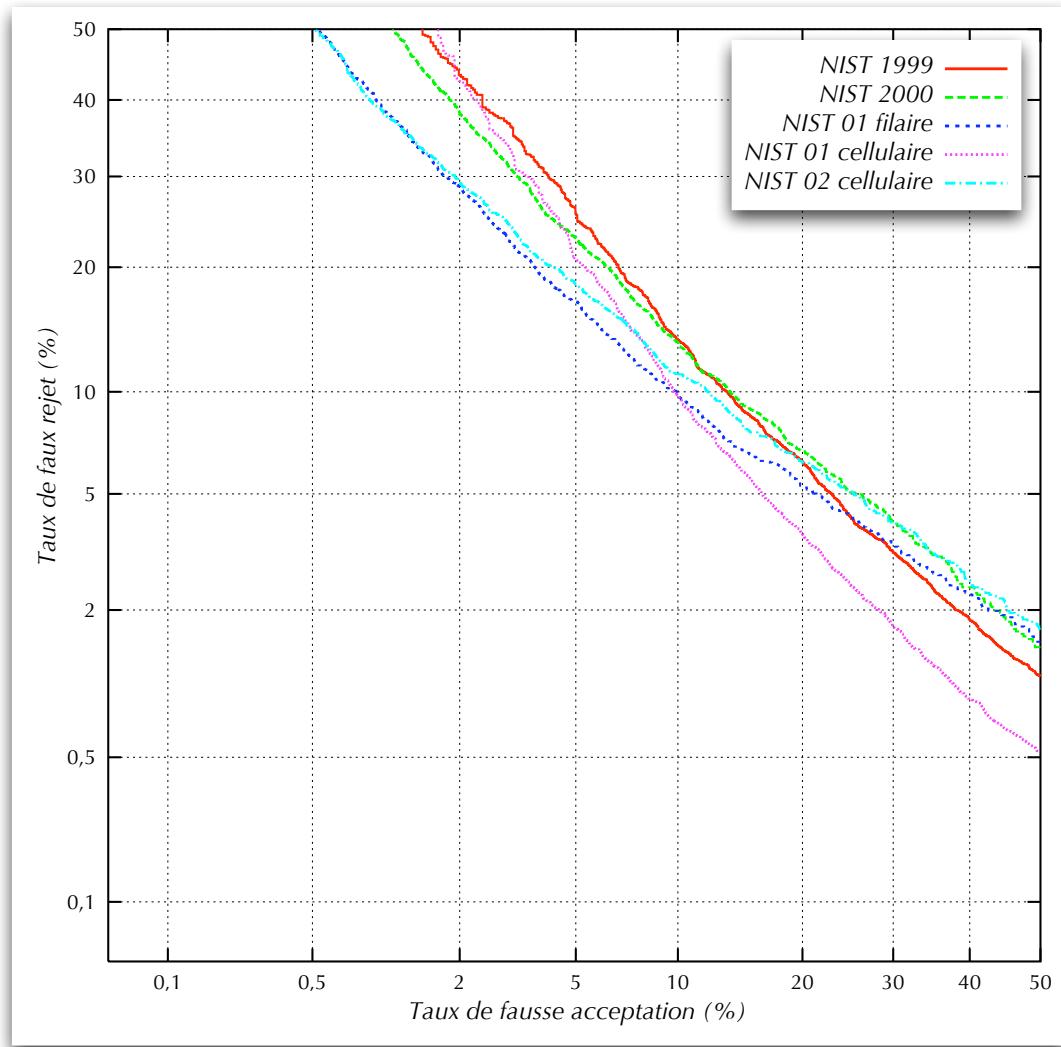


FIG. 5.5 – Évolution des résultats pour la tâche “One-Speaker Detection” des campagnes d’évaluation NIST de 1999 à 2002.

techniques de compression propres à la transmission du son sur réseau cellulaire. Les systèmes de vérification du locuteur présentés lors de la campagne d'évaluation NIST de 2002 sont bien plus performants que ceux de 1998 ou 1999, et la progression montrée ici est plus importante que ne le laisse penser l'écart observé entre les courbes de 1999 et de 2002. Le système AMIRAL à l'issue de la campagne de 2002 est un système à l'état de l'art, montrant des performances comparables à celles des meilleurs systèmes présentés à NIST en vérification du locuteur, et constituant une base solide pour les tâches multi-locuteurs (voir le chapitre 6 à ce sujet).

Cependant ce serait une erreur de ne considérer le bilan de ces années de développement et de participation régulière aux campagnes d'évaluation que du point de vue des performances du système AMIRAL. Ce bilan doit être vu également en termes d'expertise et d'expérience acquises à travers la participation à ces campagnes. L'acquisition de la méthodologie et de la rigueur nécessaires pour mener à bien les développements précédant chaque campagne, ainsi que l'expérience retirée de la confrontation régulière aux systèmes définissant l'état de l'art dans le domaine, font partie des bénéfices qu'amène, avec le temps, une participation régulière à des campagnes d'évaluation internationales.

Un exemple de l'expérience que peut apporter une telle participation est à voir dans la façon d'aborder et de tester les idées nouvelles dans le cadre du développement d'AMIRAL au LIA. Un peu anarchique les premières années, la façon d'analyser (et éventuellement rejeter ou intégrer) les pistes de recherche s'est organisée juste après la campagne NIST 2000 par l'encapsulation d'AMIRAL dans une plate-forme logicielle capable d'automatiser la génération, puis le lancement, d'une série d'expériences nécessaires pour valider un nouveau module ou analyser les effets de la variation d'une série de paramètres (*cf.* annexe B). Cette automatisation permet d'évaluer facilement le potentiel d'une nouvelle idée. Surtout, en permettant, après l'intégration de cette idée, de lancer aisément une analyse complète du fonctionnement du système, module par module, elle permet de mieux détecter l'impact négatif insoupçonné de cet ajout sur une autre partie du système. Cette rigueur est la bienvenue dans le cadre du développement d'un système complexe réalisant toutes les tâches de la reconnaissance automatique du locuteur.

L'expérience acquise au cours de la participation aux campagnes NIST est sans doute le point le plus positif de cette participation. Cette expérience a dores et déjà été mise à profit dans le cadre du projet ALIZÉ¹ du programme Technolangue² ([Bonastre 2005]). Projet de développement d'un système de reconnaissance du locuteur complet, gratuit et totalement ouvert, ALIZÉ a été réalisé en tirant profit des leçons du développement d'AMIRAL et des participations aux campagnes d'évaluations NIST.

¹<http://www.lia.univ-avignon.fr/heberges/ALIZE/index.html>

²<http://www.technolangue.net>

Chapitre 6

Les tâches multi-locuteurs lors des campagnes NIST

Sommaire

6.1 Suivi de locuteur et segmentation en locuteurs	103
6.1.1 Utilisation d'une méthode de détection de ruptures	104
6.1.2 HMM évolutif	109
6.2 Détection de locuteur dans des documents bi-locuteurs	113
6.2.1 Apprentissage sur données mono-locuteur	113
6.2.2 Apprentissage à partir de données multi-locuteurs (NIST 2002) . .	116

6.1 Suivi de locuteur et segmentation en locuteurs

Devant l'extension de la reconnaissance automatique du locuteur au traitement de documents multimédia multi-locuteurs et le besoin croissant de pouvoir évaluer les performances des nouveaux outils qui en résultent, des tâches correspondant à ces applications multi-locuteurs ont commencé à apparaître dans les campagnes d'évaluation NIST à partir de l'édition 1999, sous la forme de suivi de locuteur tout d'abord, puis de segmentation en locuteurs (*cf. chapitre 3, page 52*). Dans la définition de ces nouvelles tâches, le travail à réaliser consiste à découper les documents traités en segments correspondant chacun à l'intervention d'un locuteur. La différence essentielle tient à ce que, dans le cas du suivi de locuteur, le locuteur considéré est connu au préalable, contrairement à ce qui se passe dans le cas de la segmentation en locuteurs. Cependant, dans les deux cas, le nombre d'interventions différentes, la durée moyenne de ces interventions, et selon les cas le nombre de locuteurs, ne sont pas connus à l'avance par le système. Cette forte similarité entre les deux tâches a conduit à les traiter avec les mêmes outils.

Dans le cadre du système AMIRAL, nous avons exploré deux stratégies différentes pour élaborer des systèmes de suivi de locuteur et de segmentation en locuteurs. La première s'appuie sur un système de détection des changements de locuteurs, en amont du système de reconnaissance du locuteur proprement dit. Dans le cas du suivi de

locuteur, une fois le signal segmenté, un système classique de vérification du locuteur est mis en œuvre indépendamment sur chacun des segments obtenus pour déterminer s'il a été prononcé ou non par le locuteur cible. Dans le cas de la segmentation en locuteurs, une phase d'agrégation des segments similaires est réalisée, afin de regrouper toutes les interventions de chaque locuteur, la fin de cette phase d'agrégation permettant de déduire le nombre de locuteurs du document.

La seconde approche, modélisant les changements de locuteur par un modèle de Markov caché (HMM), intègre en un seul et même processus la détection des segments et leur regroupement en locuteur, cherchant à tirer profit à chaque instant de toutes les informations disponibles.

Ces deux approches sont présentées ci-après, ainsi que les résultats obtenus pour les tâches de suivi de locuteur et de segmentation en locuteur lors des évaluations NIST.

6.1.1 Utilisation d'une méthode de détection de ruptures

Description

Cette approche, présentée dans [Bonastre 2000b] et [Bonastre 2000a], a été réalisée en collaboration avec l'institut Eurécom (Sophia-Antipolis).

Elle consiste à réaliser une phase préliminaire de détection des changements de locuteurs, dont le but est de proposer une segmentation du message en segments homogènes (prononcés par un seul locuteur). Cette étape de pré-segmentation n'utilise aucune connaissance *a priori* des locuteurs engagés dans la conversation, même dans le cadre du suivi de locuteur. Les informations éventuellement disponibles (telles que le nombre de locuteurs ou l'identité de certains d'entre eux) ne sont exploitées que lors d'une seconde phase. Celle-ci se base sur les méthodes de reconnaissance du locuteur offertes par AMIRAL pour réaliser la tâche visée (suivi de locuteur ou segmentation en locuteurs) en prenant comme point de départ les segments obtenus précédemment.

Cette approche a été testée pour la tâche de suivi de locuteur lors des campagnes d'évaluation NIST de 1999 et 2000. En 2000, elle a également été testé pour la tâche de segmentation.

Détection des changements de locuteur Cette phase correspond à la contribution de l'institut Eurécom. Le principe n'en est que brièvement exposé ici, le lecteur étant invité à se reporter au document de thèse de Perrine Delacourt ([Delacourt 2000]) pour une présentation complète et détaillée de cette technique.

Etant données deux portions de signal paramétrisées (deux séquences de vecteurs acoustiques) $W_1 = \{x_{i-m}, \dots, x_i\}$ et $W_2 = \{x_{i+1}, \dots, x_{i+n}\}$, nous considérons le test d'hypothèses suivant pour un changement de locuteur à l'instant i :

- hypothèse H_0 : les deux portions sont relatives au même locuteur ; leur réunion $W = W_1 \cup W_2$ est modélisée par un unique processus gaussien : $\mathcal{W} = \mathcal{N}(\mu_W, \Sigma_W)$
- hypothèse H_1 : chaque portion a été prononcée par un locuteur différent et est modélisée par un processus gaussien différent : $\mathcal{W}_1 = \mathcal{N}(\mu_{W_1}, \Sigma_{W_1})$ et $\mathcal{W}_2 = \mathcal{N}(\mu_{W_2}, \Sigma_{W_2})$

Le rapport de vraisemblance généralisé, R , entre les hypothèses H_0 et H_1 est défini par :

$$R = \frac{L(W|\mathcal{W})}{L(W_1|\mathcal{W}_1) \times L(W_2|\mathcal{W}_2)}$$

La distance d_R est obtenue en prenant le logarithme de l'expression précédente : $d_R = -\log R$ (“distance” est ici un abus de langage car d_R ne vérifie pas les propriétés d'une distance).

Une valeur élevée de R (*i.e.* une faible valeur de d_R) signifie que la modélisation par une seule distribution gaussienne (hypothèse H_0) s'accorde mieux aux données. A l'opposé, une faible valeur de R (*i.e.* une forte valeur de d_R) indique que l'hypothèse H_1 , *i.e.* la modélisation par deux distributions gaussiennes, correspond mieux aux données. Dans ce cas, un changement de locuteur est détecté à l'instant i .

La distance d_R est calculée pour chaque couple de fenêtres de signal de même durée (environ 2 secondes). Ces fenêtres doivent être suffisamment longues pour estimer de manière fiable les paramètres des gaussiennes et suffisamment courtes pour faire l'hypothèse qu'elles ne contiennent les paroles que d'un seul locuteur. Ces fenêtres sont glissantes et sont déplacées à chaque itération d'un laps de temps fixe (environ 0,1 seconde) le long du signal paramétrisé, comme le montre la figure 6.1.

Les distances calculées pour chaque couple de fenêtres sont stockées pour former à la fin du processus une courbe de distances. Les pics les plus significatifs (en termes d'amplitude) de cette courbe sont alors détectés : ces pics correspondent aux changements de locuteur recherchés. Un maximum local de la courbe des distances est considéré comme significatif si les différences entre son amplitude et celle des minima situés de part et d'autre sont supérieures à un certain seuil (dépendant de la variance de la distribution des distances). Un intervalle de temps minimal entre deux changements de locuteurs consécutifs est également imposé. La détection des changements de locuteurs ne se fait donc pas en considérant l'amplitude absolue des pics mais plutôt en considérant leur facteur de forme.

Un changement de locuteur non détecté s'avère plus préjudiciable pour les tâches visées qu'un changement qui est détecté alors qu'il n'existe pas. En effet, un segment contenant les paroles de plusieurs locuteurs (résultant d'une détection manquée) sera plus difficilement identifié comme correspondant (ou non) au locuteur cible dans le cas du suivi de locuteur. Aussi, les paramètres impliqués dans la détection des changements de locuteurs ont été ajustés de manière à éviter les détections manquées au détriment du nombre de fausses alertes. Le signal est probablement sur-segmenté : les paroles consécutives d'un même locuteur sont réparties sur plusieurs segments. Cependant, la durée des segments de locuteurs obtenus est suffisamment grande pour avoir une décision de vérification fiable.

Suivi de locuteur L'utilisation d'une segmentation préliminaire en locuteurs avant le processus de vérification repose sur l'idée suivante : le score de vérification est plus fiable si les segments considérés ne comportent que des trames (vecteurs acoustiques) provenant d'un unique locuteur. De même, la longueur des segments influe très fortement sur la qualité des résultats ([Magrin-Chagnolleau 1999]). Ainsi, le but de la segmentation en locuteurs est de découper le signal de parole en plusieurs segments homogènes : chaque segment ne doit contenir des paroles prononcées par un seul locuteur.

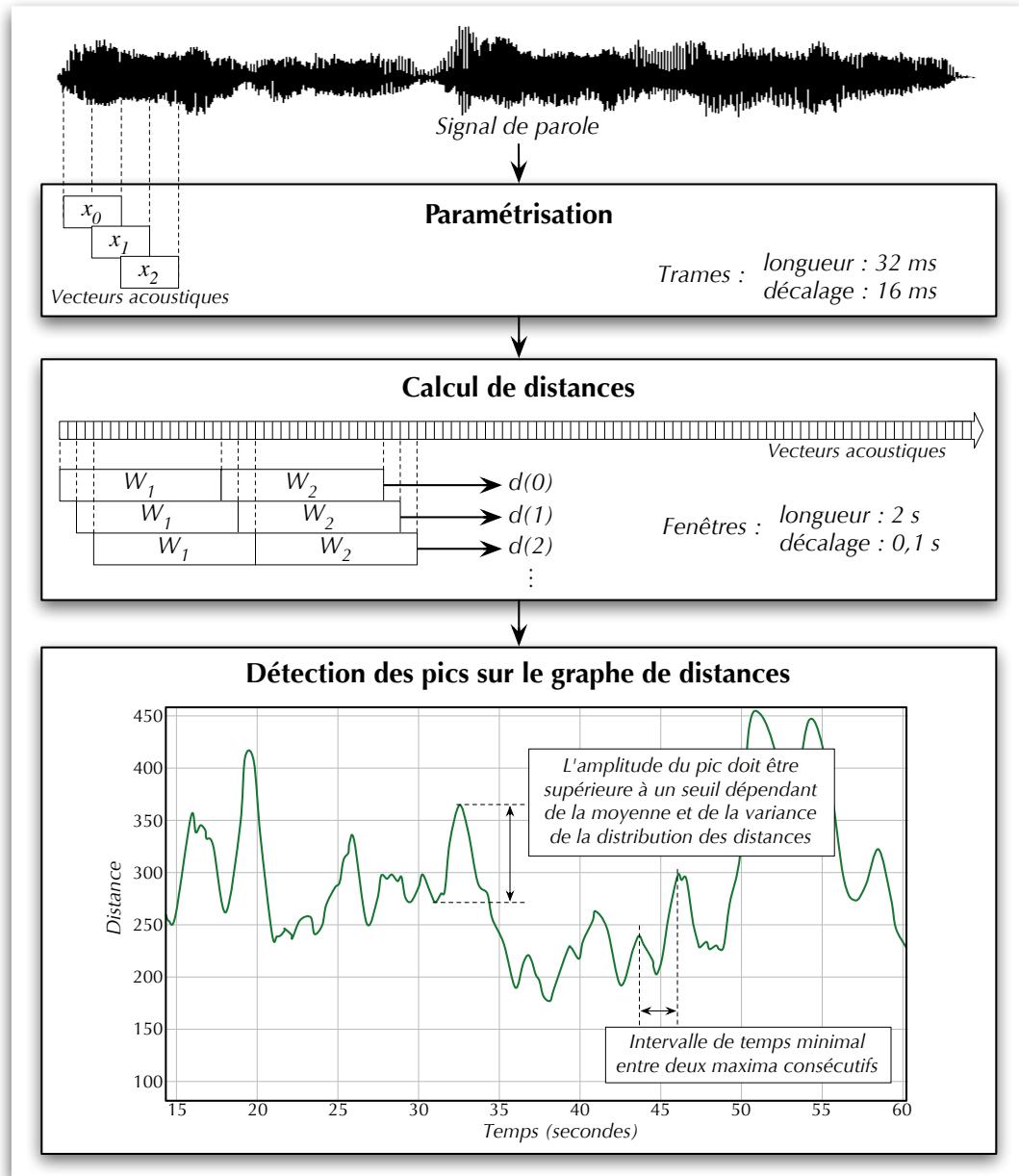


FIG. 6.1 – Détection des changements de locuteur – Calcul de distance par fenêtres glissantes.

Pour cette méthode, les segments temporels générés à l'étape précédente sont considérés comme mono-locuteur (et mono condition d'enregistrement) et le système AMIRAL est configuré en conséquence : une simple moyenne arithmétique est utilisées comme procédé de fusion temporelle (pour obtenir la mesure de similarité finale, à partir des mesures provenant de chaque bloc temporel). Un procédé de décision classique, basé sur un seuil de décision, est appliqué sur chaque segment et permet de l'attribuer (ou non) au locuteur cible.

Segmentation en locuteurs La tâche de segmentation est traitée par un processus itératif à partir des segments générés à l'étape précédente en détectant les changements de locuteur. Chacun de ces segments étant considéré comme mono-locuteur, le processus décrit ici tente de les regrouper en ensembles homogènes, chaque groupe ainsi obtenu correspondant à l'ensemble des interventions d'un même locuteur. Ce processus est illustré par la figure 6.2,

Dans un premier temps, un modèle de locuteur est estimé sur le plus long de ces segments. Ce choix du segment le plus long est justifié par la volonté d'obtenir la meilleure qualité d'estimation possible. Un test de vérification du locuteur est ensuite effectué par rapport à ce modèle sur tous les autres segments. Les segments pour lesquels la vraisemblance obtenue est supérieure au seuil de décision (le même seuil utilisé pour la tâche de vérification du locuteur) sont affectés à ce locuteur.

Le processus est ensuite réappliqué sur l'ensemble des segments encore non affectés à un locuteur : recherche du segment le plus long, estimation d'un modèle de locuteur, calcul du score des autres segments non affectés.

L'arrêt du processus se fait en fonction de trois critères :

- s'il ne reste plus aucun segment à attribuer ;
- si le nombre maximum de locuteurs fixé *a priori* est atteint ; l'ensemble des segments qui n'ont pas encore été attribués à un quelconque locuteur sont alors affectés au dernier locuteur détecté ;
- enfin, si le plus long segment non attribué est de longueur inférieure à x secondes ; il est alors considéré que les segments non attribués correspondent à des erreurs d'affectation (de type faux rejet) aux locuteurs précédemment détectés ; ces segments sont affectés au dernier locuteur détecté.

Résultats

La figure 6.3 montre les résultats obtenus par chacun des participants à la campagne NIST 99, et en particulier par le système présenté ici. Ces résultats sont fournis sous forme d'une courbe DET montrant les fausses acceptations (blocs attribués à tort au locuteur cible) en fonction des faux rejets (blocs prononcés par le locuteur cible et rejetés à tort par le système). Les résultats sont calculés à raison d'une décision tous les centièmes de seconde.

Ces résultats montrent peu d'écart entre les différents compétiteurs de la campagne NIST 99. La différence entre les systèmes est masquée d'une part par la difficulté intrinsèque de la tâche (les écarts entre les différents compétiteurs lors de la campagne NIST 99 étaient très faibles, pour le suivi de locuteur) ainsi que par le mode de calcul des performances, qui pénalise de la même manière toute erreur. En particulier, une erreur de positionnement d'une frontière de segment, de 1/100 de seconde, est comptabilisée

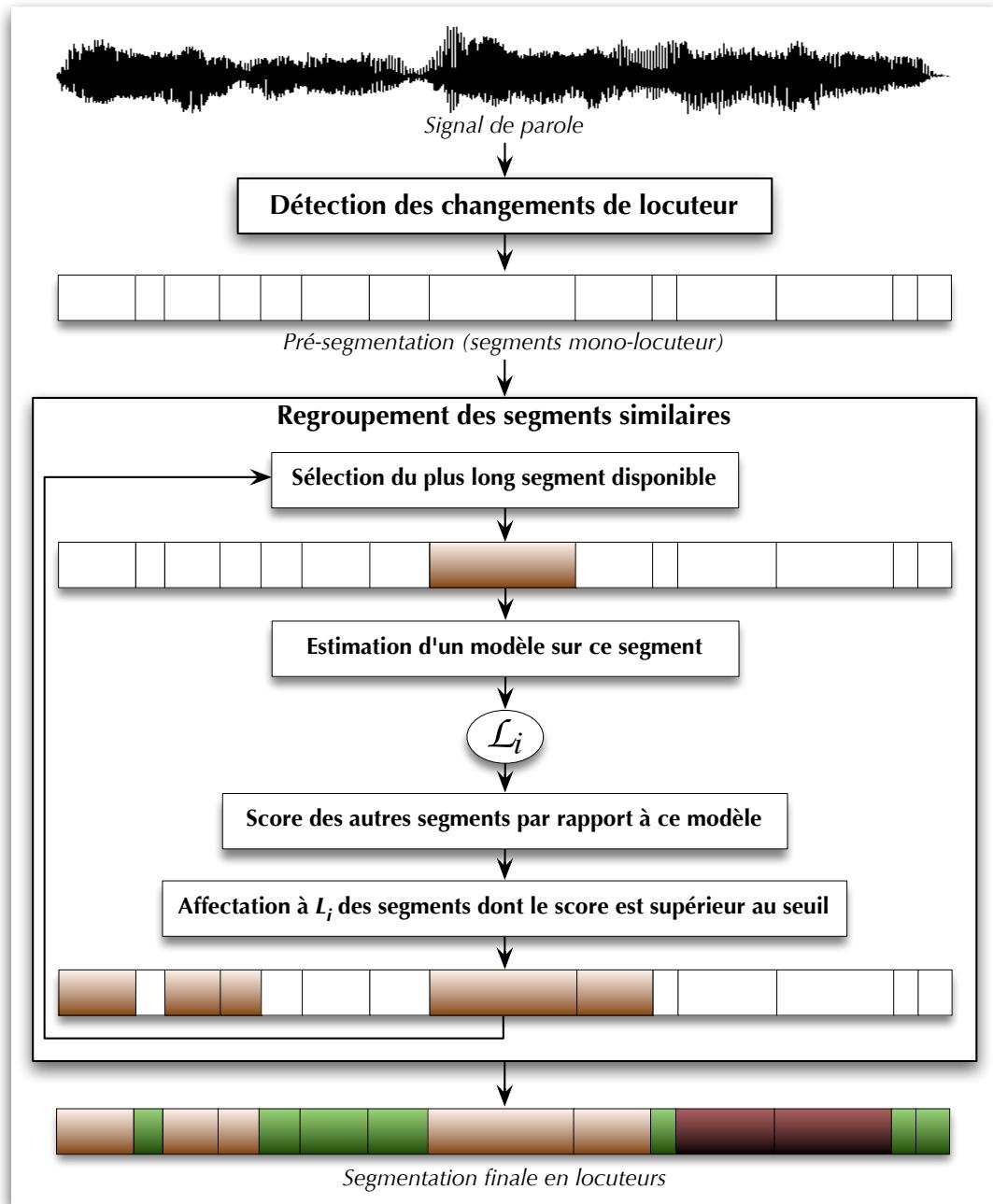


FIG. 6.2 – Segmentation en locuteurs – Illustration du processus d’agrégation des segments utilisé lors de la seconde phase de l’approche basée sur la détection de ruptures.

au même niveau qu'une fausse détection de segment.

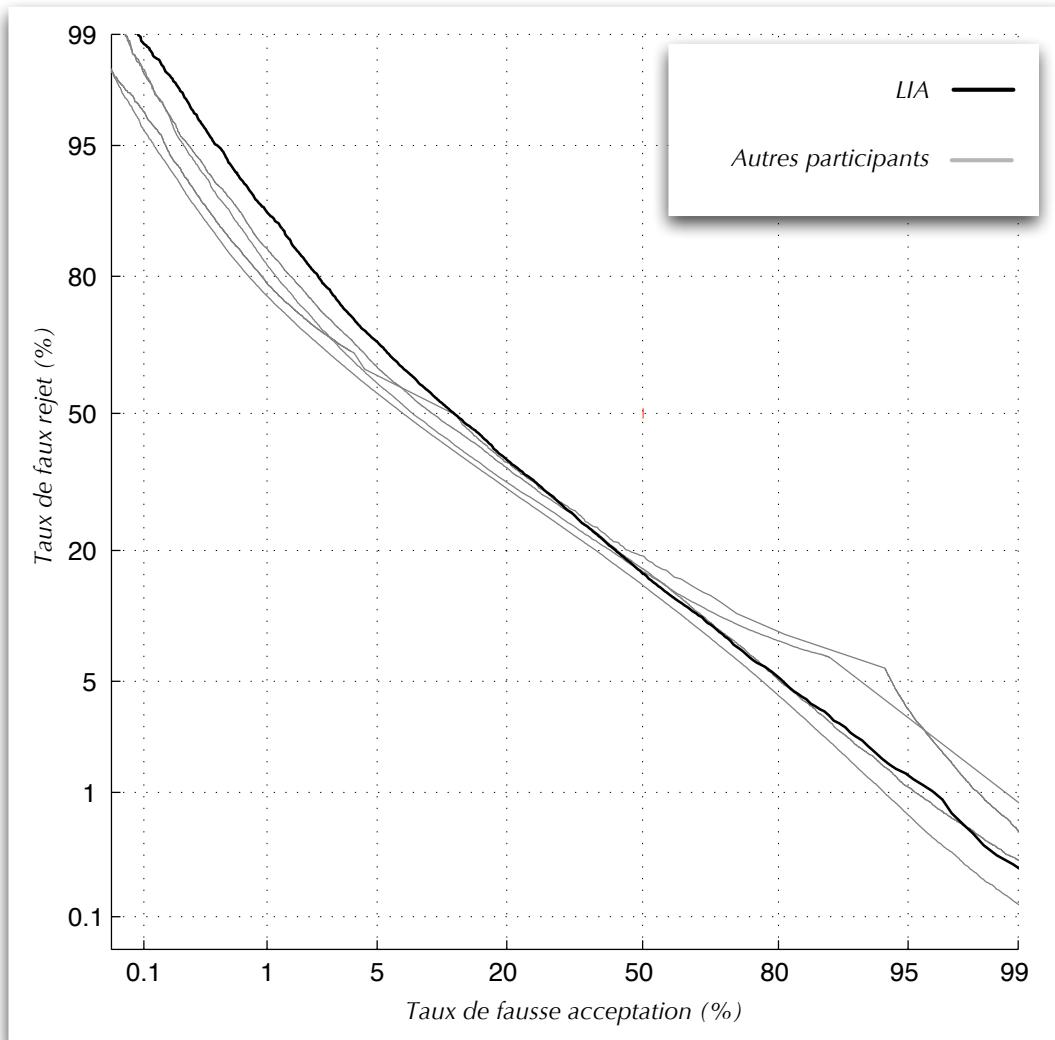


FIG. 6.3 – Suivi de locuteur – Résultats obtenus lors de la campagne d'évaluation NIST 99 par l'approche basée sur une détection de ruptures suivie d'une phase de vérification du locuteur.

6.1.2 HMM évolutif

Description

La technique exposée dans la section précédente correspond à l'approche de la segmentation en locuteurs généralement proposée dans la littérature. Elle se compose de deux étapes : une détection des ruptures dans le signal suivie d'une classification des

segments engendrés par les ruptures en classes de locuteurs. Une des limites de cette approche porte notamment sur l'incapacité de prendre en compte les informations mises en évidence lors de la classification pour aider à la détection des ruptures dans le signal.

Pour pallier ce problème de perte d'informations engendré par la méthode classique, une méthode reposant sur un modèle de conversation utilisant un modèle de Markov caché (HMM) évolutif a été proposée ([Meignier 2000], [Meignier 2001], [Meignier 2002a]). Dans cette approche, toutes les informations disponibles sont exploitées à chaque étape et remises en cause à l'étape suivante. Dans le modèle de conversation proposé, les états du HMM représentent les locuteurs du document ; les transitions entre ces états modélisent les changements de locuteur. Le processus de segmentation est itératif : les locuteurs (correspondant aux états du HMM) sont ajoutés un à un à chaque itération, d'où le terme de HMM évolutif.

La méthode proposée réalise l'ajout d'un locuteur à la segmentation en trois étapes présentées ci-après. Cet ajout est réitéré jusqu'à ce qu'aucun locuteur ne puisse plus être ajouté. L'ensemble du processus est illustré par la figure 6.4.

Initialisation Un premier modèle de locuteur, \mathcal{L}_0 , est appris sur l'ensemble des données du document. La segmentation est modélisée par un HMM contenant un seul état et l'ensemble du signal est affecté au locuteur L_0 . $\{L_0\}$ représente alors l'ensemble des locuteurs du document.

Étape 1 - Ajout d'un nouveau locuteur Un nouveau locuteur est extrait des segments libellés L_0 , qui représentent les locuteurs qui n'ont pas encore été détectés. Le modèle de ce nouveau locuteur est appris en utilisant les 3 secondes consécutives de signal qui maximisent le rapport de vraisemblance entre le modèle \mathcal{L}_0 et le modèle du monde. Cette durée de 3 secondes pour le segment initial a été fixée expérimentalement de manière à être suffisante pour que le modèle de locuteur soit robuste tout en étant assez courte pour avoir une forte probabilité de ne contenir qu'un seul locuteur. Ces conditions sont équivalentes aux hypothèses des méthodes de détection de ruptures. Cette stratégie sélectionne les données les plus proches du modèle de locuteur \mathcal{L}_0 .

Le modèle de locuteur \mathcal{L}_i correspondant à l'état L_i (i étant le numéro de l'itération) est ajouté au précédent HMM. Les probabilités de transition sont mises à jour pour vérifier les conditions suivantes :

$$\left\{ \begin{array}{l} \forall i, a_{i,i} = \gamma \\ \forall (i,j), i \neq j, a_{i,j} = \frac{1-\gamma}{N-1} \\ 0 < \gamma < 1 \end{array} \right. \quad (6.1)$$

où $a_{i,j}$ représente la probabilité de transition de l'état i vers l'état j ($a_{i,i}$ étant la probabilité de boucler sur l'état i) et N est le nombre d'états du HMM. Le poids γ permet de fixer la probabilité de bouclage sur un même état. Les probabilités de transition vers les autres états sont équiprobables et déduites de la probabilité de bouclage. Une valeur de $\gamma = 0,6$ est typiquement utilisée dans le cadre des campagnes d'évaluation NIST.

Enfin, le segment de 3 secondes (S_i sur la figure 6.4) ayant servi à l'apprentissage du modèle \mathcal{L}_i est libellé L_i au lieu de L_0 .

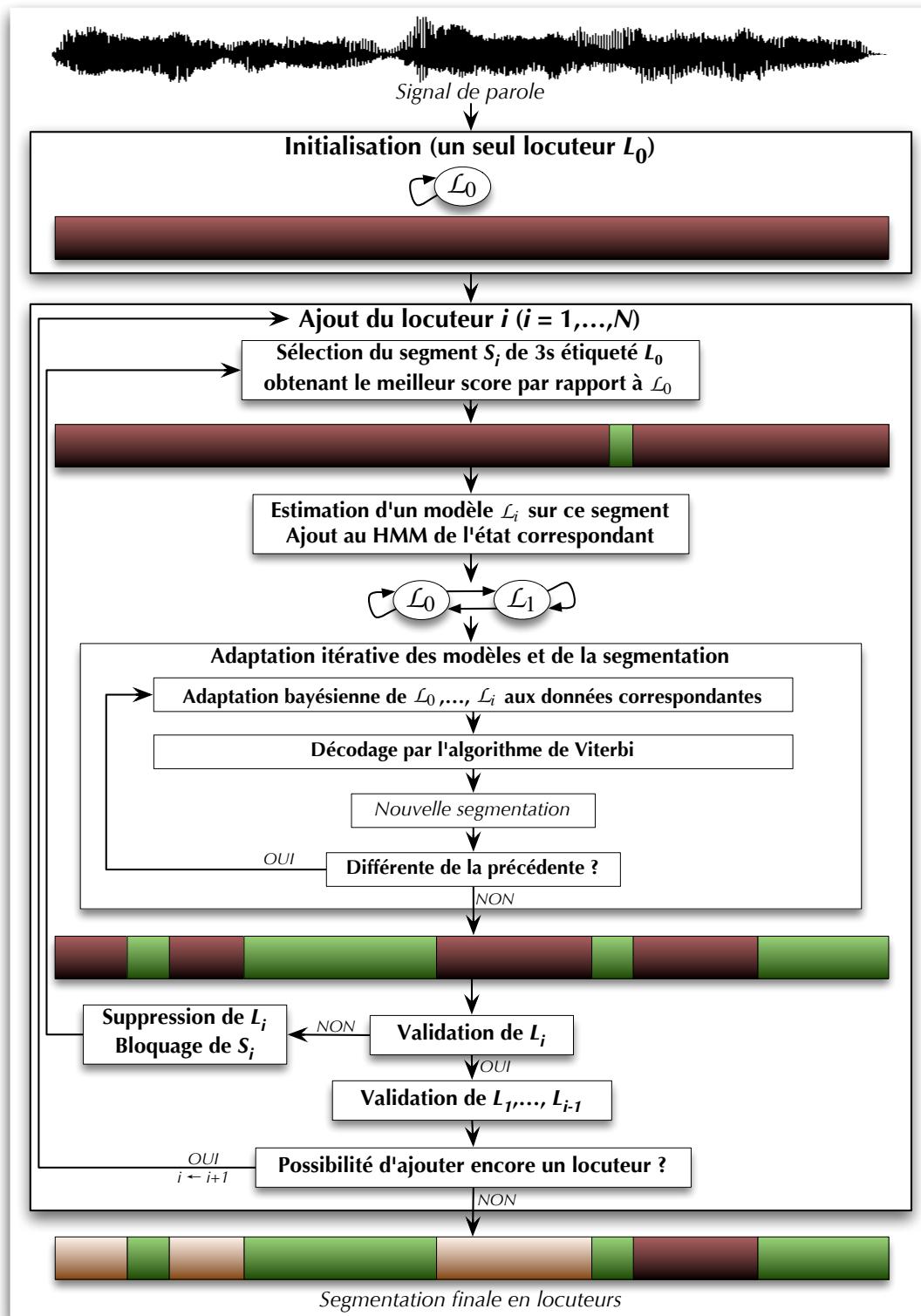


FIG. 6.4 – Algorithme de segmentation par HMM évolutif.

Étape 2 - Adaptation des modèles Cette étape permet de détecter les segments correspondant au nouveau locuteur L_i et de remettre en cause les segments des autres locuteurs. En premier lieu, tous les modèles de locuteurs sont réappris (par adaptation bayésienne du modèle du monde) aux données qui leur ont été affectées dans la segmentation courante. Puis l'algorithme de Viterbi est employé pour générer une nouvelle segmentation. L'adaptation des modèles et le décodage sont renouvelés tant qu'il existe des différences entre deux segmentations successives générées ainsi.

Étape 3 - Validation des locuteurs et test du critère d'arrêt L'ajout d'un nouveau modèle de locuteur est validé à la fin de chaque itération. Cette validation concerne le nouveau locuteur lui-même mais aussi tous les locuteurs (L_1, \dots, L_{i-1}) ajoutés avant lui. Les deux règles suivantes sont appliquées :

- Le nouveau locuteur est rejeté si la durée totale des segments de signal qui lui sont affectés est inférieure à 4 secondes. Dans ce cas, les trois secondes de signal utilisées pour l'apprentissage initial du modèle de ce locuteur ne seront plus utilisées pour cette fonction lors de l'étape 1. Le processus reprend alors à l'étape 1 avec la segmentation précédente à $i - 1$ locuteurs et le HMM correspondant.
- Les locuteurs précédemment ajoutés dont la durée totale des segments est inférieure à la durée des segments du nouveau locuteur sont retirés de la segmentation (et du HMM). Cette règle, qui force la détection en premier des locuteurs parlant le plus, est en relation avec les règles d'évaluation utilisées lors des campagnes d'évaluation NIST (cf. chapitre 3, page 56). Ces règles rendent moins coûteux, en termes d'erreur, de manquer les locuteurs s'exprimant peu dans les documents.

Dans le cas où le nouveau locuteur n'a pas été rejeté, le critère d'arrêt de la boucle d'ajout de locuteurs est testé. Ce critère repose sur l'impossibilité d'ajouter un nouveau locuteur. Il est vérifié dans l'un des cas suivants :

- il n'existe plus dans la segmentation de segment potentiel de trois secondes libellé L_0 ;
- tous les segments de trois secondes ont déjà été testés sans succès (aboutissant au rejet du locuteur correspondant).
- le nombre maximum de locuteurs a été atteint (si un tel nombre est défini ; tel est le cas dans le cadre des campagnes d'évaluation NIST).

A la différence des méthodes généralement utilisées dans la littérature, cette méthode repose sur une approche par modèle de locuteur qui permet de tirer parti des connaissances et des méthodes issues de la reconnaissance automatique du locuteur. En particulier, les méthodes de paramétrisation acoustique du signal, la modélisation de locuteurs par modèles multi-gaussiens (GMM) et les méthodes d'apprentissage (EM, adaptation bayésienne) sont utilisées ici. Ce cadre statistique permet de conserver une logique du maximum de vraisemblance au cours des différentes étapes de la segmentation.

Validation

Cette approche a été validée lors des campagnes d'évaluation NIST de 2001 à 2003. Lors de la participation à la campagne de 2002, réalisée en collaboration avec le laboratoire CLIPS¹ de Grenoble, ce système a été classé premier pour la tâche de segmen-

¹<http://www-clips.imag.fr>

tation de conversations téléphoniques (segmentation en deux locuteurs) et second sur la segmentation d'enregistrements de réunions (respectivement $\simeq 16\%$ et $\simeq 53\%$ de taux d'erreurs).

En 2003, toujours dans le cadre d'une participation commune avec le CLIPS, le système de segmentation a été classé deuxième pour la tâche de segmentation en locuteurs de journaux télévisés (taux d'erreurs d'environ 16%).

Cependant, malgré l'intérêt de la méthode, l'analyse des résultats obtenus révèle que la détection automatique du nombre de locuteurs reste problématique de même que pour les autres méthodes proposées dans la littérature.

Utilisation dans le cadre du suivi de locuteur

La technique de segmentation par HMM évolutif décrite ci-dessus a été utilisée également dans le cadre du suivi de locuteur lors de la campagne d'évaluation NIST de 2001. Toutefois, il s'agit moins d'une adaptation de la méthode à une autre problématique que d'un exemple d'application d'une technique de segmentation en aveugle dans le cadre du suivi de locuteur.

Dans l'approche proposée, la technique de segmentation est appliquée comme première phase du traitement d'un fichier de test. La tâche de suivi de locuteur définie dans les campagnes d'évaluation NIST portant sur des enregistrements bi-locuteurs, le système est réglé de façon à générer systématiquement une segmentation en deux locuteurs. Les deux ensembles de segments détectés sont comparés séparément au modèle du locuteur recherché, produisant un score pour chacun des deux locuteurs composant le fichier de test. Ce score permet de déterminer lequel de ces deux locuteurs est le plus proche du locuteur recherché. La définition de la tâche imposant de déterminer en premier lieu si le locuteur recherché est réellement présent dans l'enregistrement, une phase de décision classique correspondant à la vérification du locuteur est alors effectuée afin de déterminer si le locuteur le plus proche est réellement le locuteur recherché. En cas de réponse positive, les segments correspondants complètent la réponse du système.

6.2 Détection de locuteur dans des documents bi-locuteurs

6.2.1 Apprentissage sur données mono-locuteur

Au cours de la campagne d'évaluation NIST 2001, la tâche de détection de locuteur dans un document bi-locuteurs a été traitée en faisant appel à la technique de segmentation à base de HMM évolutif présentée à la section précédente.

L'approche retenue pour cette tâche est très similaire à celle utilisée dans le cadre du suivi de locuteur la même année, différant principalement par la phase de décision. Le fichier de test est tout d'abord séparé en deux ensembles de segments, chaque ensemble correspondant à un locuteur. Une phase de vérification du locuteur est alors appliquée à chaque ensemble de segments pour le comparer au modèle du locuteur re-

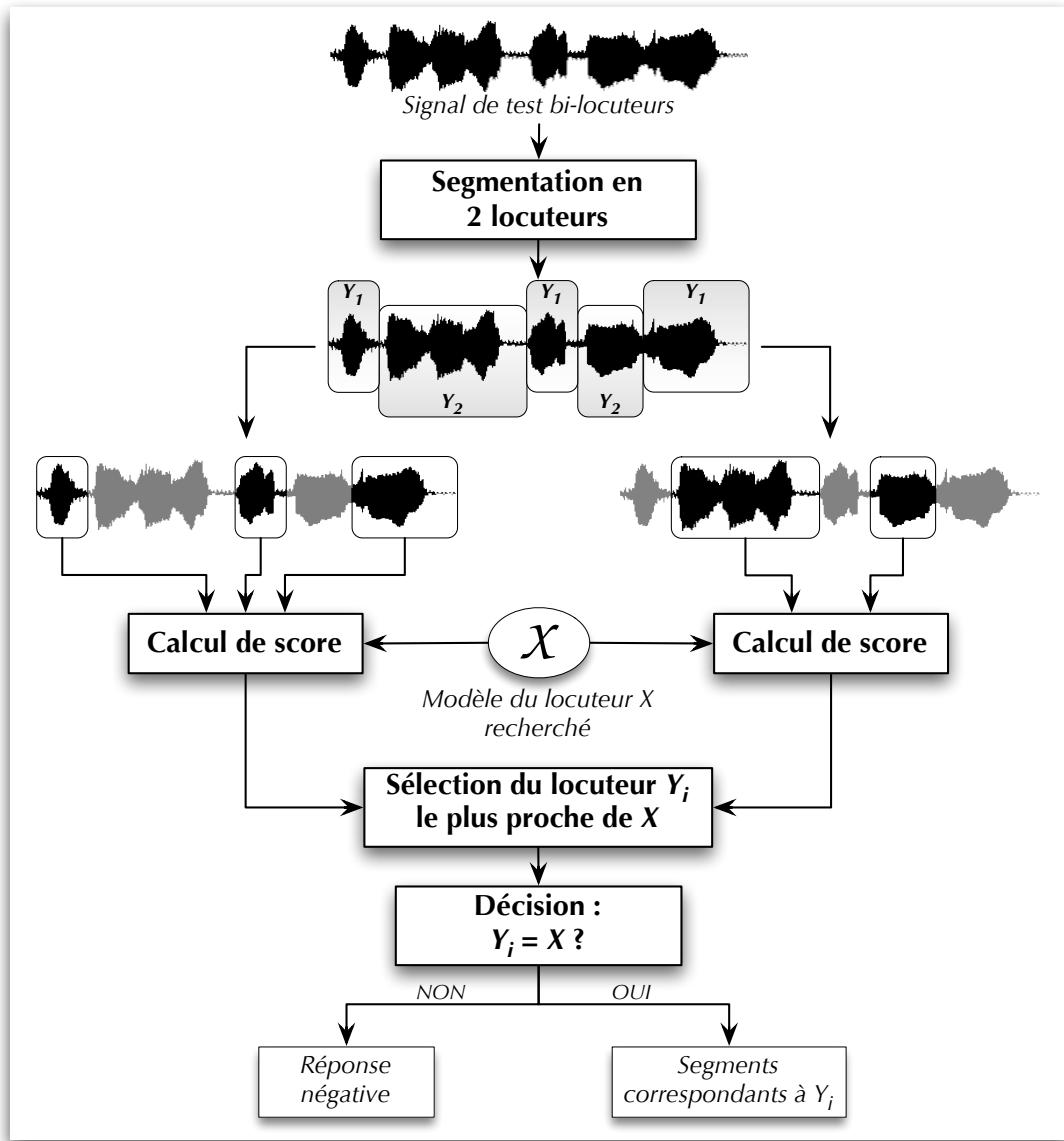


FIG. 6.5 – Suivi de locuteur — Technique à base de HMM évolutif.

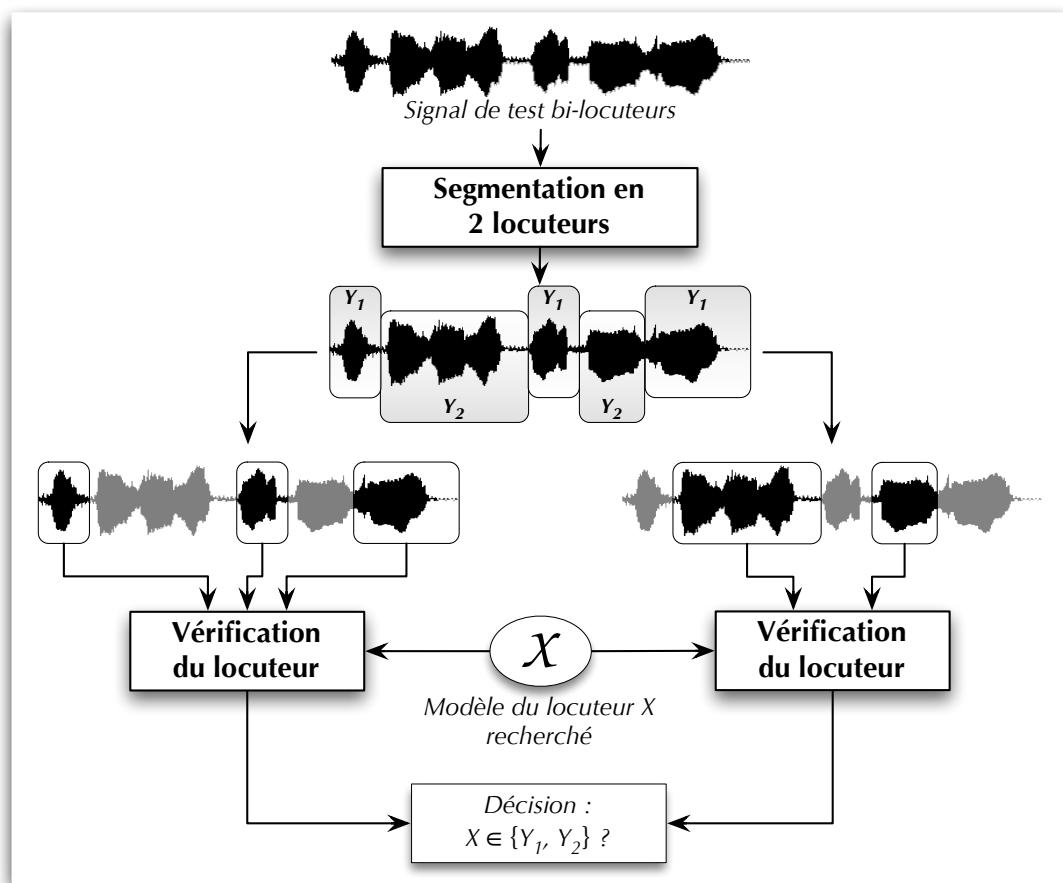


FIG. 6.6 – Détection de locuteur dans un document bi-locuteurs — La segmentation en 2 locuteurs est réalisée par la technique du HMM évolutif.

cherché, la réponse finale du système étant négative si aucun des deux ensembles ne lui correspond, positive sinon.

6.2.2 Apprentissage à partir de données multi-locuteurs (NIST 2002)

Techniques utilisées

L'approche présentée ici pour réaliser l'apprentissage d'un modèle de locuteur à partir de documents multi-locuteurs, introduite dans [Bonastre 2003b], repose sur deux phases, l'une de segmentation en locuteurs, l'autre d'appariement en locuteurs ([Meignier 2001], [Meignier 2002b]). La première phase segmente les enregistrements en deux classes correspondant chacune à un locuteur, permettant ainsi de se placer ensuite dans le cadre d'enregistrements mono-locuteur. La seconde phase consiste à sélectionner, parmi toutes les classes générées lors de la première phase, les données du locuteur cible en vue de l'apprentissage de son modèle.

Phase 1 La première phase utilise la technique de segmentation en locuteurs par HMM évolutif exposée à la section 6.1.2. Les enregistrements d'apprentissage sont découpés en segments mono-locuteur avant d'être traités dans la phase 2. Dans le cadre de la segmentation d'une conversation téléphonique, le système est réglé de façon à former au plus deux classes de locuteur.

Phase 2 La seconde phase consiste à trouver le locuteur cible présent dans les enregistrements d'apprentissage et à estimer son modèle à partir du sous-ensemble de données détecté.

La sélection des données du locuteur cible repose sur les travaux menés au LIA dans le cadre de l'appariement en locuteurs ([Meignier 2002b]). La tâche d'appariement en locuteur consiste à détecter le nombre de locuteurs intervenant dans une collection d'enregistrements et à grouper par locuteur les interventions issues des différents enregistrements. Cette phase est appliquée sur les enregistrements préalablement segmentés en locuteurs lors de la première phase. L'appariement en locuteurs, bien que correspondant à un problème de classification similaire à la segmentation en locuteur, montre un ensemble de caractéristiques spécifiques. En particulier la variabilité des conditions d'enregistrement pour un même locuteur intervenant dans plusieurs documents constitue une difficulté supplémentaire par rapport à la segmentation en locuteurs.

Comme proposé dans [Meignier 2002b], un algorithme de classification hiérarchique est appliqué pour obtenir les classes de locuteur. A chaque étape, l'algorithme groupe les deux classes les plus proches.

Le processus de classification est contraint dans le cadre de cette application. Les classes de locuteur issues de la première phase sont considérées comme sans erreur et ne sont pas remises en cause. Ainsi, seules des classes provenant de deux enregistrements différents pourront être groupées. Différents critères d'arrêt du processus de classification hiérarchique ont été évalués ; mais lors des évaluations NIST, seuls des

enregistrements à deux locuteurs sont traités, rendant trivial le critère d'arrêt. Connais-
sant le nombre de locuteurs, le processus de classification hiérarchique s'arrête après le
deuxième regroupement. La figure 6.7 illustre les étapes permettant d'obtenir le modèle
du locuteur cible dans le cadre spécifique de la campagne d'évaluation NIST 2002.

L'intérêt de l'approche présentée ici est de reposer pour chaque phase sur des méthodes qui ont été validées indépendamment dans d'autres tâches. La première phase a été validée dans la tâche de segmentation en locuteurs lors des évaluations NIST. La seconde phase a été validée en utilisant des segmentations de référence dont les locuteurs étaient connus.

Une fois l'apprentissage du modèle de locuteur réalisé, la méthode retenue pour la phase de test (la phase de détection proprement dite) est celle présentée à la section précédente, dans le cadre de l'apprentissage sur données mono-locuteur.

Résultats

L'approche proposée a été expérimentée lors de la campagne d'évaluation NIST 2002. Les données sont extraites du corpus SwitchBoard Cellular Phase II. Les enregistrements d'apprentissage et de test contiennent des conversations téléphoniques à deux locuteurs. Le genre du locuteur cible et des autres locuteurs peuvent être différents. L'ensemble des cibles est composé de 131 hommes et de 178 femmes. L'ensemble de test est composé de 1460 enregistrements d'une durée moyenne d'une minute.

Les résultats obtenus par l'approche proposée sont présentés sous forme d'une courbe DET par la figure 6.8. La comparaison avec la figure 5.4 (p. 97) fait apparaître une importante dégradation des performances entre la tâche "one-speaker detection" et la tâche "two-speaker detection" (il est important de noter la différence d'échelle entre les deux figures). Le taux d'égale erreur (EER) est approximativement le double dans le deuxième cas.

Tous les participants à cette évaluation ont constaté une perte de performance similaire. Cette perte s'explique en particulier par les différences de conditions entre les deux tâches : dans le cas de la première, le genre et le type de canal de transmission sont connus. Le processus de segmentation de la phase 1 semble suffisamment performant (taux d'erreurs de 7,5%), et l'analyse des erreurs a montré qu'elles provenaient essentiellement de la pollution de la segmentation par des zones non informatives comme des bruits. La dernière source d'erreur provient de l'appariement (phase 2). Une erreur de classification aboutit à un mauvais modèle de locuteur et pénalise le score final. Toutefois l'influence de ce type d'erreurs n'a pas pu être évaluée.

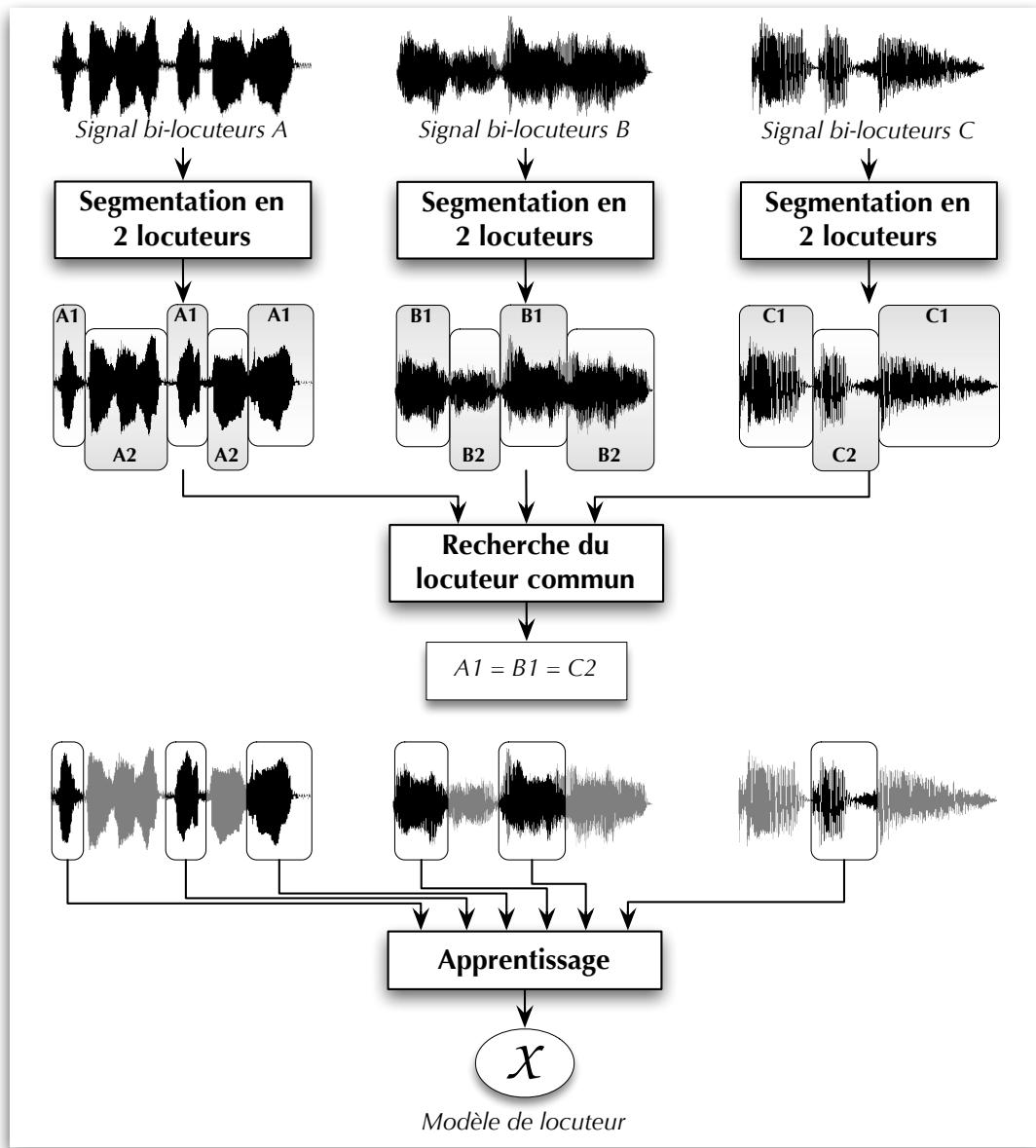


FIG. 6.7 – Apprentissage d'un modèle de locuteur à partir d'un ensemble d'enregistrements bi-locuteurs.

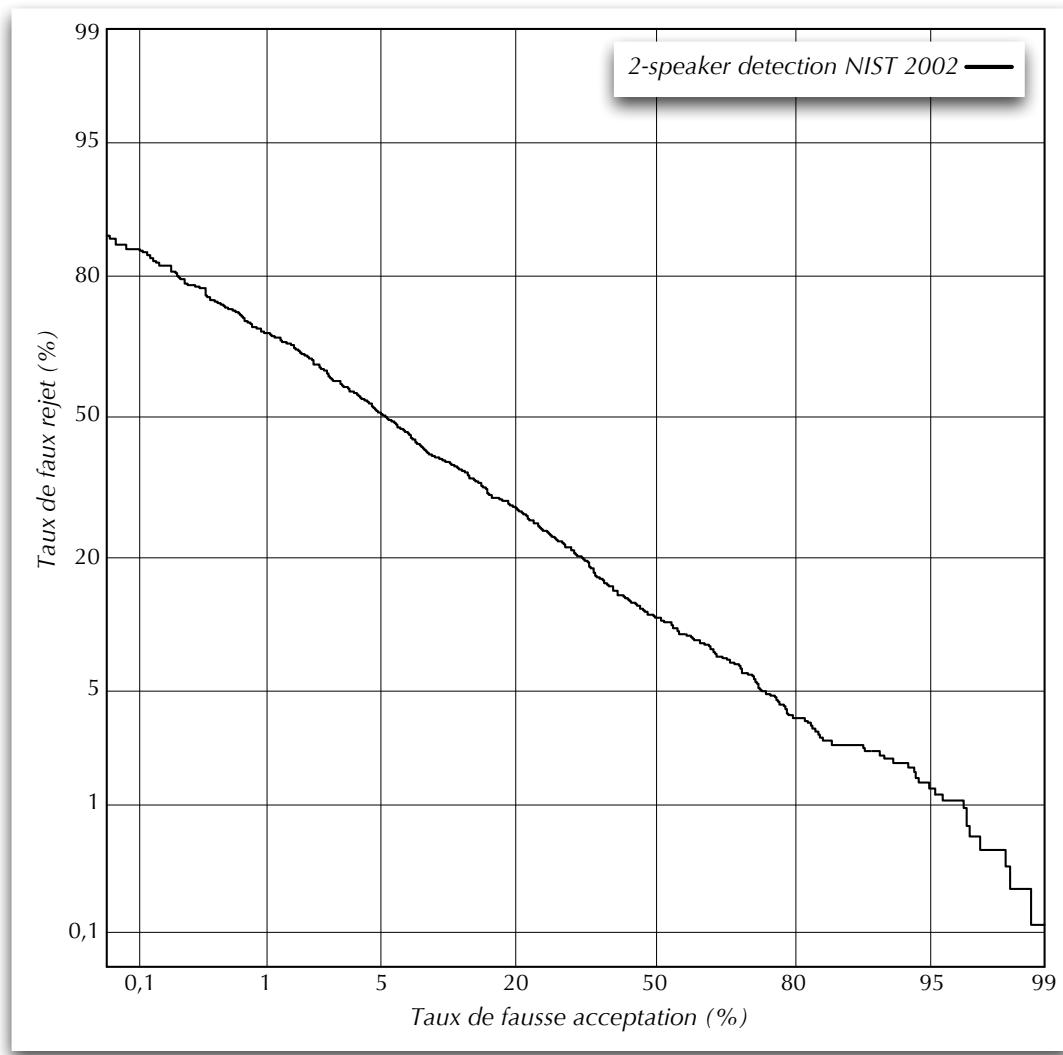


FIG. 6.8 – Détection de locuteur dans des documents bi-locuteurs — Apprentissage du locuteur cible à partir de documents multi-locuteurs — Résultats obtenus lors de la campagne d'évaluation NIST 2002.

Chapitre 7

Adaptation du système à diverses applications

Sommaire

7.1 Le projet MTM	122
7.1.1 La plateforme MTM	122
7.1.2 Problématique de l'intégration des technologies vocales	124
7.1.3 Mise en œuvre	125
7.2 La collaboration LIA-RMA	133
7.2.1 Problématique	134
7.2.2 La modalité parole	134
7.2.3 La modalité visage	135
7.2.4 Normalisation par modèle du monde pour la modalité visage	136
7.2.5 Fusion	136
7.2.6 Démonstrateur	138
7.3 Bilan	138

Le travail de thèse présenté dans ce document a été financé par l'intermédiaire de deux projets à visée applicative. Il s'agit d'une part du projet LIARMA né d'une collaboration entre le LIA et l'École Royale Militaire de Bruxelles concernant le développement d'un prototype de vérificateur d'identité biométrique et bimodal (voix et visage) et d'autre part du projet européen IST/MTM concernant l'intégration d'un module de RAL à un assistant numérique personnel. Ce chapitre rappelle les objectifs de ces deux projets, exposés aux chapitre 3, et présente la problématique associée. Les développements réalisés pour y répondre sont ensuite décrits. Ces développements représentent un exemple de transfert des technologies présentées dans les chapitres précédents ainsi que des connaissances acquises à travers le développement et l'évaluation de la plate-forme de recherche AMIRAL.

7.1 Le projet MTM

Le projet MTM¹ (*Multimedia Terminal Mobile*) du programme européen IST² a regroupé 10 partenaires industriels et universitaires, issus de 5 pays européens (Allemagne, Belgique, Espagne, France, Italie), de janvier 2000 à décembre 2001.

L'objectif du projet MTM était de définir, développer et produire un terminal multimédia complet et communiquant, intégrant caméra vidéo et accès Internet sans fil à haut débit pour offrir les fonctions d'un assistant numérique personnel combinées à celles d'un téléphone sans fil. Outre la définition de la plateforme matérielle, le projet incluait également le développement d'applications spécifiques visant à exploiter au mieux les possibilités offertes par les technologies matérielles mises à leur disposition. Dans ce cadre, les divers partenaires du projet MTM peuvent être classés en deux catégories selon leur rôle : les fournisseurs et les intégrateurs. Les premiers mettent en place les technologies et les moyens techniques utilisés par les seconds dans le développement des applications destinées au terminal.

Le Laboratoire Informatique d'Avignon a participé au projet dans la première catégorie, en la qualité d'expert dans le domaine du traitement automatique de la parole. Son rôle au sein du projet était d'offrir aux autres partenaires un accès aux technologies de reconnaissance automatique de la parole et de reconnaissance automatique du locuteur afin de les intégrer dans l'interface du système et des applications. Les technologies vocales trouvent en effet un champ naturel d'application dans l'interface d'un terminal portable, limité par nature en termes de moyens d'interaction avec l'utilisateur. Les outils vocaux apportent en effet un bénéfice non négligeable en terme d'ergonomie à l'interface utilisateur d'un tel terminal, limitée en entrée au stylet.

7.1.1 La plateforme MTM

La plateforme matérielle du projet MTM a été développée sur la base d'un matériel existant, déjà présent dans le commerce. Il s'agit d'un PDA (*Personal Digital Assistant*), le Compaq (maintenant Hewlett-Packard) iPAQ H3600, doté d'un processeur Intel StrongArm SA-1110 à 206 MHz, de 16 Mo de mémoire flash et de 32 Mo de mémoire vive. Les capacités d'extension de cet appareil, notamment par l'intermédiaire de ports PCMCIA, sont une des raisons de son choix comme point de départ des développements matériels du projet. Ces développements ont consisté en premier lieu à concevoir un nouveau boîtier pour accueillir le cœur du PDA. Ce boîtier a ensuite permis d'abriter les autres éléments complétant la plateforme matérielle MTM : une caméra vidéo destinée aux applications de visioconférence, un lecteur de cartes magnétiques et un module de communication téléphonique haut débit³ (UMTS).

D'un point de vue logiciel, la plateforme MTM, outre les outils d'organisation personnelle propres à un PDA, repose sur trois applications principales.

¹Projet IST-1999-11100 — cf. http://dbs.cordis.lu/fep-cgi/srchidadb?ACTION=D&CALLER=PROJ_IST&QF_EP_RPG=IST-1999-11100

²Présentation du programme IST : <http://www.cordis.lu/ist/>

³Le module de communication a été simulé au cours du développement par l'emploi d'une carte réseau sans fil à la norme 802.11b.



FIG. 7.1 – Projet MTM — Prototype du terminal.

Télé-médecine L’application de télé-médecine *Mobile Chili* ([Engelmann 2001]) est développée à destination des radiologues, dans le but d’assurer une meilleure communication entre plusieurs spécialistes. L’exemple type d’utilisation est la réception à distance par un radiologue, en cas d’urgence, de clichés en provenance de l’hôpital. Le radiologue sera en mesure de visualiser ces données et d’en faire une première analyse. Il pourra répondre, renvoyer un rapport, un diagnostic ou des instructions pour l’équipe médicale. Cette application prévoit également d’autres cadres d’utilisation, avec notamment une assistance aux médecins lors des visites aux patients. L’application *Mobile Chili* a remporté le prix IST 2002⁴ (récompensant les projets achevés en 2001) dans le cadre de son intégration au projet MTM.

Easy City Guide Cette application, proposée par la commune de Rome et présentée notamment aux communes de Tolède et de Madrid, a pour objectif la fourniture d’un service d’information, utilisable tant par les résidents que par les touristes, optimisant et facilitant l’accès aux informations et services fournis par la municipalité (plans, visites guidées de musées, etc.).

Enseignement à distance L’objectif de cette application, soutenue par l’université de Madrid, est de permettre aux personnes éloignées de poursuivre leur formation scolaire ou universitaire. Elle inclut également un système de visioconférence.

⁴<http://mbi.dkfz-heidelberg.de/mbi/projects/MTM/>

7.1.2 Problématique de l'intégration des technologies vocales

Le rôle du LIA en tant que fournisseur d'accès aux technologies de reconnaissance automatique de la parole et du locuteur a débuté par l'étude, en collaboration avec les partenaires appelés à les utiliser, de la stratégie d'intégration de ces technologies dans l'interface utilisateur des applications et du système. Les résultats de cette étude ont permis de cerner les besoins de la plateforme MTM en termes de technologies vocales. La problématique de l'intégration de ces technologies consiste alors à concilier ces besoins avec les contraintes imposées par la définition de la plateforme MTM (d'un point de vue technique mais aussi ergonomique) ainsi qu'avec les capacités offertes par l'état des connaissances dans le domaine (concernant la RAL, il s'agit des possibilités offertes par le système AMIRAL).

Étude des besoins

La définition des applications présentées dans la section précédente ainsi que l'utilisation des technologies vocales envisagée par l'ensemble des partenaires permet de dégager les fonctionnalités attendues de la part de ces technologies.

Pour la reconnaissance de la parole, il s'agit de la reconnaissance de commandes courtes, à vocabulaire dynamique, modifiable à volonté par les concepteurs d'applications ou l'utilisateur.

En ce qui concerne la reconnaissance du locuteur, deux types d'utilisation sont souhaités. Le premier correspond à l'utilisation explicite de la RAL, en tant que système d'authentification de l'utilisateur pour l'accès au système. Ce système d'authentification ne s'accompagne cependant pas de la recherche d'un niveau élevé de sécurité. Le second type d'utilisation correspond à une utilisation plus discrète de la RAL, en tant qu'option de confort visant à personnaliser l'usage du terminal et de ses logiciels, à travers l'adaptation automatique des réactions du système à l'identité de l'utilisateur, détectée automatiquement (notamment lorsqu'il prononce des commandes).

La capacité à répondre à ces attentes requiert un système capable de traiter les tâches d'identification mais aussi de vérification automatique du locuteur, fonctionnant en mode indépendant du texte et apte à fournir un bon niveau de performance à partir de signaux d'une durée très courte.

Contraintes

La mise en œuvre d'un système de reconnaissance du locuteur dans le cadre d'un projet applicatif tel que MTM implique de surmonter un certain nombre de problèmes auxquels ne sont habituellement pas confrontés les développements effectués dans un cadre de recherche.

Les difficultés les plus évidentes proviennent de contraintes d'ordre technique imposées par la nature même de la tâche : la reconnaissance du locuteur sur un petit terminal portatif. Au premier rang de ces contraintes, celles influençant la qualité de l'acquisition sonore ont un impact notable sur les performances de la reconnaissance (*cf. chapitre 2, p. 28*). En particulier, le microphone intégré au PDA est de type omnidirectionnel (enregistrant par conséquent tous les bruits de l'environnement) et situé

par nature à une distance variable de l'utilisateur (le mode de fonctionnement retenu pour le MTM, notamment pour *Mobile Chili*, est un fonctionnement à hauteur de poitrine, soit à 40/50 cm de la bouche). La qualité des enregistrements réalisés s'en trouve considérablement dégradée, contribuant pour beaucoup à la difficulté de la tâche. De plus, l'environnement sonore lors de l'utilisation de l'appareil est logiquement dépendant du lieu où est utilisé le terminal et donc très variable et potentiellement bruité.

Une contrainte technique supplémentaire est généralement associée au monde des systèmes embarqués : la faible quantité de ressources disponibles en termes tant d'espace mémoire que de puissance de calcul. Cette limite restreint alors fortement le choix des algorithmes utilisés pour le traitement de la parole. Même si ce fonctionnement entièrement embarqué reste possible sur le MTM, il est heureusement possible de contourner cet écueil dans le cadre de ce projet grâce à l'utilisation de la liaison sans fil à haut débit intégrée à la plateforme. En effet cette liaison est requise par plusieurs applications (notamment la télé-médecine et l'apprentissage à distance) fonctionnant en mode client-serveur. Il est dès lors envisageable de recourir à cette liaison pour déporter la plupart des calculs lourds sur un serveur distant. Ce mode de fonctionnement a été retenu comme mode principal pour la reconnaissance de la parole comme pour la reconnaissance du locuteur, autorisant l'emploi de techniques plus gourmandes en puissance de calcul que ce qu'il aurait été acceptable pour le processeur du terminal seul.

Au-delà des problèmes liés à des considérations purement techniques, d'autres contraintes sont issues d'un souci ergonomique. Il s'agit principalement de contraintes sur les délais de réponse du système, dans le but d'offrir à l'utilisateur un temps de réponse confortable. En particulier, l'apprentissage d'un nouveau modèle de locuteur doit rester dans un temps raisonnable de quelques secondes. Cette contrainte porte bien entendu sur la célérité de la modélisation mais aussi sur la durée de l'enregistrement demandé pour l'apprentissage initial d'un modèle. Peu d'utilisateurs acceptent en effet de bonne grâce de parler durant plusieurs dizaines de secondes à leur PDA pour obtenir une estimation correcte de leur modèle, même si cet effort ne leur est demandé qu'une seule fois. Le choix de la technique de modélisation des locuteurs s'en trouve orienté vers les techniques permettant une estimation efficace d'un modèle sur peu de données. Plus important encore, la phase de test doit être effectuée sur des enregistrements très courts et permettre une réponse quasiment instantanée, particulièrement lorsque la reconnaissance du locuteur se fait à l'occasion d'une commande prononcée par l'utilisateur ; si l'interprétation de la commande dépend de l'identité de l'utilisateur, cette dernière doit pouvoir être déterminée très rapidement afin de ne pas introduire de délai dans l'exécution de la commande.

Enfin une difficulté supplémentaire propre au projet MTM tient à sa nature internationale et multilingue. Le système de RAL (mais aussi celui de reconnaissance de la parole) doit de ce fait être indépendant du langage ou adapté aux langues parlées par les divers partenaires du projet.

7.1.3 Mise en œuvre

Dans le projet MTM, le choix des technologies à utiliser pour la réalisation du système de RAL a naturellement été guidé par la nécessité de répondre aux attentes des partenaires du projet tout en prenant en compte les contraintes exposées à la section précédente, certains compromis étant nécessaires pour concilier ces deux aspects. Les solutions retenues sont présentées ici, suivies d'un aperçu des développements réalisés

pour mener à bien ce projet.

Cœur du système de RAL

Ainsi qu'il en a été fait mention à la section précédente, la liaison sans fil haut débit intégrée au terminal permet de s'affranchir en grande partie des contraintes de puissance de calcul et d'occupation mémoire associées aux systèmes embarqués, en déportant sur une machine distante les calculs les plus lourds. Le terminal fonctionnant en mode connecté la plupart du temps, la décision a été prise de recourir à des techniques plus lourdes que ce qui aurait été possible en mode autonome, tout en préservant la possibilité d'un tel fonctionnement en mode autonome, sur les ressources propres du terminal (au prix d'une forte dégradation de la vitesse de réaction). La base du système de reconnaissance de locuteur du MTM repose sur une technique classique : la modélisation des locuteurs par mixtures de gaussiennes (GMM).

Ce choix a été motivé en premier lieu par les performances affichées par les systèmes reposant sur l'utilisation de GMM dans le cadre de la reconnaissance du locuteur en mode indépendant du texte (ce mode correspondant aux besoins définis pour l'utilisation de la RAL dans le cadre du projet MTM).

L'autre raison ayant conduit à choisir la modélisation par GMM est l'expertise acquise dans ce domaine à travers le développement de la plateforme de recherche AMIRAL (*cf.* chapitre 4). La perspective de réutiliser cette connaissance permet de pallier une des principales difficultés présentées par un projet comme MTM : l'absence d'enregistrements correspondant à l'utilisation du système dans son cadre d'exploitation normal. Une quantité non négligeable de telles données est pourtant nécessaire pour fixer les paramètres du système (modèle du monde, fonction de normalisation et seuil de décision dépendant de la disponibilité de telles données) et les valider par l'évaluation des performances. La solution palliative utilisée ici consiste à recourir à des paramètres établis par l'intermédiaire de la participation d'AMIRAL aux campagnes d'évaluation NIST.

Structure des modèles Les GMM utilisés dans le cadre du projet MTM reposent sur l'emploi de matrices de covariances diagonales, conformément à l'état de l'art. Comme indiqué au chapitre 4 (page81), le nombre de composantes des modèles de locuteurs est un paramètre jouant un rôle important dans la qualité de la reconnaissance. Cependant un autre aspect de ce paramètre est également crucial dans le cas de MTM : l'influence du nombre de composantes sur la vitesse d'exécution. En effet, le temps de réponse du système est directement proportionnel au nombre de composantes des modèles. Un compromis doit être trouvé entre performances et temps de réponse afin de satisfaire le besoin de réactivité du système.

La figure 7.2 illustre ce compromis en confrontant les résultats (en termes de taux d'égale erreur) obtenus sur le corpus de l'évaluation NIST 2000 pour diverses tailles de modèles à l'évolution correspondante du temps de réponse du système. Les résultats présentés ont conduit au choix de modèles GMM à 64 composantes pour le système MTM, la dégradation des performances n'étant que minime par rapport à l'emploi de modèles à 128 ou même 256 composantes pour un temps de traitement divisé par 2 ou 4 respectivement.

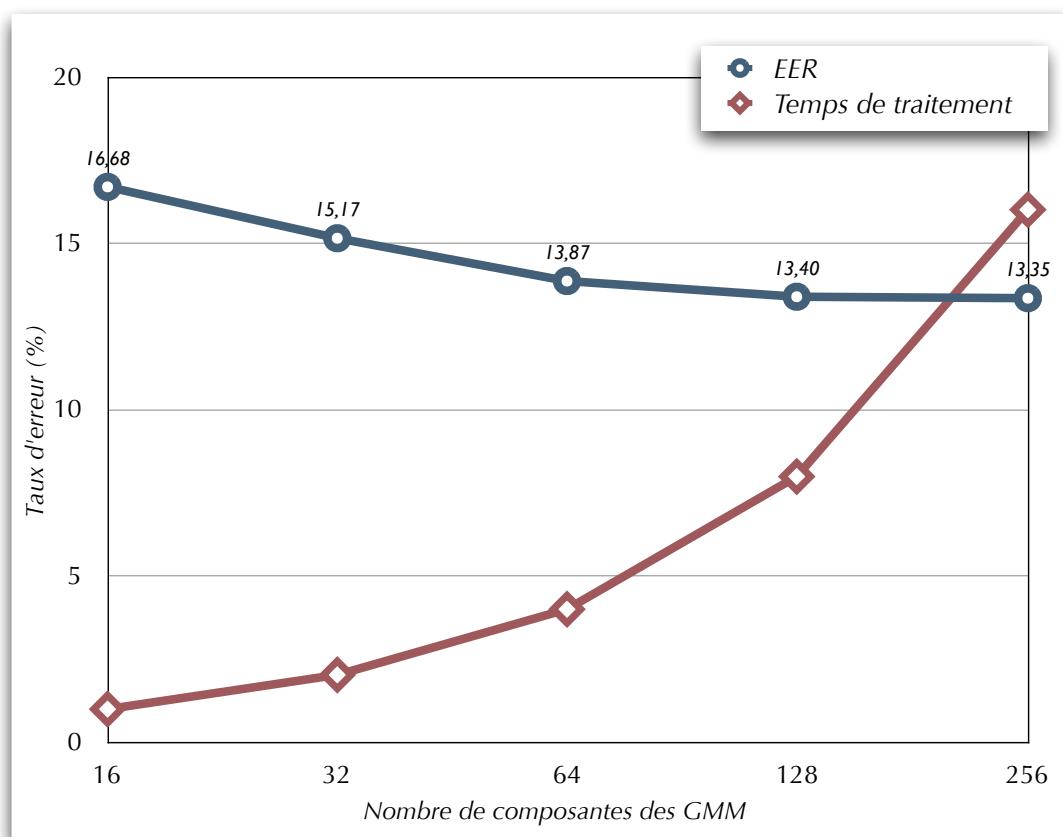


FIG. 7.2 – Choix de la taille des modèles de locuteur pour le projet MTM — Taux d'égal erreur (EER) obtenu pour diverses valeurs du nombre de composantes des GMM, confronté à l'évolution correspondante du temps de traitement.

Algorithmes d'apprentissage Les modèles de locuteurs sont estimés par adaptation bayésienne d'un modèle du monde, selon la technique présentée au chapitre 4 (page 82). L'apprentissage par adaptation présente un grand intérêt dans le cadre du projet MTM de par les bonnes performances que cette solution obtient pour l'estimation de modèles à partir de faibles quantités de données d'apprentissage, mais aussi du fait du coût réduit en termes de complexité de calcul, dès lors que seules les moyennes sont adaptées.

Comme souligné à la section 4.6, la qualité du modèle du monde, à travers la représentativité des données utilisées pour l'estimer, est un élément crucial pour les performances du système. Le modèle du monde pour le projet MTM n'est évidemment pas dépendant du type de matériel utilisé pour l'enregistrement, ce paramètre étant constant dans le cadre MTM. Il n'est pas non plus dépendant du genre, cette information étant inconnue du système qui n'intègre pas de détecteur de genre. L'estimation du modèle du monde a été faite sur un ensemble d'enregistrements réalisés directement sur le PDA afin d'être le plus proche possible des futures données d'exploitation. Ces enregistrements totalisent environ une heure de parole représentant les voix de 20 locuteurs hommes et femmes. Afin d'éviter toute sur-représentation d'un locuteur dans le modèle du monde, tous ont été enregistrés sur la même durée d'environ 3 minutes (en une ou deux sessions selon les locuteurs). Cependant, le nombre de femmes dans cet ensemble de locuteurs est inférieur au nombre d'hommes, introduisant un biais. Les éventuels effets de ce biais sur les performances de reconnaissance pour les locuteurs féminins n'ont toutefois pas pu être vérifiés, faute de base de données permettant une évaluation sérieuse du système en conditions d'exploitation.

Décimation de trames Dans la quête des temps de traitement les plus courts possibles, en particulier en phase de test, le recours à la décimation de trames décrite au chapitre 4, page 67, représente une avancée conséquente.

La figure 7.3 présente les résultats sur lesquels se base le choix de la proportion de trames à utiliser. Ces résultats représentent le taux d'égal erreur obtenu pour diverses proportions (variant de 100 % des trames à 1 trame sur 12) sur le corpus de l'évaluation NIST 2000. Le choix final s'est porté sur l'utilisation d'une trame sur huit (soit 12,5 % des trames) qui n'induit une augmentation de l'EER que de 0,16 point en échange d'une division par 8 du temps de réponse du système.

Suppression des trames non informatives Il convient de noter que, du fait du calendrier de développement du projet MTM, les technologies qui y sont intégrées correspondent à celles présentées lors de la participation à la campagne d'évaluation NIST 2000 (*cf. chapitre 5, page 94*). En particulier, la suppression des trames de basse énergie décrite au chapitre 4 (page 69) — et le gain de performances associé —, intégrée à la plateforme AMIRAL pour la participation à l'édition 2001 des campagnes d'évaluation NIST, ne fait pas partie des technologies utilisées pour le projet MTM.

Ceci explique la différence entre les résultats (EER) présentés par la figure 7.2, correspondant au système décrit ici, et ceux montrés en annexe A, page 153, obtenus avec recours à la suppression des trames de basse énergie.

Normalisation — Seuil de décision Le moteur de reconnaissance du locuteur intégré au projet MTM repose sur la technique WMAP (*cf. chapitre 4, page 75*) pour la normali-

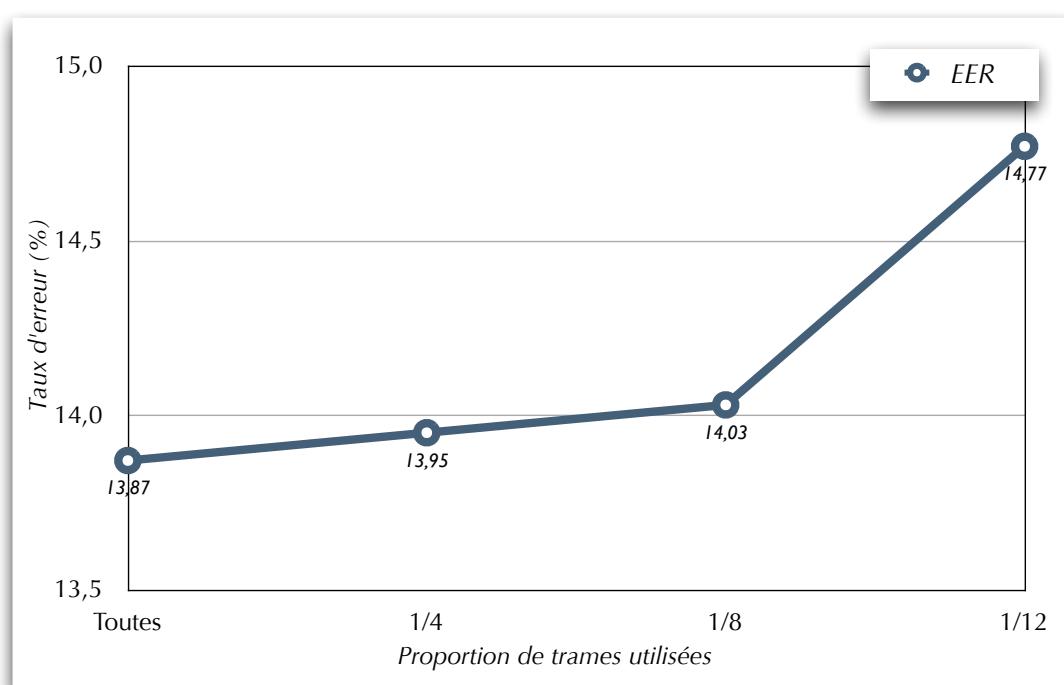


FIG. 7.3 – Décimation de trames pour le projet MTM — Taux d'égal erreur (EER) obtenus avec des modèles à 64 composantes en fonction de la quantité de trames utilisées lors du test (de toutes les trames à 1 trame sur 12).

sation des scores. Le choix de cette méthode de normalisation est motivé en partie par son faible coût en termes de puissance de calcul (lors de l'application de la normalisation au cours du test) par rapport à d'autres techniques telles Tnorm.

Toutefois l'avantage décisif de WMAP pour son utilisation dans le cadre du projet MTM tient au sens qu'elle confère à la valeur du seuil de décision. WMAP redéfinit le score d'un signal de test comme la probabilité *a posteriori* que ce signal corresponde au même locuteur que le modèle, connaissant le score obtenu et la probabilité *a priori* d'être en présence d'une tentative d'imposture. Le choix du seuil de décision en est dès lors facilité. Ce point est particulièrement important dans le cadre du projet MTM. En effet le LIA y joue un rôle de fournisseur de technologies aux développeurs d'applications. Ce sont ces derniers qui définissent le cadre d'utilisation de ces technologies et notamment le niveau de "sévérité" requis pour la RAL. Il est donc nécessaire de leur offrir l'accès à un réglage simple de ce paramètre, ne nécessitant pas une connaissance avancée du domaine de la RAL et des stratégies de décision associées. Ce contrôle est offert ici à travers le réglage du seuil de décision dans un espace probabiliste borné rendu possible par WMAP.

Malheureusement le manque de données de développement correspondant aux conditions d'exploitation de l'appareil se ressent ici aussi. Il n'a pas été possible d'établir une fonction de normalisation pour WMAP basée sur de telles données. En lieu et place, une fonction de normalisation estimée sur le corpus de données des campagnes d'évaluation NIST est utilisée. Il n'est cependant pas possible, toujours faute de données permettant l'évaluation du système⁵, de mesurer l'impact de ce choix sur les performances. Il serait pourtant intéressant de déterminer si le recours à une fonction de normalisation estimée sur un corpus différent permet de conserver le bénéfice de la normalisation ou au contraire induit une dégradation des performances.

Système de reconnaissance de la parole

Le développement du système de reconnaissance de la parole pour le projet MTM n'entre pas dans le cadre de ce travail de thèse. Néanmoins, les modules de reconnaissance du locuteur et de reconnaissance de la parole de la plateforme MTM ont été développés conjointement et sont prévus pour fonctionner ensemble, partageant certaines ressources (notamment la paramétrisation du signal). À ce titre, une description succincte des directions suivies pour la mise en œuvre de la reconnaissance de la parole au sein du projet MTM est donnée ici. Le lecteur désireux de plus de détails concernant cette partie se reportera notamment à [Lefort 2002].

Les besoins du projet MTM se traduisent en deux différents outils pour la reconnaissance de la parole :

- un système de reconnaissance de mots isolés ; les développeurs peuvent définir leur propre vocabulaire et le changer dynamiquement durant l'exécution de l'application (offrant ainsi à l'utilisateur la possibilité de définir son propre vocabulaire) ;
- un outil automatique de transcription phonétique à partir du signal de parole afin de faciliter la phonétisation et l'ajout de nouveaux mots par le développeur et/ou l'utilisateur.

⁵Il est intéressant de noter que, dans ce cas précis, si des données d'évaluation étaient disponibles en quantité suffisante, la question posée ici n'aurait plus vraiment de raison d'être (dans le cadre de MTM) puisque ces mêmes données permettraient l'estimation d'une fonction de normalisation.

L'approche retenue pour le système de reconnaissance de la parole est reposé sur des HMM avec des modèles de phonèmes. Ce choix a été préféré à des approches plus classiques dans le cadre d'un système de reconnaissance de mots isolés à petit vocabulaire, reposant sur l'algorithme de DTW ou sur des HMM avec des modèles de mots, car celles-ci se révèlent inadaptées aux fonctionnalités requises par les utilisateurs, qui imposent des modèles acoustiques indépendants du locuteur, de la langue et du vocabulaire. Les HMM avec modèles de phonèmes offrent la souplesse nécessaire, notamment pour l'aspect dynamique du vocabulaire.

Le décodage est assuré par l'algorithme de Viterbi synchrone avec *beam pruning*, suffisant pour un système de reconnaissance de mots isolés à petit vocabulaire. La modélisation acoustique est issue du système de reconnaissance de la parole continue grand vocabulaire du LIA, *Speeral* ([Nocera 2002]). Elle repose sur des HMM avec des distributions de probabilité continues et une topologie de Bakis standard avec 3 états par phonème. Néanmoins le reconnaissseur MTM n'utilise pas de phonèmes contextuels. La modélisation des phonèmes est réalisée avec des GMM à 64 composantes à matrices diagonales. L'apprentissage des modèles acoustiques a été réalisé sur un corpus de parole lue en Français. Afin de compenser la perte de performances due aux différences de microphone entre le corpus d'apprentissage et le PDA, une phase d'adaptation bayésienne (MAP) est ensuite appliquée aux modèles acoustiques, en utilisant un corpus de taille réduite mais enregistré directement sur le terminal MTM.

Bibliothèque de reconnaissance de la parole et du locuteur

Au delà des questions scientifiques posées par la mise en œuvre d'un système de reconnaissance du locuteur sur un PDA, le travail réalisé dans le cadre du projet MTM comporte également un aspect plus technique consistant à intégrer les technologies vocales proposées au sein des différentes applications. Plus précisément, le rôle du LIA dans le projet MTM consiste à définir et à réaliser des modules vocaux et à permettre un accès aisément à ces fonctionnalités pour les développeurs ou intégrateurs d'applications, non spécialistes de la parole.

Cette partie du travail a impliqué la définition d'un ensemble restreint de fonctions constituant les points d'accès des développeurs aux fonctionnalités du système de reconnaissance du locuteur. Cet accès devant présenter l'aspect le plus simple possible afin de rendre cette technologie utilisable par des développeurs non au fait des subtilités du domaine de la RAL, les fonctions définies offrent un haut niveau d'abstraction, correspondant directement aux tâches principales à accomplir, à savoir l'apprentissage de modèles de locuteurs ainsi que l'identification et la vérification du locuteur.

Un seul accès à la bibliothèque est ainsi nécessaire pour réaliser la vérification de l'identité de l'utilisateur courant par rapport à un utilisateur enregistré ; de même pour l'identification automatique (en milieu ouvert) de l'utilisateur courant.

Une unique fonction permet de réaliser à partir d'un enregistrement soit l'apprentissage initial du modèle d'un locuteur, soit la mise à jour d'un modèle existant ; si les informations d'identité fournies lors l'appel de cette fonction sont inconnues du système, un nouveau modèle de locuteur est appris et, associé à ces informations, forme une nouvelle entrée dans la base des utilisateurs connus ; dans le cas contraire, le modèle du locuteur correspondant à l'identité fournie est mis à jour, par adaptation bayésienne.

Enfin, une intégration des fonctionnalités de reconnaissance de la parole et de re-

connaissance du locuteur est proposée : une fonction permet de reconnaître un mot de commande (ou un mot de passe) et de vérifier simultanément l'identité du locuteur.

Cet accès aux fonctionnalités principales du système est complété de fonctions utilitaires facilitant la gestion d'une ou plusieurs bases de données d'utilisateurs (associant pour chacun identité, modèle de locuteur et informations diverses propres à l'application). Enfin, un accès à la valeur du seuil de décision est offert, permettant d'adapter le comportement du système de reconnaissance du locuteur aux exigences de l'application en termes de sécurité et d'ergonomie.

Outre la réalisation de cette couche d'abstraction, qui constitue le point d'accès aux technologies de reconnaissance du locuteur offert aux partenaires du projet MTM, la mise en œuvre du système de RAL pour ce projet comprend deux autres étapes majeures.

La première est l'intégration du mode client/serveur au fonctionnement du système. Ce mode, comme exposé plus haut, permet de déporter les calculs du terminal vers une machine distante, plus puissante, par l'utilisation de la liaison sans fil. Le mode client serveur repose sur un protocole de communication et une interface logicielle. Le protocole permet de transférer les données audio par blocs temporels. Cela permet de réduire le décalage temporel entre l'acquisition du signal et la réponse du système mais aussi de répartir les traitements, avec par exemple, le calcul des paramètres acoustiques sur le MTM et le reste des traitements à distance. L'interface logicielle permet un fonctionnement en mode client/serveur ou en mode autonome de façon transparente pour les couches hautes auxquelles accèdent les développeurs d'applications.

La dernière étape importante de la réalisation de ce projet est l'effort de développement consacré au cœur du système. Cet effort correspond au regroupement de nombreuses techniques développées séparément et de manière dispersée au cours de l'évolution de la plateforme de recherche AMIRAL. L'intégration de ces morceaux épars en un seul bloc logique implique également une adaptation de l'ensemble à la philosophie d'un projet applicatif, différente de celle du développement mené dans le cadre de la recherche. En particulier, de fortes contraintes existent en matière d'optimisation et de fiabilité. Contrairement à un développement réalisé pour explorer une voie de recherche et dont la durée de vie n'excède pas celle de la série d'expériences correspondante, un système appelé à fonctionner en permanence comme base des applications et de l'interface utilisateur d'un PDA se doit de montrer une stabilité exemplaire. Cet aspect du développement a par conséquent fait l'objet de beaucoup d'attention. Il est intéressant de noter que les bénéfices de cet effort dépassent le cadre du projet MTM. La plus grande partie des efforts réalisés au cours de ce projet en matière d'optimisation, d'organisation et de stabilité ont été intégré en retour dans la plateforme AMIRAL, lui offrant un gain de fiabilité et de souplesse d'utilisation très appréciables pour les activités de recherche.

Intégration des technologies vocales au système et aux applications

Le rôle du LIA au sein du projet MTM était celui d'un fournisseur de technologie, offrant aux partenaires du projet la possibilité d'intégrer la reconnaissance de la parole et du locuteur au sein leurs applications. Néanmoins, dans le but d'une part de présenter à ces partenaires une démonstration d'une telle intégration, et d'autre part de faciliter encore l'adoption des technologies vocales, ce rôle a été dépassé et deux exemples d'utilisation de ces technologies ont été réalisés.

Le choix du premier exemple a été guidé par la nature de certaines applications du projet MTM. En effet plusieurs d'entre elles (notamment les applications *Easy City Guide* et d'enseignement à distance) reposent sur l'utilisation d'un navigateur web pour leur interface utilisateur. Cette particularité a inspiré la modification d'un navigateur conçu spécifiquement pour les PDA, *Konqueror Embedded*⁶, pour y intégrer l'utilisation des technologies vocales. Cette intégration se traduit principalement par la possibilité de naviguer grâce à des commandes vocales, permettant de se déplacer au sein d'une même page ou de suivre des liens. Le recours à la reconnaissance du locuteur se fait lors de l'accès vocal aux signets. Ceux-ci sont dépendants de l'utilisateur, dont l'identité est détectée automatiquement afin de lui permettre d'accéder à son jeu de signets personnel.

L'autre exemple développé pour illustrer l'utilisation des technologies vocales est orienté plus directement vers la reconnaissance du locuteur. Il s'agit d'ajouter au PDA une gestion des sessions utilisateur (login). Le MTM est ainsi personnalisé pour chaque utilisateur. L'ouverture d'une session se fait ici en identifiant l'utilisateur par sa voix, constituant en cela une démonstration classique des tâches d'identification et de vérification du locuteur. En effet, deux modes d'ouverture de session sont possibles. Le premier implique de sélectionner une identité dans la liste des utilisateurs connus, une vérification du locuteur par rapport au modèle de l'utilisateur sélectionné étant ensuite réalisée pour autoriser l'ouverture de session. Dans le second mode, aucune sélection préalable d'identité n'est à faire et une tâche d'identification du locuteur est réalisée (en milieu ouvert, avec possibilité de rejet) pour détecter l'identité de l'utilisateur et autoriser l'ouverture de sa session.

7.2 La collaboration LIA-RMA

Dans le cadre de l'étude interne F9904, l'École Royale Militaire de Bruxelles (désignée généralement par l'acronyme anglais RMA — *Royal Military Academy*) a exploré la possibilité de réaliser un système d'identification biométrique apte à authentifier une personne par les caractéristiques de sa voix et de son visage. L'objectif poursuivi par cette étude est la diminution, par l'utilisation conjointe de ces deux modalités, de l'influence de la variabilité des caractéristiques inhérentes à chacune d'elle, avec en conséquence l'amélioration des performances globales du système du fait de leur complémentarité. Ce travail traite d'authentification et non d'identification. Ainsi, la tâche consiste ici à vérifier l'identité annoncée par un individu et non à détecter automatiquement cette identité. L'accent est porté sur l'aspect pratique de la solution envisagée ainsi que le développement d'un prototype logiciel.

Le LIA a initié une collaboration avec la RMA incluant notamment la participation conjointe à certains projets de recherche. La première manifestation de cette collaboration a été la participation à l'étude F9904. La modalité parole du reconnaisseur a été développée dans le cadre de ce travail de thèse, dont le financement a été assuré en partie par la RMA à travers cette étude. Les travaux correspondants, présentés ici, ont été réalisés à Bruxelles au sein du laboratoire *Signal and Image Centre* de l'École Royale Militaire.

⁶<http://www.konqueror.org/embedded>

7.2.1 Problématique

L'objet de l'étude F9904 est l'utilisation de techniques d'authentification des personnes par leurs caractéristiques biométriques, avec une orientation vers l'usage de caractéristiques biométriques dont l'analyse ne requiert aucun contact avec l'utilisateur. La première raison de cette orientation est la volonté de recourir à des techniques d'analyse non contraignantes pour l'utilisateur. La seconde raison provient du choix de proposer une architecture logicielle, et non matérielle, devant permettre de reposer sur un ordinateur personnel associé à des composants standards et peu coûteux pour développer un système d'authentification. Cette dernière contrainte est en effet peu compatible avec l'usage de caractéristiques biométriques dont l'analyse requiert un contact physique (telles l'analyse des empreintes digitales, de la rétine ou de la forme de la main) du fait du coût des capteurs spécifiques sur lesquels ces techniques reposent.

Cependant, les techniques d'authentification biométriques sans contact physique présentent un niveau de performances moins élevé que celles imposant un contact. Par conséquent, l'étude F9904 porte également sur les possibilités de combinaison de plusieurs modalités dans le but d'en améliorer la fiabilité.

Le choix des modalités à utiliser s'est porté sur l'image et la parole, toutes deux répondant aux critères d'analyse non contraignante pour l'utilisateur et de recours à des capteurs standards et peu coûteux disponibles sur un ordinateur personnel multimédia. Les techniques d'authentification associées sont la reconnaissance du visage et la reconnaissance du locuteur.

Le travail réalisé dans le cadre de la collaboration LIARMA comprend essentiellement le développement de l'expert vocal et son intégration dans un prototype d'authentification bimodale voix/visage sur base d'ordinateur personnel. La RMA avait plus particulièrement la responsabilité de l'expert "visage" et de la fusion des modalités, les travaux correspondants étant présentés succinctement dans les sections 7.2.3 et 7.2.5.

7.2.2 La modalité parole

Les objectifs définis dans le cadre de l'étude F9904 impliquent l'emploi de la reconnaissance du locuteur pour une tâche de vérification automatique du locuteur (*cf.* chapitre 2, p. 23). Le produit final étant une plateforme d'authentification générique, appelée à fonctionner sur un matériel quelconque dans un contexte applicatif non spécifié, aucune indication supplémentaire n'est disponible concernant les conditions d'exploitation de la reconnaissance du locuteur, au contraire du projet MTM qui fixait un cadre précis d'exploitation.

Cette spécification large dans l'utilisation introduit dans la réalisation deux contraintes fortes. La première est la capacité d'adaptation à une grande variété de conditions d'utilisation. Des facteurs importants tels que la distance au micro ou l'environnement acoustique ne sont pas maîtrisés, pouvant varier considérablement d'une application à l'autre. Le type de microphone utilisé est lui aussi susceptible de varier, étant défini simplement comme un quelconque dispositif d'acquisition sonore exploitable sur un ordinateur personnel.

La seconde contrainte importante induite par l'absence de définition précise du cadre d'exploitation est l'absence logique d'enregistrements correspondants. Le développement du système ne peut pas, par conséquent, reposer sur l'utilisation de

telles données, notamment pour la mise en œuvre des techniques de normalisation des scores.

Les caractéristiques du système de reconnaissance du locuteur défini dans le cadre de la collaboration LIARMA, ainsi que les contraintes associées, étant très génériques, peu de développements spécifiques à ce système ont été menés. En fait, l'évolution du module de RAL présenté ici correspond à l'évolution de la plateforme AMIRAL telle que présentée au chapitre 4, dont la période de développement se confond en grande partie avec la durée de la collaboration LIARMA sur l'étude F9904. Cette plateforme définissant un système générique de reconnaissance du locuteur, les avancées réalisées au cours de son développement vont dans le sens des objectifs poursuivis ici (notamment concernant la robustesse face aux variations de type de microphone) et ont été intégrées au système livré à la fin de l'étude. Ce système correspond en fait au sous-ensemble d'AMIRAL consacré à la vérification automatique du locuteur, complété de quelques développements spécifiques. Du fait de l'absence de cadre d'exploitation clairement défini et des données associées, la définition et l'optimisation des techniques utilisées reposent uniquement sur l'expérience tirée de la participation aux campagnes d'évaluation NIST.

Les développements plus spécifiques réalisés dans le cadre de cette étude concernent moins la problématique scientifique que l'aspect technique de la réalisation du démonstrateur et de l'intégration des modalités parole et visage. Cette partie technique inclut notamment l'adaptation de la structure interne d'AMIRAL aux exigences de cette intégration, impliquant le regroupement d'éléments épars sous forme d'une bibliothèque de fonctions à l'interface proprement définie et à la stabilité assurée. De plus, le matériel sur lequel repose la plateforme étant défini comme générique, une couche d'abstraction a été définie à la base d'AMIRAL pour assurer le maximum d'indépendance par rapport à l'architecture matérielle et logicielle (système d'exploitation) utilisée⁷. Ces développements à l'aspect purement technique ont une importance qui va au delà du cadre de la réalisation du démonstrateur de l'étude F9904. Intégrés à la plateforme AMIRAL, ils lui confèrent une stabilité et une facilité d'accès accrues propres à assurer un développement plus rapide des expériences menées dans le cadre de la recherche.

7.2.3 La modalité visage

Le choix de la technique utilisée pour la reconnaissance du visage a été guidé par la nécessité pour le système de fonctionner sur un ordinateur personnel de base en ayant recours à des composants standards et peu coûteux, y compris pour la caméra. La résolution réduite des images capturées par une caméra standard rendant trop imprécise une modélisation du visage par localisation des caractéristiques biométriques telles que la position des yeux, du nez, de la bouche, la hauteur et la largeur du visage, etc., la méthode retenue ici repose sur l'analyse des niveaux de luminance présents dans l'image. La reconnaissance d'un visage est réalisée par mesure de corrélation de luminance entre le visage testé et le visage de référence, sur la base des travaux présentés dans [Pigeon 1999]. Le choix de cette méthode de préférence à d'autres techniques exploitant les niveaux de luminance de l'image aux performances supérieures se justifie par sa vitesse d'exécution, adaptée à une implémentation sur un ordinateur personnel de base.

⁷AMIRAL est utilisé régulièrement sur machines Sun sous Solaris, HP sous HP-UX, Apple sous MacOS X et PC compatibles Intel sous GNU/Linux et Windows/Cygwin.

Les caractéristiques biométriques utilisées se limitent à une zone rectangulaire du visage comprenant les yeux et le nez. Ces éléments sont en effet relativement invariables et caractéristiques d'un individu. Leur pouvoir discriminant est illustré à la figure 7.4 qui présente quatre portraits à l'aspect subjectivement très différent mais dont seuls les yeux et le nez diffèrent réellement.



FIG. 7.4 – Collaboration LIARMA — Illustration de la motivation du choix des yeux et du nez comme caractéristiques biométriques exploitées par le module de reconnaissance du visage : les 4 portraits présentés ici ne diffèrent que par les yeux et le nez ; la forme du visage, les cheveux, la bouche, sont strictement identiques.

La figure 7.5 illustre le principe de reconnaissance par extraction de cette fenêtre du visage candidat pour la superposer au visage de référence. Une distance est alors calculée entre le visage issu de cette superposition et le visage de référence, la comparaison à un seuil permettant d'accepter ou de rejeter le visage candidat.

Le positionnement de la fenêtre à extraire du visage candidat est réalisé automatiquement par détection du point situé entre les deux yeux.

7.2.4 Normalisation par modèle du monde pour la modalité visage

Le travail simultané sur les deux modalités parole et visage a inspiré le transfert d'une technique de l'une vers l'autre. La normalisation des scores par rapport de vraisemblances (correspondant à un test d'hypothèse bayésien), technique couramment utilisée dans le domaine de la reconnaissance automatique du locuteur (cf. chapitre 2, p. 42), a été appliquée ici dans le cadre de la reconnaissance du visage.

Le test de l'hypothèse inverse est réalisé par comparaison un modèle du monde. Ce modèle est obtenu par la mise en correspondance d'une centaine de visages qui sont ensuite moyennés. La figure 7.6 présente ce visage moyen.

7.2.5 Fusion

La partie de l'étude F9904 consacrée à la fusion des modalités a été réalisée par la RMA avant le début de la collaboration LIARMA. Le lecteur se reportera à [Verlinde 1999]

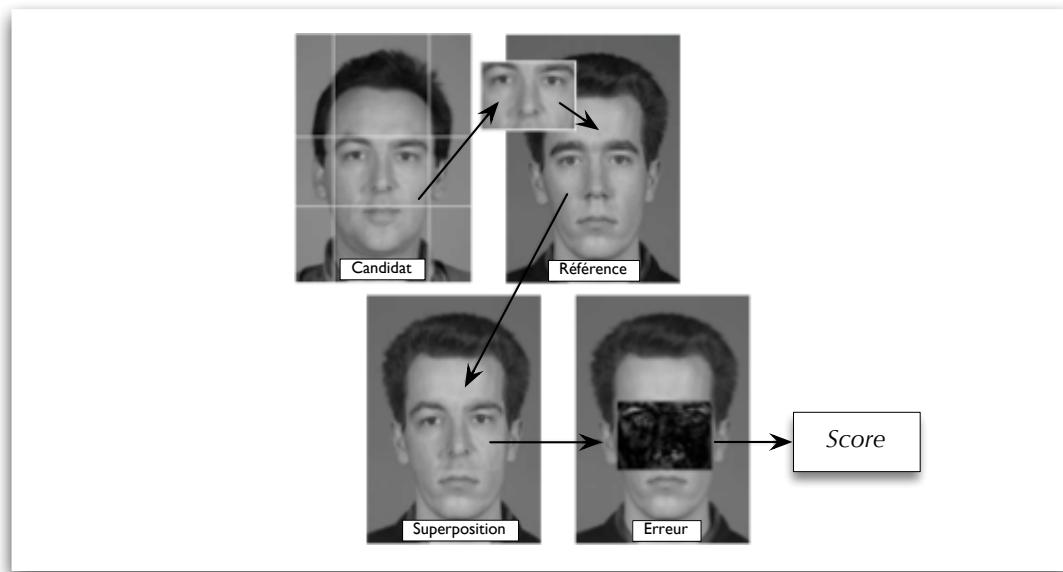


FIG. 7.5 – Collaboration LIARMA — Illustration du principe de base du module de reconnaissance de visage.

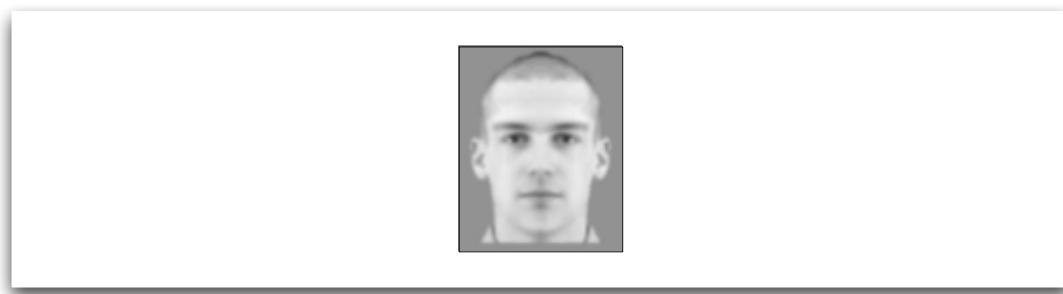


FIG. 7.6 – Collaboration LIARMA — Application de la technique de normalisation par rapport de vraisemblances à la reconnaissance de visage : visage moyen utilisé comme modèle du monde.

et [Pigeon 2000] pour une présentation complète de la technique proposée, qui repose sur le principe de la régression logistique. L'expérience menée à l'occasion de la campagne d'évaluation NIST 99 pour valider cette méthode de fusion, en l'appliquant aux réponses de l'ensemble des systèmes candidats à la tâche de vérification du locuteur, est présentée dans [Pigeon 1999]. Malgré la forte corrélation présentée par ces réponses (tous les systèmes traitant la même modalité et reposant pour la plupart sur des approches similaires), les résultats obtenus sont très satisfaisants, la fusion apportant un gain significatif par rapport au meilleur des systèmes individuels.

7.2.6 Démonstrateur

Conformément aux spécifications de l'étude F9904, un prototype de reconnaisseur bimodal a été réalisé. Ce prototype impémente les techniques présentées ici sur un ordinateur personnel standard (sous un système d'exploitation à base Linux) intégrant une carte d'acquisition sonore et une caméra de type *webcam* grand public.

Pour le module de reconnaissance du locuteur, le démonstrateur intègre l'ensemble des techniques présentées au chapitre 4. Les modèles de locuteurs sont des GMM à 128 composantes à matrices de covariance diagonales, appris par adaptation bayésienne d'un modèle du monde. Ce modèle du monde est estimé à partir de 4 heures de données issues de diverses éditions des campagnes d'évaluation NIST. Il s'agit d'un modèle de type UBM, indépendant du genre. Les enregistrements utilisés mélangeant les divers types de combinés utilisés pour les campagnes NIST. Cette variabilité du type de matériel d'enregistrement a pour but de tendre vers un modèle du monde indépendant du type de matériel utilisé, en accord avec l'indépendance au matériel définie dans les spécifications de l'étude F9904. La normalisation des scores est réalisée par la méthode WMAP, reposant sur une fonction de normalisation estimée sur un ensemble de données similaire à celui utilisé pour le modèle du monde.

Le recours à ces données pour le modèle du monde et la fonction de normalisation se justifie par l'absence de données enregistrées directement sur le prototype. Cette absence de données d'exploitation ne permet pas non plus d'envisager une évaluation des performances du démonstrateur. Une telle évaluation serait pourtant intéressante car autorisant la validation de la technique de fusion lorsqu'appliquée à deux modalités décorrélées, ainsi que du recours à une normalisation par modèle du monde pour de la reconnaissance du visage. Dans le cadre de la reconnaissance du locuteur, cette évaluation aurait permis de mesurer la pertinence du recours à des technologies développées dans le cadre de la recherche et optimisées au cours des campagnes d'évaluation NIST pour développer un système à visée applicative.

7.3 Bilan

Les deux projets exposés ici, bien que présentant des problématiques différentes, ont une caractéristique commune : la difficulté, voire l'impossibilité, de disposer au cours du développement de données représentatives des futures conditions d'exploitation du système. Cette caractéristique fait apparaître la nécessité de dériver de tels projets applicatifs de systèmes développés et optimisés dans un autre contexte, où existent les outils nécessaires à cette optimisation. Les systèmes développés dans le cadre de la recherche répondent à ce critère, étant régulièrement évalués et confrontés à l'état de l'art

en particulier dans le cadre des campagnes d'évaluation NIST.

Faute de solution plus adaptée, le choix a donc été fait de baser les développements réalisés dans le cadre de projets applicatifs sur le même noyau de techniques constituant la base de la plateforme de recherche AMIRAL, afin de profiter de l'expérience accumulée et de l'optimisation apportée à ce noyau à travers la participation régulière aux campagnes d'évaluation NIST.

Cependant, le bien-fondé de cette approche n'a pu être vérifié ici. En effet, dans le cadre du projet MTM comme dans celui de la collaboration LIARMA, la mise en exploitation du système à la fin de son développement n'est pas encore effective. Cette étape permettra de valider (ou d'invalider) l'approche décrite, par l'évaluation des performances obtenues en conditions réelles.

Toutefois, au delà de son intérêt potentiel pour la performance des systèmes applicatifs, cette approche a montré un avantage en termes de bénéfices retirés par l'outil de recherche. En effet, du fait de la nécessité d'utiliser cet outil comme base des développements réalisés dans le cadre de projets applicatifs, une partie de ces développements peut plus facilement être intégrée en retour et bénéficier à la plateforme de recherche. En particulier, les efforts réalisés en termes de stabilité du moteur de reconnaissance et de restructuration du code, indispensables dans un cadre applicatif mais fréquemment négligés dans le cadre de la recherche au profit d'un développement plus anarchique, représentent un gain pour l'effort de recherche qui bénéficie dès lors d'une base mieux définie, documentée et plus stable.

Troisième partie

Conclusion et perspectives

Chapitre 8

Conclusion et perspectives

Ce travail de thèse s'inscrit dans le cadre de la reconnaissance automatique du locuteur en mode indépendant du texte. Ce domaine, longtemps limité à la détection ou la vérification de l'identité d'une personne à partir d'un échantillon de sa voix (à travers les tâches connues dans la littérature sous les noms d'Identification du Locuteur et de Vérification du Locuteur), s'est considérablement élargi depuis quelques années suite au progrès des algorithmes utilisés ainsi que de la puissance de traitement disponible. Le champ d'application de la reconnaissance automatique du locuteur s'est notamment ouvert au traitement de documents de durée plus longue et impliquant plusieurs locuteurs. Les principales tâches portant sur ce type de documents, codifiées au fur et à mesure de leur apparition dans la littérature, sont : la détection de locuteur dans un document multi-locuteurs, le suivi de locuteur, la segmentation en locuteurs d'un document audio (ou multimédia) et enfin la recherche d'un locuteur commun à plusieurs documents (notamment en vue d'apprendre un modèle pour ce locuteur). La codification de ces tâches s'est effectuée notamment à l'occasion de leur intégration au sein de la campagne internationale d'évaluation des performances des systèmes de RAL organisée annuellement par le *National Institute of Standards and Technology* américain. Cette campagne est devenue au fil des ans le standard de fait pour l'évaluation des recherches menées dans le domaine de la reconnaissance automatique du locuteur (et aussi, de par ce rôle, un guide influençant l'orientation de la recherche en RAL, par la proposition et la définition de nouvelles tâches).

Ce document aborde le domaine de la reconnaissance automatique du locuteur à travers la présentation du développement, de la mise en œuvre et de l'évolution de la plateforme logicielle de reconnaissance du locuteur AMIRAL. Ce travail répondait à deux attentes. En premier lieu, la plateforme AMIRAL était définie comme un support pour différents travaux de recherche menés dans le domaine de la RAL. Afin de remplir ce rôle, AMIRAL devait montrer la souplesse nécessaire pour être adaptable facilement aux nouvelles directions de recherche apparaissant régulièrement dans le domaine de la RAL. Le premier objectif de cette thèse était donc la mise au point de la plateforme de recherche AMIRAL et l'évaluation régulière de ses performances en les comparant à l'état de l'art, à travers la participation aux campagnes d'évaluation NIST. Le second objectif de ce travail était lié à son financement par l'intermédiaire de deux projets à visée applicative, le projet LIARMA correspondant à une collaboration entre le LIA et l'École Royale Militaire de Bruxelles concernant le développement d'un prototype de vérificateur d'identité biométrique et bimodal (voix et visage) et le projet européen IST/MTM concernant

l'intégration d'un module de RAL à un assistant numérique personnel. Afin de répondre aux attentes définies par ces deux projets, une partie de cette thèse a été consacrée au transfert de technologies du domaine de la recherche vers le monde applicatif par l'intermédiaire du développement de systèmes de RAL sur la base de la plateforme de recherche AMIRAL.

8.1 AMIRAL, outil pour la recherche

Outil de recherche orienté initialement vers le traitement des seules tâches d'identification et de vérification du locuteur, la plateforme AMIRAL a rapidement évolué pour inclure les directions de recherche traitant des documents multi-locuteurs. Cette évolution s'est faite au rythme de l'apparition de ces nouvelles directions dans le domaine. En particulier, elle a été guidée par l'intégration des tâches correspondantes au sein des campagnes d'évaluation NIST.

Dans le but de favoriser cette évolutivité permettant l'intégration régulière du traitement de nouvelles tâches, l'architecture retenue pour le système est modulaire, privilégiant la réutilisation des connaissances et des techniques d'une tâche à l'autre. Pour ce faire, l'ensemble des développements repose sur un noyau de techniques communes (correspondant aux bases de l'approche statistique en reconnaissance automatique du locuteur). Ce noyau, correspondant pour une bonne part aux techniques utilisées pour la tâche de vérification du locuteur, a continuellement évolué pour intégrer l'état de l'art. L'évaluation de ses performances et leur optimisation a été réalisée par la participation régulière aux campagnes d'évaluation NIST pour la tâche dite de *One Speaker Detection* (vérification du locuteur). Le traitement des diverses autres tâches, étant construit sur cette base, a de ce fait profité également de cet effort d'optimisation.

Le résultat est un système complet de reconnaissance automatique du locuteur, présenté chaque année pour évaluation au cours des campagnes NIST et traitant toutes les principales tâches qui y ont été proposées (à l'exception de quelques variantes selon les années), au fur et à mesure de leur apparition. L'objectif poursuivi au fil des ans est double : intégrer les techniques représentant l'état de l'art du domaine (en particulier en ce qui concerne les techniques de base de la reconnaissance du locuteur) pour maintenir un bon niveau de performances et permettre l'exploration de nouvelles voies de recherche. Cette exploration a notamment débouché sur la mise au point d'une nouvelle méthode de normalisation des scores (WMAP). Elle a également abouti à la définition d'une nouvelle approche de la tâche de segmentation en locuteurs (HMM évolutif).

Malgré les performances légèrement en retrait de l'état de l'art obtenues par le cœur du système pour la tâche de vérification automatique du locuteur, le bilan de la mise en œuvre et de l'évolution d'AMIRAL en tant que plateforme destinée à la recherche est positif et permet d'affirmer que le premier objectif défini pour ce travail de thèse a été atteint. En effet, la plateforme AMIRAL a fait preuve de la souplesse et de l'évolutivité nécessaires à son adaptation à l'ensemble des directions de recherche explorées par le LIA dans le domaine de la RAL. Dans ce cadre, elle a servi de support au développement de solutions novatrices, telles la technique de normalisation WMAP ou la méthode de segmentation à base de HMM évolutif que ses performances ont placé à l'état de l'art (les performances obtenues par cette approche ont permis à la plateforme AMIRAL de se classer en première place pour la tâche correspondante lors de la campagne d'évaluation NIST de 2002).

De plus, au delà de ce bilan en termes de performances, la richesse de la plateforme AMIRAL est à considérer en termes d'expérience et de connaissance du domaine acquises au cours de son développement et de la participation régulière aux campagnes d'évaluation NIST. De ce point de vue, les efforts consacrés au développement de la plateforme AMIRAL ont trouvé leur aboutissement dans le projet ALIZÉ.

ALIZÉ est un projet du programme Technolangue initié par le LIA et d'autres membres du consortium ÉLISA. Il a pour but de pérenniser le savoir-faire acquis par ÉLISA, dont AMIRAL est la pièce majeure, par le développement en commun d'une plateforme de reconnaissance du locuteur *open source*. Dès la définition du projet, ALIZÉ a été conçu pour être système ouvert, extensible et utilisable par tous comme outil de recherche¹.

Bien que la structure interne sous-jacente soit nouvelle, le développement de la plateforme définie par le projet ALIZÉ a su mettre à profit l'expérience retirée de la mise en œuvre de la plateforme AMIRAL. Le développement initial, réalisé sur une période de deux ans entre 2003 et 2004, a permis de proposer un système de vérification du locuteur pour l'édition 2004 de la campagne d'évaluation NIST. Ce système a affiché lors de cette évaluation des performances à l'état de l'art. Les résultats obtenus le classent parmi les meilleurs systèmes présentés, signe de l'arrivée à maturité du projet ALIZÉ.

8.2 AMIRAL, outil de transfert de technologie

Parallèlement à son évolution en tant que plateforme de recherche, AMIRAL a servi de base à des développements à visée plus applicative, dans le cadre de divers projets auxquels a pris part le LIA. Parmi ces projets, les deux principaux ont été le projet MTM et la collaboration LIARMA, sources de financement de ce travail de thèse. Le premier comprenait l'intégration de technologies vocales (reconnaissance de la parole et du locuteur) à l'interface utilisateur d'un système multimédia embarqué sur une plateforme de type PDA, dans le cadre d'un projet européen d'une durée de deux ans impliquant une dizaine de partenaires académiques et industriels. Le second, fruit de la coopération entre le LIA et l'École Royale Militaire de Bruxelles, consistait en une étude d'un système multi-modal de vérification d'identité reposant sur la reconnaissance simultanée du visage et de la voix.

La participation à ces projets a permis d'aborder la problématique posée par le transfert de technologies du domaine de la recherche vers le monde applicatif. En effet, le développement des systèmes de reconnaissance du locuteur dans le cadre de ces projets s'est reposé sur les développements réalisés dans le cadre de la plateforme de recherche AMIRAL. La motivation à la base de cette approche était double. D'une part, la réutilisation de composants logiciels développés pour les activités de recherche permet d'assurer un transfert direct de l'expérience et des connaissances acquises dans ce cadre. D'autre part, les projets applicatifs évoqués ici présentaient comme point commun de ne mettre à disposition lors de la phase de développement aucune donnée correspondant aux futures conditions d'exploitation de l'application. Des enregistrements réalisés dans ces conditions sont pourtant nécessaires à plusieurs étapes de la mise au point et à l'optimisation d'un système de reconnaissance du locuteur, notamment pour le réglage de la normalisation des scores ou d'un seuil de décision. Devant

¹La plateforme logicielle de reconnaissance du locuteur développée dans le cadre du projet ALIZÉ est livrée au public sous licence GNU LGPL (*Library General Public License*).

l'absence de telles données, la décision de fonder sur la plateforme de recherche AMIRAL le développement des systèmes destinés à ces projets permet de recourir pour ces derniers à un ensemble de paramètres réglés et optimisés dans le cadre des campagnes d'évaluation NIST. Bien que ces campagnes d'évaluation offrent des conditions évidemment différentes des conditions d'exploitation définies pour les diverses applications considérées, elles s'attachent à recréer autant que possible les difficultés auxquelles peut être confronté un système de reconnaissance du locuteur dans un cadre applicatif réel. De ce fait, les campagnes d'évaluations NIST semblent représenter une solution viable pour pallier l'absence de données d'exploitation lors de la mise au point d'un système à visée applicative.

Sur la base de cette conclusion, l'approche consistant à baser la réalisation de projets applicatifs sur les outils développés dans le cadre de la recherche a été mise en œuvre et a permis l'aboutissement de quatre projets de ce type : MTM, LIARMA, RAVOL et Certivox. Toutefois, si l'approche décrite a effectivement permis la mise en œuvre de ces projets, sa validation n'a pu être menée à bien dans aucune de ces réalisations. En effet, la mise en exploitation du système à la fin de son développement n'est encore effective dans aucun des projets applicatifs cités. Une étude complète incluant cette étape de mise en exploitation reste donc à réaliser. Elle devra permettre de valider l'approche décrite, par l'évaluation des performances obtenues en conditions réelles.

8.3 Perspectives

L'approche mise en œuvre dans le cadre de cette thèse pour mener à bien le développement de projets applicatifs repose sur le transfert de technologies issues du domaine de la recherche. La principale motivation de cette approche est l'intuition que l'estimation des performances dans un cadre strict tel que les campagnes d'évaluation NIST, sur un corpus de taille suffisante, est suffisamment significative pour être révélatrice des performances à attendre de l'application en situation d'exploitation, en dépit des différences existant entre les conditions d'exploitation et celles de l'évaluation. Toutefois, comme il a été dit au paragraphe précédent, cette théorie n'a encore fait l'objet d'aucune validation expérimentale. Une suite à donner aux travaux présentés ici sera donc de réaliser une évaluation significative, dans le cadre d'un ou plusieurs projets applicatifs, de la corrélation entre les performances obtenues par une plateforme de recherche dans un cadre tel que les campagnes d'évaluation NIST et les performances réellement affichées par un système dérivé une fois mis en situation d'exploitation.

De même, l'idée sous-jacente qu'une optimisation apportée au noyau de base se traduira automatiquement par un gain de performances pour les développements reposant dessus, doit faire l'objet du même effort de validation. Si cet effet a bien été observé dans le cadre de la plateforme de recherche (les performances en segmentation en locuteurs, par exemple, ayant profité des améliorations apportées aux techniques de vérification du locuteur), la même relation reste à démontrer entre les performances du noyau et celles des systèmes applicatifs complets qui en sont dérivés. En effet, dans ce dernier cas, des facteurs supplémentaires peuvent entrer en jeu, comme les limites (par exemple en termes de taille des modèles) induites par les contraintes propres aux applications visées, susceptibles de bloquer l'expression au niveau applicatif des gains de performances observés sur le noyau. Une étude de ce cas sera donc intéressante, dans le but de clarifier la stratégie à adopter pour le transfert de technologies vers le

monde applicatif.

Enfin, l'intégration d'autres modalités (telles que la reconnaissance du visage, mais aussi de la signature, etc.) dans la plateforme pourrait faire l'objet d'une étude à plus long terme. Poursuivre le mouvement amorcé dans le cadre de la collaboration LIARMA par l'application à la reconnaissance d'une technique issue de la RAL (la normalisation des scores par modèle du monde) en tentant de dégager ici aussi un noyau commun de méthodes applicables à l'ensemble des diverses modalités offre une voie de recherche intéressante.

Quatrième partie

Annexes

Annexe A

Résultats divers

Les résultats d'expériences complémentaires exposés ici illustrent certains des choix techniques présentés au chapitre 4.

Tous ont été obtenus sur un sous-ensemble du corpus utilisé pour la campagne d'évaluation NIST 2001. Ce sous-ensemble correspond aux locuteurs féminins du corpus. Il se compose de 506 locuteurs, appris chacun sur 2 minutes de parole enregistrées en une session. Les signaux de test, d'une durée moyenne de 30 secondes, sont au nombre de 3026. Chacun est confronté à plusieurs identités différentes, le nombre total de tests effectués étant de 40657.

Le choix de ce sous-ensemble, motivé par le gain de temps qu'induit un corpus plus petit lors d'une expérience, est justifié par la difficulté qu'il présente. En effet, lors de l'analyse des campagnes d'évaluation NIST, la séparation des résultats par genre fait apparaître des performances de reconnaissance sur les locuteurs féminins systématiquement en retrait des performances observées pour les locuteurs masculins. Le niveau de performances atteint sur le sous-corpus choisi ici doit donc être au moins équivalent lors de l'application du système au corpus complet.

Excepté dans le cas de l'étude sur la taille des modèles, les résultats présentés ont été obtenus avec des modèles de locuteurs à 128 composantes à matrices de covariance diagonales, estimés par adaptation bayésienne à partir du modèle du monde. Le modèle du monde, dépendant du genre et du type de combiné, a été appris à partir de 2 heures de données correspondant à 310 locuteurs différents, extraites du corpus de la campagne d'évaluation NIST 1999.

La normalisation des scores se limite à un rapport de vraisemblances.

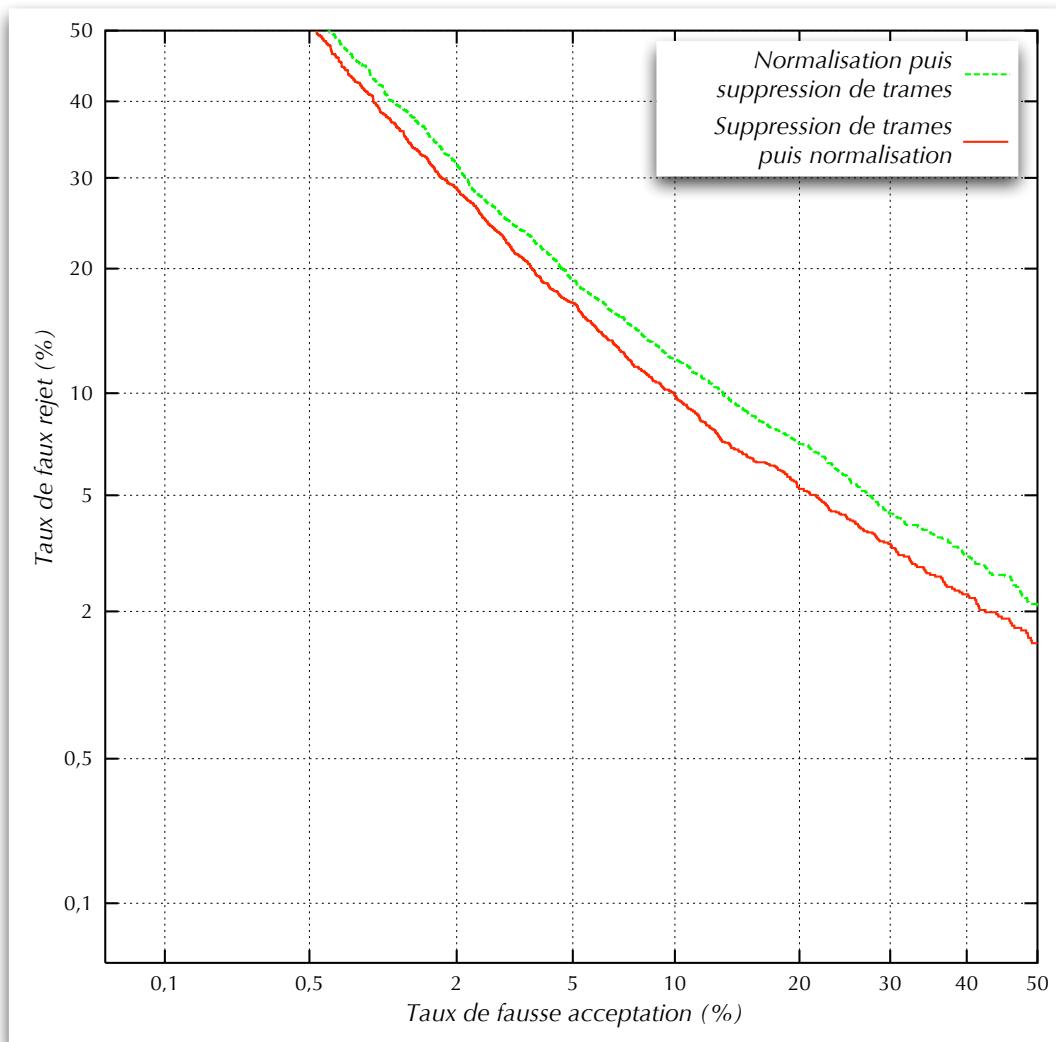


FIG. A.1 – Influence de l'ordre d'application des traitements post-paramétrisation – Comparaison des résultats obtenus sur le corpus NIST 2001 en effectuant d'abord la suppression des trames de basse énergie, puis la normalisation des vecteurs acoustiques, avec les résultats obtenus en appliquant d'abord la normalisation.

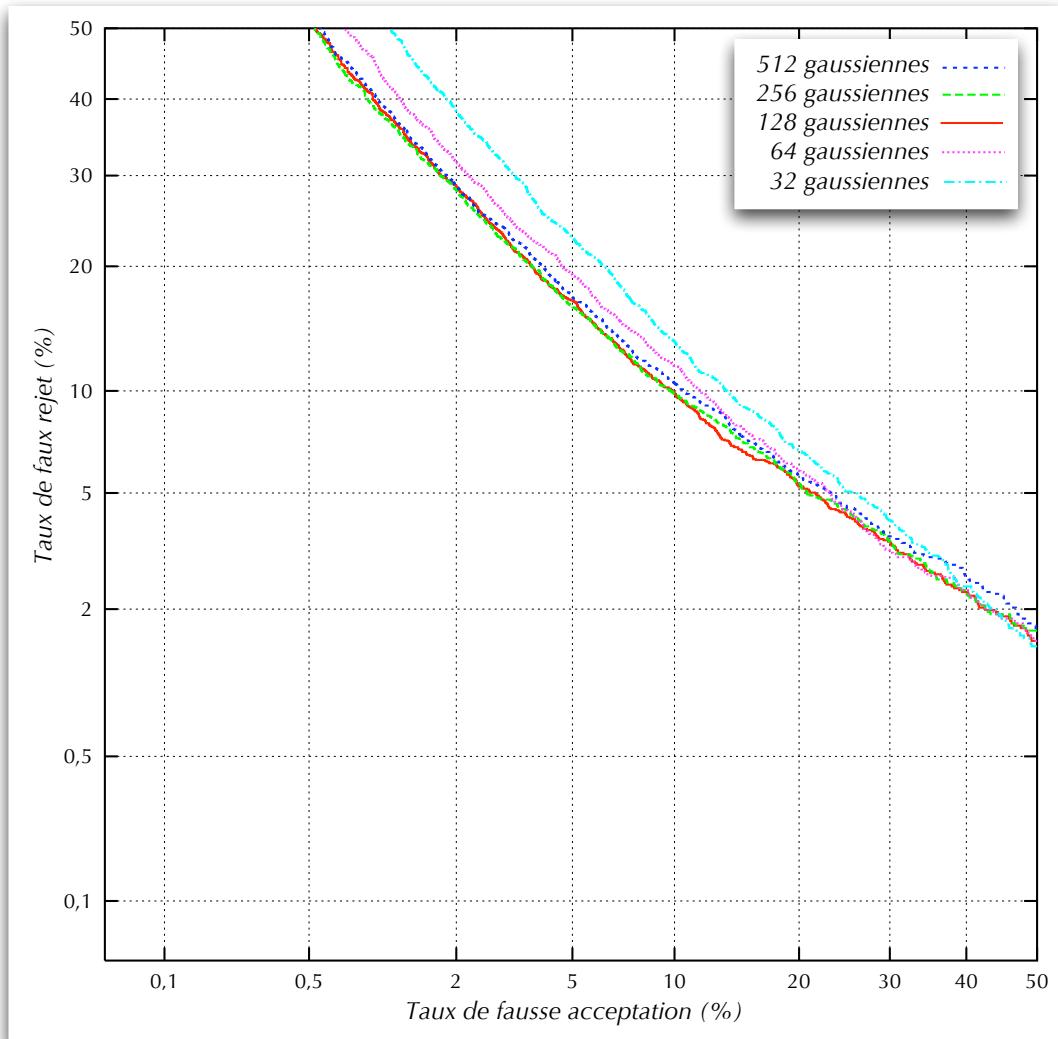


FIG. A.2 – Choix du nombre de composantes pour les modèles – Comparaison des résultats obtenus sur le corpus NIST 2001 pour des modèles matrices diagonales à 32, 64, 128, 256 et 512 gaussiennes appris par adaptation d'un modèle du monde dépendant du type de combiné.

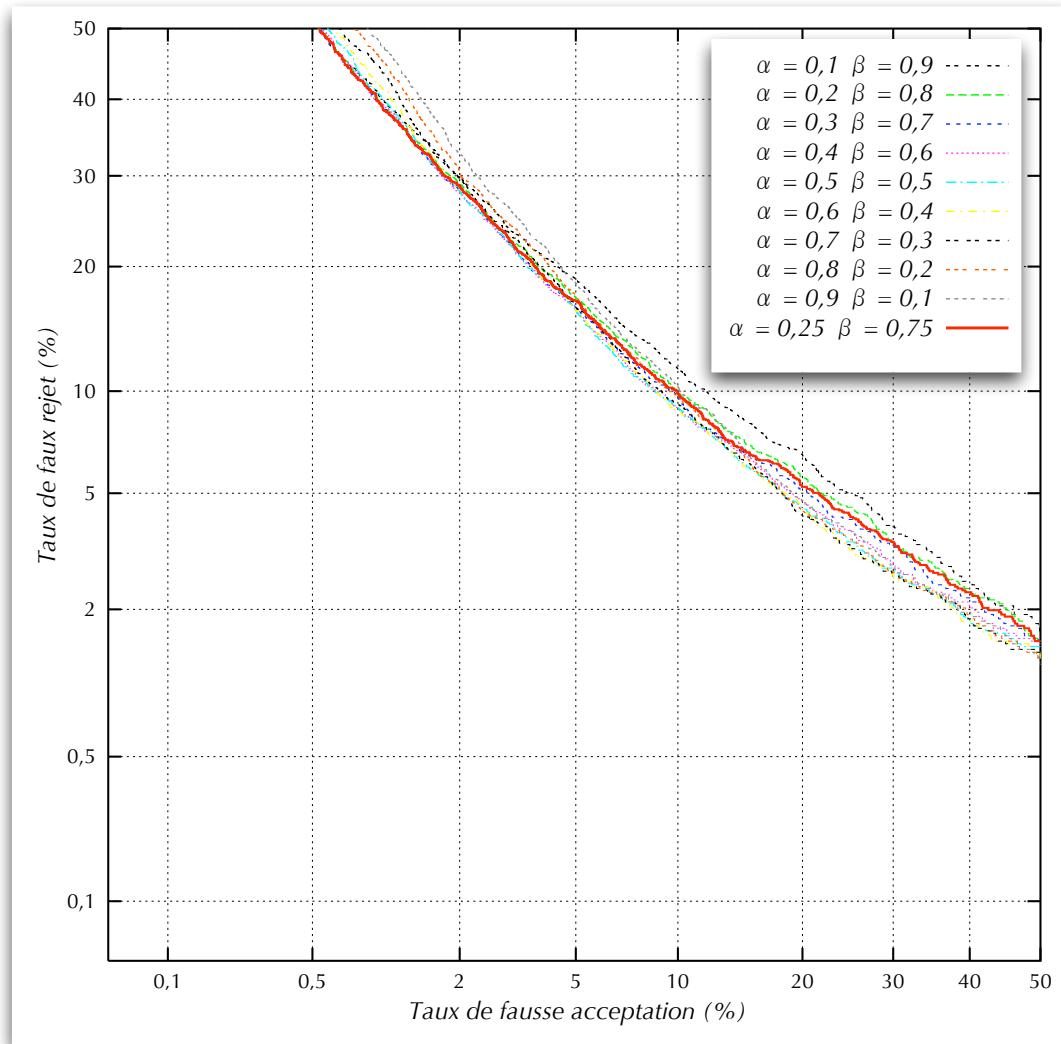


FIG. A.3 – Choix des coefficients α et β pour l'apprentissage des modèles de locuteurs par adaptation – Résultats obtenus sur le corpus NIST 2001 en faisant varier α/β de 0,1/0,9 à 0,9/ 0,1 par incrément de 0,1.

Annexe B

Structure logicielle d'AMIRAL

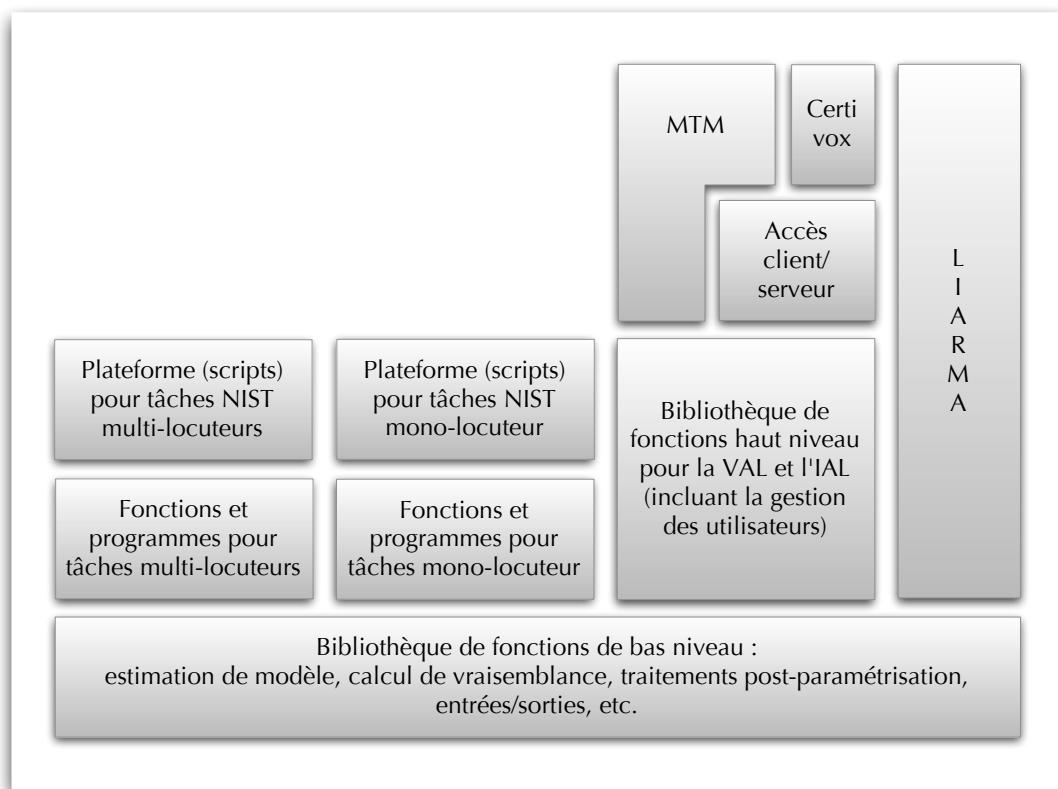


FIG. B.1 – *Structure en couche des divers blocs logiciels composant le système AMIRAL*
– Chaque bloc est conçu en tirant parti des fonctions offertes par le(s) bloc(s) de niveau directement inférieur.

Annexe C

Bibliographie personnelle

Revues internationales

Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacrétaz D. et Reynolds D. A., A tutorial on text-independent speaker verification, dans *EURASIP Journal of Applied Signal Processing, Special issue on biometric signal processing*, 2004(4), 2004.

Fredouille C., Bonastre J.-F. et Merlin T., AMIRAL : a block-segmental multirecognizer architecture for automatic speaker recognition, dans *Digital Signal Processing (DSP), a review journal — Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.

ÉLISA, The ELISA systems for NIST 99 evaluation in speaker detection and tracking, dans *Digital Signal Processing (DSP), a review journal — Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.

Conférences internationales

Bonastre J.-F., Meignier S. et Merlin T., Speaker detection using multi-speaker audio files for both enrollment and test, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, 2003.

Magrin-Chagnolleau I., Gravier G. for the ELISA consortium, Overview of the ELISA consortium research activities, dans *2001 : a Speaker Odyssey – The Speaker Recognition Workshop*, 2001.

Fredouille C., Bonastre J.-F. et Merlin T., Bayesian approach based decision in speaker verification, dans *2001 : a Speaker Odyssey – The Speaker Recognition Workshop*, 2001.

Bellot O., Matrouf D., Bonastre J.-F. et Merlin T., Additive and convolutive noise compensation for speaker recognition, dans *International Conference on Spoken Language Processing (ICSLP 2000)*, 2000.

Meignier S., Bonastre J.-F., Fredouille C. et Merlin T., Evolutive HMM for speaker tracking system, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, 2000.

Bonastre J.-F., Delacourt P., Fredouille C. et Merlin T., A speaker tracking system based on speaker turn detection for NIST evaluations, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, 2000.

Fredouille C., Bonastre J.-F. et Merlin T., Similarity normalization method based on world model and a posteriori probability for speaker verification dans *European Conference on Speech Communication and Technology (Eurospeech 99)*, 1999.

Conférences francophones

Lefort L., Merlin T., Bonastre J.-F. et Nocera P., Le projet MTM — Reconnaissance de la parole et du locuteur sur une plateforme embarquée, dans *XXIVèmes Journées d'Études sur la Parole (JEP 2002)*, 2002.

Meignier S., Bonastre J., Fredouille C. et Merlin T., Modèle de Markov évolutif pour les tâches de suivi de locuteurs, dans *XXIIIèmes Journées d'Études sur la Parole (JEP 2000)*, 2000.

Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T. et Wellekens C. J., Différentes stratégies pour le suivi du locuteur, dans *Reconnaissance des Formes et Intelligence Artificielle (RFIA 2000)*, 2000.

Workshops

Meignier S., Merlin T., Blouet R. et Bonastre J.-F., NIST 2002 speaker recognition evaluation : LIA results, dans *NIST 2002 Speaker Recognition Workshop*, 2002.

Gravier G., Kharroubi J., Chollet G., Bimbot F., Blouet R., Seck M., Bonastre J.-F., Fredouille C., Merlin T., Pigeon S., Verlinde P., Cernocky J., Petrovska D., Nedic B., Magrin-Chagnolleau I. et Durou G., The ELISA'99 speaker recognition and tracking, dans *Workshop on Automatic Identification Advanced Technologies (AutoId)*, 1999.

Fredouille C., Bonastre J.-F. et Merlin T., Segmental Normalization for Robust Speaker Verification, dans *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999.

Merlin T., Bonastre J.-F. et Fredouille C., Non Directly Acoustic Process for Costless Speaker Recognition and Indexing, dans *COST 254 Workshop*, 1999.

Autres

Pigeon S. et Merlin T., Rapport Final relatif à l'étude F9904 : Reconnaissance de Personnes Combinant les Modalités Biométriques de la Parole et de l'Image, École Royale Militaire de Bruxelles, 2002.

Merlin T. et Lefort L., MTM Project Quaterly Reports, 2000-2001.

Merlin T., Reparamétrisation du signal de parole par projection dans un espace de représentation des caractéristiques individuelles, Mémoire de DEA, Université de la Méditerranée, 1998.

Bibliographie

- [Adami 2003] Adami A., Mihaescu R., Reynolds D. et Godfrey J., Modeling prosodic dynamics for speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome IV, pages 788–791, Hong Kong, 2003.
- [Andrews 2001] Andrews W., Kohler M., Campbell J. et Godfrey J., Phonetic, idiolectical, and acoustic speaker recognition, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 55–63, Crète (Grèce), 2001.
- [Artières 1993] Artières T. et Gallinari P., Neural models for extracting speaker characteristics in speech modelization systems, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2263–2266, Berlin (Allemagne), 1993.
- [Atal 1974] Atal B. S., Effectiveness of linear predictive characteristics of the speech waves for automatic speaker identification and verification, *Journal of the Acoustical Society of America (JASA)*, 55(6), 1974.
- [Atal 1976] Atal B. S., Automatic recognition of speakers from their voices, *Proceedings of the IEEE*, 64(4) :460–475, 1976.
- [Auckenthaler 2000] Auckenthaler R., Carey M. et Lloyd-Thomas H., Score normalization for text-independent speaker verification system, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Banziger 2000] Banziger T., Klasmeyer G., Johnstone T., Kamceva T. et Scherer K. R., Améliorer les systèmes de vérification automatique du locuteur en intégrant la variabilité émotionnelle : Méthodes et premières données, dans *XXIIIèmes Journées d'Études sur la Parole (JEP)*, pages 341–344, Aussois (France), 2000.
- [Ben 2002] Ben M., Blouet R. et Bimbot F., A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando (États Unis), 2002.
- [Bennani 1994] Bennani Y. et Gallinari P., Connectionist approaches for automatic speaker recognition, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 95–102, Martigny (Suisse), 1994.
- [Bennani 1990] Bennani Y., Soulie F. F. et Gallinari P., A connectionist approach for automatic speaker identification, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 265–268, 1990.
- [Besacier 1998] Besacier L., *Un modèle parallèle pour la reconnaissance automatique du locuteur*, Thèse de doctorat, Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse, Avignon (France), 1998.

- [Besacier 2004] Besacier L., Ariyaeenia A. M., Mason J. S., Bonastre J.-F., Mayorga P., Fredouille C., Meignier S., Siau J., Evans N., Auckenthaler R. et Stapert R., Voice biometrics over the Internet in the framework of COST action 275, *EURASIP Journal of Applied Signal Processing, Special issue on biometric signal processing*, 2004(4) :466–479, 2004.
- [Besacier 2000a] Besacier L. et Bonastre J.-F., Subband approach for automatic speaker recognition, *European Journal of Signal Processing*, 2000.
- [Besacier 2000b] Besacier L., Bonastre J.-F. et Fredouille C., Localization and selection of speaker specific information with statistical modelling, dans *Speech Communication*, tome 31, pages 89–106, 2000.
- [Besacier 2000c] Besacier L., Grassi S., Dufaux A., Ansorge M. et Pellandini F., GSM speech coding and speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istamboul (Turquie), 2000.
- [Bimbot 1995] Bimbot F., Magrin Chagnolleau I. et Mathan L., Second-order statistical measures for text-independent speaker identification, dans *Speech Communication*, tome 17(1-2), pages 177–192, 1995.
- [Bimbot 1992] Bimbot F., Mathan L., De Lima A. et Chollet G., Standard ant target driven ar-vector models for speech analysis and speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5–8, San Francisco (États Unis), 1992.
- [Boë 1999] Boë L.-J., Bimbot F., Bonastre J.-F. et Dupont P., De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique, *Langues*, 2(4), 1999.
- [Boë 2001] Boë L.-J., Bonastre J.-F. et Bimbot F., Pourquoi la justice doit arrêter les expertises vocales, *Justice*, 169 :5–11, 2001.
- [Bogert 1963] Bogert B., Healy M. et Tukey J., The quefrency analysis of time series for echoes : cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, dans Rosenblatt M., rédacteur, *Symposium on Time Series Analysis*, pages 209–243, John Wiley & Sons, New York (États Unis), 1963.
- [Bonastre 2003a] Bonastre J.-F., Bimbot F., Boë L.-J., Campbell J., Reynolds D. et Magrin-Chagnolleau I., Person authentication by voice : a need for caution, dans *European Conference on Speech Communication and Technology (Eurospeech)*, Genève (Suisse), 2003.
- [Bonastre 2000a] Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T. et Wellekens C. J., Différentes stratégies pour le suivi du locuteur, dans *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 123–129, Paris (France), 2000.
- [Bonastre 2000b] Bonastre J.-F., Delacourt P., Fredouille C., Merlin T. et Wellekens C. J., A speaker tracking system based on speaker turn detection for NIST evaluations, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istamboul (Turquie), 2000.
- [Bonastre 2003b] Bonastre J.-F., Meignier S. et Merlin T., Speaker detection using multi-speaker audio files for both enrollment and test, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome II, pages 77–80, Hong Kong, 2003.
- [Bonastre 2005] Bonastre J.-F., Wils F. et Meignier S., ALIZE, a free toolkit for speaker recognition, 2005, soumission pour acceptation à ICASSP 05.

- [Booth 1993] Booth I., Barlow M. et Watson B., Enhancements to DTW and VQ decision algorithms for speaker recognition, dans *Speech Communication*, tome 13(3-4), pages 427–433, 1993.
- [Boves 1998] Boves L. et Koolwaaij J., Speaker verification in WWW applications, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 178–181, Avignon (France), 1998.
- [Campbell 2003a] Campbell J., Reynolds D. et Dunn R., Fusing high- and low-level features for speaker recognition, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2665–2668, Genève (Suisse), 2003.
- [Campbell 2002] Campbell W., Generalized linear discriminant sequence kernels for speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 161–164, Orlando (États Unis), 2002.
- [Campbell 2003b] Campbell W., Campbell J., Reynolds D., Jones D. et Leek T., Phonetic speaker recognition with support vector machines, *Advances in Neural Information Processing*, 15, 2003.
- [Campbell 2004] Campbell W., Reynolds D. et Campbell J., Fusing discriminative and generative methods for speaker recognition : experiments on Switchboard and NFI/TNO field data, dans *Odyssey 04, The ISCA Speaker and Language Recognition Workshop*, pages 41–44, Tolède (Espagne), 2004.
- [Carey 1991] Carey M. J., Parris E. S. et Bridle J., A speaker verification system using alpha-nets, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, pages 397–400, Toronto (Canada), 1991.
- [Chetouani 2004] Chetouani M., Faundez-Zanuy M., Gas B. et Zarader J., A new non-linear speaker parameterization algorithm for speaker identification, dans *Odyssey 04, The ISCA Speaker and Language Recognition Workshop*, Tolède (Espagne), 2004.
- [Cheung 1978] Cheung R. et Eisenstein B., Feature selection via dynamic programming for text independent speaker identification, *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 26 :397–403, 1978.
- [Delacourt 2000] Delacourt P., *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio*, Thèse de doctorat, Institut Eurecom, Nice (France), 2000.
- [Dempster 1977] Dempster A. P., Laird N. M. et Rubin D. B., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.
- [Doddington 2001] Doddington G., Speaker recognition based on idiolectal differences between speakers, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 4, pages 2521–2524, Aalborg (Danemark), 2001.
- [Doddington 1985] Doddington G. R., Speaker recognition. identifying people by their voices, *Proceedings of the IEEE*, 73(11) :1651–1664, 1985.
- [Dong 2002] Dong X., Zhaojun W. et Yingchun Y., Exploiting support vector machines in hidden Markov models for speaker verification, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1329–1332, Denver (États Unis), 2002.
- [Dunn 2001] Dunn R., Quatieri T., Reynolds D. et Campbell J., Speaker recognition from coded speech in matched and mismatched conditions, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 115–120, Crète (Grèce), 2001.

- [Eatoock 1994] Eatoock J. et Mason J., A quantitative assesment of the relative speaker discriminant properties of phonemes, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 133–136, Adelaide (Australie), 1994.
- [ELISA 2000] ELISA, The ELISA systems for the NIST 99 evaluation in speaker detection and tracking, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Engelmann 2001] Engelmann U., Schröter A., Borälv E., Schweitzer T. et Meinzer H., Mobile teleradiology : All images everywhere, dans *CARS 2001 : Proceedings of the 15th International Congress and Exhibition*, pages 798–803, Amsterdam, 2001.
- [Fine 2001] Fine S., Navratil J. et Gopinath R., Enhancing GMM scores using SVM "hints", dans *European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg (Danemark), 2001.
- [Frederickson 1994] Frederickson S. E. et Tarassenko L., Radial basis functions for speaker identification, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 107–110, Martigny (Suisse), 1994.
- [Fredouille 2000a] Fredouille C., *Approche statistique pour la reconnaissance automatique du locuteur : informations dynamiques et normalisation bayesienne des vraisemblances*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 2000.
- [Fredouille 1999] Fredouille C., Bonastre J.-F. et Merlin T., Similarity normalization method based on world model and a posteriori probability for speaker verification, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 2, pages 983–986, Budapest (Hongrie), 1999.
- [Fredouille 2000b] Fredouille C., Bonastre J.-F. et Merlin T., AMIRAL : a block-segmental multirecognizer architecture for automatic speaker recognition, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Fredouille 2001] Fredouille C., Bonastre J.-F. et Merlin T., Bayesian approach based decision in speaker verification, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 77–81, Crète (Grèce), 2001.
- [Furui 1977] Furui S., An analysis of long-term variation of feature parameters of speech and its application to talker recognition, *Electron. Communication*, 57 A :34–42, 1977.
- [Furui 1981a] Furui S., Cepstral analysis technique for automatic speaker verification, *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 29(2) :254–272, 1981.
- [Furui 1981b] Furui S., Comparison of speaker recognition methods using static features and dynamic features, *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 29(3) :342–350, 1981.
- [Furui 1994] Furui S., An overview of speaker recognition technology, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 1–9, Martigny (Suisse), 1994.
- [Gauvain 1994] Gauvain J. L. et Lee C. H., Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, 1994.
- [Gazit 2001] Gazit R., Metzger Y. et Toledo O., Speaker verification over cellular networks, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 125–128, Crète (Grèce), 2001.

- [Goldstein 1975] Goldstein U., Speaker-identifying features based on formant tracks, *Journal of the Acoustical Society of America (JASA)*, 59(1) :176–182, 1975.
- [Gravier 1998] Gravier G. et Chollet G., Comparison of normalization techniques for speaker recognition, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 97–100, Avignon (France), 1998.
- [Gravier 1999] Gravier G., Kharroubi J., Chollet G., Bimbot F., Blouet R., Seck M., Bonastre J.-F., Fredouille C., Merlin T., Pigeon S., Verlinde P., Cernocky J., Petrovska D., Nedic B., Magrin-Chagnolleau I. et Durou G., The ELISA'99 speaker recognition and tracking, dans *Workshop on Automatic Identification Advanced Technologies (AutoId)*, Summit (États Unis), 1999.
- [Grenier 1980] Grenier Y., Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur, dans *IXèmes Journées d'Études sur la Parole (JEP)*, pages 163–171, Strasbourg (France), 1980.
- [Griffin 1994] Griffin C., Matsui T. et Furui S., Distance measures for text-independent speaker recognition based on mar model, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 309–312, Adelaide (Australie), 1994.
- [Gu 2001] Gu Y. et Thomas T., A text-independent speaker verification system using support vector machines classifier, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1765–1769, Aalborg (Danemark), 2001.
- [Hattori 1992] Hattori H., Text-independent speaker recognition using neural networks, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 153–156, San Francisco (États Unis), 1992.
- [Heck 2000] Heck L., Konig Y., Sonmez K. et Weintraub M., Robustness to telephone handset distortion in speaker recognition by discriminative feature design, *Speech Communication*, 31(2-3) :181–192, 2000.
- [Heck 1997] Heck L. et Weintraub M., Handset-dependent background models for robust text-independent speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 2, pages 1071–1074, Munich (Allemagne), 1997.
- [Hermansky 1994] Hermansky H. et Morgan N., Rasta processing of speech, *IEEE Transactions on Speech and Audio Processing*, 2(4) :578–589, 1994.
- [Higgins 1991] Higgins A. L., Bahler L. et Porter J., Speaker verification using randomized phrase prompting, *Digital Signal Processing*, 1(2) :89–106, 1991.
- [Homayounpour 1995] Homayounpour M. M., *Vérification vocale d'identité : dépendante et indépendante du texte*, Thèse de doctorat, Université de Paris-Sud centre d'Orsay, Paris (France), 1995.
- [Jacob 2000] Jacob B., Mariéthoz J., Gravier G. et Bimbot F., Robustesse de la vérification du locuteur par un mot de passe personnalisé, dans *XXIIIèmes Journées d'Études sur la Parole (JEP)*, pages 357–360, Aussois (France), 2000.
- [Karlsson 1998] Karlsson I., Banziger T., Dankovicová J., Johnstone T., Lindberg J., Melin H., Nolan F. et Scherer K., Speaker verification with elicited speaking-styles in the verivox project, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 207–210, Avignon (France), 1998.
- [Kharroubi 2001] Kharroubi J., Petrovska-Delacrétaz D. et Chollet G., Combining GMM's with support vector machines for text-independent speaker verification, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1757–1760, Aalborg (Danemark), 2001.

- [Koolwaaij 2000] Koolwaaij J. et Boves L., Local normalization and delayed decision making in speaker detection and tracking, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 1(1-3) :113-132, 2000.
- [Lefort 2002] Lefort L., Merlin T., Bonastre J.-F. et Nocera P., Le projet MTM – reconnaissance de la parole et du locuteur sur une plateforme embarquée, dans *XXIVèmes Journées d'Études sur la Parole (JEP)*, Nancy (France), 2002.
- [Li 1988] Li K. P. et Porter J. E., Normalizations and selection of speech segments for speaker recognition scoring, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, pages 595-598, New York (États Unis), 1988.
- [Lindberg 1997] Lindberg J. et Melin H., Text-prompted versus sound prompted passwords in speaker verification system, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 22-25, Rhôdes (Grèce), 1997.
- [Magrin-Chagnolleau 2001] Magrin-Chagnolleau I., Gravier G. et the ELISA consortium, Overview of the ELISA consortium research activities, dans *2001, A Speaker Odyssey. The Speaker Recognition Workshop*, pages 67-72, Crète (Grèce), 2001.
- [Magrin-Chagnolleau 1999] Magrin-Chagnolleau I., Rosenberg A. et Parthasarathy S., Detection of target speakers in audio databases, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix (États Unis), 1999.
- [Magrin-Chagnolleau 1996] Magrin-Chagnolleau I., Wilke J. et Bimbot F., Further investigation on ar-vector models for text-independent speaker identification, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 401-404, Atlanta (États Unis), 1996.
- [Martin 2000] Martin A. et Przybocki M., The NIST 1999 speaker recognition evaluation - an overview, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Martin 1997] Martin A. F. et Przybocki M. A., The DET curve in assessment of detection task performance, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1895-1898, Rhôdes (Grèce), 1997.
- [Mason 1989] Mason J. S., Oglesby J. et Xu L., Codebooks to optimise speaker recognition, dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 267-270, Paris (France), 1989.
- [Matsui 1992] Matsui T. et Furui S., Comparison of text-independent speaker recognition methods using VQ-distortion and discrete-continuous HMMs, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 157-160, San Francisco (États Unis), 1992.
- [Matsui 1993] Matsui T. et Furui S., Concatenated phoneme models for text-variable speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, Minneapolis (États Unis), 1993.
- [Matsui 1994] Matsui T. et Furui S., Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 125-128, Adelaide (Australie), 1994.
- [McLaughlin 1999] McLaughlin J., Reynolds D. et Gleason T., A study of computation speed-ups of the GMM-UBM speaker recognition system, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 3, pages 1215-1218, Budapest (Hongrie), 1999.

- [Meignier 2002a] Meignier S., *Indexation en locuteurs de documents sonores : segmentation d'un document et appariement d'une collection*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 2002.
- [Meignier 2000] Meignier S., Bonastre J.-F., Fredouille C. et Merlin T., Evolutive HMM for speaker tracking system, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istamboul (Turquie), 2000.
- [Meignier 2001] Meignier S., Bonastre J.-F. et Igounet S., E-HMM approach for learning and adapting sound models for speaker indexing, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 175–180, Crète (Grèce), 2001.
- [Meignier 2002b] Meignier S., Bonastre J.-F. et Magrin-Chagnolleau I., Speaker utterances tying among speaker segmented audio documents using hierarchical classification : towards speaker indexing of audio databases, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 573–576, Denver (États Unis), 2002.
- [Meignier 2004] Meignier S., Moraru D., Fredouille C., Besacier L. et Bonastre J.-F., Benefits of prior acoustic segmentation for automatic speaker segmentation, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montréal (Canada), 2004.
- [Montacié 1992] Montacié C. et Le Floch J.-L., Ar-vector models for free-text speaker recognition, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 611–614, Banff (Canada), 1992.
- [Naik 1994] Naik J., Speaker verification over the telephone : databases, algorithms and performance assessment, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 31–38, Martigny (Suisse), 1994.
- [NIST 2000] NIST, The NIST 2000 speaker recognition evaluation plan, <http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm>, 2000.
- [NIST 2001] NIST, The NIST 2001 speaker recognition evaluation plan, <http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v53.%pdf>, 2001.
- [Nocera 2002] Nocera P., Linares G. et Massonié D., Principe et performances du décodeur parole continue Speeral, dans *XXIVèmes Journées d'Études sur la Parole (JEP)*, Nancy (France), 2002.
- [Oglesby 1990] Oglesby J. et Mason J. S., Optimisation of neural models for speaker identification, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 261–264, 1990.
- [Oglesby 1991] Oglesby J. et Mason J. S., Radial basis function networks for speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 393–396, Toronto (Canada), 1991.
- [Olsen 1997] Olsen J., A two-stage procedure for phone based speaker verification, dans *Audio, Video-based Biometric Person Authentication (AVBPA)*, pages 219–226, Crans-Montana (Suisse), 1997.
- [Oppenheim 1968] Oppenheim A. et Schafer R., Homomorphic analysis of speech, *IEEE Transactions on Audio and Electroacoustics*, 16(2) :221–226, 1968.
- [Oppenheim 1989] Oppenheim A. et Schafer R., *Discrete-time signal processing*, Prentice Hall, Englewood Cliffs (États Unis), 1989.
- [O'Shaughnessy 1986] O'Shaughnessy D., Speaker recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, pages 4–17, 1986.

- [Paoloni 1996] Paoloni A., Ragazzini S. et Ravaioli G., Predictive neural networks in text independent speaker verification : an evaluation on the siva database, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 2423–2426, Philadelphia (États Unis), 1996.
- [Pelecanos 2001] Pelecanos J. et Sridharan S., Feature warping for robust speaker verification, dans *2001, A Speaker Odyssey, The Speaker Recognition Workshop*, pages 213–218, Crète (Grèce), 2001.
- [Peskin 1993] Peskin B. et al., Topic and speaker identification via large vocabulary continuous speech recognition, dans *ARPA workshop on human language technology*, pages 119–124, Princeton (États Unis), 1993.
- [Peskin 2003] Peskin B., Navratil J., Abramson J., Jones D., Klusacek D., Reynolds D. et Xiang B., Using prosodic and conversational features for high-performance speaker recognition : report from JHU WS'02, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome IV, pages 792–795, Hong Kong, 2003.
- [Petrovska-Delacrétaz 2000] Petrovska-Delacrétaz D., Cernocky J., Hennebert J. et Chollet G., Segmental approaches for automatic speaker verification, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3) :198–212, 2000.
- [Pigeon 1999] Pigeon S., *Authentification multimodale d'identité*, Thèse de doctorat, Université Catholique de Louvain, Louvain-la-Neuve (Belgique), 1999.
- [Pigeon 2000] Pigeon S., Druyts P. et Verlinde P., Applying logistic regression to the fusion of the NIST99 1-speaker submissions, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3) :237–248, 2000.
- [Pruzansky 1963] Pruzansky S., Pattern matching procedure of automatic talker recognition, dans *Journal of the Acoustical Society of America (JASA)*, tome 35, pages 354–358, 1963.
- [Przybocki 2004] Przybocki M. et Martin A., NIST speaker recognition evaluation chronicles, dans *Odyssey 04, The ISCA Speaker and Language Recognition Workshop*, pages 15–22, Tolède (Espagne), 2004.
- [Przybocki 1998] Przybocki M. A. et Martin A. F., NIST speaker recognition evaluation - 97, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 120–123, Avignon (France), 1998.
- [Przybocki 1999] Przybocki M. A. et Martin A. F., Two-channel telephone data for speaker detection and speaker tracking, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 5, pages 2215–2218, Budapest (Hongrie), 1999.
- [Quatieri 2000] Quatieri T., Singer E., Dunn R., Reynolds D. et Campbell J., Speaker recognition using G.729 speech codec parameters, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istamboul (Turquie), 2000.
- [Rabiner 1989] Rabiner L. R., A tutorial on hidden markov models and selected applications in speech recognition, *IEEE Transactions on Speech and Audio Processing*, 77(2) :257–285, 1989.
- [Reynolds 1992] Reynolds D., *A Gaussian mixture modeling approach to text-independent speaker identification*, Thèse de doctorat, Georgia Institute of Technology, (États Unis), 1992.

- [Reynolds 1994] Reynolds D., Experimental evaluation of features for robust speaker identification, *IEEE Transactions on Speech and Audio Processing*, 2(2-3) :639–643, 1994.
- [Reynolds 1995] Reynolds D., Speaker identification and verification using gaussian mixture speaker models, dans *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, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, Atlanta (États Unis), 1996.
- [Reynolds 1997] Reynolds D., Comparison of background normalization methods for text-independent speaker verification, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 2, Rhôdes (Grèce), 1997.
- [Reynolds 2003a] Reynolds D., Channel robust speaker verification via feature mapping, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome II, pages 53–56, Hong Kong, 2003.
- [Reynolds 2003b] Reynolds D., Andrews W., Campbell J., Navratil J., Peskin B., Adami A., Jin Q., Klusacek D., Abramson J., Mihaescu R., Godfrey J., Jones D. et Xiang B., The SuperSID project : exploiting high-level information for high-accuracy speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome II, pages 784–787, Hong Kong, 2003.
- [Reynolds 2000] Reynolds D., Quatieri T. et Dunn R., Speaker verification using adapted gaussian mixture models, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Rosenberg 1976] Rosenberg A. E., Automatic speaker verification, a review, *Proceedings of the IEEE*, 64(4) :475–487, 1976.
- [Rosenberg 1994] Rosenberg A. E., Lee C.-H. et Soong F. K., Cepstral channel normalization techniques for HMM-based speaker verification, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1835–1838, Yokohama (Japon), 1994.
- [Rosenberg 1998] Rosenberg A. E., Magrin-Chagnolleau I., Parthasarathy S. et Huang Q., Speaker detection in broadcast speech databases, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1339–1342, Sydney (Australie), 1998.
- [Rosenberg 1996] Rosenberg A. E. et Parthasarathy, Speaker background models for connected digit password speaker verification, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, pages 81–84, Atlanta (États Unis), 1996.
- [Sambur 1975] Sambur M. R., Selection of acoustic features for speaker identification, *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 23(2) :176–182, 1975.
- [Scherer 1998] Scherer K. R., Johnstone T. et Sangsue J., L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole, dans *XXIIèmes Journées d'Études sur la Parole (JEP)*, pages 249–257, Martigny (Suisse), 1998.
- [Schmidt 1996] Schmidt M. et Gish H., Speaker identification via support vector classifiers, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 1, pages 105–108, Atlanta (États Unis), 1996.

- [Setlur 1994] Setlur A. et Jacobs T., Results of a speaker verification service trials using HMM models, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 639–642, Martigny (Suisse), 1994.
- [Siu 1992] Siu M.-H., Rohlicek R. et Gish H., An unsupervised, sequential learning algorithm for segmentation of speech waveforms with multi speakers, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco (États Unis), 1992.
- [Siu 1991] Siu M.-H., Yu G. et Gish H., Segregation of speakers for speech recognition and speaker identification, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto (Canada), 1991.
- [Sonmez 1999] Sonmez K., Heck L. et Weintraub M., Speaker tracking and detection with multiple speakers, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 5, pages 2219–2222, Budapest (Hongrie), 1999.
- [Soong 1992] Soong F. K., Rosenberg A. E., Rabiner L. R. et Juang B. H., A vector quantization approach to speaker recognition, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 387–390, Tampa (États Unis), 1992.
- [Vapnik 1995] Vapnik V., *The nature of statistical learning theory*, Springer Verlag, New York (États Unis), 1995.
- [Verlinde 1999] Verlinde P., *Contribution à la vérification multimodale d'identité en utilisant la fusion de décisions*, Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris (France), 1999.
- [Voiers 1964] Voiers W., Perceptual basis of speaker identity, *Journal of the Acoustical Society of America (JASA)*, 36 :1065–1073, 1964.
- [van Vuuren 1996] van Vuuren S., Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch, dans *International Conference on Spoken Language Processing (ICSLP)*, pages 1788–1791, Philadelphia (États Unis), 1996.
- [van Vuuren 1999] van Vuuren S., *Speaker recognition in a time-feature space*, Thèse de doctorat, Oregon Graduate Institute of Science and Technology, Portland (États Unis), 1999.
- [Wan 2003] Wan V. et Renals S., SVMSVM : Support vector machine speaker verification methodology, dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, tome 2, pages 221–224, Hong Kong, 2003.
- [Weber 2000] Weber F., Peskin B., Newman M., Corrada-Emmanuel A. et Gillick L., Speaker recognition on single- and multispeaker data, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3) :75–92, 2000.
- [Wolf 1972] Wolf J. J., Efficient acoustic parameters for speaker recognition, dans *Journal of the Acoustical Society of America (JASA)*, tome 51(6.2), pages 2044–2054, 1972.
- [Yu 1995] Yu K., Mason J. S. et Oglesby J., Speaker recognition using hidden markov models, dynamic time warping and vector quantisation, dans *IEEE vision, image, and signal processing*, Berlin (Allemagne), 1995.

Index

- algorithme d'apprentissage
 - adaptation, 83, 94, 96, 112
 - EM, 82, 90, 92, 112
- Certivox, 59
- ELISA, 50
- GMM, 39, 66, 71, 80, 81, 92, 112, 126
- HMM, 39
 - HMM évolutif, 109
- LIARMA
 - présentation de la collaboration LIA-RMA, 59
 - réalisation, 133
- MTM
 - présentation du projet, 57
 - réalisation, 122
- NIST
 - NIST 2000, 94
 - NIST 2001, 94, 112, 113
 - NIST 2002, 97, 112, 116, 117
 - NIST 2003, 98, 112
 - NIST 98, 91
 - NIST 99, 92, 107
 - participation à la tâche *One Speaker Detection*, 91
 - participation aux tâches multilocuteurs, 103
 - présentation des campagnes d'évaluation, 50
- normalisation, 41, 74, 80
 - Dnorm, 46, 80
 - Hnorm, 43, 80
 - rapport de vraisemblances, 42, 74
 - Tnorm, 43, 80, 98
 - WMAP, 47, 75, 92, 94
 - Znorm, 43, 80
- RAVOL, 59
- seuillage de la variance, 90, 92, 98
 - SVM, 40
- UBM, 90, 98, 138