N° d'ordre: 3072

THÈSE

Présentée devant

devant l'université de Rennes 1

pour obtenir

le grade de : Docteur de l'université de Rennes 1
Mention Traitement du Signal et Télécommunications

par

Mathieu Ben

Équipe d'accueil : METISS/IRISA
École doctorale : MATISSE
Composante universitaire : Structures et Propriétés de la Matière

Titre de la thèse :

Approches robustes pour la vérification automatique du locuteur par normalisation et adaptation hiérarchique

soutenue le 23 novembre 2004 devant la commission d'examen

M. :	Gérard	FAUCON	Président
MM. :	Jean-François	BONASTRE	Rapporteurs
	Gérard	CHOLLET	
MM. :	Samy	BENGIO	Examinateurs
	Delphine	CHARLET	
	John	MASON	
M. :	Frédéric	BIMBOT	Directeur de thèse

Résumé

Cette thèse est consacrée à l'élaboration et l'évaluation de techniques visant à renforcer la robustesse des systèmes de vérification automatique du locuteur. La vérification automatique du locuteur (VAL) consiste à authentifier l'identité d'une personne en analysant les caractéristiques de sa voix. Ses applications vont du contrôle d'accès à l'authentification d'enregistrements sonores, en passant par des tâches d'étiquetage automatique de documents audio en fonction des intervenants. Utilisés en situation réelle et dans des environnements perturbés, comme les applications téléphoniques notamment, les systèmes de VAL peuvent être confrontés à de fortes variations de conditions d'utilisation, entraînant une augmentation importante des erreurs de reconnaissance. Pour diminuer ce type d'erreurs, les systèmes actuels doivent intégrer des techniques de compensation dont l'objectif est d'atténuer les effets des disparités entre les données d'apprentissage et celles de test. Les approches courantes de compensation ont cependant des faiblesses qui les rendent contraignantes ou inadaptées à certaines situations applicatives. Dans cette thèse, nous développons des techniques destinées à remédier à certaines de ces limitations. Nos travaux s'inscrivent dans l'approche probabiliste pour la VAL, pour laquelle les locuteurs sont modélisés par des modèles de mélange de Gaussiennes et l'étape de décision est basée sur un test d'hypothèses Bayésien. Nous élaborons dans un premier temps de nouvelles techniques de normalisation dont le but est de renforcer la robustesse du processus de décision face aux variabilités rencontrées. Ces normalisations, basées sur l'utilisation de distances de Kullback-Leibler entre modèles de locuteurs, sont appliquées au niveau des scores de vérification et au niveau des modèles eux mêmes. Elles se distinguent des approches conventionnelles par le fait qu'elles ne font appel à aucun corpus de données réelles pour l'estimation des paramètres de normalisation. Nous formalisons également un nouveau cadre pour la vérification du locuteur dans un espace des modèles. Ce cadre conduit à un calcul simplifié des scores de vérification et autorise une manipulation efficace des modèles de locuteurs, offrant ainsi de nombreuses possibilités de normalisation. Les résultats expérimentaux montrent que les techniques proposées peuvent avantageusement remplacer certaines approches courantes, en allégeant considérablement la procédure de vérification. Nous concevons dans un deuxième temps un schéma d'adaptation Bayésienne hiérarchique qui a pour but d'améliorer l'estimation des modèles de locuteurs lorsque la quantité de données d'apprentissage est faible. La technique proposée généralise l'approche classiquement utilisée en VAL, en offrant de plus la possibilité d'intégrer des dépendances entre différentes régions acoustiques occupées par la voix d'un locuteur. La mise en pratique de cette technique est cependant délicate et les conditions dans lesquelles elle a été utilisée n'ont pas permis pour l'instant de mettre en évidence un apport décisif de la méthode. Néanmoins, le cadre théorique introduit est attrayant et offre de multiples perspectives. L'ensemble des techniques étudiées a été évalué sur des bases de données téléphoniques en parole naturelle, dans le cadre des évaluations NIST en reconnaissance du locuteur.

Table des matières

Introduction générale **1**

I Présentation générale et fondements scientifiques de la VAL **7**

1 Positionnement et aspects applicatifs de la VAL **9**
1.1 Authentification automatique des individus 10
1.1.1 Un problème de détection 10
1.1.2 Modes d'authentification 11
1.1.3 Authentification biométrique : définition et caractéristiques . . . 11
1.1.4 La VAL ou authentification biométrique vocale 13
1.2 La voix, modalité d'authentification 14
1.2.1 Caractérisation du locuteur 14
1.2.2 Facteurs limitants 17
1.2.3 Tâches et typologies des systèmes de RAL 19
1.3 Déploiement applicatif des systèmes de VAL 23
1.3.1 Phases d'utilisation et modes de fonctionnement 24
1.3.2 Profils applicatifs et difficultés sous-jacentes 25
1.4 Méthodologie d'évaluation 27
1.4.1 Facteurs de qualité d'un système de VAL 27
1.4.2 Mesure des performances 28
1.4.3 Corpus et campagnes d'évaluation 31
1.5 Approches état de l'art *vs* méthodes émergentes 33
1.6 Problématiques soulevées et orientation du travail 34

2 Approche probabiliste pour la vérification du locuteur **35**
2.1 Fondements théoriques et formalisme associé 36
2.1.1 Modélisation probabiliste des locuteurs 36
2.1.2 Estimation paramétrique des modèles 36
2.1.3 Décision par test d'hypothèses Bayésien 37
2.2 Techniques courantes des systèmes de VAL par approche probabiliste . . 38
2.2.1 Paramétrisation du signal de parole 39

2.2.2 Modélisation probabiliste pour la VAL 42
2.2.3 Processus de décision 44
2.3 Approche état de l'art : vérification du locuteur par GMMs 45
2.3.1 Modélisation par mélange de Gaussiennes 45
2.3.2 Adaptation Bayésienne des GMMs de locuteurs 46
2.3.3 Décision Bayésienne normalisée 49
2.4 Synthèse et objectifs visés 50
2.4.1 Traitement de la robustesse 50
2.4.2 Limitations des méthodes actuelles et objectifs du travail 51

II Techniques de normalisation par distances de Kullback-Leibler 53

3 État de l'art des techniques de normalisation pour la VAL 55
3.1 Problématique et objectifs 56
3.1.1 Techniques de compensation 56
3.1.2 Mise en oeuvre des techniques de normalisation 57
3.2 Normalisations dans l'espace acoustique 58
3.3 Normalisations dans l'espace des modèles 61
3.4 Normalisations dans l'espace des scores 62
3.5 Discussion sur les techniques de normalisation 65

4 Divergences KL et scores de vérification en VAL 67
4.1 Introduction : intérêt d'une mesure liée au modèle 68
4.2 Définition des divergences KL pour la VAL 69
4.3 Liens théoriques avec les scores de vérification 69
4.4 Étude expérimentale des divergences KL 72
4.4.1 Estimation des divergences KL par tirage de Monte Carlo 72
4.4.2 Corrélation avec les scores de vérification 73
4.5 Conclusion : possibilité d'exploitation des distances KL en RAL 77

5 Applications pour la VAL : normalisations de scores et de modèles 79
5.1 D-norm : une technique de normalisation des scores par distance KL . . 80
5.1.1 Rappel sur les normalisations de scores 80
5.1.2 Utilisation des distances KL pour la normalisation de score . . . 80
5.1.3 Développement sur NIST 2001 81
5.1.4 Performances sur les évaluations NIST 2002, 2003 et 2004 86
5.1.5 Conclusion sur la D-norm 89
5.2 D-MAP : normalisation entropique des modèles par adaptation contrainte 89
5.2.1 Introduction : liens entre adaptation et distribution des scores . . 89
5.2.2 Adaptation par critère MAP contraint : D-MAP 90
5.2.3 Développement expérimental 93
5.2.4 Résultats 97

5.2.5 Conclusion sur D-MAP 98
5.3 Vérification du locuteur dans un espace des modèles 99
5.3.1 Motivations : un calcul rapide de distances entre modèles 99
5.3.2 Définition d'une distance simple entre GMMs adaptés 100
5.3.3 Espace des modèles 101
5.3.4 Techniques de normalisation dans l'espace des modèles 103
5.3.5 Mise en oeuvre et résultats 104
5.3.6 Conclusion et perspectives 108

6 Contributions connexes 113
6.1 Contribution en suivi de locuteur et organisation en locuteur 114
6.1.1 Contexte : campagne d'évaluation ESTER 114
6.1.2 Suivi de locuteurs : tâche ESTER-SVL 115
6.1.3 Organisation en locuteur : tâche ESTER-SRL 120
6.2 Contribution en sélection de locuteurs 124
6.2.1 Contexte : projet NEOLOGOS 124
6.2.2 Algorithme de sélection des locuteurs 125
6.2.3 Résultats expérimentaux 127
6.3 Synthèse des travaux en reconnaissance et caractérisation du locuteur . 128

III Adaptation Bayésienne hiérarchique des modèles de locuteurs 131

7 Connaissance a priori et techniques d'adaptation 133
7.1 But recherché : estimation robuste de modèles acoustiques 134
7.2 A priori probabiliste : méthodes Bayésiennes 135
7.2.1 Adaptation MAP 135
7.2.2 Extension : utilisation de corrélations entre paramètres 138
7.3 A priori structurel : adaptation par transformation de modèles 139
7.3.1 Techniques MLLR et MAPLR 139
7.3.2 Modèles de référence et techniques "Eingenvoices" 141
7.4 A priori hiérarchique : méthodes multiéchelles 142
7.5 Synthèse sur les techniques d'adaptation et application en RAL 144

8 H-MAP : adaptation Bayésienne hiérarchique de GMMs 147
8.1 Motivations et positionnement 148
8.2 Modélisation structurelle des dépendances entre Gaussiennes 149
8.2.1 Structure de dépendances 149
8.2.2 Relations de dépendances 149
8.3 Adaptation MAP hiérarchique des moyennes d'un GMM 150
8.3.1 A priori conditionnels Gaussiens 150
8.3.2 Estimation des paramètres et propagation des dépendances . . . 151
8.4 Propriétés de H-MAP 154

8.4.1 Influence de la quantité de données : convergence de H-MAP . . 154
8.4.2 Influence des dépendances dans H-MAP 155
8.4.3 H-MAP et autres méthodes d'adaptation 156
8.5 Synthèse sur le développement théorique de H-MAP 158

9 Développement expérimental et évaluation de H-MAP 161
9.1 Protocole et système de VAL 162
9.2 Développement sur NIST 2003 162
9.2.1 Estimation de la structure de dépendance 162
9.2.2 Adaptation des modèles de locuteur 166
9.2.3 Performances sur NIST 2003 166
9.3 Évaluation sur NIST 2004 170
9.3.1 Description des données 170
9.3.2 Résultats 171
9.3.3 Discussion sur les résultats obtenus 172
9.4 Conclusion sur la mise en application de H-MAP 173

Conclusion générale 175

A Le consortium ELISA 183

B Description des systèmes ELISA/IRISA de 2001 à 2004 185
B.1 Systèmes ELISA et IRISA 2001 185
B.1.1 Système de base ELISA 2001 185
B.1.2 Variante IRISA 2001 186
B.2 Systèmes ELISA et IRISA 2002 187
B.2.1 Système de base ELISA 2002 187
B.2.2 Variante IRISA 2002 187
B.3 Système IRISA 2003 188
B.4 Système IRISA 2004 189

C Algorithme pour le tirage de Monte Carlo à partir d'un GMM 191

Table des figures

1.1 Processus d'authentification 10
1.2 Schéma de distorsion du signal de parole 18
1.3 Système de détection du locuteur 20
1.4 Système d'identification du locuteur 21
1.5 Les deux variantes d'un système de description en locuteur : suivi de locuteurs (haut) et organisation en locuteurs (bas) 22
1.6 Système de VAL en mode apprentissage 24
1.7 Système de VAL en mode test (et adaptation) 25

2.1 Architecture d'un système de VAL 38
2.2 Les différentes étapes de la paramétrisation du signal de parole. 41

3.1 Illustration en une dimension des principes des techniques de normalisation des paramètres acoustique 60

4.1 Corrélations entre divergences KL et moyennes des scores 75
4.2 Corrélations entre distances KL et moyennes des scores 76
4.3 Relation entre écarts types des scores et distances KL 77

5.1 Relation entre distances KL et moyennes des scores D-normalisés 82
5.2 Distributions des moyennes des scores imposteurs et clients, sans normalisation et après application de la D-norm 83
5.3 Relation entre distances KL et écarts types des scores D-normalisés . . . 83
5.4 Courbes DET du système IRISA/ELISA 2001 sans normalisation et avec normalisations D-norm et Z-norm 84
5.5 Courbes DET du système IRISA/ELISA : convergence de la D-norm . . 85
5.6 Courbes DET du système IRISA/ELISA 2002, sans normalisation et avec normalisations D-Norm, T-norm et DT-norm 86
5.7 Exemples de distribution des scores pour un modèle sous-adapté (bas) et un modèle sur-adapté (haut) 91
5.8 Évolution des fonctions $f_X(\beta_X)$ vis-à-vis du coefficient d'adaptation β_X, pour 10 locuteurs femmes de l'évaluation NIST'01 95
5.9 Illustration de la procédure de correction du coefficient d'adaptation β_X 96

5.10 Courbes DET du système IRISA/ELISA 2001 avec différents schémas d'estimation : ML, MAP et D-MAP 98
5.11 Représentations de GMMs : de l'espace acoustique à un espace des modèles 101
5.12 Illustration de la normalisation de modèles dans l'espace des modèles. . 103
5.13 Relation entre les mesures ${D_E}^2(X,\Omega)$ et $KL2_X$ 105
5.14 Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR et des scores basés sur la distance Euclidienne D_E, sans normalisation . 107
5.15 Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR avec normalisation D-norm, et des scores basés sur la distance Euclidienne D_E avec normalisation M-norm 107
5.16 Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR avec normalisation T-norm, et des scores basés sur la distance Euclidienne D_E avec normalisation T-norm 109
5.17 Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR avec normalisation DT-norm, et des scores basés sur la distance Euclidienne D_E avec normalisation M+T-norm 109
5.18 Identification d'une direction dépendante du canal dans l'espace des modèles (illustration) 110

6.1 Tâche ESTER-SVL : courbes DET du système de suivi de locuteur IRISA 2003 sans normalisation de scores et avec normalisation DT-norm 119

7.1 Influence du facteur de confiance τ_k dans l'adaptation MAP 136
7.2 Illustration du principe de l'adaptation MAP 137
7.3 Illustration du principe de l'adaptation par transformations 140
7.4 Arbre de classification de Gaussiennes pour une adaptation hiérarchique 142

8.1 Structure en arbre binaire et GMMs correspondants 149
8.2 Évolution de la proportion attribuée à la moyenne empirique dans l'adaptation H-MAP, en fonction du taux d'occupation de la Gaussienne . . . 155
8.3 Variation du facteur de confiance $\tau^d_{k|\pi(k)}$ en fonction du coefficient de corrélation $\rho^d_{k,\pi(k)}$ 156

9.1 Distribution des coefficients de régression estimés sur les données cellulaires NIST'01 et NIST'02. 164
9.2 Distribution des coefficients de corrélation estimés sur les données cellulaires NIST'01 et NIST'02. 165
9.3 Évolutions des points C_{det} et EER en fonction du facteur de confiance τ avec l'adaptation H-MAP 167
9.4 Courbes DET obtenues avec l'adaptation H-MAP sans normalisation : influence du facteur de confiance 167
9.5 Courbes DET du système IRISA 2003 avec adaptation MAP classique et DT-norm et avec adaptation H-MAP et T-norm 169

9.6 Courbes DET du système IRISA avec adaptation MAP classique et adaptation H-MAP, et normalisation T-norm, pour la tâche 10sec-train/10sec-test de l'évaluation NIST'04. 172

C.1 Principes du tirage aléatoire des Gaussiennes. 192

Liste des tableaux

5.1 Points de fonctionnement minC_{det} du système IRISA/ELISA pour les évaluations NIST'02, NIST'03 et NIST'04, sans normalisation, avec D-norm et avec DT-norm 88
5.2 EER du système IRISA/ELISA pour les évaluations NIST'02, NIST'03 et NIST'04, sans normalisation, avec D-norm et avec DT-norm 88
5.3 EER du système IRISA/ELISA 2001 pour différentes configurations d'adaptation : MAP, D-MAP et MAP+D-norm. 97
5.4 Points C_{det} et EER du système IRISA/ELISA 2004 avec scores LLR et scores basés sur les distances D_E, avec différents schémas de normalisation 108

6.1 Dénomination des tâches de la campagne d'évaluation ESTER 115
6.2 Tâche ESTER-SVL : min%Err et min$HTER$ du système de suivi de locuteurs IRISA 2003 avec différentes normalisations de scores 118
6.3 Tâche ESTER-SRL : taux d'erreur de classification obtenu pour chaque document et globalement sur le corpus "dev", avec segmentation manuelle et automatique 122
6.4 Tâche ESTER-SRL : taux d'erreur de classification obtenu pour chaque document et globalement sur le corpus "test", avec segmentation automatique 123
6.5 Évaluation des performances de l'algorithme de sélection de locuteurs de l'IRISA 127

9.1 Points minC_{det} et EER obtenus avec l'adaptation MAP classique et l'adaptation H-MAP, avec différentes normalisations de scores : corpus femmes NIST'03 168
9.2 Points minC_{det} et EER obtenus avec l'adaptation MAP classique et l'adaptation H-MAP : tâche 10sec-train/10sec-test, NIST'04 171

Introduction générale

La parole est un signal aléatoire complexe qui contient plusieurs types d'informations. En premier lieu, le signal de parole véhicule un message linguistique qui sert à la communication entre individus. Mais la parole transporte également des informations sur l'identité de l'individu ayant prononcé le message. Les humains se servent de ces informations pour identifier les personnes qu'ils connaissent, en particulier lorsqu'ils ne peuvent pas voir leur interlocuteur, au téléphone par exemple.

Les systèmes de Reconnaissance Automatique du Locuteur (RAL) s'intéressent précisément à ces caractéristiques particulières du signal de parole pouvant permettre de reconnaître les individus. Les applications de tels systèmes vont du contrôle d'accès à la description d'enregistrements audio en fonction des locuteurs, en passant par l'authentification de messages parlés. La tâche de Vérification Automatique du Locuteur (VAL) consiste plus spécifiquement à vérifier l'identité proclamée par une personne en analysant les caractéristiques de sa voix. Cette tâche s'inscrit dans le cadre des techniques d'authentification biométrique des individus. Elle y trouve pleinement sa place, d'une part parce que la parole est un moyen naturel et ergonomique de communiquer son identité, et d'autre part parce que la voix est parfois le seul moyen disponible pour authentifier une personne, comme dans les applications téléphoniques notamment.

Problématique

Malgré les avantages que peut procurer l'authentification biométrique vocale, les systèmes de VAL ont souffert jusqu'ici de leur faible robustesse lorsqu'ils sont utilisés en situation réelle, en particulier dans des environnements perturbés. Le signal enregistré contient dans ce cas non seulement la parole du locuteur, mais également des bruits additifs dus à l'environnement sonore de la prise de son, et des distorsions liées au microphone et éventuellement au canal de transmission. Parmi les nombreux profils applicatifs existants, les applications téléphoniques mettent à rude épreuve les systèmes de VAL. Les baisses de performance observées dans ces conditions d'utilisation ont limité le développement de ce type de systèmes au niveau du grand public. L'amélioration de leur robustesse est donc un enjeu majeur des travaux de recherche actuels en reconnaissance du locuteur. Aussi, le travail présenté dans cette thèse est-il

principalement focalisé sur les problèmes de robustesse des systèmes de VAL.

Les difficultés de reconnaissance rencontrées par les systèmes de VAL lorsque les conditions d'utilisation sont très variables viennent du fait que l'on ne dispose jamais en pratique de suffisamment de données pour apprendre à reconnaître la voix d'un individu dans toutes les situations susceptibles d'être rencontrées. Les données d'apprentissage sont nécessairement limitées et ne reflètent donc qu'une partie des conditions possibles d'environnement sonore, d'acquisition du signal et de sa transmission éventuelle. D'autre part, la voix d'un individu évolue avec le temps, ce qui rajoute encore de la variabilité perturbante pour la reconnaissance. Tout biais dans les conditions de test, par rapport aux conditions "apprises" pourra entraîner une erreur de reconnaissance. Pour diminuer ces erreurs, les systèmes de VAL état de l'art intègrent de nombreuses techniques de compensation destinées à minimiser les effets des disparités entre les données d'apprentissage et de test. Trois catégories de techniques de compensation peuvent être répertoriées :

1. des techniques de **normalisation** visant à homogénéiser certaines caractéristiques des éléments intervenant dans le système. Le but est de rendre ces caractéristiques les plus indépendantes possibles des conditions d'utilisation.
2. des techniques d'**adaptation** qui cherchent à adapter les données d'apprentissage (ou un modèle de ces données) aux conditions des données de test, ou inversement.
3. des techniques d'**incorporation de connaissances *a priori*** destinées à combler certains manques dans les données d'apprentissage.

Dans ce document, nous abordons le traitement de la robustesse des systèmes de VAL en développant de nouvelles techniques de normalisation, et une technique utilisant des connaissances *a priori* sous forme de dépendances hiérarchiques entre les régions acoustiques occupées par la voix d'un locuteur.

Cadre Théorique

Le travail que nous présentons dans cette thèse s'intègre dans l'approche probabiliste pour la VAL. Cette approche est actuellement au coeur de la plupart des systèmes, parfois associée à d'autres méthodes de classification comme les SVM (Support Vector Machines). Dans l'approche probabiliste, les locuteurs sont représentés par un modèle stochastique appelé *référence caractéristique*. La théorie des probabilités permet de calculer une mesure de similarité entre un ensemble de données de test et une référence caractéristique correspondant à l'identité proclamée. C'est à partir de cette mesure de similarité, utilisée comme score de vérification, que le système prend la décision d'accepter ou de rejeter le locuteur. Dans les systèmes état de l'art en mode indépendant du texte, les locuteurs sont représentés par des modèles de mélange de Gaussiennes (GMM) qui sont estimés au sens du critère du Maximum A Posteriori (MAP).

Principales Contributions

Les contributions de ce travail de thèse peuvent être résumées en 4 points principaux :

1. l'exploration et la mise au point d'une nouvelle technique de normalisation de score appelée "D-norm" (Distance Normalization), basée sur l'utilisation de distances de Kullback-Leibler (KL) entre modèles de locuteurs. Le principal intérêt de la D-norm d'un point de vue applicatif est qu'elle ne nécessite pas l'utilisation de données externes contrairement aux autres techniques de normalisation de scores. La D-norm a permis d'obtenir des performances équivalentes à celles que procure la Z-norm, une autre technique de normalisation de scores état de l'art.
2. l'exploration et la mise au point d'une technique de normalisation de modèles s'appuyant sur un schéma d'adaptation contrainte appelé "D-MAP" (Distance-constrained MAP). Ce schéma d'adaptation vise à homogénéiser les modèles vis-à-vis de leur distance KL par rapport au modèle du monde. Cela conduit implicitement à une homogénéisation des scores semblable à celle obtenue par la D-norm. Les performances fournies par D-MAP sont similaires à celles obtenues avec une normalisation de scores de type D-norm ou Z-norm.
3. la formalisation et le développement d'un nouveau cadre pour la vérification du locuteur dans un espace des modèles. Dans l'approche proposée, la vérification est basée sur un calcul de distances entre modèles du locuteur, du monde et du test. L'espace des modèles que nous avons défini a permis d'implémenter des techniques efficaces de normalisation et offre de nombreuses perspectives. Les résultats obtenus avec cette approche sont aussi bons, voire légèrement meilleurs que ceux fournis par l'approche conventionnelle.
4. la conception et le développement d'une technique d'adaptation Bayésienne hiérarchique des modèles de locuteurs. Cette technique a pour but d'améliorer la robustesse de l'approche classique d'adaptation MAP de modèles GMMs, vis-à-vis de faibles quantités de données d'apprentissage. Pour cela, l'approche proposée, appelée "H-MAP" (Hierarchical MAP), intègre une structure de dépendances hiérarchiques entre les Gaussiennes d'un GMM de locuteur, autorisant ainsi à certaines composantes ne recevant pas suffisamment de données d'être tout de même adaptées. D'un point de vue théorique, H-MAP permet de généraliser l'adaptation MAP classique et présente de plus certaines similitudes avec d'autres techniques d'adaptation telles que MLLR, MAPLR ou SMAP. L'approche H-MAP est cependant difficile à régler en pratique car plus complexe que le MAP classique. Elle n'a pas apportée pour l'instant d'amélioration décisive mais les résultats obtenus sont encourageants.

Les techniques mentionnées ci-dessus ont été développées et évaluées sur les bases de données des évaluations NIST en reconnaissance du locuteur.

Contexte de Travail

Cette thèse s'est déroulée sous contrat d'allocataire de recherche MENRT dans l'équipe METISS de l'IRISA. Les travaux de thèse ont été dirigés par Frédéric Bimbot, Chargé de Recherche au CNRS, Habilité à Diriger des Recherches et responsable scientifique de l'équipe METISS.

Une partie des travaux a été développée à partir d'une plate-forme de reconnaissance du locuteur commune à différents laboratoires Français regroupés au sein du consortium ELISA [1]. L'annexe A présente brièvement ce consortium et on trouvera une description détaillée des systèmes ELISA et IRISA de 2001 à 2004 dans l'annexe B.

Plan du Document

Ce document s'organise en trois parties.

La première partie permet de situer le cadre de notre travail et de présenter les fondements théoriques nécessaires à la compréhension de la suite du document. Le chapitre 1 présente les généralités sur la VAL ainsi que certains aspects applicatifs qui sont au coeur des problématiques traitées dans cette thèse. Dans le chapitre 2, l'approche probabiliste est présentée de façon formelle, et les techniques état de l'art sont décrites. Cela permet d'identifier certaines faiblesses des techniques actuelles et de définir les points particuliers que nous allons traiter.

La seconde partie est consacrée au développement de nouvelles techniques de normalisation basées sur l'utilisation de distances de Kullback-Leibler entre modèles de locuteurs. Nous y faisons d'abord, dans le chapitre 3, un panorama des techniques de normalisation développées jusqu'ici en VAL. Nous étudions ensuite dans le chapitre 4 les grandeurs sur lesquelles reposent les techniques de normalisation présentées dans la suite : les divergences et distances KL. Le chapitre 5 décrit les développements théoriques et donne les résultats expérimentaux des techniques D-norm, D-MAP et du cadre de vérification du locuteur dans un espace de modèles. Enfin, le chapitre 6 présente des contributions connexes de nos travaux en VAL dans des tâches de suivi de locuteurs et d'organisation en locuteur (dans le cadre des évaluations ESTER) ainsi que dans une tâche de sélection de locuteurs (dans le cadre du projet NEOLOGOS-TECHNOLANGUES).

La troisième partie présente les développements théoriques et expérimentaux d'une nouvelle technique d'adaptation Bayésienne hiérarchique des modèles de locuteurs : H-MAP. Nous commençons par présenter au chapitre 7 les techniques d'adaptation robuste de modèles acoustiques que l'on trouve dans la littérature. Le chapitre 8 est ensuite

1. `http://www.lia.univ-avignon.fr/heberges/ALIZE/ELISA/index.html/`

consacré aux aspects théoriques de l'adaptation H-MAP. Les résultats expérimentaux obtenus avec H-MAP sont présentés dans le chapitre 9, dernier chapitre de cette thèse.

L'ensemble du document est conclu par une récapitulation des principales contributions de cette thèse et par la présentation de quelques perspectives de développements supplémentaires de ces travaux.

Première partie

Présentation générale et fondements scientifiques de la VAL

Chapitre 1

Positionnement et aspects applicatifs de la VAL

Ce chapitre introduit les notions essentielles permettant de situer le cadre du travail présenté dans cette thèse. Il est consacré à la description de l'environnement applicatif de la VAL et des difficultés rencontrées en pratique. Ces difficultés proviennent essentiellement des différents facteurs de variabilités inhérents à la tâche : évolution de la voix, variations des conditions d'acquisition du signal de parole, etc. . En VAL, cette variabilité est particulièrement présente dans les applications téléphoniques pour lesquelles les conditions d'acquisition et de transmission du signal peuvent être très inégales. Un des enjeux majeurs de la recherche en VAL actuellement est d'améliorer la robustesse des systèmes vis-à-vis des différentes sources de variabilité rencontrées. Le renforcement de la robustesse des systèmes de VAL est également l'objectif visé du travail développé dans cette thèse.

Dans ce chapitre, nous situons dans un premier temps la VAL au sein des techniques d'authentification des personnes. En particulier, nous précisons la place et le rôle de la VAL parmi les techniques biométriques. La voix est ensuite présentée comme une modalité possible pour l'authentification des personnes. Les caractéristiques vocales justifiant la faisabilité même de la VAL sont identifiées et nous pointons les principaux facteurs de dégradation des performances liés à ce mode d'authentification. Nous situons également la VAL parmi les différentes tâches de reconnaissance automatique du locuteur. Nous discutons alors de la mise en oeuvre des systèmes de VAL et des contraintes spécifiques à certains profils d'application. Puis, nous décrivons en détail la méthodologie d'évaluation des systèmes de VAL, point essentiel pour la validation des méthodes, l'analyse des performances et la comparaison des systèmes. Enfin, nous donnons un bref aperçu des techniques état de l'art en VAL et des méthodes qui semblent émerger actuellement. Nous concluons le chapitre en faisant une synthèse des problématiques soulevées et en présentant celles qui seront abordées dans la suite du document.

1.1 Authentification automatique des individus

Il est de nos jours devenu crucial dans de nombreuses situations de pouvoir identifier les individus de façon automatique, fiable et rapide. Les retraits d'argent ou paiements automatiques, les connexions aux comptes utilisateurs sur un réseau informatique ou les accès d'individus à des locaux réservés sont autant d'exemples où l'identité des personnes doit être vérifiée quotidiennement.

Par ailleurs, les procédés d'authentification d'enregistrements audios ou vidéos sont de plus en plus utilisés par certaines organisations gouvernementales et militaires comme moyen de renseignement ou de dissuasion, dans le cadre de la lutte contre la criminalité et le terrorisme.

1.1.1 Un problème de détection

D'un point de vue technique, la tâche d'authentification automatique d'une personne peut être considérée comme un problème de détection. Il s'agit en effet de déterminer si le *matériau de test* que présente l'individu pour s'authentifier est suffisamment "concordant" avec une *référence* prédéterminée, correspondant à l'identité proclamée. Cette concordance s'exprime en général par une mesure de similarité entre le matériau de test et la référence, la détection se faisant alors par comparaison de ces deux éléments (cf figure 1.1). Ces éléments - matériau de test et référence - sont communs à tout système d'authentification automatique des personnes mais ils prennent des formes très différentes suivant le mode d'authentification utilisé. Il peut s'agir d'objets matériels (clés, cartes magnétiques,...), d'éléments symboliques (codes, mots de passe,...) ou descriptifs (informations personnelles, caractéristiques biométriques,...).

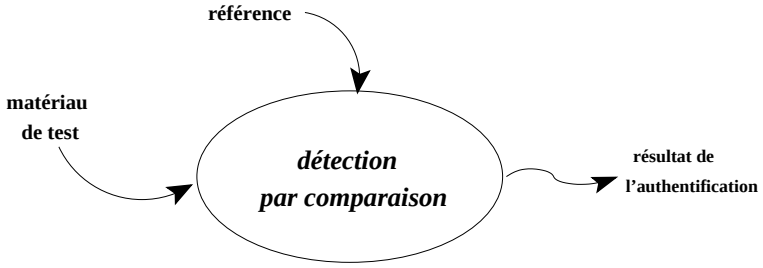


FIG. 1.1 – *Processus d'authentification*

Signalons d'autre part que pour certaines applications, notamment l'accès à des services réservés, la personne se présentant au système d'authentification doit également être identifiée parmi un groupe d'utilisateurs référencés, afin de personnaliser ses paramètres d'utilisation par exemple. Une tâche d'identification est alors couplée à la tâche de détection. On peut encore répertorier d'autres tâches connexes comme le suivi temporel des personnes, dans un contexte de surveillance par exemple, ou la description de documents audiovisuels en fonction des différents intervenants. L'ensemble de ces

tâches sera décrit de façon plus précise dans le cadre spécifique de la VAL à la section 1.2.3.

1.1.2 Modes d'authentification

Actuellement, une grande majorité des applications de contrôle d'accès et d'authentification automatiques est basée sur la présentation par l'individu soumis au contrôle de quelque chose qu'il *connaît*, comme un code ou un mot de passe, ou qu'il *détient*, une clef ou une carte magnétique par exemple. Ces deux modes peuvent être combinés pour une sécurité accrue (c'est le cas de la carte bleue).

Cependant, ces modes d'authentification n'identifient pas la *personne physique* elle-même mais un *objet matériel* ou un *identifiant*. Ainsi, toute personne ayant en sa possession cet objet ou connaissant cet identifiant sera acceptée, même si elle n'en est pas l'utilisateur légitime. En outre, les codes et mots de passe peuvent être oubliés par l'utilisateur ou devinés par quelqu'un d'autre, éventuellement à l'aide de moyens technologiques. Les clefs ou cartes magnétiques peuvent être perdues, dérobées ou falsifiées. Ce sont ici les principaux défauts des modes d'authentification à partir d'une connaissance ou d'une possession. D'autre part, ces modes d'authentification sont également mis en défaut lorsque le contrôle d'identité doit se faire à l'insu de la personne, comme dans le cas des applications de renseignement. Cela exige en effet des moyens d'authentification autres que ceux produits volontairement par l'individu.

L'authentification biométrique offre une réponse naturelle à ces problèmes. Elle propose une alternative aux précédents modes d'authentification en basant la reconnaissance sur des caractéristiques liées à la personne physique elle-même. Cependant, les propriétés intrinsèques des caractéristiques biométriques et l'imprécision de leur mesure peuvent occasionner des problèmes de reconnaissance, rendant ce mode d'authentification non-infaillible.

1.1.3 Authentification biométrique : définition et caractéristiques

1.1.3.1 Caractéristiques biométriques

Étymologiquement, le mot "biométrie" signifie "mesure du vivant". L'authentification biométrique devrait donc se limiter à désigner les techniques mettant en jeu des mesures de caractéristiques physiques humaines comme par exemple la forme du visage ou de la main, la configuration des empreintes digitales, de l'iris ou de la la rétine.

Pourtant, le terme de biométrie est communément élargi pour englober également les techniques cherchant à caractériser des traits comportementaux, tels que la parole ou l'écriture. On parle alors de biométrie "comportementale", par opposition à la biométrie "physique" (c'est-à-dire morphologique ou physiologique). La différence majeure est que les traits physiques d'un individu peuvent difficilement être changés alors que chacun peut intentionnellement modifier son comportement, soit pour tenter d'imiter celui de

quelqu'un d'autre, soit dans le but de ne pas être reconnu. D'autre part, des facteurs extérieurs ou physiologiques peuvent facilement influer sur l'attitude d'une personne.

Il en résulte que les caractéristiques physiques sont en général plus discriminantes et mènent à des performances de reconnaissance plus élevées, à condition que le captage de la caractéristique se fasse de façon précise.

On peut trouver un panorama des caractéristiques biométriques les plus usitées dans l'article introductif [Perronnin et al.02] qui renvoie également à de nombreuses références plus ciblées.

1.1.3.2 Captage et ergonomie

Un captage précis de la caractéristique biométrique se fait souvent malheureusement au détriment de l'ergonomie du système. Ainsi, la prise d'empreintes digitales ou de la forme d'une main exige un contact direct avec le capteur, ce qui est souvent mal accepté par les utilisateurs. De même pour la prise de vue de la rétine qui demande qu'un capteur soit placé tout près de l'oeil. Au contraire, les prises de vue de l'iris et du visage peuvent s'opérer à distance raisonnable et sont mieux acceptées, avec en contrepartie une plus grande variabilité des conditions de captage. La voix, l'écriture et la signature manuscrite, quant à elles, sont considérées par les individus comme des moyens d'identification naturels dont le captage n'engendre que peu de désagréments. Chacun sait cependant que ces trois éléments sont le fruit de "gestes" produits par la personne et qu'ils ont par conséquent des caractéristiques relativement variables.

1.1.3.3 Forces et faiblesses de la biométrie

"Physique" ou "comportementale", la biométrie offre quoiqu'il en soit des avantages certains par rapport aux mots de passe ou à la clef: un trait physique ou un comportement ne s'oublie pas et ne se perd pas. En outre, les caractéristiques biométriques d'un individu ne peuvent être que difficilement "volées" et ne sont en général reproductibles artificiellement qu'avec de fortes contraintes [1].

Malgré cela, la complexité technologique, les coûts de mise en oeuvre et certains problèmes de robustesse des systèmes biométriques en font encore un mode d'authentification peu développé par les industriels. Comme mentionné précédemment, certaines techniques biométriques sont encore mal acceptées par les utilisateurs car jugées trop intrusives.

Signalons enfin qu'une tendance actuelle des travaux de recherche sur les systèmes biométriques est d'associer plusieurs caractéristiques, par exemple la voix et le visage [BB et al.03], pour effectuer la reconnaissance. Cela permet en général d'améliorer la robustesse du système lorsque les caractéristiques biométriques choisies sont complémen-

1. On peut citer en contre exemple la voix, qui peut être assez facilement enregistrée et "rejouée" à l'aide d'un magnétophone. Il existe cependant des stratégies à base de texte prompté pour lutter contre ce type d'impostures.

taires (les faiblesses d'une caractéristique pouvant être compensées en partie par une autre).

Pour de plus amples informations sur la mise en oeuvre et les performances des techniques biométriques, on pourra consulter le site internet très complet du "Biometric Consortium" à l'adresse `http://www.biometrics.org`.

1.1.4 La VAL ou authentification biométrique vocale

La VAL (Vérification Automatique du Locuteur) est une tâche de reconnaissance qui consiste à vérifier l'identité proclamée d'une personne par l'analyse de sa voix. Elle est classiquement cataloguée comme une technique d'authentification biométrique comportementale. En effet, bien que les caractéristiques vocales d'un individu soient fortement contraintes par la morphologie et la mobilité de son appareil vocal, sa façon de parler et son attitude conditionnent fortement les signaux de parole émis. La parole est le résultat d'un "geste" intentionnel d'une personne. Celle-ci peut volontairement agir sur les commandes de son appareil vocal pour en modifier la configuration et donc changer les caractéristiques du signal émis. D'autre part, un état pathologique (rhume, stress, ...) ou physiologique (ivresse) peut modifier la voix d'une personne.

Pour toutes ces raisons, la voix est une caractéristique biométrique qui possède une grande variabilité. Elle est de plus non- permanente car elle évolue dans le temps, ponctuellement ou progressivement, de façon passagère ou durable.

Il est donc impropre de parler d'*empreinte vocale*, appellation trop souvent employée qui supposerait que la voix est aussi précise et constante que les empreintes digitales. Le terme de *signature vocale* est plus approprié, par analogie avec la signature manuscrite qui elle aussi varie et évolue dans le temps. Dans l'état actuel de la technique, la voix ne peut pas servir de preuve formelle pour identifier une personne au même titre que les empreintes digitales ou ADN, notamment pour les applications judiciaires. Mentionnons à ce propos la position des scientifiques français regroupés dans l'AFCP (Association Française de la Communication Parlée) qui préconisent l'arrêt des expertises vocales dans un contexte criminalistique tant que leur fiabilité n'aura pas été précisément évaluée[2]. Pour plus de précision sur le sujet nous invitons le lecteur à consulter [Boe et al.99] et [Bonastre et al.04a].

En dépit de sa forte variabilité, qui peut être sources de difficultés pour la reconnaissance d'une personne, la voix est un signal naturel à produire et partiellement contrôlable par l'individu qui l'émet. De ce fait, les utilisateurs sont en général peu réticents à fournir un échantillon de leur voix. Ce n'est pas le cas d'autres techniques biométriques comme la prise des empreintes digitales, notamment, qui est souvent associée au domaine criminalistique. La voix est aussi parfois le seul moyen disponible pour identifier un individu, en particulier pour des transactions à distance ou sur des

2. Voir le communiqué de presse consultable à l'adresse :
`http://www.afcp-parole.org/doc/LE-2002-5-communique2.pdf`.

enregistrements téléphoniques. Ces avantages font de la VAL une technique biométrique qui trouve pleinement sa place dans les applications où l'ergonomie est recherchée et dans certaines situations d'authentification à distance.

Cependant, la faible robustesse des systèmes de VAL dans des environnements perturbés a limité jusqu'ici son développement au niveau grand public. L'amélioration de cette robustesse est donc de première importance pour le déploiement d'applications de VAL en situation réelle et fait l'objet de nombreux travaux de recherche actuellement.

1.2 La voix, modalité d'authentification

Après avoir situé la place et le rôle de la VAL au sein des techniques biométriques, nous précisons dans cette section les caractéristiques vocales qui justifient la faisabilité même de ce mode d'authentification.

Nous pointons également les faiblesses intrinsèques de l'authentification des personnes à partir de leur voix, liées d'une part à la variabilité intra-locuteur du signal de parole et d'autre part aux diverses distorsions que subit ce signal lors de son acquisition et de sa transmission éventuelle. Ces facteurs de variabilité et de distortion sont une des causes principales des difficultés de reconnaissance rencontrées par les systèmes de VAL. Les performances obtenues sont donc intimement liées à la robustesse du système face à ce type de variations et de perturbations.

Nous finissons la section en présentant de façon générale les tâches de reconnaissance du locuteur en précisant les différentes typologies des systèmes et les applications correspondantes.

1.2.1 Caractérisation du locuteur

1.2.1.1 Motivations : ce que nous disent les voix

Le signal de parole ne véhicule pas seulement un message linguistique servant à la communication entre individus. Il transporte également des informations caractéristiques de la personne qui l'a émis comme le timbre de sa voix, sa façon de parler, son accent, son état émotionnel ou pathologique, etc ... De part les spécificités morphologiques et socio-culturelles de chaque individu, ces informations transportées par le signal de parole sont propres à la personne ayant prononcé le message. Les individus exploitent naturellement ces spécificités inter-locuteur pour identifier les personnes qu'ils connaissent, notamment au téléphone.

Différents niveaux d'information contenus dans le signal de parole sont utilisés de façon inconsciente par les humains pour caractériser un locuteur connu. Ces différents niveaux d'information sont également exploitables pour la reconnaissance automatique du locuteur. Nous les présentons brièvement à la section suivante en précisant les paramètres pouvant servir à caractériser un locuteur.

1.2.1.2 Paramètres caractéristiques

Le signal de parole émis par un individu est conditionné par l'ensemble des facteurs intervenant dans le processus de la communication parlée. On peut distinguer dans ce signal différents niveaux d'information. Les niveaux "bas" regroupent des informations liées principalement à des *traits physiques* de la personne (facteurs morphologiques et physiologiques). Les niveaux "hauts" sont relatifs à des informations qui représentent des *traits acquis* (facteurs socio-culturels).

Nous avons identifié ici six niveaux d'information différents que nous présentons du plus bas niveau au plus haut.

Niveau acoustique

Les paramètres acoustiques sont relatifs au contenu spectral du signal de parole et sont liés aux caractéristiques physiques de l'appareil vocal. L'enveloppe du spectre caractérise principalement la morphologie du conduit vocal du locuteur et les harmoniques reflètent la fréquence fondamentale et la forme de l'onde glottale.

Niveau prosodique

La prosodie désigne les caractéristiques d'un message parlé relatives à l'intonation, l'accentuation et les tons employés par le locuteur ainsi qu'au rythme d'élocution, aux pauses et à la durée des phonèmes. La distribution et les profils temporels de l'énergie et de la fréquence fondamentale sont par exemple caractéristiques de la prosodie employée par un locuteur.

Niveau phonétique

Les caractéristiques phonétiques du signal de parole se rattachent à la façon de prononcer les différents sons identifiables d'une langue, les phonèmes. Chaque individu a une façon de prononcer ces phonèmes qui lui est propre. Il est donc possible de caractériser certains phonèmes d'un locuteur (en général les plus fréquents) afin de les différencier des phonèmes des autres locuteurs.

Niveau idiolectal

Les caractéristiques idiolectales se rapportent aux particularités langagières propres à un individu. Il s'agit en particulier des habitudes liées à l'utilisation des mots. Certaines séquences de mots récurrentes, les "tics" de parole notamment, caractérisent la façon de parler d'un locuteur. Cependant, cette façon de parler peut être très variable en fonction du contexte dans lequel l'individu communique (interlocuteur connu ou pas, situation émotionnelle particulière,...).

Niveau dialogal

Les caractéristiques dialogales définissent la façon de communiquer d'un individu. Ainsi, des indices de fréquence et de durée des prises de parole d'un locuteur dans une conversation peuvent servir à le caractériser (personne bavarde ou non par exemple). Ce type d'informations est également très dépendant du contexte de la conversation.

Niveau sémantique

Les informations sémantiques contenues dans un message parlé ont trait au sens et à la signification de ce qui est prononcé. Les thèmes fréquemment abordés par un locuteur lors de conversations sont par exemple susceptibles de fournir une indication sur son identité. Ces informations de très haut niveau sont difficiles à caractériser et à extraire de façon automatique.

Historiquement, les paramètres de bas niveaux ont été les premiers utilisés pour caractériser les individus dans les systèmes de reconnaissance du locuteur. Cela est dû au fait qu'ils sont facilement extractibles, relativement robustes et qu'ils caractérisent bien la voix d'une personne.

Depuis quelques années cependant, les paramètres de plus haut niveau ont été explorés pour des tâches de reconnaissance du locuteur, notamment en VAL, en complémentarité avec les paramètres de bas niveau. Cette exploitation des informations haut niveau a été initiée par le travail effectué par un groupe de chercheurs internationaux lors du workshop "SuperSID" [Reynolds et al.03]. L'étude de ces paramètres et des améliorations qu'ils peuvent apporter a depuis été proposée lors des évaluations des systèmes de reconnaissance du locuteur organisées par NIST [Nist]. Notamment, l'apparition en 2002 de la tâche "Extended" a donné aux participants la possibilité d'exploiter de grandes quantités de données afin de permettre l'extraction d'informations de haut niveau.

Dans le cadre de cette thèse, nous nous limitons à l'extraction d'informations acoustiques caractérisant les locuteurs, qui est actuellement encore l'approche la plus répandue. L'extraction d'informations à des niveaux supérieurs demande en général une mise en oeuvre plus lourde, faisant appel souvent à un système de reconnaissance de la parole.

1.2.1.3 Références caractéristiques

Les paramètres extraits du signal de parole à différents niveaux d'information (cf section précédente) servent à caractériser la voix et la façon de parler d'une personne. Ils peuvent être utilisés d'une part pour constituer une référence, au moment d'une phase d'apprentissage de la voix de l'individu par le système. D'autre part, ils peuvent servir de matériau de test au moment d'une authentification.

Certains de ces paramètres, notamment les paramètres acoustiques, doivent être extraits du signal à un rythme rapide (typiquement 100 fois par seconde). Ce flot d'informations devient vite prohibitif si l'on veut par exemple stocker sur une mémoire l'ensemble des paramètres en guise de référence. Il est donc utile en pratique de construire un "résumé" de ce flot de paramètres afin de permettre un stockage moins encombrant. Ce "résumé", que l'on appelle généralement *référence caractéristique* peut par exemple décrire des caractéristiques statistiques des paramètres d'un locuteur, ou des propriétés

discriminantes par rapport aux autres locuteurs. On peut encore imaginer toutes sortes de résumés, plus ou moins condensés et prenant des formes diverses, le but commun étant de construire un modèle du locuteur.

Dans le cadre du travail présenté dans cette thèse, les références caractéristiques sont des modèles probabilistes des paramètres acoustiques émis par les locuteurs.

1.2.2 Facteurs limitants

La VAL utilise les spécificités inter-locuteurs du signal de parole pour caractériser et identifier les individus. Cette variabilité inter-locuteur permet de discriminer les personnes les unes des autres. Cependant le signal de parole qui arrive à l'entrée d'un système de VAL contient en général d'autres sources de variabilité qui sont, elles, perturbantes pour la reconnaissance. La première d'entre-elles, la variabilité intra-locuteur, est due à l'évolution temporelle des caractéristiques de la voix d'un individu. Une seconde source de variablitié vient des distorsions subies par le signal de parole lors de son acquisition et de sa transmission éventuelle. Des changements d'environnement sonore, de microphone ou de canal de transmission modifient les caractéristiques du signal de parole et peuvent fausser la reconnaissance du locuteur. Lors de la mise en oeuvre d'un système de VAL, il est crucial d'utiliser des techniques robustes à ce type de variabilités afin de minimiser la perte de performances qu'elles occasionnent.

1.2.2.1 Première limitation : la variabilité intra-locuteur

Dans une application de VAL, le système "apprend" à reconnaître la voix d'un locuteur à partir d'un ou plusieurs énoncés d'apprentissage, généralement en nombre restreint. La référence caractéristique construite à partir de cet ensemble d'apprentissages reflète les caractéristiques de la voix du locuteur aux instants précis où les énoncés ont été enregistrés.

La voix n'est cependant pas permanente. Elle évolue dans le temps, lentement et de façon progressive avec le vieillissement, ou encore brutalement et de façon passagère ou définitive en raison d'un état physiologique particulier, d'une maladie ou d'une intervention chirurgicale. D'autre part, une personne peut intentionnellement modifier sa voix pour ne pas être reconnu ou pour tenter d'imiter celle de quelqu'un d'autre. Enfin, notons qu'il est imposssible pour un individu de prononcer plusieurs fois une même phrase de façons tout à fait identiques. De légères variations aléatoires de prononciation sont toujours observées.

Pour ces raisons, une des difficultés majeures rencontrées en VAL vient du fait que les données d'apprentissage sont parfois peu représentatives des nombreuses variations possibles de la voix d'un locuteur. C'est particulièrement vrai lorsque cet apprentissage se fait à partir d'un énoncé unique, qui présente intrinsèquement peu de variabilité intra-locuteur. Pour faire face à ce type de difficultés, les sytèmes de VAL doivent intégrer des stratégies robustes de construction des références caractéristiques. Citons notamment les techniques d'enrichissement incrémental des références caractéristiques en phase opérationnelle, permettant de prendre en compte les dérives à moyen et long

termes de la voix. Un tel mode de fonctionnement, généralement appelé "adaptation incrémentale", a été étudié dans [Fredouille et al.00b] et [Barras et al.04], montrant une amélioration significative des performances lorsque le système est réglé correctement.

1.2.2.2 Seconde limitation : facteurs de distorsion

Les facteurs environnementaux et technologiques d'une application de VAL engendrent des perturbations du signal de parole sous forme notamment de bruits additifs et convolutifs. L'environnement sonore dans lequel s'effectue le captage du signal de parole détermine le type et le niveau du bruit additif. Les fonctions de transfert du microphone et du canal de transmission effectuent un filtrage convolutif du signal. Il existe également certains effets non-linéaires engendrés en particulier par les caractéristiques du capteur, par exemple des phénomènes de saturation. Le schéma de distorsion correspondant est représenté sur la figure 1.2. Des environnements très perturbés sont rencontrés notamment dans les applications téléphoniques.

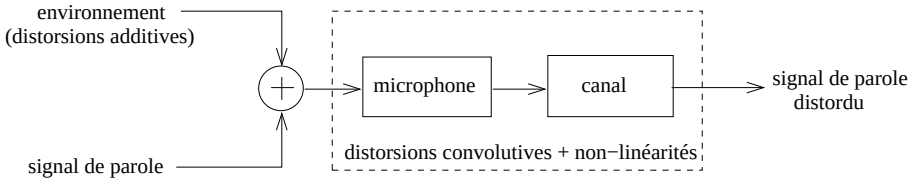


FIG. 1.2 – *Schéma de distorsion du signal de parole*

Ces perturbations subies par le signal de parole peuvent masquer certaines caractéristiques spécifiques au locuteur (niveau de bruit trop important, bande passante étroite, ...). Il est par conséquent souhaitable d'utiliser des descripteurs qui soient aussi peu représentatifs que possible des conditions environnementales et matérielles, tout en caractérisant correctement le signal de parole. D'autre part, certains traitements appliqués sur ces descripteurs peuvent aider à éliminer une partie des informations perturbantes.

Une autre source de difficulté apparaît lorsque les conditions d'aquisition et de transmission du signal de parole sont suceptibles de changer fortement entre l'apprentissage et le test. En effet, au moment de l'apprentissage le système n'apprend pas seulement les caractéristiques de la parole du locuteur mais également celles de l'environnement sonore, du microphone et du canal. Lors de la phase de reconnaisssance, la décision prise par le système peut être faussée si les conditions rencontrées lors du test sont différentes de celles de l'apprentisssage. Ce problème est lui aussi lié au manque de représentativité des données utilisées à l'apprentissage lorsque celui-ci se fait sur un nombre restreint de sessions (voir une seule). De nombreux travaux en VAL ont porté ces dernières années sur la mise au point de techniques visant à améliorer la robustesse des systèmes vis-à-vis des changements de conditions d'acquisition et de transmission du signal de parole.

Notamment, cette robustesse est primordiale pour le développement d'applications de VAL par téléphone, pour lesquelles de forts changements d'environnement peuvent être observés.

1.2.3 Tâches et typologies des systèmes de RAL

Après avoir identifié les paramètres permettant de caractériser un locuteur et les limites inhérentes à l'authentification des personnes à partir de leur voix, nous décrivons à présent les différentes tâches de la Reconnaissance Automatique du Locuteur (RAL).

La RAL est un domaine du traitement automatique de la parole qui regroupe toutes les tâches liées à l'extraction et à l'exploitation d'informations concernant l'identité des locuteurs dans un enregistrement audio. Historiquement, la vérification et l'identification du locuteur sont les premières tâches de RAL qui sont apparues, liées à des besoins de sécurisation. De nouvelles tâches ont plus récemment vu le jour en relation avec l'essor du multimédia dans notre société. Il s'agit de tâches d'extraction et d'exploitation d'informations relatives aux locuteurs dans des bases de données audio ou multimédias. Notons enfin l'utilisation grandissante des techniques de RAL dans des tâches de transcriptions orthographiques de documents audio. Les informations liées aux locuteurs peuvent en effet grandement aider à améliorer les performances des systèmes de reconnaissance de parole.

Dans la présentation qui suit, nous avons distingué trois tâches principales de la RAL en fonction du type d'information qu'elles permettent d'obtenir, c'est-à-dire en fonction de la nature de la sortie fournie par le système :

- la détection du locuteur qui fournit une sortie binaire du type détection/non-détection;
- l'identification du locuteur dont la sortie est un identifiant de locuteur;
- la description en locuteur de documents audio qui fournit une liste descriptive des interventions des locuteurs intervenant dans le document.

A l'intérieur de ces tâches de base, nous présentons des variantes liées aux conditions applicatives, ou des sous-tâches intermédiaires. Nous décrivons les différents types de systèmes d'un point de vue fonctionnel, en identifiant les entrées/sorties et les données disponibles pour effectuer la tâche considérée. Nous présentons également les applications principales de chacune des tâches.

1.2.3.1 Détection du locuteur

Le but de la **détection du locuteur** est de déterminer si un locuteur donné est présent ou non sur un enregistrement audio. Les entrées d'un système de détection du locuteur en mode de test sont : une entrée "signal" où arrivent les échantillons de l'enregistrement audio, et une entrée "sélection" permettant de sélectionner le locuteur à détecter parmi un ensemble de locuteurs référencés. La sortie du système est de type

binaire indiquant la détection (présence) ou la non-détection (absence) du locuteur cible sur l'enregistrement de test. Le schéma d'un tel système est donné sur la figure 1.3.

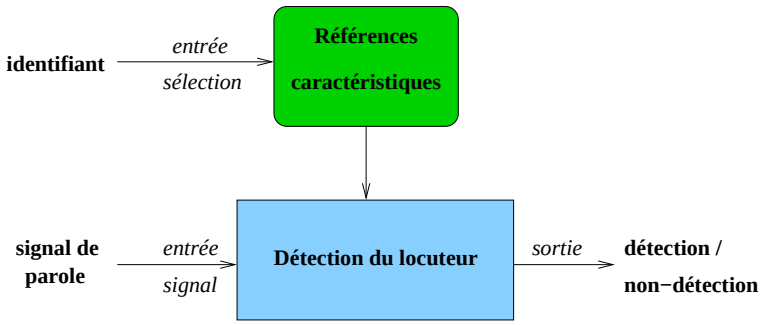


FIG. 1.3 – *Système de détection du locuteur*

La **vérification du locuteur** est un cas particulier de détection du locuteur où l'on fait l'hypothèse que la parole contenue dans l'enregistrement ne provient que d'un unique locuteur. Dans ce cas l'identité proclamée est présentée sur l'entrée "sélection" du système et la sortie est de type acceptation/rejet.

Actuellement, la détection de locuteur est principalement utilisée pour l'authentification d'enregistrements audio, soit dans un contexte de sécurisation de services, soit pour des applications criminalistiques. D'autres utilisations de la détection du locuteur sont liées aux besoins grandissants d'étiquetage de documents audio et multimédias, à des fins de navigation et de recherche rapide dans des bases de données importantes. Ainsi, la détection de locuteurs peut permettre de sélectionner tous les documents d'une base de données contenant la parole d'un locuteur particulier.

On pourra consulter les thèses françaises récentes de Corinne Fredouille [Fredouille00], Raphaël BLOUET [Blouet02] et Yassine Mami [Mami03] consacrées (en partie pour certaines) à la vérification du locuteur.

1.2.3.2 Identification du locuteur

A partir d'un ensemble de locuteurs référencés, l'**identification du locuteur** consiste à déterminer celui qui a prononcé un énoncé de test. Si l'on fait l'hypothèse que seuls les locuteurs connus peuvent accéder au système, on parle d'identification en ensemble fermé. Dans le cas contraire on parle d'identification en ensemble ouvert et le système doit pouvoir rejeter l'énoncé, indiquant ainsi qu'il n'a été prononcé par aucun des locuteurs référencés. L'unique entrée du système est l'entrée "signal", la sortie est l'identifiant du locuteur reconnu ou un indicateur de rejet si l'on travaille en ensemble ouvert (cf figure 1.4).

En pratique, l'énoncé de test doit être comparé à chacune des références caractéristiques des locuteurs connus pour déterminer laquelle en est la plus proche. Les temps de

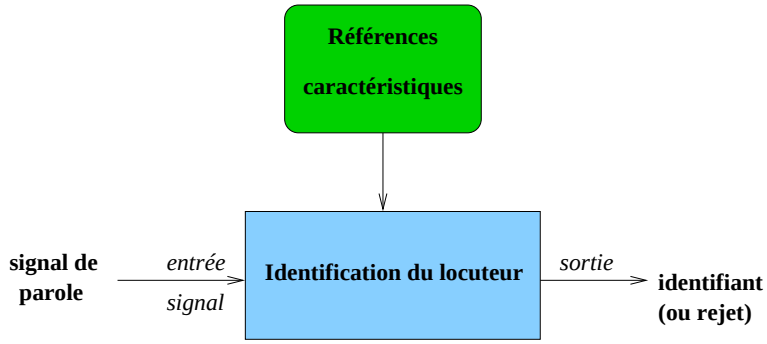


FIG. 1.4 – *Système d'identification du locuteur*

calculs pour effectuer la totalité de ces comparaisons deviennent prohibitifs lorsque le nombre de locuteurs référencés est trop important. De plus, les performances d'identification diminuent lorsque l'on augmente le nombre de locuteurs référencés. Ceci limite les applications de l'identification du locuteur à des ensembles relativement restreints d'individus. En ensemble fermé, elle peut être utilisée pour personnaliser des paramètres d'un groupe d'utilisateurs de services ou d'appareils. En ensemble ouvert, elle permet par exemple de limiter à une population déterminée l'accès à des lieux ou à des services. Elle trouve également des applications en adaptation au locuteur des systèmes de reconnaissance de la parole et peut aussi être utilisée comme sous-tâche d'une tâche de RAL plus complexe.

Des travaux de recherche récents en identification (et vérification) du locuteur peuvent être trouvés dans la thèse de Yassine Mami [Mami03]. On trouvera également des travaux consacrés à l'identification du locuteur dans la thèse de Corinne Fredouille [Fredouille00].

1.2.3.3 Description en locuteur de documents audio

La tâche de **description en locuteur** a pour objectif de structurer un enregistrement audio en fonction des locuteurs intervenants sur cet enregistrement. L'unique entrée du système est l'entrée "signal" où arrive le flux audio et la sortie est une description des interventions des locuteurs, c'est-à-dire une liste d'intervalles temporels étiquetés en fonction des différents intervenants (cf figure 1.5).

D'un point de vue général, la tâche de description en locuteur comprend deux sous-tâches : une tâche de **segmentation en locuteur**[3] du flux audio, c'est-à-dire une détermination des tours de parole, et une tâche de reconnaissance appliquée sur chaque segment. Suivant les algorithmes utilisés, ces deux sous-tâches peuvent être traitées conjointement ou séquentiellement, voire même itérativement.

3. la segmentation en locuteur peut être vue comme une tâche de description qui fournit des intervalles de temps non-étiquetés correspondant aux tours de parole

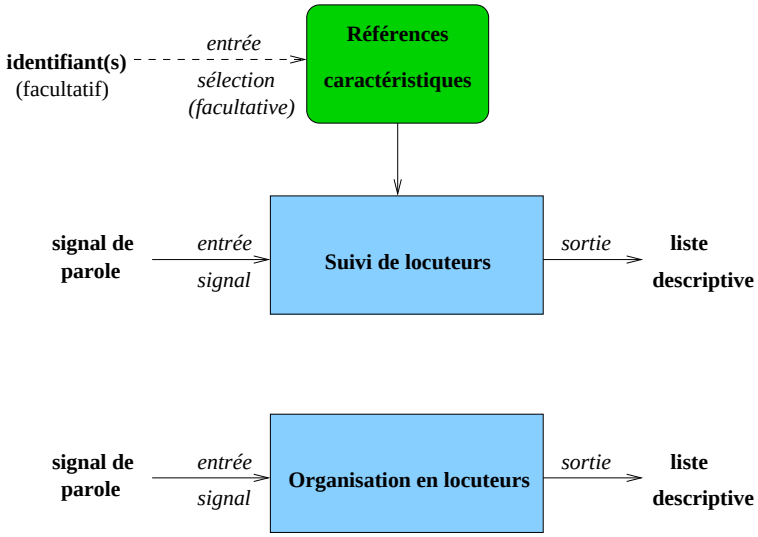


FIG. 1.5 – *Les deux variantes d'un système de description en locuteur : suivi de locuteurs (haut) et organisation en locuteurs (bas)*

Deux variantes de la tâche de description en locuteur existent en fonction des informations *a priori* dont on dispose pour structurer le document.

Dans la première configuration, la description est effectuée à partir d'un ensemble de références caractéristiques de locuteurs connus. On parle alors de **suivi de locuteurs** et le document est structuré en fonction des tours de parole des locuteurs référencés. Une entrée de sélection du ou des locuteurs à suivre peut éventuellement être utilisée.

Dans la deuxième configuration, la tâche de description en locuteur doit se faire sans aucune connaissance *a priori* : on ne possède pas de références caractéristiques des locuteurs et on ne connaît pas non plus le nombre de locuteurs intervenants dans l'enregistrement. Le but est alors de segmenter le document en tours de parole et d'étiqueter ces tours de parole en fonction des différents locuteurs intervenants. Nous identifierons cette tâche sous l'appellation d'**organisation en locuteurs** (en anglais "speaker diarization") d'un document audio. La sortie du système est une liste descriptive qui représente temporellement les interventions des différents locuteurs "supposés" qui sont alors identifiés de façon arbitraire et répertoriés dans l'ensemble du document.

Les applications de la description en locuteur sont principalement liées au domaine du multimédia, en plein essor actuellement. La description en locuteur peut servir à indexer un document sonore ou audiovisuel afin de permettre une navigation rapide à l'intérieur de celui-ci pour retrouver les interventions de tel ou tel locuteur. La description en locuteur peut également servir d'aide à la transcription manuelle d'enregistrement audio, par exemple des comptes-rendus de réunion. Dans ce cas, la liste descriptive fournie par le système permet au transcripteur d'identifier plus rapidement les différents intervenants au cours de la réunion. Enfin, les systèmes de transcription

orthographique automatique de documents audio peuvent utiliser la description en locuteur pour enrichir la transcription ou pour adapter le système de reconnaissance de parole au locuteur.

On pourra trouver des études détaillées de la tâche de description en locuteur et de ses sous-tâches dans les travaux de thèse de Sylvain Meignier [Meignier02] (organisation en locuteurs et segmentation), de Mouhamadou Seck [Seck01] (segmentation et suivi de classes de sons) et de Perrine Delacourt [Delacourt00] (segmentation et regroupement pour l'organisation en locuteur).

1.2.3.4 Interactions entre les tâches de RAL

La principale interaction entre les différentes tâches de RAL vient du fait qu'elles partagent pour l'essentiel les mêmes bases scientifiques. Par conséquent, des transferts de méthodes peuvent facilement être effectués entre ces tâches, et des avancées dans l'une d'entre elles profitent souvent rapidement aux autres.

D'autre part, les tâches basiques de RAL peuvent être utilisées comme sous-tâches de tâches plus complexes. Par exemple l'identification du locuteur en ensemble ouvert est couramment traitée comme une tâche d'identification en ensemble fermé suivie d'une tâche de vérification. La tâche de suivi de N locuteurs connus peut être traitée comme une tâche de segmentation suivie, sur chaque segment, d'une tâche de détection de chacun des N locuteurs connus. Si l'on fait l'hypothèse qu'il n'y a qu'un seul locuteur présent sur chaque segment, ces N tâches de détection peuvent être remplacées par une tâche d'identification en ensemble ouvert des N locuteurs connus [4]. Enfin, en dernier exemple, la tâche de segmentation en locuteur peut utiliser des informations fournies par les tâches de reconnaissance des locuteurs afin d'affiner la détection des frontières de segments. Dans ce cas, il est possible d'alterner segmentation et reconnaissance afin d'améliorer de façon itérative la description en locuteurs d'un document audio.

1.3 Déploiement applicatif des systèmes de VAL

Nous avons pour l'instant présenté les techniques biométriques, et la VAL en particulier, d'un point de vue général. Nous nous intéressons maintenant à la mise en oeuvre de tels systèmes. La façon de procéder à cette mise en oeuvre est globalement commune à toutes les techniques biométriques. Nous la décrivons dans le cadre de la VAL et présentons ensuite les contraintes particulières rencontrées lors du déploiement des systèmes de VAL en situation réelle.

4. ces manières de procéder ne sont pas les seules possibles, la segmentation et la reconnaissance du locuteur pouvant être menées conjointement

1.3.1 Phases d'utilisation et modes de fonctionnement

Du point de vue d'un utilisateur il existe deux phases distinctes d'utilisation d'un système de VAL :

1. la phase d'apprentissage durant laquelle le système doit *acquérir* un ou plusieurs exemples de la voix de l'utilisateur. Ces exemples servent à *construire la référence caractéristique* de l'individu qui sera utilisée comme point de comparaison pour l'authentification d'énoncés de test.
2. la phase de test où le système doit *authentifier* tout nouvel énoncé.

Les points mis en évidence ci-dessus font apparaître les modules nécessaires à la mise en oeuvre d'un système de VAL :

- un module d'*acquisition* chargé de mesurer le signal de parole à l'aide d'un microphone, et d'en extraire éventuellement des descripteurs.
- un module de *construction des références caractéristiques* de chaque individu, associé à une zone de stockage où elles sont mémorisées.
- un module d'*authentification* chargé de comparer un énoncé de test à une ou plusieurs références caractéristiques, et de fournir une décision.

Le système peut alors être configuré suivant les deux modes de fonctionnement "apprentissage" (cf figure 1.6) ou "test" (cf figure 1.7).

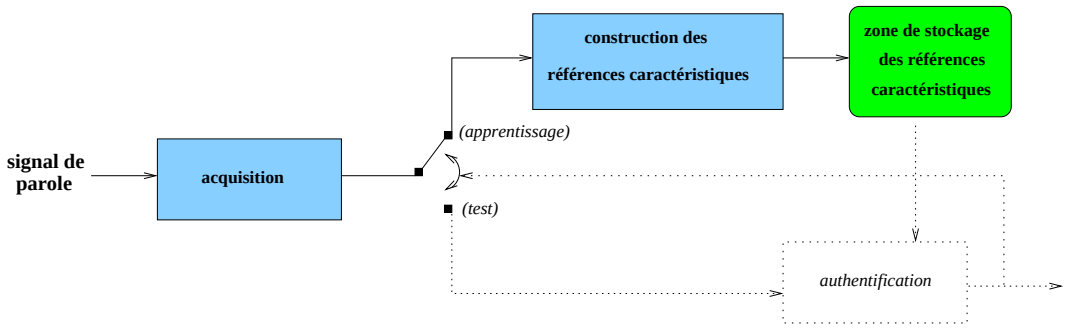


FIG. 1.6 – *Système de VAL en mode apprentissage*

Un troisième mode de fonctionnement est parfois utilisé : le mode *adaptation*. Il permet d'adapter de façon incrémentale les références caractéristiques à de nouvelles acquisitions de test. Si le système en mode test fournit une décision positive à une référence donnée, il bascule en mode apprentissage et utilise l'énoncé de test pour mettre à jour cette référence.

Du point de vue de la mise au point et du déploiement du système en situation réelle, on distingue également deux phases. Durant la première phase, la phase de développement, les paramètres du système sont réglés sur une base de données *a priori*, afin d'optimiser les performances sur cette base et pour l'application visée. Ensuite, le système est placé en fonctionnement réel avec les réglages déterminés lors de la première phase ; c'est la phase opérationnelle. Pour que le passage de la phase

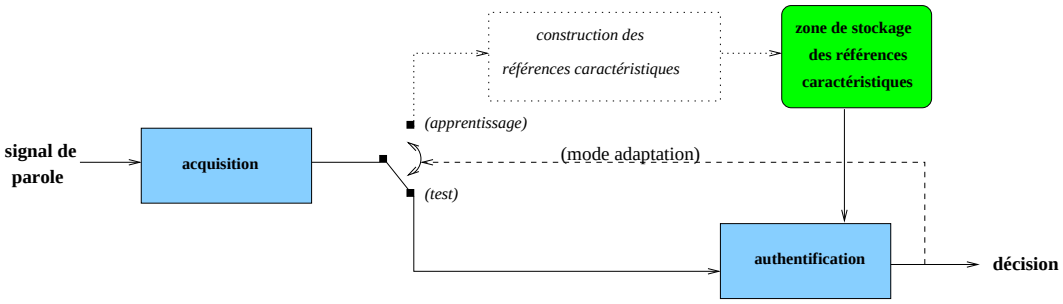


FIG. 1.7 – *Système de VAL en mode test (et adaptation)*

de développement à la phase opérationnelle se fasse correctement, il faut adopter une méthodologie d'évaluation permettant de tirer des conclusions significatives lors de la phase de développement. En premier lieu, les réglages trouvés lors de cette phase ne seront utilisables que si la base de données *a priori* reflète au mieux les conditions rencontrées en phase opérationnelle. D'autre part, ils ne seront pertinents que si l'évaluation des performances du système sur cette base a une précision suffisante. Il faut pour cela que l'évaluation soit faite sur un nombre suffisamment important de tests. Cette analyse qualitative sera précisée de façon quantitative dans la section 1.4 qui traite de la méthodologie d'évaluation des systèmes de VAL en détail.

1.3.2 Profils applicatifs et difficultés sous-jacentes

1.3.2.1 Problématique générale

Lors du déploiement d'un système de VAL pour une application donnée, de nombreux facteurs liés aux conditions d'utilisation du système vont affecter ou conditionner les énoncés recueillis lors de l'apprentissage et du test. Ces facteurs peuvent provenir de considérations ergonomiques ou de contraintes géographiques et technologiques inhérentes à l'application considérée.

D'un point de vue général, ces contraintes vont rendre chaque énoncé spécifique des conditions ponctuelles de l'environnement et de l'instant d'acquisition. Cela pose encore le problème de la représentativité des données utilisées pour construire les références caractéristiques des locuteurs. Dans l'idéal, ces données devraient représenter toutes les configurations susceptibles d'être rencontrées en phase opérationnelle. Cela englobe toutes les variabilités possibles dans les conditions de prise de son et de transmission du signal, d'environnement sonore, de contenu linguistique des énoncés ainsi que la prise en compte des dérives intrinsèques de la voix des locuteurs. En pratique, des considérations ergonomiques ou des contraintes inhérentes à la tâche limitent la durée et le nombre des sessions d'apprentissage à quelques sessions - voire une seule - de quelques minutes tout au plus. Des disparités au niveau des conditions d'acquistion et du contenu linguistique des énoncés risquent donc d'être rencontrées entre l'apprentissage et le test, entraînant des difficultés de reconnaissance.

Dans la suite de cette section, nous présentons les principaux facteurs applicatifs influant sur les performances d'un système de VAL. Nous distinguons les facteurs liés au contenu linguistique des énoncés de ceux liés à l'environnement physique et technologique du système.

1.3.2.2 Dépendance au texte et contenu linguistique

Les contraintes sur le contenu linguistique des énoncés concernent en premier lieu le degré de dépendance au texte du système. On distingue ainsi les systèmes dits dépendants du texte des systèmes indépendants du texte. Pour les premiers, le contenu linguistique des énoncés d'apprentissage et de test est fixé et connu à l'avance, alors que pour les seconds, aucune contrainte linguistique particulière n'est imposée. Cette dépendance au texte est dans la plupart des cas intentionnelle car elle permet de supprimer les disparités de contenu linguistique entre les énoncés d'apprentissage et de test. De ce fait, les erreurs de reconnaissance dues à ce type de variabilité sont fortement diminuées. En pratique, cette contrainte sur le texte peut avoir différents degrés de dépendance. On peut citer notamment, par ordre décroissant du degré de dépendance au texte, les systèmes à mot de passe individuel ou commun à un groupe d'utilisateurs, les systèmes à texte prompté, les systèmes dépendants d'un événement phonétique particulier, et les systèmes à texte libre.

Théoriquement, plus la contrainte sur le texte est forte, plus les performances obtenues sont bonnes. Les systèmes fortement dépendants du texte sont par conséquent privilégiés pour des applications à haut niveau de sécurisation. Cependant les systèmes à mot de passe par exemple peuvent être sujets à des tentatives d'imposture à base de parole enregistrée, ce qui n'est pas le cas des systèmes à texte prompté. Ceux-ci sont en effet associés à un système de reconnaissance de parole afin de vérifier que ce qui a été prononcé par l'utilisateur correspond bien à la phrase prescrite.

Quoiqu'il en soit, le choix de la dépendance au texte est très souvent dicté par des considérations ergonomiques et par le contexte d'utilisation du système. En général, un système à texte libre est plus souple à utiliser qu'un système à texte contraint. D'autre part, pour certaines applications la reconnaissance du locuteur doit se faire sur de la parole spontanée, par exemple des conversations téléphoniques. Le système doit alors impérativement fonctionner en mode indépendant du texte. D'une manière générale, pour les applications de VAL indépendante du texte, c'est essentiellement la quantité de données disponibles qui détermine la richesse linguistique des ensembles d'apprentissage et de test. Plus cette quantité est réduite, plus les disparités de contenu linguistique entre l'apprentissage et le test peuvent être grandes. Les performances du système sont alors en général directement liées à cette quantité de données.

1.3.2.3 Contraintes environnementales

Les facteurs environnementaux influant sur les performances d'un système de VAL proviennent de l'implantation géographique de l'interface homme/machine et des li-

mites technologiques des éléments intervenants dans le captage et la transmission du signal de parole. Ces contraintes d'environnement sont imposées par le type d'application visé et par des considérations ergonomiques et économiques. Il existe par exemple de grandes différences d'environnement sonore entre des applications de VAL implantées dans un lieu public et dans un bureau. D'autre part, la qualité du capteur sera généralement moins bonne pour des applications grand public que pour celles exigeant un haut niveau de sécurisation. Enfin, le signal de parole doit parfois être transmis, par le canal téléphonique par exemple, avant d'effectuer la reconnaissance du locuteur, d'où une qualité moindre à la réception. Bien entendu, plus la qualité du signal se dégrade, plus les performances de reconnaissance diminuent.

Parmi les nombreuses configurations applicatives possibles, les applications de VAL par téléphone sont de celles qui subissent le plus de contraintes environnementales. Dans ces applications, le signal de parole est fortement conditionné par le type de capteur utilisé dans les combinés et par les caractéristiques de la transmission. De plus, elles englobent souvent de nombreuses conditions différentes d'utilisation, notamment en ce qui concerne la téléphonie mobile où l'environnement sonore, le codage du signal et la configuration du canal de transmission sont intrinsèquement très variables. C'est dans ces environnements particulièrement perturbés que NIST propose d'évaluer les systèmes de VAL lors de ses campagnes annuelles, incitant ainsi les participants à développer des systèmes robustes aux variabilités rencontrées. C'est également le cadre que nous avons choisi pour valider et évaluer le travail présenté dans ce document, lui aussi consacré à l'étude d'approches robustes pour la VAL.

1.4 Méthodologie d'évaluation

1.4.1 Facteurs de qualité d'un système de VAL

La qualité d'un système de VAL s'évalue en fonction de différents facteurs. En premier lieu bien sûr, ce sont les performances de vérification du système, en termes de taux d'erreur, qui vont déterminer sa qualité. Mais dans les applications réelles, d'autres facteurs liés à des problèmes d'utilisation du système doivent être pris en compte. En effet, pour être en mesure d'effectuer une vérification il faut au préalable avoir réussi à construire la référence caractéristique du locuteur d'une part, et à acquérir l'énoncé de test d'autre part. Il convient donc de tenir compte des échecs à l'apprentissage et des échecs à l'acquisition des tests pour évaluer la qualité globale d'un système. Ce type d'échecs détermine notamment la perception qu'ont les utilisateurs de l'ergonomie de l'application. Un système qui demande aux utilisateurs de répéter plusieurs fois les sessions d'apprentissage ou de test sera en général mal perçu, même si ses performances de vérification sont correctes.

La proportion d'échecs à l'apprentissage et au test ne peut être évaluée qu'en situation réelle d'exploitation d'un système de VAL. Notre travail de recherche s'effectuant en laboratoire à partir de bases de données pré-enregistrées, nous nous limitons par

conséquent à l'évaluation de la qualité des systèmes qu'au travers de l'étude des performances de vérification.

1.4.2 Mesure des performances

1.4.2.1 Critère d'évaluation en phase opérationnelle

Les performances d'un système de VAL en situation réelle s'évaluent en fonction des taux d'erreur qu'il engendre. Les deux types d'erreurs que l'on rencontre sont les *fausses acceptations* d'une part, et les *faux rejets* d'autre part. Les premières correspondent à l'acceptation à tort d'un imposteur, les secondes ont lieu lorsque le locuteur de test est rejeté alors qu'il s'agit bien du locuteur légitime (locuteur "client"). Les taux de fausses acceptations T_{FA} et de faux rejets T_{FR} déterminent le point de fonctionnement du système. Ils sont calculés de la manière suivante :

$$T_{FA} = \frac{Nombre\ de\ fausses\ acceptations}{Nombre\ d'accès\ imposteurs} \ , \tag{1.1}$$

$$T_{FR} = \frac{Nombre\ de\ faux\ rejets}{Nombre\ d'accès\ clients} \ . \tag{1.2}$$

Ces deux taux approximent, sur l'ensemble des tests d'évaluation, les probabilités de fausses acceptations P_{FA} et de faux rejets P_{FR} intrinsèques au système de VAL. Ils varient toujours à l'inverse l'un de l'autre en fonction des réglages effectués au niveau du module chargé de prendre la décision. L'évaluation quantitative des performances s'effectue alors par une mesure pondérée de ces deux taux d'erreur, appelée *fonction de coût de détection* (DCF : Detection Cost Function). Les coefficients de pondérations des deux taux d'erreur sont déterminés en fonction de facteurs relatifs au contexte applicatif, à savoir les coûts attribués à chacun des deux types d'erreurs possibles et les probabilités *a priori* d'accès au système de locuteurs imposteurs et de locuteurs clients. Cette DCF s'exprime sous la forme :

$$DCF = C_{FR}.P_{cl}.T_{FR} + C_{FA}.P_{imp}.T_{FA} \ , \tag{1.3}$$

où :
C_{FR} est le coût associé à un faux rejet;
C_{FA} est le coût associé à une fausse acceptation;
P_{cl} est la probabilité *a priori* d'un accès client;
P_{imp} est la probabilité *a priori* d'un accès imposteur.

La DCF est parfois normalisée par le terme $(C_{FR}.P_{cl} + C_{FA}.P_{imp})$ ce qui permet d'obtenir des valeurs comprises entre 0 et 1.

Plus la valeur de DCF est petite, meilleur est le système pour les conditions applicatives spécifiées. C'est cette mesure qui est préconisée pour évaluer les performances d'un système en situation réelle. Elle rend compte des capacités de reconnaissance du

système mais aussi de la qualité des réglages effectués au niveau du module de décision. En effet, si ces réglages sont de mauvaise qualité, la valeur de la DCF s'écarte de sa valeur optimale (valeur minimale de la DCF que l'on peut atteindre pour le système considéré).

En règle générale, la décision lors d'un accès de test s'effectue en comparant un score de vérification à un seuil θ. Le réglage de ce seuil détermine les taux de fausses acceptations $T_{FA}(\theta)$ et de faux rejets $T_{FR}(\theta)$ du système. Il doit donc être optimisé de façon à minimiser la DCF définie à l'équation 1.3. Cette optimisation ne peut être effectuée que sur une population *a priori* pendant la phase de développement du système. La valeur obtenue est ensuite utilisée sur le système en phase opérationnelle, c'est-à-dire en exploitation réelle. Il est donc très important que la valeur optimale du seuil soit peu variable entre la phase de développement et la phase opérationnelle. Les systèmes de VAL actuels utilisent en général des techniques permettant d'améliorer la robustesse de ce seuil.

Signalons enfin qu'il existe des tests statistiques permettant de définir un intervalle de confiance autour d'une valeur mesurée de DCF, pour un corpus d'évaluation donné (voir par exemple [Bengio et al.04]). Ces tests servent également à déterminer si les performances obtenues par deux systèmes sont statistiquement différentes.

1.4.2.2 Critères d'analyse en phase de développement

Pendant la phase de développement, il est souvent pratique d'utiliser des critères permettant de comparer différents systèmes de façon directe (c'est-à-dire sur la population de développement elle-même), à des fins d'analyse. Les critères utilisés sont des grandeurs qui rendent compte des performances du système après optimisation du seuil de décision *a posteriori*. Ils ne tiennent donc pas compte de l'étape de portage du seuil de la phase de développement vers la phase opérationnelle. Ils permettent cependant d'alléger grandement l'évaluation de nouvelles techniques. En outre, si la base de données utilisée lors du développement est suffisamment conséquente et si elle reflète bien les conditions rencontrées en phase opérationnelle, le portage du seuil se fait en général sans perte importante de performance.

Dans ce cadre d'analyse des systèmes en phase de développement, trois critères sont principalement utilisés :

- le point DCF_{min}. C'est le point de fonctionnement pour lequel la DCF est minimale :

$$DCF_{min} = \min_{\theta} C_{FR}.P_{cl}.T_{FR}(\theta) + C_{FA}.P_{imp}.T_{FA}(\theta) \ .$$

- l'$HTER_{min}$. C'est le point de fonctionnement obtenu en minimisant l'HTER (Half Total Error Rate), moyenne des taux d'erreur :

$$HTER_{min} = \min_{\theta} \frac{T_{FR}(\theta) + T_{FA}(\theta)}{2} \ .$$

Il correspond à une configuration particulière du point DCF_{min} pour laquelle on a $C_{FR}.P_{cl} = C_{FA}.P_{imp} = \frac{1}{2}$.

- l'EER (Equal Error Rate). C'est le point de fonctionnement à taux d'erreur égaux, c'est-à-dire pour lequel on a $T_{FR} = T_{FA}$. Ce point de fonctionnement n'a pas d'interprétation directe en terme de coût mais est très souvent utilisé comme une première indication des performances du système.

Dans le cadre de notre travail, les critères d'analyse que nous utilisons sont l'EER et le point DCF_{min} avec les configurations de coûts et de probabilités *a priori* définies par NIST dans le cadre de leurs évaluations [Nist]. La fonction de coût correspondante est traditionnellement noté C_{det} par NIST. Sa définition est la suivante :

$$C_{det} = 0,99 \times T_{FA} + 0,1 \times T_{FR} \ . \tag{1.4}$$

Cette fonction attribue donc presque 10 fois plus de coût aux fausses acceptations qu'aux faux rejets, repoussant ainsi les points de fonctionnement à minimum de coût vers les régions à faibles taux de fausses acceptations. Nous insistons sur le fait que les critères EER et DCF_{min} servent à l'analyse des systèmes en phase de développement et ne prennent pas en compte tous les facteurs influant sur les performances du système en phase opérationnelle, notamment l'utilisation d'un seuil réglé *a priori*.

1.4.2.3 Courbes DET

Outre les critères d'analyse présentés ci-dessus qui évaluent les performances en un point particulier de fonctionnement, il est également intéressant de pouvoir visualiser en même temps les performances du système pour différents points de fonctionnement, en particulier lorsque l'on étudie un système sans qu'il soit destiné à une application précise. Pour cela, il est possible de tracer des courbes représentant l'évolution des couples (T_{FA}, T_{FR}) lorsque le seuil de décision varie. Les plus utilisées en VAL sont les courbes DET (Detection Error Tradeoff) [Martin et al.97], variantes des courbes ROC (Receiver Operating Charasteristic) mais dont les échelles d'axes suivent l'évolution de la fonction de répartition d'une loi normale. Sous l'hypothèse de gaussianité des scores obtenus pour des accès imposteurs d'une part et pour des accès clients d'autre part, la courbe DET représentant l'évolution des points de fonctionnement d'un système est une droite. Elles sont très utilisées pour comparer différents systèmes car elles permettent de visualiser leurs performances en diverses zones de points de fonctionnement. L'EER apparaît également directement sur ces courbes comme le point d'intersection de la courbe DET avec la première bissectrice.

Les courbes DET seront très souvent utilisées dans ce document pour illustrer les apports des méthodes développées.

1.4.3 Corpus et campagnes d'évaluation

1.4.3.1 Corpus de données pour l'évaluation des systèmes

Comme nous l'avons déjà évoqué précédemment, les corpus de données utilisés pour le développement et le réglage d'un système de VAL jouent un rôle essentiel quant à la portabilité des réglages en phase opérationnelle. En effet, les conclusions tirées d'une évaluation et notamment la mesure des performances n'ont de sens que dans un contexte applicatif similaire. Les données de développement doivent donc ressembler au mieux aux données rencontrées en phases opérationnelles.

D'autre part, l'évaluation des performances d'un système sur un corpus donné se fait par l'intermédiaire des taux d'erreur T_{FA} et T_{FR} qui sont des approximations des probabilités réelles de fausses acceptations P_{FA} et de faux rejets P_{FR} du système. L'inexactitude de ces approximations fait que la valeur de DCF trouvée est elle-même entachée d'erreur. Cette erreur sera d'autant plus faible que les taux d'erreur sont correctement estimés, ce qui implique que le nombre d'accès considérés pour chacun d'entre eux soit suffisamment important. En résumé, plus le corpus d'évaluation est important, plus grande est la précision sur les taux d'erreur et sur la DCF.

Il n'existe pas de règle formelle pour fixer la taille minimale d'un corpus en fonction de la précision désirée sur la mesure des taux d'erreur. Une règle communément utilisée, connue sous le nom de "règle de Doddington" et justifiée dans [Porter97], stipule que le système doit se tromper au moins 30 fois pour être sûr à 90% que le taux d'erreur mesuré est dans un intervalle de 30% autour de la probabilité d'erreur réelle. Cette règle, qui a au moins l'avantage d'exister, ne permet cependant que de vérifier *a posteriori* si le corpus utilisé répond bien au critère, après lecture des taux d'erreur obtenus. Par exemple, l'évaluation d'un système ayant des taux d'erreur T_{FA} et T_{FR} de 1% n'est fiable que si le corpus d'évaluation comporte au moins 3000 accès clients et 3000 accès imposteurs. En général, c'est le nombre d'accès clients qui fait défaut dans les corpus de données d'évaluation.

En résumé, plus le système est performant et plus le corpus d'évaluation doit être important. Concrètement, les développeurs d'un système de VAL doivent déterminer la quantité de données nécessaires à l'évaluation fiable de leur système en estimant les performances maximales qu'ils peuvent espérer atteindre dans les conditions applicatives qui les intéressent. Ils peuvent par exemple se baser sur les performances atteintes par les systèmes état de l'art dans les mêmes conditions. Ils doivent ensuite essayer de réunir suffisamment de données représentatives des conditions opérationnelles, soit lors d'une phase de récolte de données soit en utilisant des corpus de données enregistrées dans des conditions similaires.

Ces contraintes fortes sur la quantité et la qualité des données nécessaires à la calibration fiable d'une application de VAL fixent les limites de ce mode d'authentification pour les contextes où les conditions applicatives ne sont pas connues *a priori*, comme c'est le cas très souvent dans les expertises criminalistiques. Dans un tel contexte, le degré de fiabilité du résultat fourni par un système de VAL ne peut être évalué de

façon précise. Notons cependant qu'une campagne d'évaluation des systèmes de VAL a récemment été organisée dans un contexte de données récoltées lors d'investigations criminalistiques [vL et al.04].

1.4.3.2 Campagnes d'évaluation NIST

Chaque année, des campagnes d'évaluation des systèmes de reconnaissance du locuteur sont organisées par l'organisme américain NIST (National Institue of Standard and Technology) auxquelles peuvent participer les laboratoires scientifiques du monde entier. Ces campagnes jouent un rôle moteur dans le développement de nouvelles technologies en proposant un cadre d'évaluation bien calibré et l'accès à des corpus de données conséquents.

Le plan d'évaluation publié chaque année avant le début de la campagne (voir par exemple [Przybocki et al.04]) définit les règles de l'évaluation et les différentes configurations des tâches proposées. Actuellement, ces tâches sont toutes des tâches de détection du locuteur, sous différents conditionnement des données (quantité de données variable à l'apprentissage et/ou au test, énoncés mono-locuteurs ou multi-locuteurs). Les bases de données utilisées sont issues des corpus collectés par le LDC (Linguistic Data Consortium [Ldc]) et contiennent essentiellement des conversations téléphoniques en anglais américain (téléphonie filaire et mobile).

La première campagne a eu lieu en 1996 et l'IRISA y a participé tous les ans depuis 1997, permettant ainsi au laboratoire de valider de nouveaux développements, de suivre l'évolution de l'état de l'art et de confronter nos systèmes à ceux des autres laboratoires participants.

En outre, ces évaluations ont permis d'identifier de nouvelles problématiques dans le domaine de la reconnaissance du locuteur en pointant notamment les principales difficultés rencontrées par les systèmes de VAL. Ces difficultés sont principalement liées à un manque de robustesse des systèmes vis-à-vis de changements des conditions d'enregistrement, de niveau de bruit et des quantités de données disponibles. Les corpus de données proposés par NIST permettent de mettre à rude épreuve les systèmes de VAL, en particulier ceux contenant des données de téléphonie mobile pour lesquelles le niveau de bruit ambiant, les caractéristiques du microphone et du canal de transmission ainsi que le processus de codage du signal de parole, sont très variables. La confrontation des systèmes de VAL à ce type de données a incité les laboratoires participant aux évaluations NIST à développer des techniques visant à améliorer leur robustesse.

Ces évaluations sont également un cadre applicatif fort de nos travaux. Ceux-ci s'orientent également autour de méthodes robustes pour la VAL et nous avons utilisé les corpus de données NIST pour développer et évaluer les techniques que nous proposons.

Pour plus de détails sur les évaluations NIST, on pourra consulter le site internet `http://www.nist.gov/speech/tests/spk/`

1.5 Approches état de l'art *vs* méthodes émergentes

Depuis maintenant plus de 10 ans, les approches probabilistes ont dominé l'état de l'art des systèmes de reconnaissance du locuteur. Elles proposent un cadre puissant pour prendre en compte un certain nombre des variabilités contenues dans le signal de parole. En outre, elles permettent de définir de façon précise une mesure de similarité entre un ensemble de données de test et une référence caractéristique, représentée dans ce cas par un modèle probabiliste du locuteur. Le succès des méthodes probabilistes a engendré une multitude de travaux. En mode indépendant du texte, la technique de modélisation qui prédomine est basée sur des modèles de mélanges de Gaussiennes. De nombreuses techniques destinées à renforcer la robustesse des systèmes basés sur cette approche ont été développées, laissant parfois peu de place au développement d'autres méthodes.

Citons cependant certaines méthodes alternatives, comme par exemple les réseaux de neurones probabilistes [Ganchev et al.03] ou les classifieurs linéaires polynomiaux [Campbell et al.02], qui ont obtenu de bonnes performances en VAL.

Plus récemment, les approches par SVM (Support Vector Machine) ont fait une percée parmi les méthodes les plus performantes en VAL.

Les travaux de thèse de Vincent WAN [Wan03] notamment ont marqué un pas dans l'avancement de ces techniques en développant des systèmes SVM rivalisant avec les systèmes basés sur l'approche probabiliste. Depuis, les SVM ont été utilisés dans le cadre des évaluations NIST [Campbell03]. De nouvelles méthodes visant à renforcer la robustesse de ce type de systèmes ont été développées [Solomonoff et al.04], amenant les performances des systèmes SVM à un niveau équivalent, voire parfois supérieur, à celui des systèmes état de l'art utilisant l'approche probabiliste. En outre, le caractère fondamentalement différent de la méthode de classification par SVM (modèles discriminants) par rapport aux approches probabilistes (modèles génératifs) a permis d'obtenir des gains significatifs des performances lorsque ces deux types d'approches sont utilisés conjointement, dans le cadre d'une fusion de plusieurs systèmes [Campbell et al.04].

Signalons enfin une tendance générale des systèmes de VAL à utiliser de plus en plus d'informations contenues dans le signal de parole, en particulier lorsqu'on dispose d'une grande quantité de données d'apprentissage. Aux paramètres acoustiques classiquement utilisés viennent maintenant s'ajouter des paramètres de plus haut niveau :

- des paramètres *prosodiques* ont été utilisés pour la VAL dans [Adami et al.03]);
- un exemple d'utilisation de paramètres *phonétiques* pour la VAL peut être trouvé dans [Andrews et al.02, Klusacek et al.03];
- une étude des possibilités d'exploitation d'informations idiolectales pour des tâches de reconnaissance du locuteur a été menée par George Doddington dans [Doddington01];
- enfin, des indices de fréquence et de durée des prises de parole dans une conversation (niveau dialogal) ont été utilisés par le système du laboratoire MIT-LL lors des évaluations NIST 2002 dans la tâche "Extended" [Nist].

A notre connaissance, les informations sémantiques contenues dans le signal de parole n'ont pas encore été exploitées à ce jour pour des tâches de reconnaissance du locuteur.

L'ensemble de ces paramètres est utilisé de façon conjointe par les sytèmes de VAL, par l'intermédiaire de techniques de fusion d'informations [Campbell et al.03]. Cela permet d'améliorer significativement les performances lorsque la quantité de données est suffisante pour extraire de façon fiable les paramètres de hauts niveaux.

1.6 Problématiques soulevées et orientation du travail

Dans ce chapitre nous avons ciblé les principales difficultés auxquelles sont confrontés les systèmes de VAL. Ces difficultés viennent d'une part de la variabilité intrinsèque de la voix d'un locuteur et d'autre part des conditions d'acquisition et de transmission du signal de parole et du degré de variation de ces conditions. Bien que les sources de variabilité soient différentes (variabilité intra-locuteur d'un côté et variabilité d'environnement sonore et de matériel de l'autre), les problèmes engendrés viennent d'une même cause : la quantité de données d'apprentissage est toujours limitée. Du fait de cette limitation, les références caractéristiques construites au moment de l'apprentissage ne sont représentatives que d'une partie restreinte des conditions d'acquisition et de transmission du signal de parole. De même elles ne prennent pas en compte toutes les variations et dérives possibles de la voix du locuteur. Si les conditions rencontrées lors du test sont différentes de celle de l'apprentissage, la reconnaisssance du locuteur risque d'être faussée par cette disparité.

Il est donc primordial de prendre en compte les effets de ces disparités dans les applications de VAL, en particulier celles opérant dans des environements perturbés. Trois types de stratégies peuvent être envisagés à cet effet :

1. On peut d'une part chercher à atténuer autant que possible les effets des sources de variabilité en divers points du système de VAL, en utilisant des techniques de *normalisation*.
2. D'autre part, afin de minimiser les disparités entre conditions d'apprentissage et de test, des techniques d'*adaptation* peuvent être utilisées. Il s'agit alors soit d'adapter la référence caractéristique aux conditions des données de test, soit inversement de transformer les données de test pour qu'elles soient mieux adaptées aux conditions représentées par la référence caractéristique.
3. Enfin, une dernière approche consiste à *utiliser des informations a priori* extérieures aux données d'apprentissage pour tenter de combler certains manques dans ces données.

L'ensemble de ces approches a pour but d'améliorer la robustesse des systèmes de VAL face aux changements d'environnements et de matériel ou face à une quantité de parole insuffisante pour caractériser correctement la voix d'un locuteur.

Dans cette thèse nous attaquons les problèmes de robustesse par les stratégies 1 et 3 mentionnées ci-desssus. La mise en oeuvre des techniques correspondantes s'intègre dans le cadre théorique de l'approche probabiliste pour la VAL, qui est décrite au chapitre suivant.

Chapitre 2

Approche probabiliste pour la vérification du locuteur

L'aspect hautement aléatoire du signal de parole impose que les techniques dédiées à son traitement puissent prendre en compte de façon efficace de nombreuses variabilités. L'approche probabiliste est bien appropriée pour gérer, avec un minimum de robustesse, ces configurations variées du signal. En outre, elle permet de définir une mesure de similarité entre un ensemble de données de test et un modèle de référence correspondant à une classe donnée. Cette mesure est utile dans tous les problèmes de classification où l'on doit prendre une décision à partir d'un nombre variable d'échantillons de test.

En VAL, l'approche probabiliste a constitué l'état de l'art depuis plus de 10 ans maintenant. On la trouve dans la quasi-totalité des systèmes actuelles, parfois associée à d'autres méthodes de classification. La modélisation des locuteurs en mode indépendant du texte repose actuellement sur les modèles de mélanges de Gaussiennes (GMM) et la technique d'estimation qui prédomine est l'adaptation Bayésienne. Autour de cette approche de modélisation se sont greffées de multiples techniques visant à améliorer la robustesse des systèmes, en particulier dans des environnements perturbés.

Ce chapitre présente tout d'abord les fondements théoriques de l'approche probabiliste pour la VAL. Ensuite, les principales techniques utilisées au sein des systèmes actuels sont décrites. En particulier, l'adaptation Bayésienne des modèles GMM de locuteurs est présentée plus en détail. Enfin, nous identifions certaines faiblesses des techniques actuelles, permettant ainsi de définir les objectifs principaux de ce travail de thèse.

2.1 Fondements théoriques et formalisme associé

2.1.1 Modélisation probabiliste des locuteurs

L'approche probabiliste repose sur l'hypothèse qu'une classe de son X, dans notre cas un locuteur, peut être représentée par un modèle stochastique, que l'on notera également X. Ce modèle probabiliste décrit la distribution statistique des observations acoustiques issues de la classe X. En pratique ce modèle "vrai" est inconnu et inaccessible à la mesure dans sa globalité. Des choix de structure et de représentation de la classe X doivent donc être effectués et on associe à cette classe une fonction de vraisemblance $p(y|X)$ qui approxime la densité de probabilité réelle des observations y de la classe X. Si la fonction de vraisemblance a une forme paramétrique, la classe X est décrite avec un ensemble fini de paramètres que l'on notera [1] Λ_X. Dans la suite du document, nous écrirons de façon équivalente $p(y|X)$ ou $p(y|\Lambda_X)$ pour se référer à la fonction de vraisemblance paramétrique de la classe X. D'autre part, par souci de simplicité nous parlerons souvent de "modèle du locuteur X" à la place de "fonction de vraisemblance associée au locuteur X", en gardant à l'esprit que cette appellation ne désigne en aucun cas le modèle statistique réel et complet de X.

2.1.2 Estimation paramétrique des modèles

L'estimation des paramètres du modèle d'un locuteur X à partir d'un ensemble de données d'apprentissage $\mathcal{Y}_X$ est basée sur une étape d'optimisation. Elle consiste à maximiser un critère de modélisation par rapport aux paramètres Λ_X du modèle, étant donné l'ensemble $\mathcal{Y}_X$.

Des critères bien connus sont par exemple le critère du maximum de vraisemblance (ML : Maximum Likelihood) ou du maximum *a posteriori* (MAP). Pour ce dernier la règle d'optimisation des paramètres Λ_X s'écrit

$$\Lambda_X^{(MAP)} = \arg\max_{\Lambda} p(\mathcal{Y}_X|\Lambda)p(\Lambda) \ . \tag{2.1}$$

Dans cette équation, $p(\Lambda)$ désigne la distribution *a priori* des paramètres Λ_X. Le critère ML peut être vu comme un cas particulier du critère MAP où la distribution *a priori* $p(\Lambda)$ est non-informative et disparaît donc de l'équation 2.1. L'étape d'estimation consiste dans ce cas à trouver l'ensemble de paramètres $\Lambda_X^{(ML)}$ qui maximise la fonction de vraisemblance $p(\mathcal{Y}_X|\Lambda)$.

Si l'on dispose d'informations *a priori* suffisamment fiables sur la distribution des paramètres du modèle, le critère MAP conduit en général à une estimation plus robuste que celle obtenue par ML. Les valeurs estimées des paramètres sont en effet régularisées par leur valeur *a priori*, ce qui réduit le problème de sur-adaptation du modèle aux données d'apprentissage. C'est particulièrement le cas lorsque celles-ci sont limitées

1. l'indice de classe X dans Λ_X sera parfois omis par la suite lorsque l'on se référera à un ensemble de paramètres non spécifiquement associé à une classe

ou éparses. Des alternatives aux critères ML et MAP existent, notamment les critères MMI (Maximum Mutual Information) et MCE (Minimum Classification Error) qui conduisent à un apprentissage discriminant. Néanmoins, ce type de critère est souvent plus délicat à mettre en oeuvre que les critères ML ou MAP.

L'estimation des paramètres dans le cas de modèles à variables cachées comme les modèles de Markov cachés (HMM : Hidden Markov Models) ou les modèles de mélange de Gaussiennes (GMM : Gaussian Mixture Models) peut être menée de façon itérative par l'algorithme EM (Expectation-Maximization) [Dempster et al.77] qui garantit une convergence vers un optimum local. Les détails de l'estimation des paramètres d'un GMM au sens du critère MAP via l'algorithme EM sont présentés à la section 2.6.

2.1.3 Décision par test d'hypothèses Bayésien

La VAL est une tâche de reconnaissance qui doit fournir une sortie binaire, acceptant ou rejetant l'identité proclamée de l'énoncé de test. Pour un énoncé de test $\mathcal{Y}$ et une identité proclamée X, le processus de décision du système de VAL doit faire un choix entre les deux hypothèses H_X et $H_{\bar{X}}$ suivantes :

H_X : "l'énoncé $\mathcal{Y}$ a été prononcé par le locuteur X".

$H_{\bar{X}}$: "l'énoncé $\mathcal{Y}$ a été prononcé par un autre locuteur que X".

Dans l'approche probabiliste, l'hypothèse H_X est représentée par la fonction de vraisemblance $p(\mathcal{Y}|X)$ du locuteur X et l'hypothèse $H_{\bar{X}}$ par la fonction de vraisemblance $p(\mathcal{Y}|\bar{X})$ associée à la classe $\bar{X}$, représentant tous les locuteurs autres que X. La décision doit se prendre en fonction de la vraisemblance des deux hypothèses concurrentes, mais aussi des coûts applicatifs associés au choix à tort de chacune de ces hypothèses (i.e. coût de fausse acceptation C_{FA} et coût de faux rejet C_{FR}). Le problème se résout dans le cadre de la théorie de la décision Bayésienne en formant le rapport de vraisemblance des deux hypothèses qui est comparé à un seuil de décision θ :

$$\frac{p(\mathcal{Y}|X)}{p(\mathcal{Y}|\bar{X})} > \theta \longrightarrow \text{l'hypothèse } H_X \text{ est acceptée;}$$
$$\frac{p(\mathcal{Y}|X)}{p(\mathcal{Y}|\bar{X})} < \theta \longrightarrow \text{l'hypothèse } H_X \text{ est rejetée.}$$

La valeur théoriquement optimale du seuil de décision θ est donnée par la théorie Bayésienne comme le rapport des probabilités *a priori* des deux hypothèses, pondérées par leur coût d'erreur respectif :

$$\theta = \frac{p(\bar{X}).C_{FA}}{p(X).C_{FR}} \quad . \tag{2.2}$$

Cependant, ce seuil n'est optimal que si les fonctions de vraisemblance représentent les modèles exacts des classes X et $\bar{X}$. Ce n'est jamais le cas en pratique et le seuil doit être réajusté pour chaque locuteur considéré. De manière équivalente, le rapport de vraisemblance obtenu lors d'un test peut être normalisé de façon dépendante de l'identité proclamée afin de garder le seuil de décision fixe pour tous les locuteurs. Le

but de la normalisation des scores est alors que le seuil optimal soit le plus stable possible d'un accès à un autre. Cela permet d'avoir une procédure de recherche d'un seuil unique optimisant les performances du système pour l'ensemble des locuteurs, facilitant ainsi l'ajout de nouveaux utilisateurs.

2.2 Techniques courantes des systèmes de VAL par approche probabiliste

Nous avons posé les bases théoriques de la VAL par approche probabiliste. Les techniques qui ont été développées dans ce cadre pour la VAL sont nombreuses. Nous n'en ferons pas ici une description exhaustive mais présenterons les méthodes les plus couramment utilisées.

Le schéma de fonctionnement d'un système de VAL par approche probabiliste est représenté sur la figure 2.1, rappelant l'architecture déjà présentée à la section 1.3.1

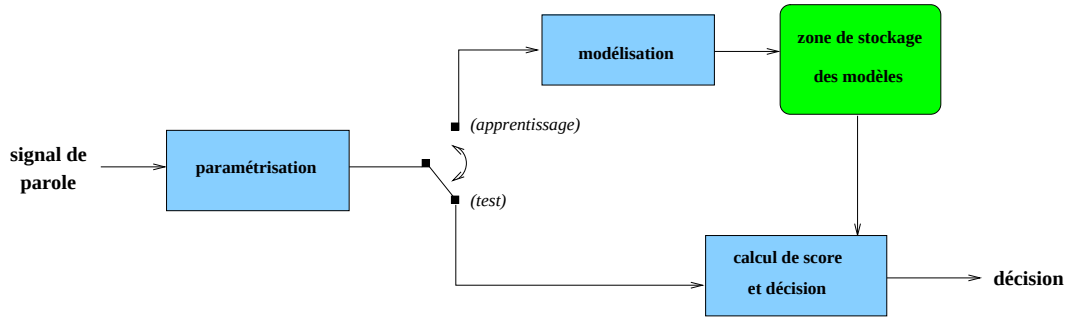


FIG. 2.1 – *Architecture d'un système de VAL*

L'appellation des différents modules intervenant au long de la chaîne de vérification a été adaptée au cadre probabiliste :

- le module de *paramétrisation* a pour but de fournir une représentation du signal de parole adaptée à une tâche de reconnaissance du locuteur.
- le module de *modélisation* est chargé d'estimer les modèles probabilistes servant de références caractéristiques des locuteurs. Le fonctionnement de ce module est basé sur les développements théoriques décrits aux sections 2.1.1 et 2.1.2.
- le module de *calcul de score et décision* chargé d'accepter ou de rejeter l'identité proclamée d'un énoncé de test. Son fonctionnement est fondé sur la théorie de la décision Bayésienne décrite en section 2.1.3.

Pour chacun de ces modules nous décrivons les principales méthodes actuellement utilisées.

2.2.1 Paramétrisation du signal de parole

2.2.1.1 Propriétés recherchées

Le signal de parole qui arrive en entrée d'un système de VAL est conditionné par les contraintes applicatives liées à son acquisition (cf section 1.3.2). Il peut être perturbé par des bruits environnants, limité dans sa bande de fréquences par la bande passante du microphone et du canal de transmission ou encore affecté par un processus de codage. De ce signal distordu, le module de paramétrisation doit extraire des paramètres caractéristiques du locuteur qui permettent de le discriminer des autres individus et qui soient aussi stables que possible d'une session d'acquisition à une autre. En d'autres termes, afin que les différentes classes de locuteur soient facilement séparables dans l'espace des paramètres, ceux-ci devraient idéalement avoir une forte variabilité inter-locuteur, et une faible variabilité intra-locuteur. D'autre part il est souhaitable que ces paramètres soient peu affectés par les perturbations citées ci-dessus afin que celles-ci ne faussent pas le processus de reconnaissance du locuteur. Des paramètres satisfaisant l'ensemble de ces conditions n'existent pas en pratique mais le but de l'étape de paramétrisation est de s'en approcher autant que possible afin de faciliter la vérification d'identité et de rendre le système plus robuste aux perturbations.

Les différents types de paramètres exploitables pour caractériser un locuteur ont été présentés à la section 1.2.1.2. Nous nous limitons ici à la description de l'extraction des paramètres acoustiques, qui sont actuellement les plus répandus au sein des systèmes de VAL et qui sont ceux que nous utilisons dans le cadre du travail présenté dans ce document.

2.2.1.2 Analyse acoustique

En VAL, de nombreux procédés d'analyse acoustique ont été mis au point et évalués. On peut trouver une comparaison de certains d'entre eux dans [Bimbot et al.00, Reynolds94, MC et al.96].

Actuellement, l'approche la plus couramment adoptée consiste à extraire du signal de parole une suite de vecteurs de coefficients qui caractérisent l'enveloppe spectrale[2] à court terme du signal. Pour cela, après une étape de pré-traitements (pré-accentuation des hautes fréquences, filtrage passe-bande,...), le signal de parole est découpé en trames de taille fixe (typiquement 25 ms) à un taux de 100Hz environ, et sur-lesquelles il est supposé avoir des caractéristiques stationnaires. Pour chacune de ces trames, un vecteur de coefficients cepstraux est extrait. Ceux-ci correspondent aux coefficients de décomposition en série de Fourrier du logarithme de la densité spectrale de puissance (DSP) du signal [Bogert et al.63].

2. Nous rappelons que cette enveloppe spectrale reflète les caractéristiques morphologiques du conduit vocal du locuteur

Extraction des coefficients cepstraux

Deux approches coexistent actuellement pour l'extraction des coefficients cepstraux.

La première approche est basée sur une étape d'estimation spectrale qui utilise un calcul de FFT (Fast Fourrier Transfrom) sur la trame en cours, suivie d'une analyse en banc de filtre fournissant des coefficients d'énergie par bandes de fréquences. Ces bandes de fréquences peuvent être réparties suivant une échelle linéaire pour le calcul de coefficients LFCC (Linear Frequency Cepstral Coefficient), ou suivant une échelle Mel [Davis et al.80] pour le calcul de coefficients MFCC (Mel Frequency Cepstral Coefficient). Une transformation DCT (Discret Cosine Transform) est ensuite appliquée au logarithme des coefficients d'énergie issus de l'analyse en banc de filtre afin d'obtenir les coefficients cepstraux. Cette transformation DCT a pour effet de décorréler les coefficients cepstraux ce qui amène a une meilleure représentation du signal, plus "unique" et plus compacte. D'autres techniques plus complexes d'analyse de données ont été testées pour tenter de décorréler au mieux les coefficients de représentation, notamment l'analyse en composante principale (ACP) [MC et al.00], mais sans qu'il n'en ressorte un avantage suffisamment important pour que ces techniques se généralisent.

La seconde approche pour l'extraction des coefficients cepstaux est basée sur l'analyse LPC (Linear Predictive Coding) qui fournit un ensemble de coefficients de prédiction linéaire. Les coefficients cepstraux LPCC (Linear Predictive Cepstral Coefficient) sont ensuite directement calculables à partir des coefficients LPC par un simple algorithme récursif [Atal74].

Quelle que soit l'approche utilisée pour les extraire, seuls les premiers coefficients cepstraux sont retenus (typiquement une quinzaine).

Ajout d'informations dynamiques

On ajoute en général aux coefficients cepstraux "statiques" des coefficients représentatifs de la dynamique du signal de parole. Ces coefficients dynamiques sont très souvent des estimations locales (sur quelques trames) des dérivés premières (coefficients Δ) et parfois secondes (coefficients $\Delta\Delta$) des coefficients cepstraux "statiques". D'autres paramètres ont été étudiés pour la représentation d'informations dynamiques (voir par exemple [Fredouille00, MC97]) mais l'approche la plus répandue aujourd'hui reste l'utilisation des coefficients Δ et $\Delta\Delta$. Des coefficient énergétiques tels que la log-énergie et sa dérivée première (Δ-log-énergie) peuvent également être ajoutés aux autres coefficients pour former le vecteur acoustique caractérisant la trame en cours.

Il n'existe pas actuellement d'étude ayant mis en évidence l'avantage d'une des approches citées ci-dessus de façon systématique. Chacune de ces approches a montré de légers avantages dans certaines situations particulières mais sans que cela ne se généralise à toutes les conditions. On trouve ainsi de multiples configurations au sein des systèmes actuels : extraction de LFCC, de MFCC ou de LPCC, utilisation ou non des coefficients $\Delta\Delta$, de la log-énergie et/ou de la Δ-log-énergie, etc...

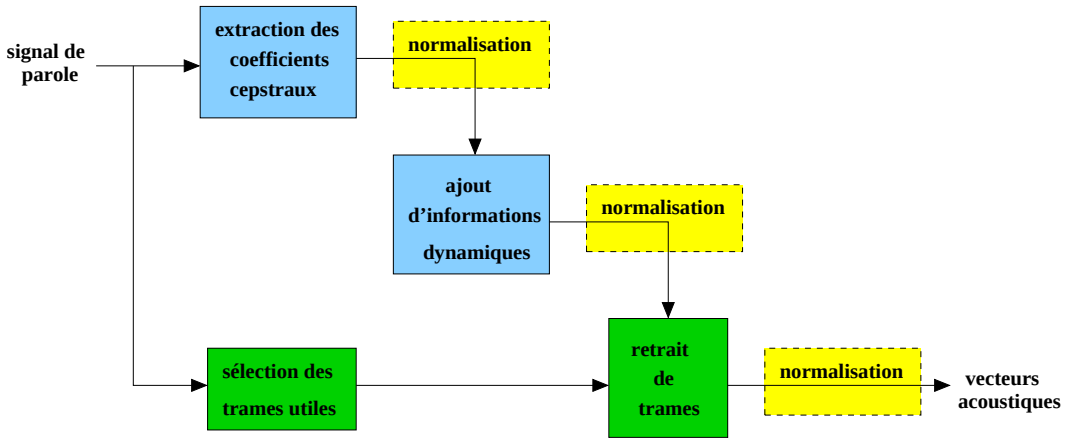


FIG. 2.2 – *Les différentes étapes de la paramétrisation du signal de parole.*

2.2.1.3 Traitements supplémentaires

L'extraction des vecteurs acoustiques est en général associée, dans le module de paramétrisation, à d'autres procédés ayant pour but de renforcer la robustesse des descripteurs utilisés vis-à-vis des perturbations (bruit, canal...) :

- un procédé de **sélection des trames utiles**. Ce procédé est très souvent basé sur une modélisation bi-Gaussienne de l'énergie des trames d'un énoncé, les trames utiles (parole ou parole+bruit) étant supposées appartenir à la Gaussienne de haute énergie, les autres (silence, bruit) à la Gaussienne de basse énergie. La sélection des trames peut alors se faire à partir d'un seuil dépendant des paramètres de la Gaussienne de haute énergie (voir par exemple [MC et al.01]) ou par classification au maximum de vraisemblance. On peut encore imaginer des procédés plus élaborés utilisant une modélisation plus complexe du profil d'énergie ou basés sur un modèle de Markov à 2 états parole/silence, etc...

- une étape de **normalisation des vecteurs acoustiques**. Cette étape a pour but d'atténuer les effets liés aux perturbations engendrées par les bruits additifs (environnement sonore) et convolutif (caractéristiques du microphone, distorsion de canal). Certaines propriétés des coefficients cepstraux ont permis la mise au point de techniques de normalisation très efficaces dans l'espace acoustique. Ce type de normalisation fait l'objet d'une section dans le chapitre 3, où elles seront présentées plus en détail.

Les deux types de traitements mentionnés ci-dessus, bien qu'apparemment secondaires, sont en réalité très importants pour garantir une certaine robustesse du système aux changements de conditions d'utilisation et aux perturbations. Notons enfin que des normalisation peuvent être appliquées en divers endroits du module de paramétrisation (cf figure 2.2) : après l'extraction des coefficients cepstraux statiques, après l'ajout des

coefficients dynamiques ou en dernière étape du module de paramétrisation, après le retrait de trames.

2.2.2 Modélisation probabiliste pour la VAL

La plupart des approches développées jusqu'ici pour la VAL peuvent être exprimées ou reformulées dans un formalisme probabiliste. Une grande partie d'entre elles entre dans le cadre général des modèles de Markov cachés (HMMs) à lois d'émission multi-Gaussiennes. Les modèles Gaussiens, multi-Gaussiens (GMMs), les HMMs ergodiques ou gauche-droite à loi d'émission Gaussienne ou multi-Gaussiennes, s'expriment directement comme des configurations particulières de HMMs à loi d'émission multi-Gaussiennes.

Mais d'autres approches moins récentes telle que la DTW (Dynamic Time Warping) ou la quantification vectorielle (VQ) peuvent également être reformulées dans ce cadre. La DTW peut être vue, moyennant quelques restrictions, comme une version particulière de HMM gauche-droite. Quant à la VQ, elle est ni plus ni moins équivalente à un GMM dont les composantes Gaussiennes ont des covariances toutes égales.

Dans la suite de ce document, nous nous placerons exclusivement dans le cadre des approches probabilistes s'exprimant dans le formalisme des HMMs à loi d'émission multi-Gaussiennes.

2.2.2.1 Modèles de locuteurs : cadre des HMMs

Le choix d'une configuration donnée d'un modèle de locuteur dans le cadre des HMM est principalement guidé par le type de parole que l'on rencontre dans l'application de VAL. Ainsi les HMMs gauche-droite sont utilisés pour les systèmes dépendants du texte car ils intègrent la notion de succession temporelle des événements acoustiques. En mode indépendant du texte, les HMMs ergodiques et les GMMs sont privilégiés car la succession temporelle des événements acoustiques est supposée non-contrainte.

D'autre part, la quantité de données disponibles pour l'estimation des modèles guide souvent le choix d'un type de lois d'émission. Plus cette quantité est importante, plus les lois d'émission peuvent être complexes car un nombre plus important de paramètres peut être estimé de façon fiable.

Les modèles mono-Gaussiens, ou MSSO (Modèles Statistiques du Second Ordre) [Bimbot et al.95] peuvent être utilisés, d'une part pour leur simplicité, et d'autre part lorsque la quantité de données est très limitée. En effet, le faible nombre de paramètres du modèle fait que l'estimation de ces paramètres reste relativement robuste même si l'on ne dispose que d'une faible quantité de données d'apprentissage.

Cependant, lorsque la quantité de données augmente, les performances des modèles mono-Gaussiens sont vite surpassées par des modèles plus complexes comme les GMMs. Ceux-ci permettent en effet de mieux prendre en compte la complexité des distributions

réelles des vecteurs acoustiques. Si l'estimation des paramètres d'un GMM est en théorie peu robuste à de faible quantité de données, des techniques d'adaptation à partir d'un modèle *a priori* permettent d'améliorer cette robustesse. Ce type de méthodes constitue actuellement l'état de l'art des systèmes de VAL par approche probabiliste en mode indépendant du texte [Reynolds et al.00].

Le travail présenté dans ce document s'inscrit également dans le cadre de cette approche. Elle sera décrite plus en détail à la section 2.3.

2.2.2.2 Modèles de non-locuteurs

Le modèle de non-locuteur représente l'hypothèse alternative $H_{\bar{X}}$ dans le processus de décision Bayésienne. Si le modèle d'un locuteur X s'estime naturellement à partir des données d'apprentissage disponibles pour ce locuteur, l'estimation du modèle de non-locuteur associé à X est plus délicate. Ce modèle de non-locuteur est censé représenter l'ensemble des autres locuteurs que X. Bien entendu, des données ne sont pas disponibles pour représenter l'ensemble de ces locuteurs et certaines stratégies ont été adoptées en VAL pour approximer au mieux le modèle de non-locuteur pour chaque locuteur X.

La première approche consiste à sélectionner pour chaque locuteur X un ensemble de locuteurs $\{\bar{X}_1, ..., \bar{X}_N\}$, appelé *cohorte* [Rosenberg et al.92]. Ces locuteurs sont sélectionnés en fonction de certaines caractéristiques qu'ils ont vis-à-vis du locuteur X (par exemple des caractéristiques proches). L'hypothèse $H_{\bar{X}}$ peut alors être représentée par un modèle unique $p(y|\{\bar{X}_n\})$ appris sur toutes les données des locuteurs de la cohorte. Une autre solution est de représenter $H_{\bar{X}}$ par l'ensemble des modèles $\{p(y|\bar{X}_1), ..., p(y|\bar{X}_N)\}$ appris sur les données de chaque locuteur, auquel cas la moyenne des vraisemblances de ces modèles est utilisée dans le test d'hypothèses Bayésien. Dans les deux cas, le modèle unique de non-locuteur ou la cohorte de modèles de non-locuteur est dépendante du locuteur X.

La seconde approche consiste à utiliser un modèle unique de non-locuteur pour tout locuteur X [Carey et al.91, Rosenberg et al.96]. Ce modèle est appris sur de la parole concaténée provenant d'un nombre important de locuteurs imposteurs représentant suffisamment bien toute la variété possible de la parole. Le modèle obtenu est un modèle générique de parole souvent appelé "modèle du monde" ou "UBM" (en anglais : Universal Background Model). Ce modèle, que l'on notera $p(y|\Omega)$, est utilisé comme représentation commune de l'hypothèse $H_{\bar{X}}$ pour l'ensemble des locuteurs référencés.

Des deux approches mentionnées ci-dessus, la deuxième est la plus largement utilisée actuellement au sein des systèmes de VAL. Elle possède en effet l'avantage de ne nécessiter l'estimation et la mémorisation que d'un seul modèle de non-locuteur pour l'ensemble des locuteurs d'une application. De plus elle n'engendre en général pas de perte significative des performances lorsque la quantité de données utilisée pour son estimation est suffisamment importante.

2.2.3 Processus de décision

2.2.3.1 Scores de vérification

D'après la théorie Bayésienne de la décision, le rapport de vraisemblance des hypothèses H_X et $H_{\bar{X}}$ comparé à un seuil permet d'obtenir les performances de classification optimales. C'est donc naturellement que cette statistique s'est imposée comme score de vérification pour la VAL[3]. Pour des raisons liées à la précision des calculs et à la maniabilité des valeurs, c'est en général le logarithme de ce rapport de vraisemblance (LLR : Log Likelihood Ratio) qui est utilisé. D'autre part, l'hypothèse d'indépendance temporelle des observations acoustiques est communément adoptée, ce qui permet d'écrire le LLR d'un énoncé complet $\mathcal{Y} = \{y_1, ..., y_N\}$ comme la somme des LLRs pour chaque trame de l'énoncé. La valeur de cette somme présente cependant le désavantage d'être très dépendante du nombre de trames. On normalise par conséquent cette somme par le nombre de trames N, de telle sorte que le score de vérification $S_X(\mathcal{Y})$ corresponde au final à la moyenne des LLRs par trame :

$$S_X(\mathcal{Y}) = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n|X)}{p(y_n|\bar{X})} \quad . \tag{2.3}$$

En théorie, c'est ce score brut qui devrait être comparé à un seuil unique pour prendre la décision. Mais en pratique, certains ajustements sont nécessaires pour compenser des écarts au cadre théorique de la décision Bayésienne.

On trouve également dans la littérature quelques approches utilisant des règles de décisions différentes de la simple comparaison d'un score à un seuil. Dans [Bengio et al.01], Bengio et Mariéthoz utilisent un vecteur de scores formé par les deux composantes $\log p(\mathcal{Y}|X)$ et $\log p(\mathcal{Y}|\bar{X})$ du LLR. La fonction de classification client/imposteur de ces vecteurs de scores est ensuite apprise à l'aide d'un SVM (Support Vector Machine).

2.2.3.2 Ajustement du seuil *vs* normalisation des scores

En pratique, les données d'apprentissage des modèles de locuteur sont toujours limitées. Par conséquent les modèles estimés ne sont que des approximations des densités de probabilité exactes et l'on sort du cadre théorique de la décision Bayésienne. Notamment, les disparités entre les conditions d'acquisition des données d'apprentissage et celles de test entraînent des biais dans le LLR. Cela veut dire qu'un seuil unique de décision conduit à des performances de classification sous-optimales et il est nécessaire de réajuster ce seuil en fonction de chaque accès (couple modèle/énoncé de test) considéré afin de compenser les effets liés aux variabilités rencontrées.

3. notons cependant que les premières approches n'utilisaient que la vraisemblance associée à H_X, ce qui revient à considérer la vraisemblance de $H_{\bar{X}}$ comme constante dans le rapport de vraisemblance.

De manière équivalente, il est possible de normaliser les scores en fonction du locuteur proclamé et/ou de l'énoncé de test d'un accès. Cet ajustement des scores a pour but de rendre le seuil optimal de décision plus stable d'un accès à un autre et d'améliorer ainsi la robustesse du processus de vérification vis-à-vis des changements de conditions d'utilisation. Dans ce cas le score normalisé de l'accès est une fonction $\mathcal{F}(\cdot)$ du score brut, du locuteur proclamé et/ou de l'énoncé de test. On cherche alors à déterminer une fonction $\mathcal{F}(\cdot)$ telle que l'application d'un seuil unique de décision θ conduise à des performances optimales. La règle de décision s'exprime alors de la manière suivante :

$$S_X^{norm}(\mathcal{Y}) = \mathcal{F}(S_X(\mathcal{Y}), X, \mathcal{Y}) \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta \ . \tag{2.4}$$

Les techniques correspondantes sont appelées "normalisations de scores" . Une partie de nos travaux s'inscrit dans le cadre de ces techniques. Elles seront présentées en détail au chapitre 3 à la section 3.4.

2.3 Approche état de l'art : vérification du locuteur par GMMs

Les vecteurs acoustiques issus de la phase de paramétrisation du signal de parole ont une distribution complexe. Les modèles simples mono-Gaussiens sont insuffisants pour prendre en compte tous les détails de cette distribution. Des modèles plus complexes ont donc été étudiés pour approximer plus finement la distribution des vecteurs acoustiques. En particulier les modèles de mélanges de Gaussiennes (GMMs) ont apporté des gains de performance importants grâce à leur capacité à modéliser des formes complexes de distribution. Cependant, le nombre important de paramètres d'un GMM exige l'utilisation de méthodes robustes d'estimation des modèles de locuteurs, en particulier lorsque la quantité de données d'apprentissage est restreinte. La méthode d'estimation la plus répandue actuellement est l'adaptation Bayésienne, présentée dans cette section.

2.3.1 Modélisation par mélange de Gaussiennes

L'utilisation des GMMs pour la modélisation des locuteurs a été initiée par les travaux de thèse de Douglas Reynolds [Reynolds92]. Cette approche a donné, depuis plus de 10 ans maintenant, les meilleures performances pour les systèmes de reconnaissance du locuteur en mode indépendant du texte basés sur l'approche probabiliste. La plupart des systèmes état de l'art actuels utilisent une modélisation des locuteurs par GMM (parfois associée à d'autres approches).

La densité de probabilité $p(y|\Lambda)$ d'un GMM à K composantes pour des vecteurs

acoustiques de dimension D est définie de la façon suivante :

$$p(y|\Lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(y|m_k, S_k) \quad . \tag{2.5}$$

Dans cette équation, $\mathcal{N}(\cdot|m_k, S_k)$ est une fonction Gaussienne avec un vecteur moyenne m_k de dimension D, une matrice de covariance S_k de dimension $D \times D$. Le paramètre w_k est le poids de la Gaussienne $\mathcal{N}(\cdot|m_k, S_k)$ dans le mélange, avec les contraintes :

$$\sum_{k=1}^{K} w_k = 1, \quad \text{et} \quad w_k \geq 0 \quad \forall k \ .$$

Un GMM est donc défini par l'ensemble des paramètres $\Lambda = (\{w_k\}, \{m_k\}, \{S_k\}$. Cette famille de modèles est bien adaptée pour approximer les densités de probabilités réelles multidimensionnelles. En effet, en augmentant le nombre K de composantes Gaussiennes, un GMM peut théoriquement modéliser n'importe quelle loi probabiliste. Cependant, l'augmentation du nombre de Gaussiennes intensifie la complexité des calculs, à la fois pour l'étape d'estimation et pour le calcul de scores lors d'une vérification. D'autre part, cela pose également le problème de la robustesse de l'estimation des paramètres du modèle. Il faut en effet davantage de données pour estimer de façon fiable un nombre plus important de paramètres.

En pratique, la quantité de données d'apprentissage est toujours limitée, parfois même très restrictive pour certaines tâches de VAL. Dans ce cas, l'estimation au maximum de vraisemblance d'un GMM de locuteur conduit en général à un modèle sur-adapté aux données, se généralisant mal à d'autres accès clients. L'utilisation du critère MAP permet d'éviter cette sur-adaptation en introduisant de l'information *a priori* sur la distribution des paramètres. Si cette information *a priori* est correctement définie, l'estimation des modèles qui en découle est généralement plus robuste au manque de données d'apprentissage.

Les principes de l'estimation au sens du critère MAP (aussi appelée adaptation Bayésienne) des paramètres d'un GMM ont été développés dans [Gauvain et al.94] et appliqués pour les systèmes de VAL en mode indépendant du texte dans [Reynolds et al.00]. Cette approche a permis une amélioration conséquente des performances pour les tâches de VAL où la quantité de données d'apprentissage est limitée. Nous rappelons brièvement les principes de cette approche dans la section suivante et donnons les formules d'estimations correspondantes, dans le cadre de l'algorithme EM.

2.3.2 Adaptation Bayésienne des GMMs de locuteurs

La règle d'estimation des paramètres d'un GMM au sens du critère MAP s'écrit

$$(\hat{\mathbf{w}}, \hat{\mathbf{m}}, \hat{\mathbf{S}}) = \arg \max_{(\mathbf{w},\mathbf{m},\mathbf{S})} \sum_{k=1}^{K} w_k . \mathcal{N}(\mathcal{Y}|m_k, S_k) p(\mathbf{w}, \mathbf{m}, \mathbf{S}) \ , \tag{2.6}$$

où $(\mathbf{w}, \mathbf{m}, \mathbf{S})$ est l'ensemble des poids, vecteurs moyennes et matrices de covariance du GMM. Dans [Gauvain et al.94], Gauvain et Lee font l'hypothèse que les paramètres de deux Gaussiennes différentes dans le mélange sont indépendants. Ils définissent la densité *a priori* conjointe $p(\mathbf{w}, \mathbf{m}, \mathbf{S})$ des paramètres du modèle comme le produit d'une distribution de Dirichlet, loi *a priori* des poids, et d'une distribution Normal-Wishart inverse, loi *a priori* conjointe des matrices de covariance et des vecteurs moyennes. Le choix de ce type d'*a priori* (*a priori* "conjugués") est motivé en premier lieu par les avantages qu'ils procurent en terme de calcul des formules d'estimations des paramètres. Toutefois, en dehors de ces avantages calculatoires, on peut trouver une justification de ces formes d'*a priori* pour les paramètres d'un GMM dans [Geiger et al.98].

2.3.2.1 Estimation des paramètres par l'algorithme EM

Le calcul détaillé des formules d'estimation des paramètres d'un GMM au sens du critère MAP peut être trouvé dans [Gauvain et al.94]. Nous donnons ici des formes simplifiées de ces formules d'estimation proposées par Reynolds *et al.* dans [Reynolds et al.00].

Étant donné un énoncé de test $\mathcal{Y} = \{y_1, ..., y_T\}$ de durée T, les équations à utiliser lors de la phase M de l'algorithme EM pour la mise à jour des paramètres d'un GMM estimé au sens du critère MAP sont les suivantes :

$$\hat{w}_k \propto \alpha_k^w(\gamma_k/T) + (1-\alpha_k^w)\pi_k \ , \tag{2.7}$$

$$\hat{m}_k = \alpha_k^m \overline{y}_k + (1-\alpha_k^\mu)\mu_k \ , \tag{2.8}$$

$$\hat{S}_k = \alpha_k^S \overline{y^2}_k + (1-\alpha_k^S)(\Sigma_k + \mu_k^2) - \hat{m}_k^2 \ . \tag{2.9}$$

Les paramètres π_k, μ_k et Σ_k sont respectivement les poids, vecteur moyenne et matrice de covariance *a priori* pour la Gaussienne k. Les statistiques γ_k , $\overline{y}_k$ et $\overline{y^2}_k$ sont respectivement le taux d'occupation global des données pour la Gaussienne k et les moments d'ordre 1 et 2 des données, calculés selon la Gaussienne k. Ces statistiques sont calculées lors de la phase E de l'algorithme EM à partir des estimations $\{\hat{w}_k^{(-)}, \hat{m}_k^{(-)}, \hat{S}_k^{(-)}\}$ des paramètres obtenues à l'itération précédente, selon les formules :

$$\gamma_k = \sum_{t=1}^{T} \gamma_k(t) \ , \tag{2.10}$$

$$\overline{y}_k = \frac{1}{\gamma_k} \sum_{t=1}^{T} \gamma_k(t) y(t) \ , \tag{2.11}$$

$$\overline{y^2}_k = \frac{1}{\gamma_k} \sum_{t=1}^{T} \gamma_k(t) y(t) y(t)^* \ , \tag{2.12}$$

avec :

$$\gamma_k(t) = \frac{\hat{w}_k^{(-)} \mathcal{N}(y(t)|\hat{m}_k^{(-)}, \hat{S}_k^{(-)})}{\sum_{j=1}^{K} \hat{w}_j^{(-)} \mathcal{G}_j(y(t)|\hat{m}_j^{(-)}, \hat{S}_j^{(-)})} \ , \tag{2.13}$$

et où * désigne la transposition matricielle.

Les coefficients de pondération $\{\alpha_k^w, \alpha_k^m, \alpha_k^S\}$ entre les paramètres *a priori* et ceux estimés sur les données (correspondant aux estimations ML) sont définis par :

$$\alpha_k^\rho = \frac{\gamma_k}{\gamma_k + \tau_k^\rho} \ , \tag{2.14}$$

avec $\rho \in \{w, m, S\}$ et ou τ_k^ρ est un facteur de confiance à l'*a priori* qui contrôle le degré d'adaptation aux données du paramètre ρ dans la Gaussienne k. Ce facteur, qui est homogène à γ_k, peut être vu comme un nombre de trames *a priori* attribuées à la Gaussienne k pour l'estimation du paramètre ρ.

2.3.2.2 Propriétés de l'adaptation MAP

Les formules d'estimations des paramètres au sens du critère MAP (équations 2.7, 2.8 et 2.9) permettent d'effectuer de façon automatique une balance entre les valeurs *a priori* et l'estimation ML des paramètres, en fonction du taux d'occupation de chaque Gaussienne par les données d'apprentissage. Lorsque la Gaussienne k reçoit une grande quantité de données ($\gamma_k \to \infty$), l'estimation des paramètres correspondants est dominée par les statistiques calculées sur les données. Au contraire, lorsque la Gaussienne k reçoit peu de données, ($\gamma_k \to 0$) les paramètres estimés sont dominés par leur valeur *a priori*. En résumé, seules les Gaussiennes appartenant à des régions de l'espace acoustiques explorées par les données d'apprentissage sont modifiées, les paramètres des autres Gaussiennes restant inchangés par rapport à leur valeur *a priori*. Cela empêche que le modèle soit sur-adapté aux données contrairement à l'estimation ML pour laquelle certaines composantes peuvent se sur-spécialiser sur des données singulières de l'ensemble d'apprentissage. Si les valeurs *a priori* des paramètres sont fiables, l'adaptation Bayésienne permet d'améliorer la robustesse de l'étape de modélisation d'un système de VAL face à un manque de données d'apprentissage.

En pratique, des simplifications peuvent être adoptées dans le schéma d'adaptation en considérant par exemple des facteurs de confiance τ_k^ρ indépendants du type de paramètre, voir même indépendants de la Gaussienne considérée. D'autre part, des expériences rapportées dans [Reynolds et al.00] ont montré qu'en adaptant seulement les moyennes du GMM, la perte de performance est négligeable par rapport à l'adaptation beaucoup plus coûteuse de l'ensemble des paramètres (poids, moyennes, covariances).

2.3.2.3 Estimation des paramètres *a priori*

L'adaptation Bayésienne permet d'obtenir des estimations plus robustes des modèles de locuteurs par rapport à l'estimation ML, à condition que les informations *a priori* intégrées dans le schéma d'estimation soient fiables. Une approche communément adoptée

pour déterminer les paramètres *a priori* est de fixer leurs valeurs à celles des paramètres d'un modèle générique (modèle du monde) appris sur une grande quantité de données, avec une grande variété de locuteurs et une bonne représentativité des conditions d'enregistrement susceptibles d'être rencontrées. Cette approche repose sur l'hypothèse que le modèle du monde capture les valeurs moyennes des paramètres de chaque Gaussienne. On considère alors ces valeurs moyennes comme valeurs *a priori* des paramètres. Ce modèle du monde, qui est utilisé comme initialisation de l'algorithme EM dans l'étape d'estimation, a pour but de combler certains manques dans la distribution des données d'apprentissage. L'efficacité de l'adaptation Bayésienne repose en grande partie sur la qualité du modèle du monde en terme de représentativité des conditions pouvant être rencontrées. Si ce modèle du monde est appris sur un large ensemble de données représentatives, les valeurs estimées de ses paramètres sont en général fiables. Son utilisation comme modèle *a priori* dans l'adaptation Bayésienne permet alors d'améliorer significativement la robustesse du processus d'estimation des GMMs de locuteurs.

2.3.3 Décision Bayésienne normalisée

Les systèmes état de l'art en VAL utilisent tous des techniques de normalisations de scores destinées à renforcer leur robustesse vis-à-vis des changements de conditions d'acquisition, particulièrement significatives pour les applications par téléphone par exemple. Ces techniques seront présentées en détail à la section 3.4 du chapitre 3. Nous résumons ici leurs objectifs et les principes d'implémentation de ces techniques.

D'un point de vue général, les techniques de normalisation des scores cherchent à compenser des biais introduits dans le processus de vérification du fait de disparités entre les conditions d'acquisition des données d'apprentissage et celles de test. Ces biais dépendants des conditions d'apprentissage et de test impliquent que le processus de décision optimale ne peut pas être basé sur un seuil unique, comme le prédit pourtant la théorie Bayésienne de la décision. Chaque type de conditions engendre des biais spécifiques dans les scores.

Les techniques actuelles de normalisation des scores sont basées sur l'étude de ces biais au travers d'une modélisation de la distribution des scores engendrés par tel modèle ou tel énoncé de test. Elles utilisent pour cela un ensemble d'accès réels dont elles tirent une série de scores servant à estimer certains paramètres (moyenne et variance) de la distribution des scores. Le score de l'accès de test est ensuite normalisé en utilisant les paramètres de normalisation.

En général, seule la distribution des scores imposteurs est modélisée car on ne dispose pas d'un nombre d'accès clients suffisant pour modéliser la distribution des scores clients. Néanmoins, les techniques de normalisation basées sur l'étude des distributions des scores imposteurs permettent de réduire considérablement la variabilité des scores. Elles améliorent de façon significative les performances de reconnaissance lorsqu'on utilise un seuil de décision unique.

2.4 Synthèse et objectifs visés

Un des cadres applicatifs forts du travail effectué en reconnaissance du locuteur au sein de l'équipe METISS dans laquelle s'est déroulée cette thèse est constitué par les évaluations organisées annuellement par NIST. Ces évaluations proposent de tester et de comparer les systèmes de VAL dans des conditions particulièrement variables en terme d'environnement d'acquisition et de durée des énoncés. Dans ce cadre, c'est naturellement que nous nous sommes intéressés à la robustesse des système de VAL à ce type de variabilité.

Dans cette section, nous identifions tout d'abord les différents niveaux où peut intervenir le traitement de la robustesse des systèmes de VAL. Puis, nous présentons quelques limitations des méthodes actuelles destinées à améliorer cette robustesse. Ces limitations nous permettent ensuite de définir les objectifs de notre travail, destinés à remédier à certains des problèmes soulevés.

2.4.1 Traitement de la robustesse

Les problèmes de robustesse des systèmes de VAL sont principalement liés aux différentes variabilités rencontrées dans ce type d'applications. Ces variabilités proviennent de l'évolution intrinsèque de la voix du locuteur, des différences de contenus linguistiques des énoncés et des changements de conditions d'utilisation du système. La quantité toujours limitée de données disponibles à l'apprentissage fait que la référence caractéristique d'un locuteur est toujours spécifique d'un état ponctuel de la voix du locuteur, d'un contenu linguistique particulier (lorsqu'on travaille en mode indépendant du texte) et de conditions environnementales et technologiques particulières. En raison de cette spécificité, de fortes disparités peuvent être rencontrées entre les données de test et celle d'apprentissage, entraînant des difficultés de reconnaissance. Pour tenter de minimiser les effets de ces disparités, plusieurs approches ont été considérées en VAL.

Des techniques de normalisations ont été développées à différents niveaux d'un système de VAL pour tenter d'éliminer les biais causés par les sources de variabilité. Le but est alors de rendre les éléments intervenant dans la chaîne de vérification les plus indépendants possibles des informations perturbantes tout en préservant au mieux les caractéristiques spécifiques au locuteur. Les techniques de normalisation des vecteurs acoustiques et des scores de vérification entrent dans ce cadre.

Pour minimiser les disparités entre les données de test et d'apprentissage, des techniques d'adaptation peuvent être utilisées. Dans ce cas on peut soit modifier la référence caractéristique du locuteur en l'adaptant aux conditions de test, soit transformer les données de test pour qu'elles correspondent mieux aux conditions d'apprentissage.

Enfin, il est possible de combler certains manques dans les données d'apprentissage en utilisant de l'information *a priori*. La prise en compte de ces informations *a priori* intervient en particulier au niveau de la modélisation. L'adaptation Bayésienne des modèles de locuteurs à partir d'un modèle *a priori* fait typiquement partie de ce type de méthodes.

Ces trois approches différentes visant à améliorer la robustesse des systèmes de

VAL ont montré chacune leur intérêt pour l'amélioration des performances pour les tâches soumises à de fortes variabilités. Cependant elles possèdent chacune certaines limitations qui les rendent parfois difficilement utilisables pour certaines applications ou inadaptées à certaines situations.

2.4.2 Limitations des méthodes actuelles et objectifs du travail

Dans cette thèse nous avons abordé le traitement de la robustesse en développant tout d'abord de nouvelles techniques de normalisation.

Certaines des techniques courantes de normalisation, en particulier les normalisations de scores très utilisées actuellement, nécessitent l'utilisation d'un corpus de données externes servant à estimer les paramètres de normalisation. Par conséquent, ces techniques ne peuvent être appliquées dans certains cas applicatifs où un corpus externe, représentatif de l'application, n'est pas disponible. Dans la deuxième partie de ce document, nous explorons les possibilités de mise en oeuvre de technique de normalisation ne nécessitant pas de données additionnelles. Les nouvelles techniques de normalisation que nous proposons interviennent aux niveaux des scores et des modèles. Ces normalisations ne sont pas tributaires d'un corpus de données externes et peuvent donc être appliquées plus facilement.

Dans un deuxième temps nous étudions un schéma d'adaptation MAP hiérarchique destiné à renforcer la robustesse de l'adaptation Bayésienne classique lorsque l'on dispose de très peu de données d'apprentissage.

L'adaptation Bayésienne des modèles de locuteurs a permis une amélioration des performances considérable dans les tâches de RAL où la quantité de données disponibles pour l'apprentissage des modèles est limitée. Cette amélioration est en partie due au fait que les composantes d'un modèle sont contraintes par les lois *a priori* des paramètres, les empêchant ainsi de se sur-spécialiser sur des données éparses (donc peu fiables) de l'ensemble d'apprentissage. En contrepartie, les composantes ne recevant que très peu de données ne sont pas modifiées, de par le fait de l'hypothèse d'indépendance entre différentes Gaussiennes du GMM. Cela peut conduire à des modèles faiblement adaptés et donc peu représentatifs si la tâche considérée est très restrictive en terme de données d'apprentissage. Il est clair cependant que l'utilisation d'un simple modèle générique de parole en terme de modèle *a priori* de locuteur, n'exploite pas toute l'information que l'on peut trouver dans un ensemble d'énoncés de locuteurs. En particulier, l'hypothèse d'indépendance des Gaussiennes d'un GMM est certainement approximative et certains liens peuvent exister entre les différentes régions acoustiques occupées par la voix d'un locuteur. La méthode d'adaptation que nous proposons a pour but de capturer ces dépendances au sein d'une structure hiérachique et d'exploiter cette structure pour adapter des composantes Gaussiennes ne recevant que très peu, ou pas du tout, de données d'apprentissage. On espère ainsi pouvoir obtenir de meilleures estimations des modèles de locuteur lorsque la quantité de données d'apprentissage est très réduite.

Deuxième partie

Techniques de normalisation par distances de Kullback-Leibler

Chapitre 3

État de l'art des techniques de normalisation pour la VAL

Ce chapitre montre un panorama des techniques de normalisation qui ont été développées pour la VAL, dans le cadre de l'approche probabiliste. Ces techniques ont pour principal objectif d'améliorer la robustesse du système en tentant de neutraliser les variabilités perturbantes.

Après avoir identifié les éléments sur lesquels il est possible d'agir pour traiter cette robustesse, nous présentons les techniques mises en place au niveau :

- des paramètres acoustiques;
- des modèles;
- des scores de vérification.

Nous concluons le chapitre par une discussion qui fait une synthèse des techniques de normalisation présentées et dresse un bilan de leurs avantages et inconvénients.

3.1 Problématique et objectifs

Dans les applications réelles de VAL, l'estimation imparfaite des modèles de locuteurs et de non-locuteurs, due au volume restreint des données d'apprentissage, implique que le cadre de la théorie Bayésienne de la décision ne soit pas strictement respecté. Un seuil unique de décision ne mène donc généralement pas à des performances optimales. Des biais provenant de disparités entre les contextes d'apprentissage et de test apparaissent dans les valeurs du rapport de vraisemblance, qui présentent alors une grande variabilité d'un accès à un autre. Par conséquent, le seuil de décision optimal pour chaque accès dépend à la fois du contexte de l'ensemble d'apprentissage et de celui de l'énoncé de test.

Le contexte d'un ensemble de données, qu'il soit destiné à l'apprentissage ou au test, regroupe les conditions d'acquisition du signal (bruits ambiants, prise de son, transmission), le contenu linguistique du message et l'état du locuteur. Ces contextes peuvent être très variables, notamment dans les applications de VAL par téléphonie (terrestre et/ou mobile) et en mode indépendant du texte. C'est le cas des évaluations NIST des dernières années qui proposent des conditions d'utilisation très variables, mettant à rude épreuve la robustesse des systèmes de VAL.

Pour renforcer la robustesse des systèmes face à de telles variations de conditions, de nombreuses techniques de compensation ont été mises au point afin de minimiser la variabilité des scores de vérification. Différentes approches ont été adoptées, le but commun étant d'atténuer les effets des disparités entre les données d'apprentissage et de test.

3.1.1 Techniques de compensation

Dans cette section nous avons identifié trois types de techniques de compensation :

1. **Par incorporation de connaissances *a priori*.**
   Un premier type de techniques consiste à utiliser de l'information *a priori* lors de l'estimation des modèles afin de combler certains manques dans les données d'apprentissage, au niveau du contenu linguistique ou de la variété des conditions d'acquisition. Ce sont les techniques d'estimation de modèles à partir d'un modèle *a priori*, notamment l'adaptation Bayésienne et les méthodes à base de transformation des paramètres du modèle. Ces méthodes seront présentées plus en détail au chapitre 7.

2. **Par adaptation.**
   Un second type de techniques cherche à adapter des éléments du processus de vérification aux conditions particulières d'un autre élément, afin d'homogénéiser les conditions au sein d'un même accès. Ces techniques sont principalement destinées à minimiser des disparités de matériel d'acquisition (microphone) et de transmission (canal). Il peut s'agir d'adapter les modèles aux conditions particulières du test [Teunen et al.00, Beaufrays et al.97] ou de transformer les vec-

teurs acoustiques de test pour qu'ils correspondent mieux aux "conditions" [1] du modèle de locuteur [Mak et al.02]. Les techniques de sélection d'un modèle du monde (ou d'une cohorte d'imposteurs) dépendant des conditions du modèle du locuteur entrent également dans ce cadre [Heck et al.97].

3. **Par normalisation.**
   Le troisième type de techniques consiste à projeter certaines grandeurs intervenant dans le processus de vérification dans des espaces indépendants des conditions d'acquisition des données d'apprentissage et/ou de test. Ce type de techniques est communément appelé "normalisation", même si elles n'effectuent pas une normalisation au sens mathématique ou probabiliste du terme. Elles cherchent à homogénéiser certaines caractéristiques statistiques des éléments intervenant dans la vérification, afin de les rendre indépendantes des conditions d'apprentissage ou de test. Nous réservons le terme de "normalisation" aux techniques qui homogénéisent de façon directe les grandeurs sur lesquelles elles agissent.

Nous nous intéressons dans la suite de ce chapitre à la mise en oeuvre et à l'état de l'art des techniques de normalisation.

3.1.2 Mise en oeuvre des techniques de normalisation

La mise en oeuvre des techniques de normalisation nécessite d'une part de pouvoir mesurer certaines caractéristiques statistiques de la grandeur à normaliser, et d'autre part de définir des caractéristiques cibles "normales" de ces statistiques. Suivant les cas, ces caractéristiques cibles peuvent être choisies arbitrairement ou sur la base de connaissances *a priori*. La normalisation consiste ensuite à contraindre ou à transformer la grandeur à normaliser afin que sa distribution suive les caractéristiques cibles.

Les techniques de normalisation peuvent intervenir à différents niveaux de la chaîne de vérification. En effet, chaque élément intervenant dans le calcul du rapport de vraisemblance (modèle du locuteur, modèle du non-locuteur et vecteurs acoustiques de test) est susceptible d'entraîner un biais spécifique dans le score. Il est donc possible d'agir au niveau de chacun de ces éléments pour réduire la variabilité des scores. On peut également agir directement au niveau des scores eux-mêmes.

Dans la suite de ce chapitre, nous présentons les principales techniques de normalisations qui ont été développées pour la VAL, au niveau des paramètres acoustiques (section 3.2), au niveau des modèles (section 3.3) et au niveau des scores (section 3.4).

1. il s'agit en fait de la configuration particulière du modèle de locuteur, induite par les conditions spécifiques des données d'apprentissage

3.2 Normalisations dans l'espace acoustique

Les distorsions du signal de parole liées aux conditions d'acquisition affectent directement les vecteurs acoustiques issus de la phase de paramétrisation. Leur distribution statistique se retrouve alors biaisée et déformée de façon dépendante de ces conditions. Les techniques de normalisation des paramètres acoustiques cherchent à homogénéiser les caractéristiques de ces distributions afin d'éliminer autant que possible les informations spécifiques des conditions d'acquisition. Elles doivent cependant veiller à conserver au mieux les informations caractéristiques du locuteur.

Historiquement, les bases de données sur lesquelles se sont développés les systèmes de VAL ont présenté de plus en plus de variabilité dans les conditions d'acquisition du signal de parole. Parallèlement, les techniques de normalisation des paramètres acoustiques ont évolué pour prendre en compte cette variabilité grandissante. Nous présentons ici cette évolution par ordre croissant de complexité des techniques.

Retrait de la moyenne cepstrale

Il a tout d'abord été constaté que le type de canal de transmission affectait la moyenne des coefficients cepstraux [Furui81]. Cela est dû au fait qu'un filtrage convolutif du signal de parole se retrouve au niveau des coefficients cepstraux comme une composante additive. Si l'on suppose que le canal de transmission est stationnaire sur l'ensemble d'un énoncé, l'effet linéaire du filtrage de canal se retrouve donc dans la moyenne des coefficients cepstraux. La CMS (Cepstral Mean Substraction : soustraction de la moyenne cepstrale) est une technique de normalisation qui consiste à retirer à chaque coefficient cepstral sa moyenne calculée sur toute la durée de l'énoncé, afin d'en éliminer la composante liée au canal de transmission. La CMS peut également être appliquée sur une fenêtre glissante pour prendre en compte des variations lentes du canal au cours de l'énoncé. Cette normalisation permet d'améliorer significativement les performances dans des applications présentant de fortes disparités de canal, comme cela a pu être observé durant les évaluations NIST notamment. Notons cependant que l'action de la CMS se limite à la compensation des effets linéaires introduits par le canal. Les effets non-linéaires comme par exemple des saturations dues au microphone, ne sont pas compensés.

Filtrage des trajectoires cepstrales

Une autre technique de normalisation appelée filtrage RASTA (RelAtive SpecTrAl) [Hermansky et al.94] consiste à éliminer par filtrage passe-bande toutes variations des coefficients cepstraux qui soient trop lentes ou trop rapides pour être dues aux variations du signal de parole. En particulier, ce filtrage a pour effet de retirer la moyenne cepstrale, et conduit à des performances très similaires à celles obtenues par une CMS glissante. Ces deux techniques sont efficaces pour compenser les biais liés au canal de transmission mais sont assez peu robustes face aux bruits additifs.

Centrage et réduction

Si les bruits convolutifs (distorsion de canal) affectent la moyenne des coefficients

cepstraux, il a été observé que des bruits additifs stationnaires ont pour effet de diminuer leur variance. Une explication de ce phénomène peut être trouvée dans [Pelecanos et al.01]. Certaines techniques cherchent donc non seulement à retirer la moyenne des coefficients cepstraux mais également à normaliser leur variance. Pour cela, on effectue un centrage et une réduction des coefficients cepstraux [Viikki et al.98], sur l'ensemble d'un énoncé ou sur une fenêtre glissante. Ce type de normalisation permet de diminuer les disparités liées à la fois à différents canaux de transmission et à différents environnements sonores.

Gaussianisation

La moyenne et la variance des coefficients cepstraux ne sont pas les seuls moments affectés par les distorsions du signal de parole. C'est toute la distribution de ces coefficients qui est déformée par des bruits additifs ou convolutifs et par les effets non-linéaires de la chaîne de captage et de transmission. Cela veut dire que les normalisations de la moyenne et de la variance des coefficients cepstraux ne suffisent pas nécessairement à compenser toute la complexité des effets liés aux conditions d'acquisition. Des techniques plus récentes de normalisation se sont attachées à normaliser la distribution complète des vecteurs acoustiques. Elles utilisent pour cela une distribution "cible" qui peut être choisie arbitrairement ou construite à partir d'un ensemble de données *a priori* sélectionnées.

Le"feature warping" [Pelecanos et al.01] est une normalisation qui consiste à transformer indépendemment chaque dimension des vecteurs acoustiques de façon à ce que la distribution marginale à moyen terme de chaque coefficient acoustique devienne Gaussienne, centrée et réduite. La distribution réelle de chaque coefficient est analysée sur une fenêtre glissante (typiquement 3 secondes) et la transformation du coefficient central de la fenêtre est déterminée grâce à une table de correspondance. Cette table fait correspondre la répartition réelle des coefficients acoustiques avec la répartition "cible" donnée par une loi Gaussienne centrée réduite. Pour plus d'information sur la mise en oeuvre de cette technique, nous invitons le lecteur à consulter [Pelecanos et al.01]. Le "feature warping" effectue implicitement un centrage et une réduction, sur une fenêtre glissante, de la distribution marginale de chaque coefficient acoustique. Mais cette technique normalise en plus la forme de ces distributions minimisant ainsi encore les disparités causées par des différences de conditions d'acquisition. Les performances obtenues par le "feature warping" sont légèrement meilleures que celles obtenues avec un simple centrage/réduction [Barras et al.03], mais la mise en oeuvre de la technique est beaucoup plus lourde dans le premier cas. Une variante du "feature warping" effectue une Gaussianisation conjointe des coefficients acoustiques [Xiang et al.02]. Pour cela, l'étape de Gaussianisation (qui est identique à celle du "feature warping") est précédée d'une étape d'Analyse en Composantes Indépendantes (ACI) qui a pour but de rendre indépendants les différents coefficients de chaque vecteur acoustique.

Projection dans un espace indépendant du canal

La dernière technique de normalisation que nous présentons dans cette section a pour but de projeter les paramètres acoustiques dans un espace indépendant du canal de transmission. Cette technique appelée "feature mapping" [Reynolds03] utilise une

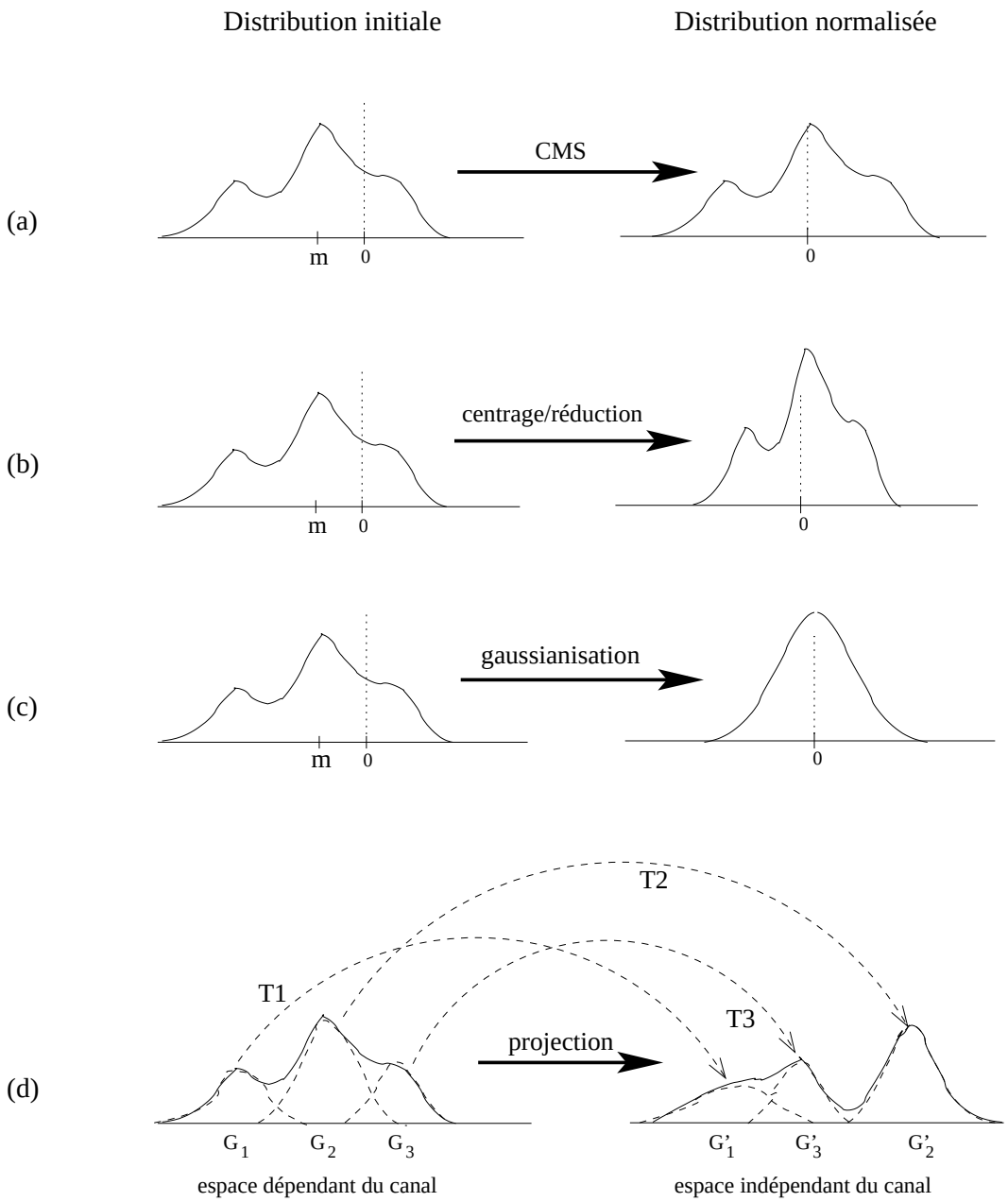


FIG. 3.1 – *Illustration en une dimension des principes des techniques de normalisation des paramètres acoustique- (a) CMS - (b) centrage/réduction - (c) Gaussianisation - (d) projection dans un espace indépendant du canal.*

distribution "cible" indépendante du canal qui est modélisée par un GMM universel entraîné sur une grande quantité de données provenant de canaux différents. Des modèles GMM de chaque canal sont ensuite estimés par adaptation du modèle indépendant à des données spécifiques des canaux. Pour chaque Gaussienne d'un modèle de canal, une transformation linéaire est apprise permettant de passer à la Gaussienne correspondante dans le modèle indépendant du canal. L'étape de normalisation des vecteurs acoustiques consiste tout d'abord à identifier au maximum de vraisemblance le modèle de canal de l'énoncé. Ensuite, pour chaque trame, la Gaussienne la plus vraisemblable est identifiée et la transformation correspondante vers l'espace indépendant du canal est appliquée au vecteur acoustique en cours. Le "feature mapping" n'effectue pas une normalisation complète de la forme de la distribution des vecteurs acoustiques, comme le fait le "feature warping". Elle projette ces vecteurs acoustiques vers un espace commun dans lequel leurs statistiques locales sont supposées indépendantes de la chaîne d'acquisition et de transmission. Cette technique a montré de très bonnes capacités à compenser les variabilités de canal. Elle est cependant tributaire des données disponibles pour construire les différents modèles de canal qui sont nécessaires à l'apprentissage des transformations des vecteurs acoustiques.

Le principe des différentes techniques de normalisation décrites ci-dessus est illustré sur la figure 3.1, pour des vecteurs acoustiques à une dimension.

Ces techniques apportent des améliorations significatives des performances lorsque les conditions d'acquisition présentent une forte variabilité car elles compensent une partie des biais introduits par le canal, le microphone ou les bruits ambiants. De plus elles peuvent être associées en cascade pour améliorer encore la robustesse du système (par exemple un filtrage RASTA, suivi du "feature warping" suivi du "feature mapping"). Cependant, ces techniques de normalisation retirent également des informations spécifiques du locuteur contenus dans les vecteurs acoustiques originaux. Par conséquent, lorsque le signal de parole est peu perturbé, ou que les conditions d'acquisition sont peu variables l'application de ces techniques de normalisation des vecteurs acoustiques peut occasionner une perte de performance.

3.3 Normalisations dans l'espace des modèles

A notre connaissance, peu de techniques de normalisation opérant au niveau des modèles ont été développées pour la VAL. Les techniques de compensation mises en place à ce niveau sont en général des techniques d'adaptation qui visent à minimiser certaines disparités entre les caractéristiques d'un modèle et les conditions d'acquisition des données de test. La normalisation des paramètres d'un modèle se fait en réalité de façon indirecte lorsque l'on utilise une technique de normalisation des paramètres acoustiques (cf. section précédente).

Notons toutefois que le processus d'estimation lui-même peut occasionner certains biais dans les modèles qu'il serait souhaitable de compenser par des techniques de nor-

malisation. Dans [Bonastre et al.04b], les auteurs proposent à cet effet une méthode de déformation des modèles ("model warping"). Le but de cette technique est de compenser les disparités entre la moyenne et la variance globales d'un modèle GMM (estimées par agglomération des composantes Gaussiennes), et la distribution normalisée centrée/réduite des données sur-lesquelles il a été entraîné. Une transformation des paramètres du modèle est apprise afin d'obtenir une moyenne globale nulle et une variance globale unitaire. Cette technique récente n'a pas apporté d'amélioration significative pour l'instant.

Une des raisons pour lesquelles peu de techniques de normalisation de modèles ont été développées vient certainement du fait qu'on ne sait pas bien quelles sont les grandeurs liées au modèle qu'il convient de normaliser et la manière de le faire. En outre, l'application d'une normalisation au niveau des modèles doit garantir que ceux-ci restent cohérents avec les données utilisées, faute de quoi les vraisemblances calculées pourraient perdre toute signification.

Dans le chapitre suivant nous présentons une technique originale de normalisation qui agit spécifiquement au niveau des modèles. Cette normalisation appelée D-MAP a pour but de compenser des biais provenant de disparités dans la qualité des estimations des modèles. Nous utilisons pour cela les mesures de distances de Kullback-Leibler (KL) entre les modèles de locuteurs et le modèle du monde. Dans la méthode proposée, ce sont les distances KL qui sont définies comme grandeurs à normaliser. La normalisation est effectuée grâce à un schéma d'adaptation Bayésienne contrainte. Les résultats obtenus avec D-MAP (Distance constrained MAP) montrent que cette technique permet d'éliminer certains biais liés aux disparités entre modèles, de façon équivalente à une normalisation de score de type Z-norm.

3.4 Normalisations dans l'espace des scores

D'une manière générale, les techniques de normalisation ont pour but de réduire la variabilité des scores. La manière la plus directe de réduire cette variabilité est de normaliser les scores eux-mêmes. Comme nous l'avons déjà mentionné à la section 2.2.3.2, cette normalisation directe au niveau des scores est équivalente, du point de vue de la décision, à la recherche d'un seuil optimal adapté, pour chaque accès, au locuteur proclamé et/ou à l'énoncé de test. Cependant, la normalisation des scores permet d'avoir une procédure plus souple d'optimisation d'un seuil unique.

L'optimisation d'un seuil de décision pour un accès donné (couple identité proclamé/énoncé de test) - ou de façon équivalente, l'estimation des paramètres de normalisation - ne peut être effectuée que si l'on connaît la distribution exacte des scores imposteurs et des scores clients relatifs à cet accès. En pratique, on ne connaît pas ces distributions exactes, mais on peut chercher à les modéliser à l'aide d'accès réels clients et imposteurs. Les bases de données actuelles ou les contraintes applicatives ne permettent pas en général de disposer de suffisamment d'accès clients pour modéliser la distribution des scores clients. Très souvent, seule la distribution des scores imposteurs

peut être modélisée, conduisant à une détermination sous-optimale du seuil de décision, ou des paramètres de normalisation[2]. Néanmoins, les techniques de normalisation de score utilisant ce principe ont montré des gains de performance significatifs.

Les normalisations de scores actuelles reposent sur l'hypothèse que les scores imposteurs sont distribués suivant une Gaussienne dont les paramètres moyenne m_S et écart type σ_S dépendent du locuteur et/ou de l'énoncé de test considéré. Après estimation de ces paramètres à partir d'accès imposteurs réels, la normalisation d'un score brut $S_X(\mathcal{Y})$ s'effectue alors de la manière suivante :

$$S_X^{norm}(\mathcal{Y}) = \frac{S_X(\mathcal{Y}) - m_S}{\sigma_S} \ . \tag{3.1}$$

Cette transformation a pour objectif de rendre la distribution des scores imposteurs centrée et réduite, quel que soit le locuteur et/ou l'énoncé de test. Si tel est le cas, la valeur du seuil de décision détermine alors directement le taux de fausses acceptations du système.

Normalisation Z-norm

La Z-norm (zero normalization) [Li et al.88, Reynolds et al.00] est une normalisation qui estime les paramètres m_S et σ_S vis-à-vis du modèle du locuteur proclamé X. Pour cela, un ensemble important d'énoncés imposteurs est nécessaire. Soit $\{\mathcal{Y}_{I_n}\}_{n=1,\ldots,N}$ cet ensemble d'énoncés imposteurs. Les paramètres m_S et σ_S sont estimés selon :

$$m_S(X) = \frac{1}{N}\sum_{n=1}^{N} S_X(\mathcal{Y}_{I_n}) \ , \tag{3.2}$$

$${\sigma_S}^2(X) = \frac{1}{N}\sum_{n=1}^{N} S_X(\mathcal{Y}_{I_n})^2 - m_S(X)^2 \ . \tag{3.3}$$

La Z-norm a pour but de compenser les biais dépendants des locuteurs engendrés par les différences de qualité entre les modèles estimés. Les paramètres de normalisation peuvent être calculés juste après la phase d'apprentissage, en dehors de la phase de test (normalisation "off-line"). En général, l'effet constaté de la Z-norm est d'améliorer les performances pour les points de fonctionnement à faible taux de faux rejets.

Normalisation T-norm

La T-norm [Auckenthaler et al.00] est une autre normalisation de score pour laquelle les paramètres m_S et σ_S sont estimés vis-à-vis de l'énoncé de test $\mathcal{Y}$. Elle nécessite l'utilisation d'un ensemble de modèles imposteurs $\{p(y|I_n)\}_{n=1,\ldots,N}$ servant à calculer

2. On pourra toutefois consulter l'article [Bimbot et al.00] présentant une étude sur l'optimisation des seuils à partir de données clients

la moyenne et l'écart type des scores imposteurs engendrés par l'énoncé $\mathcal{Y}$. On a alors :

$$m_S(\mathcal{Y}) = \frac{1}{N}\sum_{n=1}^{N} S_{I_n}(\mathcal{Y}) \ , \tag{3.4}$$

$$\sigma_S{}^2(\mathcal{Y}) = \frac{1}{N}\sum_{n=1}^{N} S_{I_n}(\mathcal{Y})^2 - m_S(\mathcal{Y})^2 \ . \tag{3.5}$$

Cette normalisation sert quant à elle à compenser les biais engendrés par les conditions particulières du test. Elle est particulièrement utilisée lors des évaluations NIST car elle a pour effet d'améliorer significativement les performances pour les points de fonctionnement à faible taux de fausse acceptation, qui est la zone privilégiée par la fonction de coût définie par NIST. On peut trouver une étude approfondie de l'influence qu'a cette normalisation sur la distribution des scores dans [Navratil et al.03]. Un défaut majeur de la T-norm est que les paramètres de normalisation doivent être calculés au moment de l'accès au système (normalisation "on-line"), ce qui rallonge significativement les temps de test.

Variantes et associations de normalisations de scores

Il existe également des variantes de la Z-norm et de la T-norm qui cherchent à compenser les biais liés spécifiquement aux microphones ou aux canaux de transmission. Pour cela, un couple de paramètre (m_S, σ_S) est estimé pour chaque type de microphone ou de canal. Au moment du test, le type de microphone ou de canal est détecté et les paramètres de normalisation correspondant sont utilisés. Parmi les variantes de la Z-norm on trouve la H-norm (handset normalization) [Reynolds et al.00] et la C-norm (channel normalization). La HT-norm [Auckenthaler et al.00] est quant à elle une version de la T-norm dépendante du type de microphone.

Enfin, différents types de normalisation de score peuvent être combinés pour compenser à la fois les biais liés au locuteur et à l'énoncé de test (ZT-norm, TZ-norm par exemple).

Les techniques de normalisation de score présentées dans cette section ont montré leur efficacité, notamment au cours des évaluations NIST où elles ont permis des gains de performance significatifs. Cependant, elles sont toutes tributaires d'un ensemble de données externes servant à l'estimation des paramètres de normalisation. Si ces données externes ne sont pas disponibles, aucune de ces normalisations ne peut s'appliquer.

Dans le chapitre 5 de cette thèse nous présentons le développement d'une nouvelle technique de normalisation de score, la D-norm (section 5.1). Contrairement aux normalisations de scores présentées ci-dessus, la D-norm estime les paramètres de normalisation sans nécessiter de données externes. Elles s'appuie pour cela sur des mesures directement estimées sur les modèles eux-mêmes, les distances de Kullback-Leibler. Cette technique s'inscrit dans le cadre des normalisations de score visant à compenser les biais provenant des différences de qualité dans l'estimation des modèles, au même titre que la Z-norm.

3.5 Discussion sur les techniques de normalisation

Toutes les techniques de normalisation présentées dans ce chapitre ont pour but de minimiser les effets venant des disparités entre conditions de test et d'apprentissage. L'atténuation de ces effets permet de réduire la variabilité des scores, ce qui autorise l'utilisation d'un seuil unique de décision quel que soit le locuteur et l'énoncé de test considéré. Pour compenser ces biais dans les scores, les techniques de normalisation agissent dans différents espaces : espace acoustique, espace des modèles ou espace des scores. Des techniques appliquées dans des espaces différents ont montré des performances similaires (par exemple le "feature mapping" et la H-norm [Reynolds03]), indiquant ainsi que des équivalences existent entre les normalisations.

En réalité, toute l'information sur les conditions d'acquisition est contenue dans les vecteurs acoustiques. Les normalisations intervenant dans l'espace acoustique agissent donc sur la source même des biais apparaissant dans les scores, de façon indépendante des locuteurs et de la technique de modélisation. Cependant il est souvent plus pratique d'agir dans l'espace des scores. En effet, l'espace acoustique est de grande dimension et engendre donc un nombre important de paramètres de normalisation qu'il faut être capable d'estimer. L'espace des scores est quant à lui de dimension 1 et n'engendre que peu de paramètres de normalisation.

Les différentes techniques de normalisation ne sont pas équivalentes d'un point de vue applicatif. Certaines d'entre elles nécessitent l'utilisation de données externes spécifiques pour le calcul des paramètres de normalisation. C'est le cas notamment des normalisations de scores qui ont besoin d'énoncés imposteurs (Z-norm) ou de modèles imposteurs (T-norm). Les normalisations de score H-norm et C-norm ainsi que la technique "feature mapping" agissant dans l'espace acoustique nécessitent quant à elles des données étiquetées suivant le microphone ou le canal. Ces contraintes sur la disponibilité de données externes spécifiquement dédiées à ce type de normalisation peut être un frein à leur application dans certaines situations réelles. Les techniques de normalisation CMS, filtrage RASTA, centrage/réduction et "feature warping" agissent quant à elles directement sur les données de test, sans données extérieures.

Une autre distinction entre les techniques de normalisation vient du fait que certaines d'entre elles nécessitent de calculer des paramètres de normalisation au moment du test même (normalisations "on line"). C'est le cas de la CMS, du filtrage RASTA, du centrage/réduction, du "feature warping" et de la T-norm. Du point de vue applicatif, cela peut rallonger significativement le temps de test pour certaines d'entres elles, lorsque le calcul des paramètres de normalisation est coûteux. Les techniques "feature mapping" et toutes les variantes de Z-norm (H-norm, C-norm) estiment quant à elles les paramètres de normalisation avant la phase de test (normalisations "off line") ce qui permet de ne rallonger le temps de test que de façon négligeable.

Dans ce manuscrit, nous proposons plusieurs méthodes de normalisation qui ne nécessitent pas de données externes spécifiques. Elles sont donc plus facilement applicables. De plus, aucune d'entre elles ne rallonge le temps de test de façon significative. Ces techniques de normalisation pour la VAL sont présentées au chapitre 5.

Chapitre 4

Divergences KL et scores de vérification en VAL

Une des limitations les plus contraignantes des normalisations de score actuelles vient du fait qu'elles sont tributaires d'un ensemble de données externes servant à l'estimation des paramètres de normalisation. Cependant, ces données externes servent seulement à mesurer de façon indirecte des caractéristiques intrinsèques du modèle ou des données de test. Une mesure directe de ces caractéristiques pourrait permettre d'obtenir des informations pertinentes sur la distribution des scores, sans avoir recours à des accès réels. On pourrait alors envisager de mettre en oeuvre des techniques de normalisation ne nécessitant pas de données externes.

Les divergences de Kullback-Leibler (KL) sont des mesures liées aux modèles probabilistes couramment utilisées comme informations de similarité entre les modèles. Elles pourraient fournir une information sur le comportement des modèles de locuteurs. En outre, ces divergences sont intimement liées au rapport de vraisemblance, de par leur définition mathématique.

Dans ce chapitre, nous définissons des divergences et distances KL dans le cadre de la VAL et étudions leurs liens théoriques avec les scores de vérification. Nous menons ensuite une série d'expériences faisant apparaître des corrélations fortes entre les divergences et distances KL et la moyenne des scores imposteurs. Pour cette étude expérimentale, les divergences KL sont estimées par une méthode de Monte Carlo. Les relations mises en évidence nous ont permis de mettre en oeuvre de nouvelles techniques de normalisation qui sont présentées au chapitre suivant.

4.1 Introduction : intérêt d'une mesure liée au modèle

Les nombreuses sources de variabilité et de perturbation rencontrées dans les applications de VAL réelles se répercutent au niveau des scores par des biais qui sont spécifiques à chaque accès. Ces biais dépendent d'une part de la configuration du modèle de l'identité proclamée[1] et d'autre part des conditions d'acquisition (bruit ambiant, type de microphone) et de transmission (canal) des données de test ainsi que de leur contenu linguistique. Les variations de l'ensemble de ces conditions d'utilisation font que l'on observe une grande variabilité dans les valeurs des scores, cause de problèmes de robustesse des systèmes de VAL.

Les techniques de normalisation employées dans les systèmes de VAL (cf. chapitre précédent) sont destinées à renforcer cette robustesse. Parmi ces techniques, les normalisations de scores courantes analysent directement les biais engendrés au niveau des scores par tel modèle ou tel énoncé de test. Les paramètres de normalisation sont estimés à partir d'un ensemble d'accès réels. Ces normalisations de scores nécessitent par conséquent une base de données externe qu'il n'est pas toujours facile de se procurer.

On peut cependant raisonnablement penser que les valeurs des paramètres de normalisation ne sont en réalité que le reflet de caractéristiques intrinsèques des éléments intervenant dans le calcul d'un score LLR, à savoir : le modèle du locuteur proclamé, le modèle du non-locuteur et l'énoncé de test. Des mesures directement effectuées au niveau de ces éléments pourraient donc fournir des informations sur la distribution des scores qu'ils sont susceptibles d'engendrer. On sait notamment que des biais dépendant de la qualité d'estimation des modèles de locuteurs existent dans les scores. Des modèles estimés de façon hétérogène entraînent alors des valeurs de scores également hétérogènes. C'est ce type de biais que cherche à compenser la normalisation Z-norm. Une mesure directement liée aux modèles de locuteurs pourrait également permettre de corriger ces biais.

Nous proposons d'utiliser les divergences et distances KL entre les modèles de locuteurs et le modèle du monde, comme mesure d'information de la "qualité" d'estimation des modèles. Pour cela, nous étudions l'influence des divergences et distances KL sur le comportement du système, notamment sur la distribution des scores de vérification. La mise en évidence de liens entre les divergences ou distances KL et les scores laisserait envisager des possibilités de techniques de normalisation ne nécessitant pas de données externes.

La définition des divergences et distances KL que nous étudions dans le cadre de la VAL est donnée à la section 4.2. Une étude théorique de ces grandeurs et de leurs liens avec les scores de vérification est effectuée à la section 4.3 et leur étude expérimentale est menée à la section 4.4. Nous concluons ensuite sur les possibilités d'exploitation des divergences et distances KL pour la RAL.

1. Cette configuration du modèle reflète les conditions d'acquisition/transmission et le contenu linguistique des données d'apprentissage

4.2 Définition des divergences KL pour la VAL

En théorie de l'information, les divergences de Kullback-Leibler définissent une mesure d'entropie relative entre deux densités de probabilité (d.d.p). La divergence KL entre la d.d.p. $p_1(y)$ et la d.d.p $p_2(y)$ s'écrit

$$KL(p_1(y)||p_2(y)) = E_{p_1}[\log \frac{p_1(y)}{p_2(y)}] = \int p_1(y) \log \frac{p_1(y)}{p_2(y)} dy \ , \tag{4.1}$$

où $E_{p_1}[\cdot]$ est l'espérance mathématique calculée selon la d.d.p $p_1(y)$. De façon générale, ces divergences peuvent fournir une mesure de similarité entre deux modèles probabilistes. Ce sont des mesures non-symétriques, c'est-à-dire qu'elles dépendent de celui des deux modèles qui est considéré pour le calcul de l'espérance (dans l'équation 4.1 c'est $p_1(y)$ qui est utilisé pour ce calcul).

Dans un contexte de VAL, nous définissons deux divergences KL entre un modèle estimé de locuteur $p(y|X)$ et le modèle estimé du non-locuteur $p(y|\bar{X})$:

$$KL_X = E_X[\log \frac{p(y|X)}{p(y|\bar{X})}] \ , \tag{4.2}$$

$$KL_{\bar{X}} = E_{\bar{X}}[\log \frac{p(y|\bar{X})}{p(y|X)}] \ . \tag{4.3}$$

$E_X[\cdot]$ et $E_{\bar{X}}[\cdot]$ désignent les espérances mathématiques calculées selon les lois $p(y|X)$ et $p(y|\bar{X})$ respectivement. Nous définissons également de façon classique une mesure symétrisée entre les deux modèles, appelée distance de Kullback-Leibler, et qui sera notée $KL2_X$. Elle est définie comme la somme des deux divergences KL duales :

$$KL2_X = KL_X + KL_{\bar{X}} \ . \tag{4.4}$$

Les grandeurs KL_X, $KL_{\bar{X}}$ et $KL2_X$ sont positives et ne s'annulent que si et seulement si les modèles $p(y|X)$ et $p(y|\bar{X})$ sont strictement identiques.

Dans le cadre d'un système de VAL de type UBM-GMM, le modèle du non-locuteur $p(y|\bar{X})$ est commun à tous les locuteurs X et représenté par l'UBM. Les divergences et distances KL telles que définies ci-dessus, fournissent alors un ensemble de mesures de similarité entre chacun des modèles locuteurs et un modèle commun servant de référence : l'UBM.

4.3 Liens théoriques avec les scores de vérification

Dans cette section nous montrons que les rapports de vraisemblances utilisés comme scores de vérification en VAL sont asymptotiquement équivalents (c'est-à-dire en moyenne probabiliste) à une différence entre deux divergences KL. Ces deux divergences sont définies, pour la première entre le modèle "vrai" des données de test et le modèle estimé du non-locuteur, et pour la seconde, entre le modèle "vrai" des données de test et le modèle estimé du locuteur. Cette relation laisse supposer que les divergences et

distances KL sont des mesures liées aux modèles qui sont susceptibles de fournir une information sur la distribution des scores.

Soit $\mathcal{Y} = \{y_1, \ldots, y_N\}$ un énoncé de test composé de N observations, et X le locuteur correspondant à l'identité proclamée. Le score de vérification $S_X(\mathcal{Y})$ obtenu est la moyenne des log-rapports de vraisemblance pour chaque trame :

$$S_X(\mathcal{Y}) = \frac{1}{N} \sum_{n=1}^{N} log \frac{p(y_n|X)}{p(y_n|\bar{X})} \ . \tag{4.5}$$

La classe dont est issu l'énoncé de test $\mathcal{Y}$ sera notée Y et $\mathcal{M}_Y(y)$ désignera le modèle "vrai" (et inconnu en pratique) de cette classe. La moyenne probabiliste du score $S_X(\mathcal{Y})$ par rapport aux énoncés de test est alors l'espérance mathématique du log-rapport de vraisemblance des classes X et $\bar{X}$ calculée selon $\mathcal{M}_Y(y)$:

$$E[S_X(\mathcal{Y})] = E_{\mathcal{M}_Y}[log \frac{p(y|X)}{p(y|\bar{X})}] \ . \tag{4.6}$$

Cette expression peut également s'écrire comme la différence de deux divergence KL :

$$\begin{aligned} E[S_X(\mathcal{Y})] &= E_{\mathcal{M}_Y}[log \frac{\mathcal{M}_Y(y)}{p(y|\bar{X})}] - E_{\mathcal{M}_Y}[log \frac{\mathcal{M}_Y(y)}{p(y|X)}] \ , & (4.7) \\ &= KL(\mathcal{M}_Y(y)||p(y|\bar{X})) - KL(\mathcal{M}_Y(y)||p(y|X)) \ . & (4.8) \end{aligned}$$

Nous allons maintenant interpréter cette relation dans le cas d'accès clients d'une part et d'accès imposteurs d'autre part.

Divergences KL et scores "clients"

Dans le cas d'un accès client, l'énoncé de test $\mathcal{Y}$ est issu de la classe du locuteur X. En désignant par $\mathcal{M}_X(y)$ la densité de probabilité exacte de cette classe, la moyenne probabiliste des scores clients $S_X(\mathcal{Y}|X)$ par rapport aux énoncés de test peut s'écrire sous la forme :

$$E[S_X(\mathcal{Y}|X)] = KL(\mathcal{M}_X(y)||p(y|\bar{X})) - KL(\mathcal{M}_X(y)||p(y|X)) \ . \tag{4.9}$$

Cette expression fait apparaître un premier terme qui correspond à la divergence KL entre le modèle vrai du locuteur et le modèle estimé du non-locuteur. Il exprime "l'éloignement" du modèle locuteur X par rapport au modèle non-locuteur (estimé) : c'est le terme discriminant du score client. Il est indépendant de l'estimation du modèle locuteur, notamment de la qualité de cette estimation (mais il est dépendant de l'estimation du modèle non-locuteur).

Le second terme, qui est soustrait à la première divergence, correspond à la divergence KL entre le modèle "vrai" du locuteur et son modèle estimé. Ce terme rend compte de la qualité de l'estimation du modèle locuteur et est indépendant de la

classe non-locuteur. Il ajoute au terme discriminant un biais négatif qui correspond à "l'éloignement" du modèle estimé du locuteur par rapport à son modèle exact, repoussant ainsi la distribution des scores clients vers les valeurs négatives. C'est par exemple ce qui est observé en pratique lorsque l'estimation du modèle d'un locuteur est sur-adapté aux données d'apprentissage. Le modèle estimé est alors trop spécialisé et ne se généralise pas bien aux autres accès du locuteur, entraînant des valeurs de scores clients faiblement positives voir même négatives.

Notons que dans le cas théorique d'une estimation parfaite du modèle locuteur, ce deuxième terme s'annule et la moyenne des scores clients est donnée par la valeur de la divergence KL_X définie en 4.2 :

$$p(y|X) = \mathcal{M}_X(y) \longrightarrow KL_X = E[S_X(\mathcal{Y}|X)] \ . \tag{4.10}$$

Divergence KL et scores "imposteurs"

Le cas d'un accès "imposteur" peut être étudié de façon similaire à celle d'un accès client. En notant $\mathcal{M}_{\bar{X}}(y)$ la densité de probabilité exacte de la classe non-locuteur, la moyenne probabiliste des scores "imposteurs" $S_X(\mathcal{Y}|X)$ fait également apparaître une différence de deux termes :

$$E[S_X(\mathcal{Y}|\bar{X})] = KL(\mathcal{M}_{\bar{X}}(y)||p(y|\bar{X})) - KL(\mathcal{M}_{\bar{X}}(y)||p(y|X)) \ . \tag{4.11}$$

Le premier terme est la divergence KL entre le modèle "vrai" et le modèle estimé du non-locuteur. Il rend compte de la qualité de l'estimation du modèle non-locuteur et ajoute un biais positif pénalisant aux scores imposteurs. Le second terme est la divergence KL entre le modèle vrai du non-locuteur et le modèle estimé du locuteur. C'est le terme discriminant des scores imposteurs. Dans le cas idéal d'une estimation parfaite du modèle non-locuteur, l'opposé de la divergence $KL_{\bar{X}}$ de l'équation 4.3 est égal à la moyenne des scores "imposteurs" :

$$p(y|\bar{X}) = \mathcal{M}_{\bar{X}}(y) \longrightarrow -KL_{\bar{X}} = E[S_X(\mathcal{Y}|\bar{X})] \ . \tag{4.12}$$

En résumé, les divergences KL définies à la section 4.2 sont directement liées aux scores de vérification dans le cas idéal (théorique) d'une estimation parfaite des modèles. Sous cette hypothèse, la valeur moyenne des scores clients est donnée par la divergence KL_X, la valeur moyenne des scores imposteurs est donnée par la divergence $KL_{\bar{X}}$. Notons également que l'écart entre ces deux moyennes est dans ce cas donné par la distance $KL2_X$:

$$p(y|X) = \mathcal{M}_X(y) \text{ et } p(y|\bar{X}) = \mathcal{M}_{\bar{X}}(y) \longrightarrow E[S_X(\mathcal{Y}|\mathcal{X})] - E[S_X(\mathcal{Y}|\bar{X})] = KL2_X \ . \tag{4.13}$$

En réalité cependant, l'estimation imparfaite des modèles entraîne une déviation des moyennes des scores par rapport à ces valeurs théoriques.

4.4 Étude expérimentale des divergences KL

Après l'étude théorique des divergences KL en VAL menée aux sections 4.2 et 4.3, nous étudions ces grandeurs de façon expérimentale. Nous décrivons dans un premier temps la méthode d'estimation des divergences que nous avons utilisée car celles-ci ne sont pas calculables directement en pratique dans le cas d'une modélisation par GMM. Nous mettons ensuite en évidence de façon expérimentale, l'existence de corrélations entre ces divergences et les scores obtenus lors d'une évaluation réelle.

4.4.1 Estimation des divergences KL par tirage de Monte Carlo

Dans le cas d'une modélisation des locuteurs par mélange de Gaussiennes, le calcul direct des divergences KL 4.2 et 4.3 est irréalisable en pratique. En effet, celles-ci ne peuvent pas être exprimées sous une forme simple comme dans le cas de modèles Gaussiens par exemple. En outre, le caractère multidimensionnel des données et la complexité des modèles rendent très compliqué, voire impossible, le calcul de l'intégrale intervenant dans l'expression des divergences KL (cf équation 4.1). Nous avons donc opté pour l'utilisation d'une méthode de type "Monte-Carlo" [Roberts et al.99] afin d'obtenir des estimations de ces divergences. Le principe est simple: des données "artificielles" sont générées à partir du modèle considéré et l'espérance mathématique est approximée par une moyenne sur un grand nombre d'échantillons. Ce principe utilise la propriété de convergence de la moyenne statistique vers l'espérance mathématique. Pour toute fonction $\mathcal{F}(y)$ d'une variable aléatoire y de d.d.p. $p(y)$ on a:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{F}(y_n) = \int p(y)\mathcal{F}(y)dy = E_{p(y)}[\mathcal{F}(y)] \ , \tag{4.14}$$

où les échantillons $\{y_n\}$ suivent la loi $p(y)$. Les formules d'estimation des divergences KL_X et $KL_{\bar{X}}$ et de la distance $KL2_X$ sont donc:

$$\widehat{KL}_X = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n|X)}{p(y_n|\bar{X})} \quad \text{avec} \quad y_n \sim p(\cdot|X) \ , \tag{4.15}$$

$$\widehat{KL}_{\bar{X}} = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n|\bar{X})}{p(y_n|X)}, \quad \text{avec} \quad y_n \sim p(\cdot|\bar{X}) \ , \tag{4.16}$$

$$\widehat{KL2} = \widehat{KL}_X + \widehat{KL}_{\bar{X}} \ , \tag{4.17}$$

où $y_n \sim p(\cdot|X)$ signifie que les échantillons y_n sont générés suivant le modèle du locuteur X et $y_n \sim p(\cdot|\bar{X})$ qu'ils sont générés suivant le modèle du non-locuteur. Ces estimateurs sont sans biais et l'erreur quadratique d'estimation décroît en $1/N$. Pour obtenir les estimations 4.15, 4.16 et 4.17, il faut générer des données artificielles suivant un modèle de mélange de Gaussiennes. Le tirage de chaque échantillon y_n issu du GMM est effectué en deux étapes:

1. tirage aléatoire d'une des composantes parmi les Gaussiennes du mélange, en fonction de leur poids respectif.

2. génération d'une donnée Gaussienne suivant la moyenne et la matrice de covariance de la composante tirée au sort à l'étape 1.

Ces deux étapes sont répétées autant de fois que l'on désire d'échantillons artificiels. Le détail de l'algorithme implémenté pour le tirage de Monte-Carlo est donné en annexe C.

La conséquence de cette estimation par tirage de Monte Carlo est que les valeurs obtenues pour les divergences KL d'un GMM donné varient d'un tirage à l'autre, du fait de l'aspect aléatoire de la méthode. Ces variations seront d'autant plus faibles que le nombre d'échantillons utilisés pour chaque estimation sera grand. Il y a là un compromis à trouver entre la précision des estimations et le temps nécessaire au tirage de l'ensemble de données $\{y_n\}$. Ce temps de calcul est proportionnel au nombre N d'échantillons générés et à la dimension des données. Il est également lié à la complexité du GMM, c'est-à-dire au nombre K de composantes Gaussiennes.

4.4.2 Corrélation avec les scores de vérification

L'étude théorique menée à la section 4.3 indique qu'il existe un lien entre les divergences KL étudiées et les scores de vérification. Pour mettre en évidence ce lien nous effectuons une évaluation complète et comparons les scores obtenus aux divergences correspondantes.

4.4.2.1 Protocole expérimental

La base de données utilisée pour les expériences présentées ici est un sous-ensemble de la base NIST'01 [Nist] (ensemble des locuteurs femmes, 506 individus). Cette base contient des conversations téléphoniques en anglais américain avec des microphones de type électret. La durée des énoncés d'apprentissage est d'environ 2 minutes et les énoncés de test durent de 15 à 45 secondes.

Le système de VAL utilisé est le système de base IRISA/ELISA 2001 [MC et al.01]. L'analyse acoustique de ce système fournit des vecteurs de dimension 32 composés des 16 premiers coefficients cepstraux et de leurs 16 coefficients Δ respectifs. Ces vecteurs acoustiques sont normalisés par CMS. Les modèles de locuteurs sont des GMM à 128 composantes avec matrices de covariances diagonales. Ils sont estimés par adaptation de modèles du monde dépendant du genre qui ont été entraînés sur les données de l'évaluation NIST'99. Seuls les paramètres moyennes sont adaptés, les autres paramètres (poids, covariances) gardant leur valeur *a priori*.

Pour chaque locuteur X, la moyenne des scores clients m_{S_X} d'une part et celle des scores imposteurs $m_{S_{\bar{X}}}$ d'autre part sont calculées en utilisant tous les accès clients et imposteurs correspondant à ce locuteur dans la base de données NIST'01. Ces moyennes

sont calculées selon les équations suivantes :

$$m_{S_X} = \frac{1}{N_X} \sum_{i=1}^{N_X} S(\mathcal{Y}_i|X) \ , \tag{4.18}$$

$$m_{S_{\bar{X}}} = \frac{1}{N_{\bar{X}}} \sum_{i=1}^{N_{\bar{X}}} S(\mathcal{Y}_i|\bar{X}) \ , \tag{4.19}$$

$$\tag{4.20}$$

où N_X est le nombre d'accès clients et $N_{\bar{X}}$ est le nombre d'accès imposteurs.

Ces moyennes sont des estimations des espérances mathématiques $E[S(\mathcal{Y}|X)]$ et $E[S(\mathcal{Y}|\bar{X})]$ respectivement. La base de données NIST'01 contient en moyenne 50 accès imposteurs et 5 accès clients par locuteur. Par conséquent, la moyenne m_{S_X} sera en général une mauvaise estimation de l'espérance mathématique $E[S(\mathcal{Y}|X)]$ des scores clients, alors que $m_{S_{\bar{X}}}$ sera une estimation plus raisonnable de l'espérance mathématique $E[S(\mathcal{Y}|\bar{X})]$ des scores imposteurs. Les divergences KL_X, $KL_{\bar{X}}$ et la distance $KL2_X$ sont également estimées pour chacun des locuteurs X de l'évaluation, en utilisant la méthode de Monte Carlo décrite à la section 4.4.1 avec 1000 échantillons synthétiques générés pour chaque estimations.

4.4.2.2 Observation des corrélations

Pour mettre en évidence le lien existant entre les divergences KL_X et $KL_{\bar{X}}$ et les moyennes des scores m_{S_X} et $m_{S_{\bar{X}}}$, les nuages de points (KL_X, m_{S_X}) ('+' bleus) et $(KL_{\bar{X}}, m_{S_{\bar{X}}})$ ('•' rouges) ont été représentés sur la figure 4.1. Cette figure fait clairement apparaître une corrélation entre la moyenne des scores imposteurs $m_{S_{\bar{X}}}$ et la divergence $KL_{\bar{X}}$. Le coefficient de corrélation correspondant est de -0,8. Par contre, aucune corrélation évidente n'apparaît concernant les accès clients entre m_{S_X} et la divergence KL_X (coefficient de corrélation de 0,03). Ce résultat est cependant peu significatif en raison du faible nombre d'accès clients disponibles pour l'estimation de m_{S_X}.

Les droites en pointillés sur la figure représentent les droites de régression des deux nuages de points. Nous avons forcé ces droites à passer par le point origine $(0, 0)$ qui est un point théorique de passage. En effet, pour des valeurs nulles des divergences KL_X et $KL_{\bar{X}}$, les modèles du locuteur et du non-locuteur sont strictement identiques. Par conséquent les scores obtenus sont nécessairement nuls eux aussi.

En théorie et dans le cas hypothétique d'une estimation parfaite des modèles, les pentes de ces droites devraient être de -1 pour les accès imposteurs et de $+1$ pour les accès clients (cf section 4.3). Sur l'exemple réel de la figure 4.1, la pente de la droite concernant les accès imposteurs est de $-1,5$ et celle concernant les accès clients est de $+8,1$. Cela confirme la déviation des deux types de scores mise en évidence à la section 4.3 par rapport à leurs distributions théoriques "idéales" : les scores imposteurs sont légèrement repoussés vers les valeurs positives et les scores clients sont fortement repoussés vers les valeurs négatives (les résultats concernant les accès clients restent

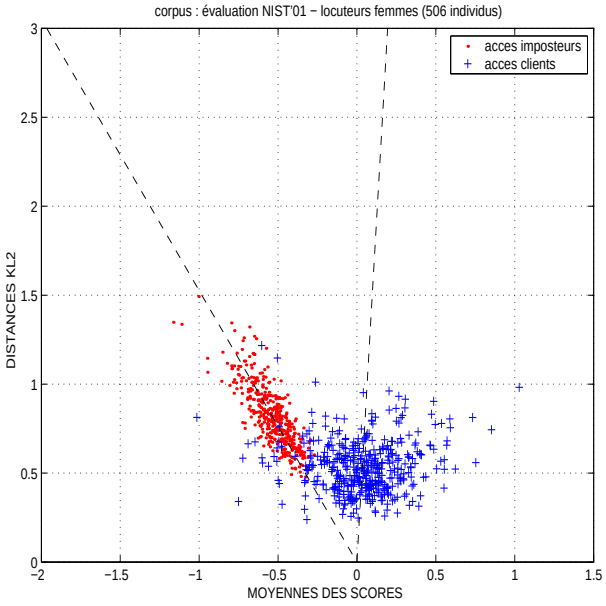


FIG. 4.1 – *Corrélations entre divergences KL et moyennes des scores. Distributions des points* (KL_X, m_{S_X}) *et* $(KL_{\bar{X}}, m_{S_{\bar{X}}})$.

cependant peu significatifs pour les raisons évoquées ci-dessus). Dans le cas des accès imposteurs où une forte corrélation est observée, la déviation des scores vers les valeurs positives semble être linéairement proportionnelle à la valeur de la divergence $KL_{\bar{X}}$. Cela permet de conserver une relation quasi-linéaire entre $m_{S_{\bar{X}}}$ et $KL_{\bar{X}}$:

$$KL_{\bar{X}} \approx -a.m_{S_{\bar{X}}} \quad \text{avec } a = 1,5 \ . \tag{4.21}$$

Ainsi, à un facteur de proportionnalité près, la divergence $KL_{\bar{X}}$ permet de prédire approximativement la valeur moyenne des scores imposteurs.

Sur la figure 4.2, le même type de diagramme a été tracé en utilisant la distance $KL2_X$ à la place des divergences $KL_{\bar{X}}$ et KL_X.

Cette figure fait également apparaître une forte corrélation entre $KL2_X$ et $m_{S_{\bar{X}}}$. Le coefficient de corrélation entre ces deux grandeurs est de $-0,9$, soit légèrement supérieur en valeur absolue à celui qui relie $KL_{\bar{X}}$ et $m_{S_{\bar{X}}}$. L'utilisation de la divergence KL_X en complément de $KL_{\bar{X}}$ dans la distance $KL2$ semble donc renforcer la dépendance avec $m_{S_{\bar{X}}}$. La moyenne des scores imposteurs $m_{S_{\bar{X}}}$ et la distance $KL2_X$ sont également liées par une relation quasi-linéaire du type :

$$KL2_X \approx -a.m_{S_{\bar{X}}} \quad \text{avec } a = 2,5 \ . \tag{4.22}$$

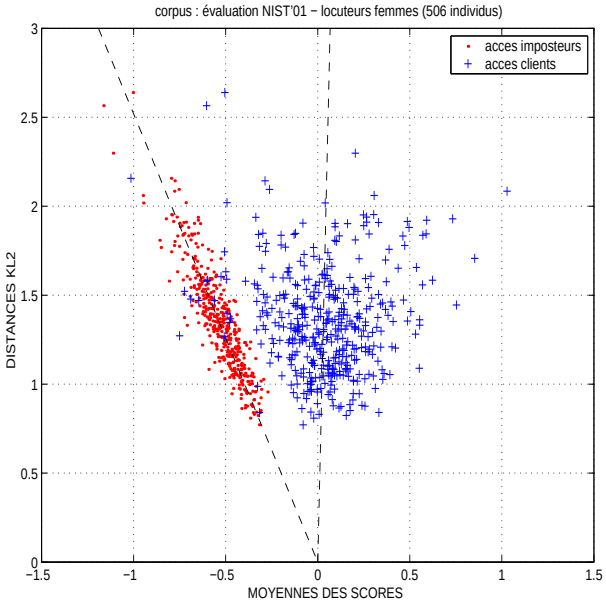


FIG. 4.2 – *Corrélations entre distances KL et moyennes des scores. Distributions des points* $(KL2_X, m_{S_X})$ *et* $(KL2_X, m_{S_{\bar{X}}})$.

De la même manière que la divergence $KL_{\bar{X}}$, la distance $KL2_X$ permet de prédire approximativement la valeur moyenne des scores imposteurs, avec une précision légèrement meilleure (corrélation légèrement plus forte).

Pour conclure cette analyse expérimentale des liens existant entre les divergences et distances KL et les scores de vérification, nous avons tracé sur la figure 4.3 le nuage de points représentant l'écart type $\sigma_{S_{\bar{X}}}$ des scores imposteurs en fonction de la distance $KL2_X$. L'écart type $\sigma_{S_{\bar{X}}}$ a été estimé pour chaque locuteur en utilisant tous leurs accès imposteurs respectifs disponibles dans l'évaluation NIST01.

Il semble d'après cette figure que l'écart type des scores imposteurs $\sigma_{S_{\bar{X}}}$ est globalement croissant lorsque la distance $KL2_X$ augmente. Cependant, en raison de la forte dispersion des points, la relation qui relie ces deux grandeurs est loin d'apparaître clairement. Malgré tout, la distance $KL2_X$ semble également pouvoir fournir une indication sur la dispersion des scores imposteurs, en plus de la relation quasi-linéaire qui la relie à la moyenne de ces scores.

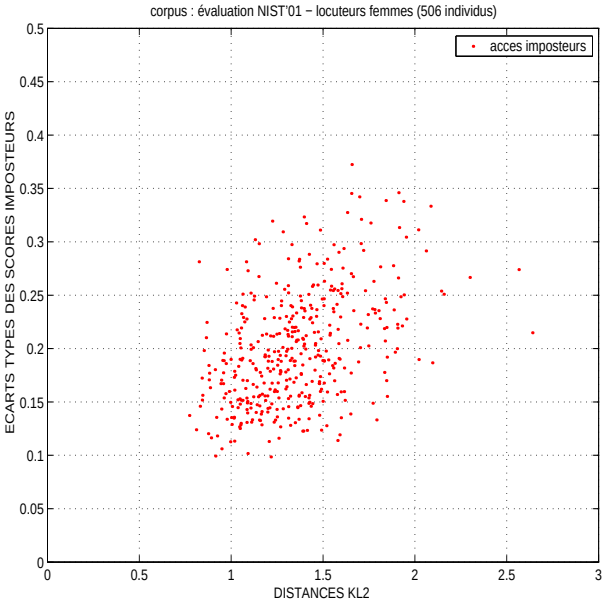


FIG. 4.3 – *Relation entre écarts types des scores et distances KL. Distributions des points* $(KL2_X, \sigma_{S_{\bar{X}}})$.

4.5 Conclusion : possibilité d'exploitation des distances KL en RAL

D'une manière générale, une mesure de similarité entre modèles probabilistes trouve de multiples applications dans les problèmes de classification. Pour les tâches de RAL, l'état de l'art des systèmes utilisant l'approche probabiliste est basé sur une modélisation des locuteurs par GMM. La méthode d'estimation des divergences KL entre GMM proposée à la section 4.4.1 permet d'obtenir un ensemble de mesures de similarité entre modèles de locuteurs. Les possibilités d'exploitation de ces mesures sont multiples pour toutes les tâches de RAL utilisant des méthodes de classifications basées sur des similarités entre modèles, par exemple la segmentation et le regroupement en locuteur.

En VAL plus spécifiquement, nous avons mis en évidence l'existence d'une forte corrélation entre les distances KL et la moyenne des scores imposteurs. Cela signifie que ces distances, estimées uniquement à partir des modèles des locuteurs, fournissent une information très pertinente sur la distribution des scores imposteurs. Sans même devoir effectuer d'évaluation avec des tests réels, elles permettent de prédire quelle sera approximativement la distribution des scores imposteurs obtenus pour chacun des locuteurs.

Dans le chapitre suivant, nous utilisons ces relations entre distances KL et scores

imposteurs dans de nouvelles techniques de normalisation pour la VAL.

Le chapitre 6 est quant à lui consacré à la description de contributions connexes dans d'autres tâches de reconnaissance et de caractérisation du locuteur.

Chapitre 5

Applications pour la VAL : normalisations de scores et de modèles

Dans ce chapitre nous présentons des techniques de normalisation pour la VAL exploitant les relations entre distances KL et scores imposteurs mises en évidence au chapitre précédent.

La première technique présentée à la section 5.1 est une méthode de normalisation de scores basée sur les distances KL. Elle se distingue des techniques courantes de normalisation de scores par le fait qu'elle ne nécessite aucune donnée supplémentaire pour estimer les paramètres de normalisation.

La seconde technique exposée à la section 5.2 réalise une normalisation des modèles de locuteur. Elle a pour but d'homogénéiser les modèles vis-à-vis de leur distance KL par rapport au modèle du monde et s'appuie sur un schéma d'adaptation contrainte.

Enfin, une distance Euclidienne dérivée des distances KL entre modèles, nous a permis de mettre en oeuvre un nouveau cadre pour le calcul des scores en VAL. Ce cadre permet notamment un calcul très efficace des paramètres de normalisation T-norm. Nous le présentons à la section 5.3.

5.1 D-norm : une technique de normalisation des scores par distance KL

5.1.1 Rappel sur les normalisations de scores

Les techniques de normalisation de scores développées pour la VAL et présentées à la section 3.4 visent à homogénéiser la distribution des scores imposteurs vis-à-vis des différentes variabilités rencontrées. Le but est de rendre plus robuste le processus de décision à partir d'un seuil unique.

Un premier type de variabilité des scores provient par exemple du comportement hétérogène des modèles eux-mêmes. C'est ce que cherche à compenser la Z-norm. La T-norm tente d'atténuer les variabilités liées à l'hétérogénéité des énoncés de test. Enfin la H-norm et la C-norm compensent des biais provenant de disparités entre conditions d'apprentissage et de test (disparité de microphone pour la H-norm et de canal pour la C-norm).

Les normalisations de scores état de l'art utilisent des statistiques calculées sur un ensemble d'accès imposteurs réels afin d'estimer leurs paramètres de normalisation.Elles nécessitent l'utilisation d'une base de données externe de laquelle sont issus des énoncés de tests imposteurs (pour la Z-norm) ou des modèles d'imposteurs (pour la T-norm). Cette contrainte empêche l'application de telles normalisations si une base de données externe suffisamment conséquente n'est pas disponible.

Nous étudions dans cette section une possibilité de normalisation de scores sans donnée additionnelle, en utilisant les divergences et distances KL entre les modèles des locuteurs et le modèle du monde.

5.1.2 Utilisation des distances KL pour la normalisation de score

L'étude théorique et expérimentale des distances KL développée au chapitre précédent montre que ces mesures peuvent fourni une information *a priori* très pertinente sur la distribution des scores imposteurs engendrée par chaque modèle de locuteur. Nous utilisons cette relation pour implémenter une technique de normalisation des scores, la D-norm ("Distance normalization").

Modélisation des scores imposteurs

La méthode de normalisation proposée fait l'hypothèse classique d'une distribution Gaussienne des scores imposteurs. Contrairement aux techniques courantes, les paramètres moyenne et variance de cette distribution ne sont pas estimés à partir d'un ensemble de tests réels. Ils sont approximés en utilisant les relations mises en évidence à la section 4.4 entre les distances KL et les scores imposteurs.

L'étude expérimentale de la distance $KL2_X$ a montré que cette mesure est très corrélée à la moyenne $m_{S_{\bar{X}}}$ des scores imposteurs. Ces deux grandeurs sont liées par une relation approximativement linéaire (cf équation 4.22). D'autre part, d'après la fi-

gure 4.3, il semble que l'écart type $\sigma_{S_{\bar{X}}}$ des scores imposteurs soit globalement croissant lorsque la distance $KL2_X$ croit. Faute de pouvoir identifier plus finement la relation qui lie ces deux grandeurs, nous la supposerons approximativement linéaire.

Pour un modèle de locuteur X, un score imposteur $S_X(\mathcal{Y}|\bar{X})$ est alors modélisé de la façon suivante :

$$S_X(\mathcal{Y}|\bar{X}) = -a.KL2_X + b.KL2_X.\varepsilon \ , \tag{5.1}$$

où ε est une v.a. Gaussienne centrée réduite indépendante de $KL2_X$. Les coefficients a et b sont supposés constants, positifs et indépendants du locuteur considéré X.

Normalisation D-Norm

La modélisation simple des scores imposteurs proposée ci-dessus permet d'appliquer une normalisation directe des scores en utilisant la distance $KL2_X$. A partir de l'estimation de cette distance, un score brut $S_X(\mathcal{Y})$ est normalisé de la façon suivante :

$$S_X^{dnorm}(\mathcal{Y}) = \frac{S_X(\mathcal{Y})}{KL2_X} \ . \tag{5.2}$$

D'après l'équation 5.1, les scores imposteurs normalisés sont à présent Gaussiens de moyenne $-a$ et d'écart type b :

$$S_X^{dnorm}(\mathcal{Y}|\bar{X}) = -a + b.\varepsilon \ \sim \ \mathcal{N}(\cdot; -a, b) \ . \tag{5.3}$$

Sous les hypothèses considérées, la normalisation D-norm conduit donc à des distributions homogènes des scores imposteurs quel que soit le locuteur X. Notons que pour appliquer cette normalisation, les coefficients a et b n'ont pas besoin d'être estimés, ils sont simplement supposés constants. Enfin, il est important de préciser que la normalisation de score D-norm a pour but de compenser des biais qui sont liés aux caractéristiques du modèle uniquement, comme le fait la Z-norm. Elle ne compense pas les biais liés aux disparités entre conditions d'apprentissage et de test, comme le font la H-norm, la C-norm ou la T-norm.

5.1.3 Développement sur NIST 2001

La D-norm a été développée sur la base de donnée de l'évaluation NIST 2001 selon le protocole et avec le système décrits à la section 4.4.2.1. Le principe de la D-norm est validé sur cette base de données et ses performances sont comparées à celles obtenues avec la Z-norm.

5.1.3.1 Validation expérimentale

La figure 5.1 illustre l'effet de la D-norm sur la moyenne des scores. Elle représente les points $(m_{S_{\bar{X}}}, KL2_X)$ après avoir appliqué la D-norm.

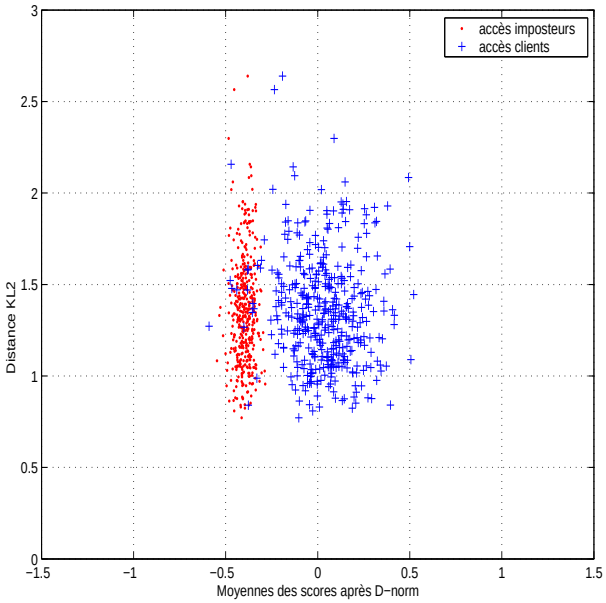


FIG. 5.1 – *Relation entre distances KL et moyennes des scores D-normalisés. Distributions des points* $(KL2_X, m_{S_X^{dnorm}})$ *et* $(KL2_X, m_{S_{\bar{X}}^{dnorm}})$ *(corpus femmes NIST'01)*

Sur cette figure, on n'observe plus de corrélation entre la moyenne des scores normalisés et la distance $KL2_X$, ce qui indique que la D-norm a effectivement compensé les biais dépendants de cette distance qui apparaissent dans les scores.

La figure 5.2 représente les distributions des moyennes des scores imposteurs et des scores clients obtenues lorsqu'on n'utilise pas de normalisation de scores (haut) et après avoir appliqué la D-norm (bas). Lorsque la D-norm est appliquée, la distribution des moyennes des scores imposteurs est plus concentrée, relativement à celle des moyennes des scores clients. Cela présage que l'on pourra mieux séparer les accès imposteurs des accès clients et donc diminuer les taux d'erreurs.

Sur la figure 5.3, l'écart type des scores imposteurs pour chaque locuteur est représenté en fonction de la distance $KL2_X$, après application de la D-norm. D'après cette figure, il semble que l'écart type des scores imposteurs normalisés soit devenu indépendant de la distance $KL2_X$, alors qu'il était globalement croissant avec cette distance avant application de la D-norm (cf section 4.4.2, figure 4.3). On observe toutefois quelques écarts types plus importants pour les distances les plus faibles. Cela peut indiquer que l'hypothèse de linéarité de l'écart type en fonction de la distance $KL2_X$ n'est pas tout à fait vérifiée dans ce cas. La rareté de ces cas et la forte dispersion des points en général ne permet cependant pas de conclure fermement sur le degré d'exactitude de cette hypothèse.

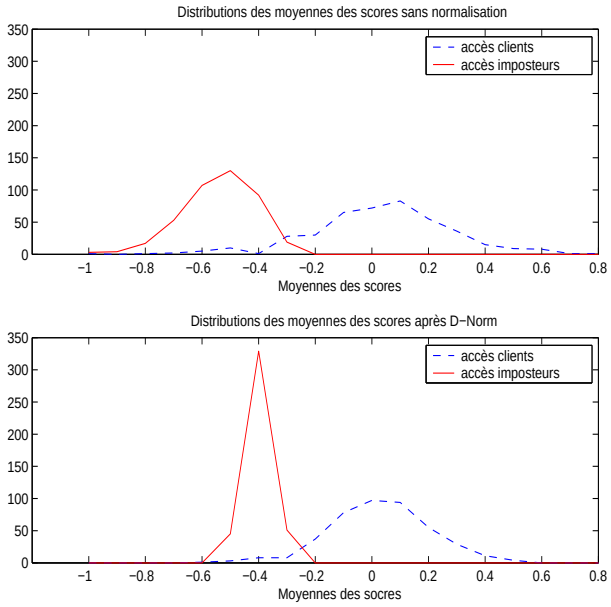


FIG. 5.2 – *Distributions des moyennes des scores imposteurs (lignes continues) et clients (pointillés) sans normalisation (haut) et après application de la D-norm (bas)*

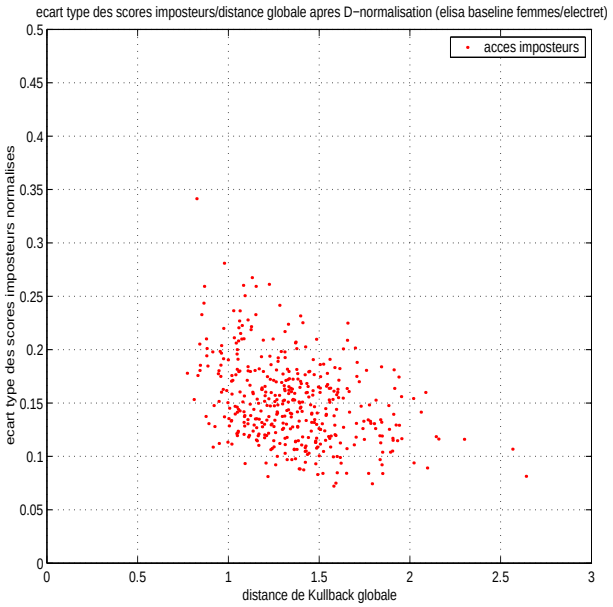


FIG. 5.3 – *Relation entre distances KL et écarts types des scores D-normalisés. Distributions des points* $(KL2_X, \sigma_{S_X^{Dnorm}})$ *et* $(KL2_X, \sigma_{S_{\bar{X}}^{Dnorm}})$ *(corpus femmes NIST'01)*

5.1.3.2 Analyse des performances

L'influence de la normalisation de score D-norm sur les performances du système est illustrée sur la figure 5.4. Elle représente les courbes DET du système IRISA/ELISA obtenues pour les locuteurs femmes de l'évaluation NIST 2001.

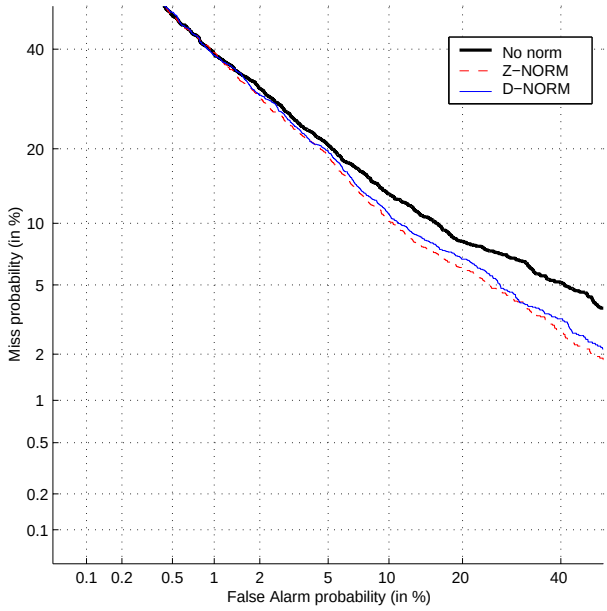


FIG. 5.4 – *Courbes DET du système IRISA/ELISA 2001 sans normalisation (No norm) et avec normalisations D-norm et Z-norm (corpus femmes NIST'01)*

La courbe DET du système avec application de la D-norm est comparée à celles obtenues sans appliquer de normalisation de score (No norm) et en utilisant la Z-norm. Ces courbes montrent que la D-norm et la Z-norm conduisent à des performances similaires et qu'elles améliorent significativement les performances pour les points de fonctionnement à faible taux de faux rejets jusqu'à l'*EER*. Pour les points à faible taux de fausses acceptations, ces deux normalisations n'apportent pas d'amélioration significative.

Ces résultats confirment les observations faites à la section précédente concernant les distributions des moyennes des scores. La D-norm rend les scores imposteurs et clients mieux séparables par un seuil unique de décision, diminuant ainsi les taux d'erreurs pour certaines régions de fonctionnement. Cela valide le rôle de normalisation de score de la technique D-norm, dont l'effet est de compenser les biais introduits par les disparités de qualités des modèles de locuteur, au même titre que la Z-norm.

5.1.3.3 Convergence de la D-norm

Les paramètres de normalisation utilisés par la D-norm sont les distances $KL2_X$. Ces distances sont estimées par une méthode de Monte Carlo (cf section 4.4.1) en générant des échantillons synthétiques de vecteurs acoustiques. Le nombre d'échantillons générés détermine la précision des estimations des distances et influence donc indirectement les performances obtenues avec la D-norm. La figure 5.5 illustre cette influence. Elle montre la convergence des performances de la D-norm au travers des courbes DET obtenues avec différentes quantités de vecteurs acoustiques (V.A) artificiels générés pour l'estimation des distances $KL2_X$ (2 V.A., 10 V.A, 100 V.A, 1000 V.A. et 10000 V.A.).

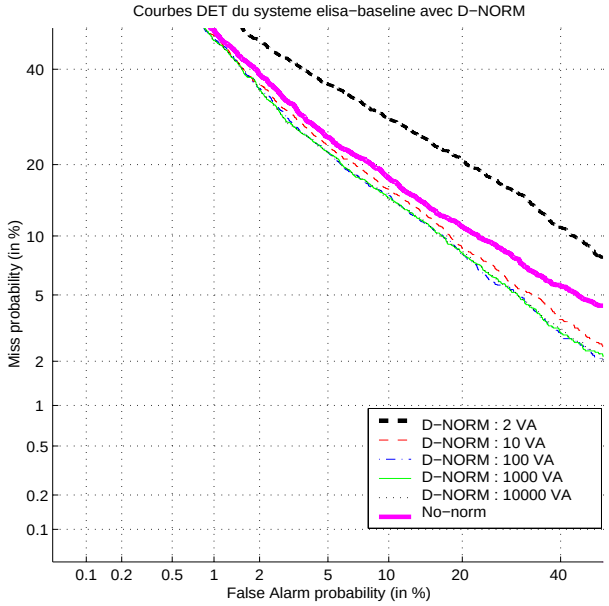


FIG. 5.5 – *Courbes DET du système IRISA/ELISA : convergence de la D-norm*

Pour seulement 2 échantillons générés, l'application de la D-norm détériore les performances du système de base (No norm), l'estimation des distances $KL2_X$ étant dans ce cas trop aléatoire. Cependant, à partir de 10 échantillons, la D-norm améliore déjà les performances. Au delà de 100 échantillons, les courbes DET obtenues se confondent ce qui indique que les performances de la D-norm ont convergé.

L'analyse de ces courbes DET montre que les performances de la D-norm convergent rapidement en fonction du nombre d'échantillons générés pour l'estimation des paramètres de normalisation (distances $KL2_X$). Le temps d'estimation de ces paramètres peut être par conséquent très court, en comparaison du temps nécessaire au calcul des paramètres dans le cas de la Z-norm (réduction d'un facteur 10 à 100 si on utilise respectivement 1000 ou 100 échantillons).

5.1.4 Performances sur les évaluations NIST 2002, 2003 et 2004

5.1.4.1 Normalisations D-norm et DT-norm

La D-norm a été utilisée aux cours des évaluations NIST 2002, 2003 et 2004 avec les systèmes IRISA/ELISA correspondants à ces années (voir [Ben et al.02a, Ben et al.03b, Ben et al.04b]). Pour ces évaluations, la D-norm a été associée à la T-norm (DT-norm) afin de compenser à la fois les biais liés au modèle et à l'énoncé de test. Pour appliquer la DT-norm, le score d'un énoncé de test est tout d'abord D-normalisé. Les paramètres de la T-norm sont également estimés avec un ensemble de scores D-normalisés, issus de modèles imposteurs. Ces paramètres servent ensuite à transformer le score D-normalisé du test, de la façon suivante :

$$S_X^{dtnorm}(\mathcal{Y}) = \frac{S_X^{dnorm} - \mu_{S^{dnorm}}(\mathcal{Y})}{\sigma_{S^{dnorm}}(\mathcal{Y})} \quad . \tag{5.4}$$

L'effet de la DT-norm est illustré sur la figure 5.6 sur laquelle sont tracées les courbes DET du système IRISA/ELISA pour l'évaluation NIST 2002.

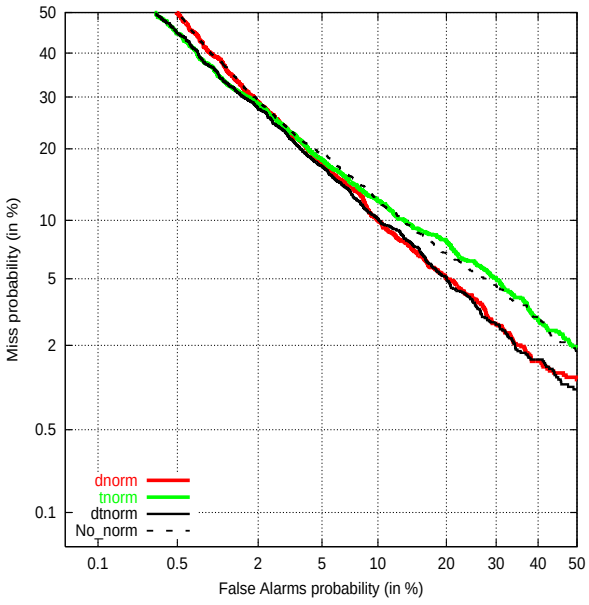


FIG. 5.6 – *Courbes DET du système IRISA/ELISA 2002, sans normalisation (No norm) et avec normalisation D-Norm, T-norm et DT-norm (corpus NIST'02)*

Les quatre courbes représentées correspondent aux performances du système lorsque l'on n'applique pas de normalisation de score (No_norm), lorsque l'on applique la D-norm seule (D-norm), lorsque l'on applique la T-norm seule (T-norm) et lorsque l'on associe la D-norm à la T-norm (DT-norm). L'effet de la D-norm pour l'évaluation NIST 2002 est similaire à celui observé sur la base NIST 2001 : elle améliore les performances

pour les points de fonctionnement à faible taux de faux rejets. La T-norm quand à elle améliore les performances pour les points à faible taux de fausse acceptations (zone privilégiée par la fonction de coût définie par NIST). On constate que l'association de ces deux normalisations permet de tirer partie des avantages de chacune, en améliorant les performances pour tous les points de la courbe DET.

5.1.4.2 Résultats

Le tableau 5.1 donne les points de fonctionnement à fonction de coût C_{det} minimum des systèmes IRISA/ELISA pour les évaluation NIST des années 2002, 2003 et 2004, et pour les différentes tâches auxquelles l'IRISA a participé. Le tableau 5.2 donne l'EER de ces systèmes dans les mêmes conditions.

Les données d'évaluation utilisées en 2002 et 2003 sont issues de la base SWITCHBOARD Cellular [Ldc]. Le système IRISA/ELISA utilisé pour l'évaluation 2002 a été développé sur les données "cellular" et "landline" de l'évaluation NIST 2001. Pour le système IRISA/ELISA de 2003, les données de développement sont exclusivement des données "cellular" issues des évaluations NIST 2001 et 2002.

En 2004, NIST a proposé d'utiliser des données issues d'une nouvelle base récoltée par le LDC [Ldc] : la base MIXER. Cette base contient des données principalement téléphoniques (filaires et cellulaires), avec des conditions très variables. Les conditions rencontrées dans MIXER sont parfois assez différentes de celles que l'on peu trouver dans la base SWITCHBOARD. Cette dernière a été utilisée pour le développement du système IRISA/ELISA 2004 (au travers des données des évaluations NIST 2001, 2002 et 2003). D'autre part, il faut noter que les pré-traitements appliqués par NIST sur les données MIXER de l'évaluation 2004 sont différents de ceux appliqués les années précédentes (notamment, il n'y a pas eu de retrait automatique des silences en 2004). Cette hétérogénéité entre les données de développement et celles d'évaluation explique en grande partie les pertes de performance observées en 2004 pour le système IRISA/ELISA (et pour tous les systèmes participants également).

Dans les tableaux 5.1 et 5.2, les performances pour les systèmes de base sans normalisation de score (baseline) sont rapportées, ainsi que celles obtenues lorsque l'on applique la D-norm seule et la DT-norm. Les résultats confirment les observations effectuées auparavant sur la base NIST'02, à savoir que la D-norm n'apporte pas d'amélioration au point de fonctionnement à minimum de coût C_{det} mais améliore en général les performances à l'EER. L'amélioration à l'EER est moins sensible en 2003 et 2004 qu'en 2002 où la D-norm permet un gain absolu de performance de $1,3\%$ (gain relatif d'environ 10%). Cela peut être dû à de nouveaux traitements utilisés à partir de 2003, au niveau de la paramétrisation (nouveau procédé de sélection de trame) et à un meilleur réglage du système de base sur les données téléphoniques cellulaires.

La DT-norm quant à elle permet d'améliorer significativement les performances au point min C_{det}, grâce à l'utilisation de la T-norm, et conserve les améliorations apportées par la D-norm au point EER.

	$\min C_{det}$ ($\times 10^{-3}$)				
évaluation	NIST'02	NIST'03	NIST'04	NIST'04	NIST'04
condition	1sp (primary)	1sp (primary)	1sp 1sd/1sd (all)	1sp 1sd/30s (all)	1sp 10s/10s (all)
baseline	47.9	37.9	65.7	72.4	99.9
D-norm	47.7	37.5	66.6	72.7	99.9
DT-norm	44.9	31.7	62.1	70.2	94.3

TAB. 5.1 – *Points de fonctionnement* $\min C_{det}$ *du système IRISA/ELISA pour les évaluations NIST'02, NIST'03 et NIST'04, sans normalisation (baseline), avec D-norm et avec DT-norm*

	EER (%)				
évaluation	NIST'02	NIST'03	NIST'04	NIST'04	NIST'04
condition	1sp (primary)	1sp (primary)	1sp 1sd/1sd (all)	1sp 1sd/30s (all)	1sp 10s/10s (all)
baseline	11.35	8.52	16.38	18.71	30.91
D-norm	10.04	8.50	15.68	17.81	30.39
DT-norm	10.12	7.85	15.61	18.11	30.15

TAB. 5.2 – *EER du système IRISA/ELISA pour les évaluations NIST'02, NIST'03 et NIST'04, sans normalisation (baseline), avec D-norm et avec DT-norm*

5.1.5 Conclusion sur la D-norm

La mise en oeuvre de la D-norm a permis de mettre en évidence que les normalisations de scores ne nécessitent pas forcément de corpus externe pour estimer leurs paramètres de normalisation. L'information utile à la normalisation est en réalité présente en partie dans les éléments intervenant dans le processus de vérification. Dans le cas de la D-norm, nous avons utilisé les distances KL entre chaque modèle de locuteurs et le modèle du monde comme mesures d'information liées aux modèles. Ces mesures ont montré une forte corrélation avec la moyenne des scores imposteurs obtenus pour chaque modèle de locuteur, et nous avons exploité ces corrélations pour implémenter la technique de normalisation de scores D-norm. Cette normalisation peut avantageusement remplacer la Z-norm au sein des systèmes de VAL car elle permet d'obtenir des performances équivalentes, sans nécessiter l'utilisation d'un corpus de données externe.

La D-norm, qui est une normalisation basée sur le comportement des modèles, ne permet pas d'obtenir les effets d'une normalisation de test, comme la T-norm. Nous pensons cependant que l'information nécessaire à la normalisation des scores vis-à-vis des conditions particulières des données de test est présente directement dans ces données. Une mesure directe de ces informations pourrait permettre d'implémenter une normalisation de test, sans nécessiter de modèles imposteurs externes comme c'est le cas pour la T-norm. Ces suppositions ont été confirmées lors de l'implémentation d'une nouvelle technique de normalisation appliquée à des modèles appris sur les données de test. Cette technique de normalisation de modèle est présentée à la section 5.3.

Nos travaux sur la D-norm ont donné lieu à une publication en conférence [Ben et al.02b].

5.2 D-MAP : normalisation entropique des modèles par adaptation contrainte

L'étude des divergences et distances KL a montré que ces mesures liées à des caractéristiques intrinsèques des modèles influent sur la distribution des scores obtenus pour chacun des modèles. Dans la section précédente, nous avons utilisé cette relation pour mettre au point une technique de normalisation au niveau des scores, la D-norm. Nous examinons à présent la possibilité d'intervenir directement au niveau des modèles afin d'homogénéiser leur distance KL. Cette homogénéisation des distances KL pour l'ensemble des locuteurs peut être interprétée comme une normalisation des modèles locuteurs vis-à-vis de leurs entropies relatives (ou divergences KL) par rapport au modèle du monde.

5.2.1 Introduction : liens entre adaptation et distribution des scores

Pour introduire les principes de la méthode de normalisation des modèles que nous proposons, nous discutons ici de façon qualitative des liens existant entre le "degré" d'adaptation d'un modèle locuteur par rapport au modèle du monde, et la distribution

des scores obtenus pour ce modèle.

Dans les schémas d'adaptation Bayésienne utilisés en VAL, les GMMs de locuteur sont estimés à partir d'un modèle du monde, servant de modèle *a priori*, et dont les Gaussiennes sont adaptées aux données d'apprentissage des locuteurs. Dans cette procédure d'estimation, des paramètres permettent de régler le "degré" d'adaptation du modèle aux données d'apprentissage. Ces paramètres de réglage s'interprètent comme des facteurs de confiance attribués aux valeurs *a priori* données par les paramètres du modèle du monde. Des valeurs élevées [1] de facteurs de confiance mèneront à des modèles de locuteur faiblement adaptés aux données d'apprentissage. Au contraire, des facteurs de confiance faibles donneront des modèles très adaptés aux données d'apprentissage et proches de leur estimation au maximum de vraisemblance.

Lorsque le modèle est faiblement adapté aux données d'apprentissage, il reste "proche" du modèle du monde et est par conséquent peu discriminant. Les scores obtenus pour ce modèle seront proches de zéro (cf figure 5.7, bas) que ce soit pour des accès clients ou imposteurs.

Un modèle fortement adapté sera par contre très spécifique de l'ensemble d'apprentissage. Il discriminera en général assez bien les accès imposteurs ce qui veut dire que les scores imposteurs seront fortement négatifs en moyenne. Cependant, il peut aussi mal se généraliser aux accès clients s'il est sur-adapté aux données d'apprentissage. Dans ce cas, les scores clients seront en moyenne faiblement positifs voir même négatifs (cf figure 5.7, haut).

Cette analyse qualitative montre que le "degré" d'adaptation des modèles de locuteur influe sur la distribution des scores engendrés par chacun de ces modèles. Il apparaît clairement que lorsque les modèles sont adaptés de façon hétérogène, avec notamment certains modèles sous-adaptés et d'autres sur-adaptés, un seuil unique de décision sera en général très sous-optimal. En outre, l'influence du "degré" d'adaptation sur la distribution des scores indique qu'il devrait être possible de modifier ces distributions, et en particulier les homogénéiser, en jouant sur les paramètres de l'adaptation de façon dépendante du locuteur.

5.2.2 Adaptation par critère MAP contraint : D-MAP

5.2.2.1 Distances KL et facteur d'adaptation

Nous proposons de normaliser les modèles vis-à-vis de leur distance KL en jouant sur les paramètres de l'adaptation Bayésienne. Cette normalisation des modèles devrait implicitement entraîner une normalisation des scores imposteurs étant données les relations existant entre ces grandeurs, mises en évidence à la section 4.4.2.

1. Dans le schéma d'adaptation Bayésienne présenté à la section 2.3.2, les valeurs "élevées" ou "faibles" des facteurs de confiance sont considérées relativement aux taux d'occupation des Gaussiennes

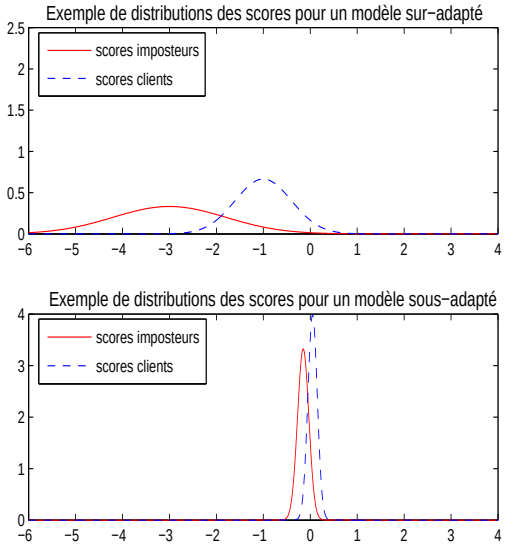


FIG. 5.7 – *Exemples de distribution des scores pour un modèle sous-adapté (bas) et un modèle sur-adapté (haut)*

Nous considérons le schéma d'adaptation Bayésienne de modèles GMM de locuteurs décrit dans [MC et al.01][2]. Dans ce schéma d'adaptation, un modèle du monde est utilisé comme modèle *a priori* des GMM de locuteurs. Les paramètres du modèle du monde servent d'initialisation à la première itération de l'algorithme EM. Seules les moyennes des GMMs sont adaptées dans la procédure considérée, les poids et matrice de covariance étant fixés à leur valeur *a priori*. L'estimation d'une moyenne m_k est obtenue par combinaison linéaire des moyennes *a priori* μ_k et empirique $\overline{y}_k$ selon l'équation (cf section 2.3.2) :

$$\hat{m}_k = \frac{\gamma_k}{\gamma_k + \tau_k}\overline{y}_k + \frac{\tau_k}{\gamma_k + \tau_k}\mu_k \quad . \tag{5.5}$$

Dans le schéma proposé, le facteur de confiance τ_k est déterminé de la manière suivante :

$$\tau_k = \frac{\beta}{1-\beta}.w_k.T \quad , \tag{5.6}$$

où w_k est le poids de la Gaussienne k dans le modèle du monde et T est le nombre de trames dans l'ensemble d'apprentissage. Le coefficient β contrôle l'adaptation du modèle de locuteur, avec $\beta \in [0, 1]$.

2. les principes de la méthode proposée peuvent s'étendre à tout type de schémas d'adaptation Bayésienne

La méthode de normalisation des modèles que nous proposons consiste à rechercher une valeur de β dépendante de chaque locuteur et qui permet d'homogénéiser les distances KL pour l'ensemble des modèles.

Le coefficient β_X pour l'adaptation d'un modèle de locuteur X est relié à la distance $KL2_X$ de la manière suivante :

- si $\beta_X = 0$ (c.à.d. $\tau_k = 0$), l'adaptation maximale des moyennes aux données d'apprentissage est effectuée. Les moyennes sont alors estimées au maximum de vraisemblance. On notera la distance KL correspondante $KL2_X^{(ML)}$.
- si $\beta_X = 1$ (c.à.d. $\tau_k = \infty$), les moyennes du modèle X ne sont pas adaptées du tout. Dans ce cas le modèle X est strictement identique au modèle du monde et on a $KL2_X = 0$.

D'après ces considérations, $KL2_X$ et β_X sont donc liées par une relation du type

$$KL2_X = f_X(\beta_X).KL2_X^{(ML)} \quad , \tag{5.7}$$

où $f_X(\cdot)$ est une fonction de β_X qui dépend de X et avec les contraintes $f_X(0) = 1$ et $f_X(1) = 0$. Cette fonction est donnée par l'évolution du rapport $\frac{KL2_X}{KL2_X^{(ML)}}$ en fonction de β_X.

5.2.2.2 Formulation générale du critère

Une distance "cible" de référence $KL2_{ref}$ étant définie, le coefficient β_X pour l'adaptation de chaque modèle de locuteur X est déterminé de façon à avoir $KL2_X = KL2_{ref}$ quel que soit X. Cette contrainte permet d'obtenir la valeur de β_X correspondant, en inversant la relation 5.7 :

$$\beta_X = f_X^{-1}\left(\frac{KL2_{ref}}{KL2_X^{(ML)}}\right) \quad . \tag{5.8}$$

Cette dernière équation définit un facteur d'adaptation qui est dépendant du locuteur et permettant d'obtenir un modèle qui est à une distance $KL2_{ref}$ du modèle du monde. Le schéma d'adaptation correspondant est appelé D-MAP ("Distance constrained MAP").

5.2.2.3 Liens avec la D-norm

Le but de l'adaptation D-MAP est d'homogénéiser les modèles locuteurs vis-à-vis de leur distance KL par rapport au modèle du monde, afin de normaliser les scores de façon indirecte. Cette normalisation implicite des scores s'appuie sur le fait que les paramètres moyenne $m_{S_{\bar{X}}}$ et écart type $\sigma_{S_{\bar{X}}}$ de la distribution des scores imposteurs pour chaque locuteur X sont liés à la distance $KL2_X$ (cf chapitre 4, section 4.4.2). En approximant les relations correspondantes par des fonctions déterministes du type $m_{S_{\bar{X}}} = m(KL2_X)$ et $\sigma_{S_{\bar{X}}} = s(KL2_X)$ et en supposant que la distribution des scores

imposteurs est Gaussienne, un score imposteur $S_X(\mathcal{Y}|\bar{X})$ peut être modélisé de la façon suivante :

$$S_X(\mathcal{Y}|\bar{X}) = m(KL2_X) + s(KL2_X).\varepsilon \quad \text{avec} \quad \varepsilon \sim \mathcal{N}(\cdot; 0, 1) \ . \tag{5.9}$$

Dans le cadre de la D-norm, les fonctions $m(KL2_X)$ et $s(KL2_X)$ ont été modélisées comme de simples relations linéaires, ce qui a permis d'implémenter une normalisation de score par simple division par $KL2_X$.

Dans le cadre de D-MAP, les distances $KL2_X$ sont normalisées à la valeur de référence $KL2_{ref}$. On a donc :

$$S_X^{dmap}(\mathcal{Y}|\bar{X}) = m(KL2_{ref}) + s(KL2_{ref}).\varepsilon \ . \tag{5.10}$$

Dans ce cas, les fonctions $m(KL2_X)$ et $s(KL2_X)$ n'ont pas besoin d'être identifiées formellement. Il suffit de supposer qu'elles sont bijectives et indépendantes du locuteur X considéré. Sous ces hypothèses, les scores imposteurs $S_X^{dmap}(\mathcal{Y}|\bar{X})$ sont distribués suivant une Gaussienne de moyenne constante $m(KL2_{ref})$ et d'écart type constant $s(KL2_{ref})$, quel que soit le locuteur X :

$$S_X^{dmap}(\mathcal{Y}|\bar{X}) \sim \mathcal{N}(\cdot; m(KL2_{ref}), s(KL2_{ref})), \quad \forall X \ . \tag{5.11}$$

Ces hypothèses moins restrictives sur la modélisation des scores imposteurs sont un avantage potentiel de l'adaptation D-MAP par rapport à la normalisation de scores D-norm. Néanmoins, ces deux techniques de normalisation reposent sur un même principe, qui est d'utiliser les dépendances entre les distances KL et les scores imposteurs pour normaliser la distribution des scores.

5.2.2.4 Discussion sur la mise en application de D-MAP

Au regard de l'équation 5.8 qui détermine la valeur du coefficient β_X pour un modèle de locuteur X, D-MAP implique de connaître la distance $KL2_X^{(ML)}$ pour chaque locuteur. Une estimation au maximum de vraisemblance des moyennes des GMMs doit donc être effectuée avant d'appliquer D-MAP. D'autre part l'équation 5.8 suppose que la fonction $f_X(\beta_X)$ est connue et inversible. En réalité, cette fonction est inconnue et ne peut être déterminée de façon explicite dans le cas d'une modélisation par GMM. La mise en application de D-MAP exige donc l'utilisation d'une procédure dichotomique afin de trouver une valeur de β_X pour laquelle la distance correspondante $KL2_X$ approxime la référence $KL2_{ref}$ avec une précision suffisante. En pratique cependant, il est possible d'accélérer cette procédure dichotomique en approximant $f_X(\cdot)$ par une fonction simple de β_X.

5.2.3 Développement expérimental

L'adaptation D-MAP a été développée sur un sous-ensemble de l'évaluation NIST'01 avec le protocole et le système décrits à la section 4.4.2.1.

5.2.3.1 Approximation de la fonction $f_X(\beta_X)$

Dans le cas de modèles GMM de locuteurs, la distance $KL2_X$ ne s'exprime pas sous une forme explicite, en fonction des paramètres des modèles. On ne peut dons pas relier de façon simple la distance $KL2_X$ au coefficient β_X. Pour des modèles Gaussiens par contre, cette relation s'exprime simplement en fonction des paramètres du modèle de locuteur et du modèle du monde. Nous nous inspirons de ce cas basique pour trouver une fonction facilement inversible permettant d'approximer la relation entre $KL2_X$ et β_X, de manière empirique, dans le cas de GMM.

Cas mono-Gaussien

Le modèle mono-Gaussien est un cas particulier de GMM pour lequel le nombre K de composantes est égal à 1. Le poids w de l'unique gaussienne est égal à 1 et son taux d'occupation γ correspond au nombre de trames total T de l'ensemble d'apprentissage ($w_k = 1$ et $\gamma_k = T$ dans les équations 5.5 et 5.6). L'équation d'estimation de la moyenne m_X du modèle mono-Gaussien du locuteur X au sens du critère MAP se réduit donc à :

$$\hat{m_X} = (1 - \beta_X).\overline{y_X} + \beta_X.\mu \ . \tag{5.12}$$

Si l'on considère un modèle de locuteur dont seule la moyenne est adaptée d'un modèle mono-Gaussien du monde, on peut alors montrer que la distance $KL2_X$ est liée à β_X par une relation quadratique :

$$\begin{aligned} KL2_X &= [m_X - \mu]^* \Sigma^{-1} [m_X - \mu] \ , \\ &= [\overline{y}_X - \mu]^* \Sigma^{-1} [\overline{y}_X - \mu].(1 - \beta_X)^2 \ , \\ &= KL2_X^{(ML)}.(1 - \beta_X)^2 \ . \end{aligned} \tag{5.13}$$

Dans cette équation, * désigne la transposition, Σ est la matrice de covariance commune au modèle de locuteur et au modèle du monde, et $KL2_X^{(ML)}$ est la distance KL entre ces deux modèles lorsque la moyenne du modèle de locuteur est estimée au maximum de vraisemblance.

Approximation empirique dans le cas multi-Gaussien

Dans le cas multi-Gaussien, on ne peut pas trouver de forme simple reliant $KL2_X$ à β_X. Néanmoins, en s'inspirant du cas mono-Gaussien, on peut chercher à approximer de manière empirique cette relation par une fonction simple de la forme :

$$KL2_X \approx KL2_X^{(ML)}.(1 - \beta_X)^n \ , \tag{5.14}$$

soit :

$$f_X(\beta_X) \approx (1 - \beta_X)^n \ . \tag{5.15}$$

Le coefficient de puissance n peut être estimé sur un ensemble de développement.

Cette relation permet d'obtenir une première estimation $\beta_X^{(0)}$ du coefficient β_X pour l'adaptation D-MAP :

$$\beta_X^{(0)} = 1 - \left(\frac{KL2_{ref}}{KL2_X^{(ML)}} \right)^{\frac{1}{n}} . \tag{5.16}$$

Remarquons que, pour garantir que $\beta_X^{(0)}$ soit compris entre 0 et 1, il faut choisir la valeur cible $KL2_{ref}$ telle que l'on ait $KL2_{ref} \leq KL2_X^{(ML)}$ quel que soit le locuteur X considéré.

Pour les expériences présentées dans cette section nous avons choisi $n = 4$ dans l'équation . Cette valeur s'est avérée être un bon compromis pour approximer le comportement global des fonction $f_X(\beta_X)$ sur l'ensemble de locuteurs considéré.

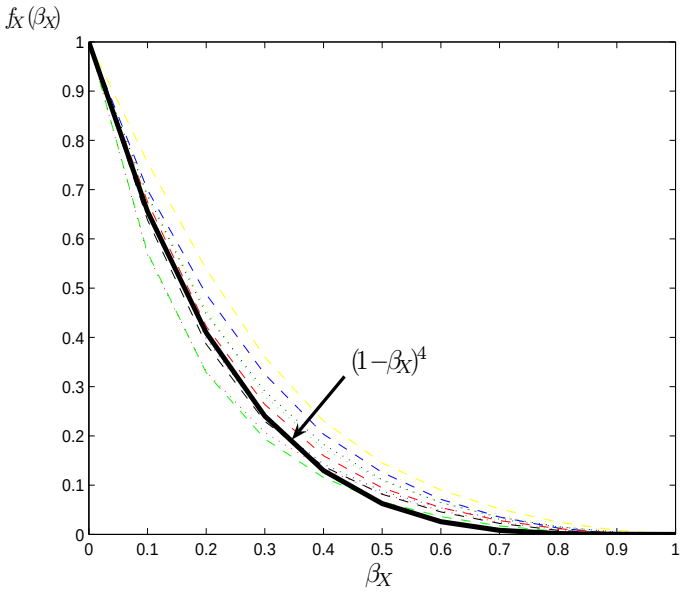


FIG. 5.8 – *Évolution des fonctions $f_X(\beta_X)$ vis-à-vis du coefficient d'adaptation β_X, pour 10 locuteurs femmes de l'évaluation NIST'01*

La figure 5.8 représente les variations des fonctions $f_X(\beta_X)$ (donnée par l'évolution du rapport $KL2_X/KL2_X^{(ML)}$) en fonction de $\beta_X \in [0, 1]$, pour 10 locuteurs femmes de l'évaluation NIST'01, choisis au hasard. D'après cette figure, il apparaît que les fonctions $f_X(\beta_X)$ diffèrent sensiblement d'un locuteur à un autre, en ayant néanmoins un comportement global similaire. La fonction $(1 - \beta_X)^4$ est également tracée sur la figure, montrant que ce choix permet d'approximer le comportement général des fonction $f_X(\beta_X)$ mais qu'il sera nécessaire de réajuster le coefficient β_X par dichotomie pour un grand nombre de locuteurs.

5.2.3.2 Accélération de la procédure dichotomique

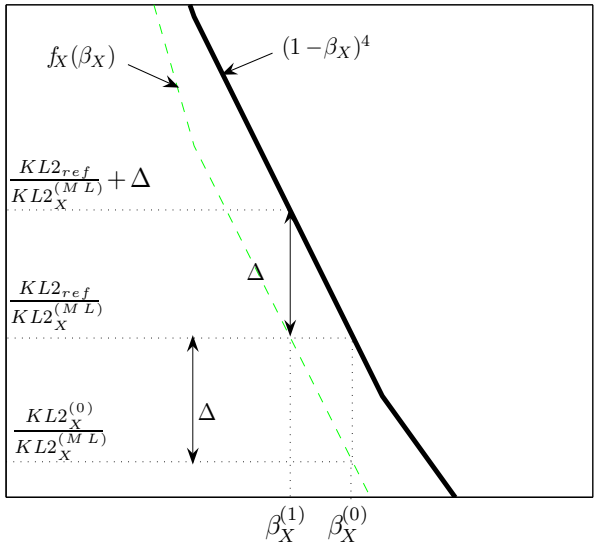


FIG. 5.9 – *Illustration de la procédure de correction du coefficient d'adaptation β_X*

La procédure dichotomique d'affinement du paramètre β_X peut être grandement accélérée, d'une part grâce à la première estimation $\beta_X^{(0)}$, et d'autre part en utilisant une procédure de correction de ce coefficient basée sur la relation $f_X(\beta_X) \approx (1-\beta_X)^4$.

On considère pour cela que les fonctions $f_X(\beta_X)$ et $(1-\beta_X)^4$ sont représentées par des courbes parallèles aux alentours de la première estimation $\beta_X^{(0)}$ (cf figure 5.9). La distance KL réellement donnée par cette estimation sera notée $KL2_X^{(0)}$.

Il est alors possible de corriger la valeur de β_X pour obtenir une nouvelle estimation $\beta_X^{(1)}$, en utilisant la formule :

$$\beta_X^{(1)} = 1 - \left(\frac{KL2_{ref}}{KL2_X^{(ML)}} + \Delta \right)^{\frac{1}{4}} \quad , \tag{5.17}$$

où le biais Δ est défini par :

$$\Delta = \frac{KL2_{ref}}{KL2_X^{(ML)}} - \frac{KL2_X^{(0)}}{KL2_X^{(ML)}} \quad .$$

Cette procédure de correction est illustrée sur la figure 5.9. Elle peut être appliquée itérativement de façon à converger rapidement vers $KL2_{ref}$.

En pratique, lorsque l'on veut atteindre la valeur cible $KL2_{ref}$ avec une précision donnée (par exemple 5%), cette procédure de correction peut ne pas converger pour

certains locuteurs, qui restent rares cependant. Dans ce cas, on termine le raffinement du coefficient α_X par une recherche dichotomique classique. Dans la plupart des cas, il s'est avéré que la procédure converge rapidement et la valeur $KL2_{ref}$ est atteinte avec une précision de 5% en seulement une ou deux itérations.

5.2.4 Résultats

L'adaptation D-MAP a été évaluée sur la base de données de l'évaluation NIST'01, et comparée à l'adaptation MAP implémentée dans le système de base IRISA/ELISA2001 [MC et al.01]. Les *EERs* obtenus par le système de base sont donnés dans le tableau 5.3 pour différentes valeurs du coefficient β qui contrôle l'adaptation. Les performances de ce système sont globalement optimisées à l'EER pour la valeur $\beta = 0.4$ ($EER = 11,2\%$).

MAP					
β	0	0.25	0.4	0.6	0.8
EER(%)	13.7	12.05	11.2	11.25	11.45

D-MAP				**MAP0.4**
$KL2_{ref}$	0.25	0.5	0.8	**+D-Norm**
EER(%)	10.45	10.4	10.4	10.75

TAB. 5.3 – *EER du système IRISA/ELISA 2001 pour différentes configurations d'adaptation : adaptation MAP classique avec différents coefficients d'adaptation β, adaptation D-MAP, et adaptation optimale MAP0.4 avec normalisation D-Norm*

Le tableau 5.3 rapporte également l'EER obtenu par le système utilisant l'adaptation D-MAP, avec différentes valeurs de la distance de référence $KL2_{ref}$. D'après les résultats, le choix de cette distance de référence ne semble pas être déterminant pour les performances de D-MAP. L'EER obtenu avec l'adaptation D-MAP est de $10,4\%$, surpassant ainsi l'adaptation MAP classique avec réglage optimal du coefficient d'adaptation β.

La figure 5.10 montre les courbes DET obtenues par le système IRISA/ELISA2001 avec l'adaptation MAP classique optimisée (MAP0.4), et lorsque l'adaptation D-MAP est utilisée (D-MAP) avec $KL2_{ref} = 0,8$. La courbe DET du système lorsque les moyennes des GMM sont estimées au maximum de vraisemblance est également tracée (MAP0), comme courbe de référence.

D'après ces courbes DET, les approches par critère MAP pour l'adaptation des moyennes des GMMs de locuteurs confirment leur supériorité par rapport à l'estimation ML, sur le type de données rencontrées dans les évaluations NIST. L'adaptation D-MAP permet d'obtenir des performances supérieures à celles obtenues avec l'adap-

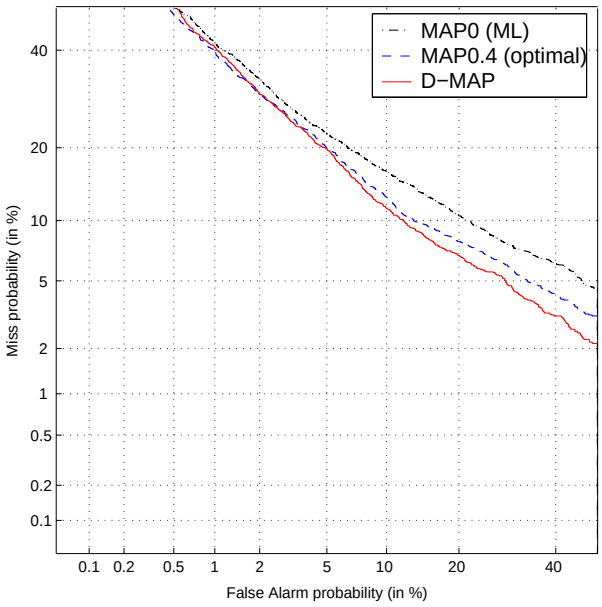


FIG. 5.10 – *Courbes DET du système IRISA/ELISA2001 avec différents schémas d'estimation : estimation ML, adaptation MAP classique optimisée (MAP0.4), et adaptation D-MAP (corpus femmes NIST'01)*

tation MAP classique optimisée. L'amélioration est significative depuis les points de fonctionnement à faibles taux de faux rejets jusqu'à l'EER. Pour les points à faibles taux de fausses acceptations, les deux approches conduisent à des performances similaires.

Le tableau 5.3 donne également l'EER obtenu avec le système MAP optimisé (MAP0.4) et lorsqu'on utilise la normalisation de score D-norm. Les performances ($EER = 10,75\%$) sont proches de celles fournies par le système D-MAP sans normalisation de score. D'autre part, des expériences complémentaires ont mis en évidence que les normalisations de scores basées sur le comportement des modèles (Z-norm et D-norm) n'ont plus aucune influence sur les performances du système lorsqu'on utilise l'adaptation D-MAP. Cela indique que ce schéma d'adaptation compense implicitement les biais engendrés par les différences de qualité entre les modèles estimés, et qu'une normalisation des scores cherchant à compenser ces biais ne semble plus nécessaire en aval.

5.2.5 Conclusion sur D-MAP

Tout comme la D-norm, la technique D-MAP a confirmé que certains biais présents dans les scores proviennent de particularités des modèles. Ces particularités peuvent

être mesurées au travers des distances KL. La technique d'adaptation D-MAP permet d'homogénéiser les modèles vis-à-vis de ces distances KL en jouant sur les paramètres de l'adaptation. La conséquence est que les scores se trouvent implicitement normalisés eux aussi, vis-à-vis des biais liés aux modèles.

Ces résultats indiquent qu'un schéma d'adaptation adéquat peut remplacer une normalisation de score en agissant plus en amont dans le système. Les normalisations de scores correspondantes ne sont alors plus nécessaires.

Dans la section suivante, nous montrons qu'il est possible également de compenser certains biais liés aux conditions des données de test, par l'intermédiaire de modèles de test que nous normalisons de la même manière que les modèles de locuteur.

La technique d'adaptation D-MAP a donné lieu à une publication en conférence [Ben et al.03a].

5.3 Vérification du locuteur dans un espace des modèles

5.3.1 Motivations : un calcul rapide de distances entre modèles

Les distances de Kullback-Leibler fournissent des mesures de similarité entre modèles probabilistes. En VAL, leurs liens avec les scores de vérification nous ont permis d'implémenter des techniques de normalisation, au niveau des scores d'une part, avec la D-norm, et au niveau des modèles avec l'adaptation D-MAP.

Dans cette section nous étudions la possibilité d'utiliser des distances entre les modèles de locuteurs, le modèle du monde et des modèles de test pour calculer des scores de vérification. Ce type de score pourrait avoir plusieurs avantages si le calcul des distances entre modèle est simple et rapide.

Premièrement, certaines normalisations de scores coûteuses dans l'espace des rapports de vraisemblance, comme la T-norm notamment, pourraient être grandement allégées avec un calcul de scores rapide basé sur des distances entre modèles (même s'il faut préalablement estimer un modèle du test).

Deuxièmement, la vérification du locuteur dans un espace des modèles supprimerait complètement l'utilisation de données acoustiques du calcul des scores. Les énoncés de test et d'apprentissage seraient alors représentés tous les deux par des éléments homogènes, des modèles probabilistes, que l'on pourrait manipuler de façon identique. En particulier, le même type de normalisations pourrait être appliqué sur les modèles de test et sur les modèles de locuteur.

Enfin, un calcul allégé du score de vérification pourrait permettre d'effectuer ce calcul sur une architecture matérielle à capacité limitée, comme une carte à puce par exemple. L'estimation du modèle de test devrait dans ce cas être déléguée à un terminal hôte plus puissant, seul le calcul du score devant être effectué impérativement sur la carte elle-même.

Pour effectuer de façon efficace cette vérification du locuteur dans un espace des

modèles, il faut cependant disposer de distances entre modèles qui soient simplement et rapidement calculables. Les distances KL entre GMMs présentent l'inconvénient de devoir être estimées par une méthode de tirage de Monte Carlo qui peut être relativement coûteuse. Une distance directement calculable à partir des paramètres du modèle serait plus appropriée.

5.3.2 Définition d'une distance simple entre GMMs adaptés

En utilisant les principes de l'adaptation Bayésienne et une propriété générale des divergences KL, nous définissons une distance entre GMMs dont l'expression est une fonction simple des paramètres du modèle.

Soient p et $\tilde{p}$ deux modèles GMM à K composantes et de paramètres $(\{w_k\}, \{m_k\}, \{S_k\})$ et $(\{\tilde{w}_k\}, \{\tilde{m}_k\}, \{\tilde{S}_k\})$ respectivement. On peut montrer (voir [Do03]) que la divergence $KL(p||\tilde{p})$ entre ces deux modèles est majorée de la façon suivante :

$$KL(p||\tilde{p}) \leq KL(\mathbf{w}||\tilde{\mathbf{w}}) + \sum_{k=1}^{K} w_k . KL(\mathcal{N}_k||\tilde{\mathcal{N}}_k) \ . \tag{5.18}$$

Le terme $KL(\mathbf{w}||\tilde{\mathbf{w}})$ est la divergence KL entre les distributions de masses de probabilité $\mathbf{w} = (w_1, ..., w_K)$ et $\tilde{\mathbf{w}} = (\tilde{w}_1, ..., \tilde{w}_K)$ et $KL(\mathcal{N}_k||\tilde{\mathcal{N}}_k)$ est la divergence KL entre la Gaussienne k du modèle p et la Gaussienne k du modèle $\tilde{p}$.

Cette inégalité est vérifiée pour tout alignement arbitraire des indices k des Gaussiennes des deux modèles. Dans le cas de GMMs de locuteurs adaptés d'un modèle du monde, cet alignement entre Gaussiennes est implicitement donné par le principe même de l'adaptation : la Gaussienne k de n'importe quel GMM de locuteur a été adaptée de la Gaussienne k du modèle du monde.

Dans ce travail, nous considérons uniquement le cas de GMMs de locuteurs dont seules les moyennes sont adaptées, les poids et matrices de covariance étant fixés à leur valeur *a priori* donnée par le modèle du monde. Dans ce cas, les distributions de masses de probabilité $\mathbf{w}$ et $\tilde{\mathbf{w}}$ sont identique et le terme $KL(\mathbf{w}||\tilde{\mathbf{w}})$ est nul dans l'inégalité .

D'autre part, en considérant des matrices de covariance diagonales, cette inégalité peut s'exprimer sous la forme explicite :

$$KL(p||\tilde{p}) \leq \frac{1}{2} \sum_{k=1}^{K} \sum_{d=1}^{D} w_k . \frac{(m_{k,d} - \tilde{m}_{k,d})^2}{\sigma_{k,d}^2} \ , \tag{5.19}$$

où $m_{k,d}$ et $\tilde{m}_{k,d}$ sont les coefficients d des vecteurs moyennes des Gaussiennes $\mathcal{N}_k$ et $\tilde{\mathcal{N}}_k$ respectivement, et $\sigma_{k,d}^2$ est l'élément $d \times d$ de leur matrice de covariance commune S_k.

En remarquant que le terme de droite de l'inégalité ci-dessus est symétrique suivant les paramètres des modèles p et $\tilde{p}$, on peut majorer la distance KL (somme des deux divergences KL duales) comme suit :

$$KL2(p, \tilde{p}) \leq \sum_{k=1}^{K} \sum_{d=1}^{D} w_k . \frac{(m_{k,d} - \tilde{m}_{k,d})^2}{\sigma_{k,d}^2} \ . \tag{5.20}$$

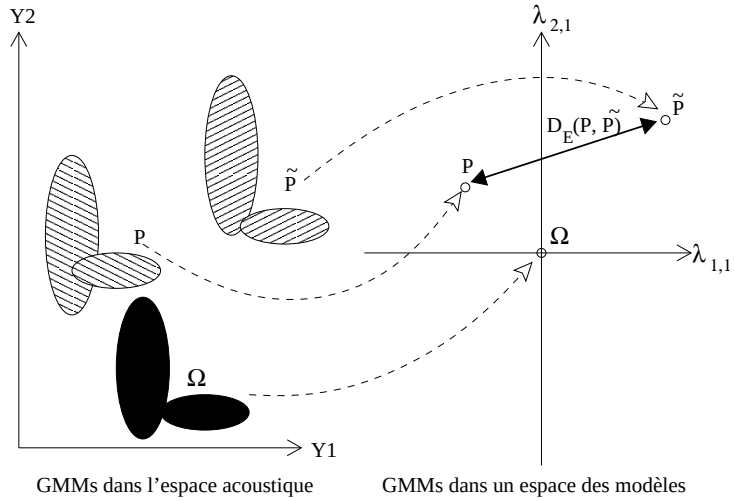


FIG. 5.11 – *Représentations de GMMs : de l'espace acoustique (gauche) à un espace des modèles (droite).*

Le terme de droite de cette inégalité définit une mesure de similarité entre des GMMs dont les moyennes ont été adaptées d'un modèle du monde. Il correspond à une somme pondérée des écarts quadratiques entre les coefficients de même indice k et de même dimension d des vecteurs moyennes des deux GMMs. La contribution de chaque terme dans la somme est proportionnelle au poids w_k de la Gaussienne k et inversement proportionnelle à la variance du coefficient de dimension d de la Gaussienne k. En particulier, cette somme est nulle lorsque les deux modèles sont identiques (c'est-à-dire lorsque leurs moyennes sont identiques).

Dans la section suivante, nous montrons que ce terme s'interprète comme le carré d'une distance Euclidienne dans un espace formé par les paramètres des modèles.

5.3.3 Espace des modèles

5.3.3.1 Définition de l'espace des modèles

Le terme de droite de l'inégalité 5.20 est homogène au carré d'une distance Euclidienne entre deux points dans un espace défini par les paramètres $\{\lambda_{k,d}\}$, tels que :

$$\lambda_{k,d} = \sqrt{w_k}\frac{\Delta m_{k,d}}{\sigma_{k,d}} \quad . \tag{5.21}$$

Le terme $\Delta m_{k,d} = m_{k,d} - m_{k,d}^{(\Omega)}$ est un biais relatif du coefficient d de la moyenne k par rapport au coefficient correspondant dans le modèle du monde Ω.

Dans l'espace ainsi défini, un GMM est représenté par un simple point et le modèle du monde Ω correspond au point origine (cf figure 5.11). Pour simplifier, nous appellerons cet espace "l'espace des modèles". Nous noterons $D_E(p, \tilde{p})$ la distance Euclidienne

entre les points correspondant aux modèles p et $\tilde{p}$. Cette distance est donc définie par :

$$D_E(p,\tilde{p}) = \sqrt{\sum_{k=1}^{K}\sum_{d=1}^{D} w_k.\frac{(m_{k,d}-\tilde{m}_{k,d})^2}{\sigma_{k,d}^2}} \ , \tag{5.22}$$

et le carré de cette distance correspond à la borne majorante de la distance $KL2$ dans l'inégalité 5.20.

L'avantage de cette métrique est qu'elle peut être calculée directement à partir des paramètres des GMMs, avec un coût de calcul faible. Cependant, lorsqu'elle est calculée entre un modèle de test et un modèle de locuteur, il faut au préalable avoir estimé le modèle du test.

5.3.3.2 Distances et scores dans l'espace des modèles

Nous montrons maintenant comment les distances Euclidiennes définies à la section précédente peuvent être utilisées pour le calcul d'un score de vérification en VAL.

Au chapitre 4 nous avons montré que le log-rapport de vraisemblance calculé lors d'un accès de test est asymptotiquement équivalent à une différence entre deux divergences KL. Nous rappelons ici cette relation :

$$E_{\mathcal{M}_Y}[\log\frac{p(y|X)}{p(y|\Omega)} = KL(\mathcal{M}_Y(y)||p(y|\Omega)) - KL(\mathcal{M}_Y(y)||p(y|X)) \ . \tag{5.23}$$

Dans cette équation, $\mathcal{M}_Y$ désigne le modèle "vrai" du test, $p(y|X)$ est le modèle estimé du locuteur X correspondant à l'identité proclamée et $p(y|\Omega)$ est le modèle estimé du non-locuteur, donné par le modèle du monde Ω.

De façon similaire, nous définissons un score de vérification $S_X(\mathcal{Y})$ en utilisant les distances Euclidiennes D_E de la manière suivante :

$$S_X(\mathcal{Y}) = D_E^2(p(y|Y)||p(y|\Omega)) - D_E^2(p(y|Y)||p(y|X)) \ . \tag{5.24}$$

La première distance est calculée entre le modèle $p(y|Y)$ estimé sur l'énoncé de test $\mathcal{Y}$ et le modèle du monde $p(y|\Omega)$. La deuxième est calculée entre le modèle estimé du test $p(y|Y)$ et le modèle estimé $p(y|X)$ du locuteur proclamé. Dans la suite du document, afin d'alléger les écritures nous noterons $D_E(X_1,X_2)$ la distance Euclidienne calculée entre les modèles estimés $p(y|X_1)$ et $p(y|X_2)$ des locuteurs X_1 et X_2.

Le score $S_X(\mathcal{Y})$ défini à l'équation 5.24 possède un lien fort avec le rapport de vraisemblance, de part les relations 5.23 et 5.19. Il est simple à calculer et n'utilise plus les données de test directement puisque son calcul est uniquement basé sur des paramètres de modèles. De plus, il fait intervenir des paramètres oeuvrant dans un espace Euclidien ce qui permet de manipuler les modèles grâce à des transformations géométriques simples sur leurs paramètres. Nous montrons dans la section suivante comment utiliser cette propriété pour implémenter des techniques de normalisation efficaces.

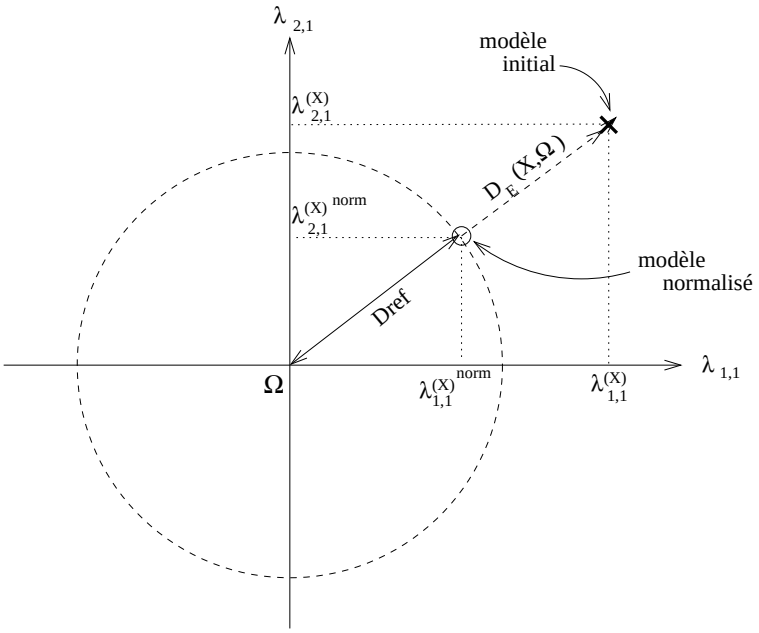


FIG. 5.12 – *Illustration de la normalisation de modèles dans l'espace des modèles.*

5.3.4 Techniques de normalisation dans l'espace des modèles

5.3.4.1 Normalisations de scores

Les techniques de normalisation de scores telles que la Z-norm ou la T-norm sont directement applicables aux scores calculés dans l'espace des modèles. Il suffit simplement d'estimer les paramètres moyennes m_S et écart type σ_S nécessaires à ces normalisations, à partir des scores définis à la section précédente.

5.3.4.2 Normalisations de modèles

En s'inspirant de la technique de normalisation D-MAP (cf. section 5.2) nous proposons une méthode pour normaliser les modèles vis-à-vis de leur distance $D_E(X, \Omega)$ par rapport au modèle du monde. Pour cela nous définissons arbitrairement [3] une distance "cible" D_{ref} (par exemple $D_{ref} = 1$) à laquelle doivent être normalisées toutes les distances $D_E(X, \Omega)$ pour l'ensemble des locuteurs. Cette normalisation équivaut à une projection des points représentant les GMMs dans l'espace des modèles sur une hyper-sphère de rayon D_{ref} (voir figure 5.12).

En appliquant le théorème de Thalès, les nouveaux paramètres $\{\lambda_{k,d}^{(X)^{norm}}\}$ pour un

3. La valeur de cette distance "cible" n'a théoriquement aucune influence sur le procédé de vérification dans l'espace des modèles

locuteur X s'expriment en fonction des anciens paramètres $\{\lambda_{k,d}^{(X)}\}$ de la façon suivante :

$$\lambda_{k,d}^{(X)^{norm}} = \frac{D_{ref}}{D_E(X,\Omega)} \lambda_{k,d}^{(X)} \quad . \tag{5.25}$$

Ces paramètres normalisés correspondent à un nouveau modèle GMM du locuteur X dont les moyennes sont données par :

$$m_k^{(X)^{norm}} = \frac{D_{ref}}{D_E(X,\Omega)} m_k^{(X)} + \left(1 - \frac{D_{ref}}{D_E(X,\Omega)}\right) m_k^{(\Omega)} \quad , \tag{5.26}$$

en considérant les poids et matrices de covariance fixes.

Le schéma de normalisation proposé permet de façon très simple d'homogénéiser les modèles de locuteurs vis-à-vis de leur distance $D_E(X,\Omega)$, comme l'adaptation D-MAP le fait vis-à-vis des distances $KL2_X$. Elle peut s'interpréter comme un "effacement" de l'information d'éloignement des modèles par rapport au modèle du monde. Elle repose sur l'hypothèse que cette information d'éloignement est perturbante pour la vérification, car trop dépendante du contenu informationnel des énoncés, et que l'information caractéristique du locuteur est essentiellement présente dans la "direction" que prend le modèle dans l'espace des modèles.

Notons que cette normalisation peut également s'appliquer sur les modèles appris sur les tests, ce qui pourrait engendrer un effet de type "T-norm" sur les performances du système. Dans la suite nous nommerons cette normalisation des modèles la "M-norm" (Model normalization).

5.3.5 Mise en oeuvre et résultats

5.3.5.1 Corrélation entre D_E^2 et $KL2$

La figure 5.13 montre la corrélation entre $D_E^2(X,\Omega)$ et $KL2_X$ pour les locuteurs femmes de l'évaluation NIST 2004 (condition "1side train" [Przybocki et al.04]), avec des modèles GMMs adaptés d'un modèle du monde (moyennes seulement) et non-normalisés. Elle fait apparaître une très forte corrélation entre ces deux grandeurs (coefficient de corrélation supérieur à 0,99).

Le tracé de la première diagonale sur ce graphique (droite en pointillés) montre que les mesures D_E^2 sont toujours supérieures aux distances $KL2$, confirmant l'inégalité théorique de l'équation 5.20. La relation quasi-déterministe de proportionnalité existant entre ces deux grandeurs indique que la mesure D_E^2 peut remplacer avantageusement la distance $KL2$ comme mesure de similarité entre des GMMs dont les moyennes sont adaptées d'un modèle du monde. Notamment D_E^2 peut être utilisée en remplacement de $KL2$ dans les techniques de normalisation D-norm et D-MAP.

5.3.5.2 Discussion sur les temps de calcul

Nous discutons dans ce paragraphe des temps de calcul nécessaires pour un accès, lorsque l'on utilise le cadre de calcul des scores défini aux sections précédentes.

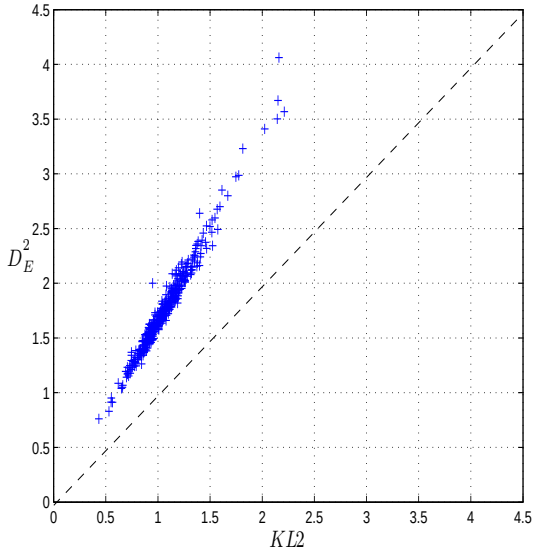


FIG. 5.13 – *Relation entre les mesures* $D_E{}^2(X,\Omega)$ *et* $KL2_X$ *(corpus femmes NIST'04 1side-train)*

Calcul des scores bruts

Le calcul des scores $S_X(\mathcal{Y})$ dans l'espace des modèles nécessite l'estimation des GMMs de locuteurs $p(y|X)$ mais aussi d'un modèle GMM $p(y|Y)$ pour chaque énoncé de test. Nous estimons ces modèles de tests de la même manière que les modèles de locuteurs, par adaptation Bayésienne des moyennes du modèle du monde sur les données correspondantes. Comme dans [Reynolds et al.00], nous faisons une seule itération de l'algorithme EM pour adapter les modèles. Il a été observé que cela n'occasionne pas de perte de performance par rapport à une procédure d'adaptation des modèles où l'algorithme EM est itéré plusieurs fois. En outre cela permet d'obtenir le score $S_X(\mathcal{Y})$ plus rapidement.

Le temps nécessaire à l'obtention de ce score correspond approximativement au temps d'estimation du modèle de test $p(y|Y)$. En effet, une fois ce modèle obtenu, le temps de calcul des mesures $D_E^2(Y,\Omega)$ et $D_E^2(Y,X)$ est négligeable. Pour estimer le modèle $p(y|Y)$ avec une itération d'EM, il est nécessaire de calculer la vraisemblance de chaque trame pour chaque Gaussienne du modèle du monde. Lorsque ces vraisemblances sont obtenues, le temps de calcul des statistiques nécessaires à la mise à jour des paramètres du modèle est négligeable.

En résumé, le temps d'estimation de $p(y|Y)$ (avec une itération d'EM), et donc d'obtention du score $S_X(\mathcal{Y})$, est approximativement équivalent à la moitié du temps nécessaire au calcul d'un score LLR lors d'un test (calcul de la vraisemblance de chaque

trame pour chaque Gaussienne dans le modèle du monde *et* le modèle du locuteur). Si ce score LLR est estimé avec la technique N-Best (voir [Reynolds et al.00]), le temps d'obtention de $S_X(\mathcal{Y})$ dans l'espace des modèles est environ équivalent au temps de calcul d'un score LLR/N-best.

Calcul de la T-norm

L'application de la T-norm lors d'un test implique de calculer un grand nombre de scores (une cinquantaine au moins) donnés par l'énoncé de test pour un ensemble de modèles imposteurs. Lorsque l'on utilise le score LLR, le temps d'estimation des paramètres de normalisation rallonge de façon non-négligeable le temps nécessaire à la vérification, même lorsqu'on utilise la technique N-Best. Dans le cadre de calcul de scores que nous proposons, ce temps d'estimation des paramètres de normalisation T-norm est très court. En effet, une fois que le modèle du test $p(y|Y)$ est obtenu, les seuls calculs nécessaires sont des calculs de distances D_E entre ce modèle et les modèles imposteurs de la T-norm. Le calcul de ces distances étant très rapide, l'application de la T-norm dans ce cadre ne rallonge que de façon négligeable le temps de test.

5.3.5.3 Performances

Nous comparons les performances obtenues en utilisant le cadre de calcul des scores dans l'espace des modèles et celles données par le système de base utilisant des scores LLR. La comparaison est effectuée sur un sous-ensemble (locuteurs femmes) de l'évaluation NIST 2004, dans la condition "1side-train/1side-test", c'est-à-dire avec une conversation téléphonique complète à l'apprentissage et une conversation téléphonique complète pour le test.

La figure 5.14 représente les courbes DET du système de base (*scores_LLR*) et du système utilisant le nouveau cadre de calcul des scores (*scores_*D_E) lorsqu'aucune normalisation n'est utilisée (No_norm). Le système utilisant les scores basés sur la distance D_E semble mal se comporter dans ce cas-là, engendrant une forte perte de performance. Ces résultats peuvent peut-être s'expliquer par la remarque faite à la section 5.3.4.2, à savoir que les différences de contenu des énoncés sur lesquels sont appris les modèles de locuteurs et de tests entraînent des disparités d'éloignement de ces modèles par rapport au modèle du monde. Cela peut occasionner des erreurs de reconnaissance, lors d'accès clients, spécifiquement dues à ce type de disparités.

La figure 5.15 représente les courbes DET du système à *scores_LLR* utilisant la D-norm et du système *scores_*D_E avec application de la M-norm. Les deux normalisations utilisées ici ne nécessitent aucune donnée spécifique externe. D'après ces courbes, l'application de la M-norm sur le système *scores_*D_E amène un gain de performance considérable. Les deux systèmes se comportent de façon comparable pour les points de fonctionnement à faible taux de faux rejets. Par contre, pour les points à faible taux de fausse alarme, le système utilisant le calcul des scores dans l'espace des modèles et la normalisation de modèles M-norm surpasse le système de base. Cela confirme l'effet "T-norm" que nous envisagions, apporté par l'application de la normalisation M-norm sur les modèles de test.

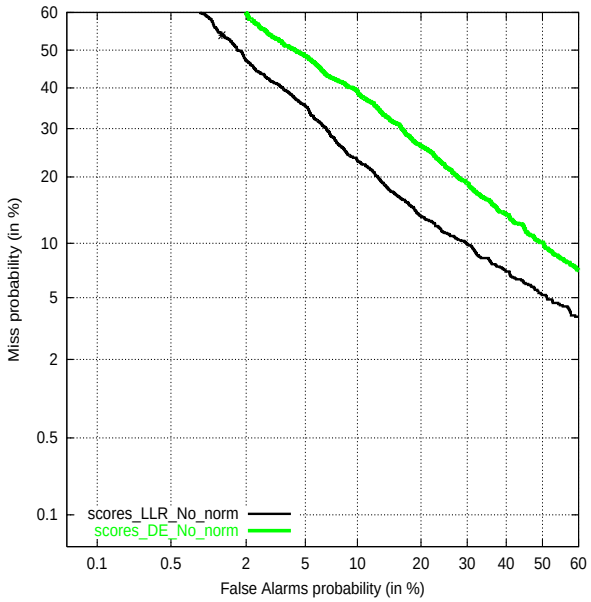


FIG. 5.14 – *Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR (scores_LLR) et des scores basés sur la distance Euclidienne D_E (scores_D_E), sans normalisation*

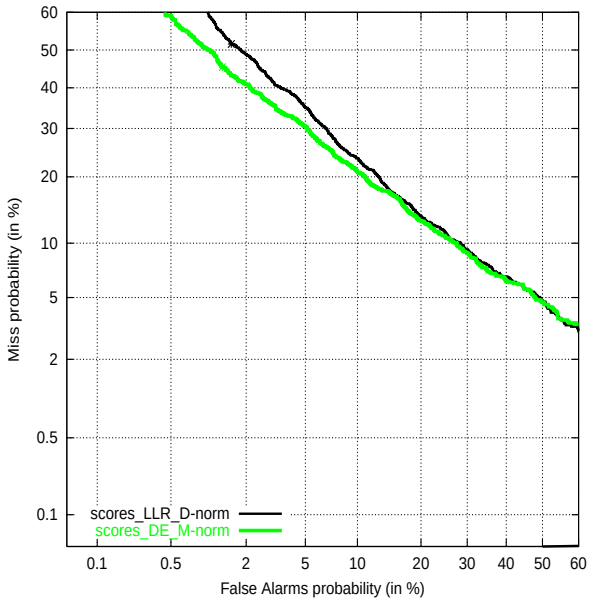


FIG. 5.15 – *Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR (scores_LLR) avec normalisation D-norm, et des scores basés sur la distance Euclidienne D_E (scores_D_E) avec normalisation M-norm*

Sur la figure 5.16 sont tracées les courbes DET des systèmes *scores_LLR* et *scores_*D_E avec application de la normalisation de score T-norm, sans D-norm ni M-norm. L'application de la T-norm sur le système *scores_*D_E permet d'améliorer les résultats par rapport à la courbe DET de la figure 5.14, mais ne suffit pas à rattraper l'écart de performance constaté sur cette figure entre les deux systèmes.

Enfin, la figure 5.17 représente les courbes DET du système *scores_LLR* avec application de la DT-norm, et du système *scores_*D_E utilisant la normalisation de modèles M-norm, suivie d'une normalisation de scores T-norm (M+T-norm). Sur cet exemple, le système *scores_*D_E utilisant la M+T-norm obtient des performances légèrement meilleures que le système *scores_LLR* avec DT-norm.

	$\min C_{det}$ ($\times 10^{-3}$)		EER (%)	
	scores_LLR	*scores_*D_E	*scores_LLR*	*scores_*D_E
No-norm	67,09	76,78	16,35	23,53
D-norm	67,10	-	16,29	-
M-norm	-	58,75	-	16,17
T-norm	56,86	67,53	15,61	20,58
DT-norm	60,27	-	15,91	-
M+T-norm	-	**54,61**	**15,04**	-

TAB. 5.4 – *Points* $\min C_{det}$ *et EER du système IRISA/ELISA 2004 avec scores LLR et scores basés sur les distances* D_E*, avec différents schémas de normalisation*

L'ensemble des points $\min C_{det}$ et EER pour les systèmes présentés ci-dessus sont résumés dans le tableau 5.4. Un point remarquable dans ce tableau est que les performances du système *scores_*D_E avec la normalisation de modèle M-norm, c'est-à-dire n'utilisant pas de données externes, sont proches de celles du système *scores_LLR* avec utilisation de la T-norm, qui quant à elle nécessite l'utilisation d'un ensemble de modèles imposteurs. La normalisation M-norm, qui est appliquée sur les modèles de locuteurs et de test, semble donc compenser certains biais liés aux données de test elles-mêmes. Cela indique qu'il est possible de faire de la normalisation de test sans utiliser de données additionnelles, et que les informations nécessaires à cette normalisation sont contenues dans le modèle de test lui-même.

5.3.6 Conclusion et perspectives

Le nouveau cadre de calcul des scores en VAL que nous proposons a montré des performances et des propriétés prometteuses. Ce calcul de scores intervient dans un espace défini par les paramètres des modèles eux-mêmes. Cet espace est Euclidien

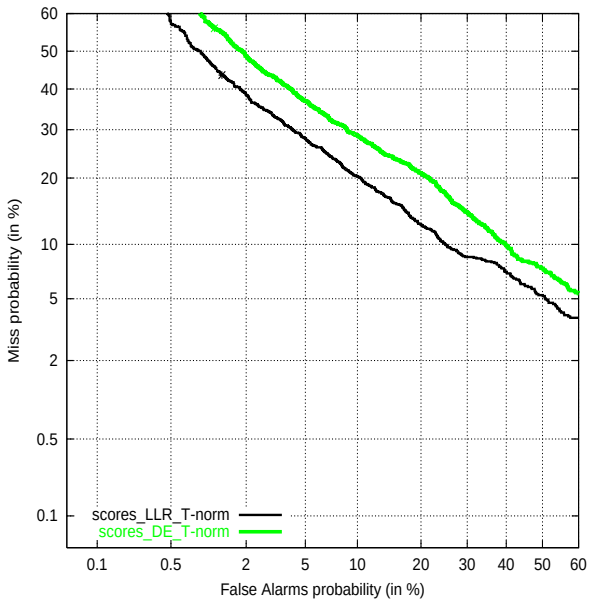


FIG. 5.16 – *Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR (scores_LLR) avec normalisation T-norm, et des scores basés sur la distance Euclidienne D_E (scores_D_E) avec normalisation T-norm*

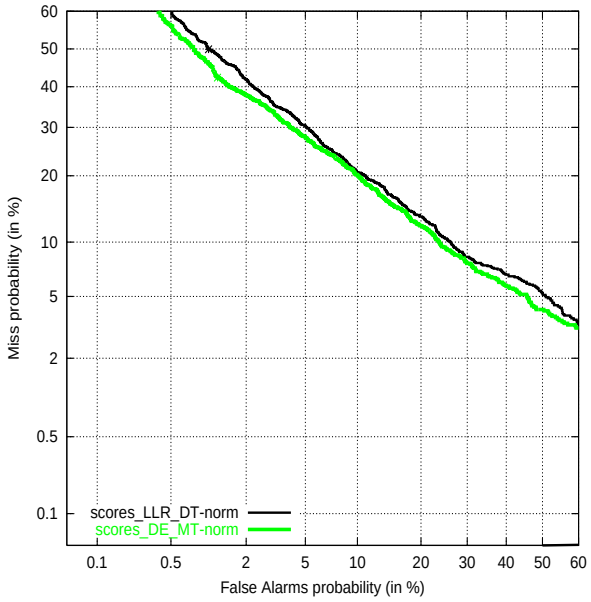


FIG. 5.17 – *Courbes DET du système IRISA/ELISA 2004 utilisant des scores LLR (scores_LLR) avec normalisation DT-norm, et des scores basés sur la distance Euclidienne D_E (scores_D_E) avec normalisation M+T-norm*

et autorise donc des manipulations des modèles par des transformations simples de géométrie. Cela a permis de mettre en place la M-norm, une technique de normalisation de modèles de locuteurs et de tests qui ne nécessite pas de données additionnelles. Dans ce contexte (absence de données externes) le système utilisant le calcul des scores dans l'espace des modèles a obtenu des performances meilleures que le système de base pour la région des points de fonctionnement à faible taux de fausses acceptations.

D'autre part, lorsque des modèles imposteurs sont disponibles pour appliquer la normalisation T-norm (en complément de la M-norm), le système utilisant les scores calculés dans l'espace des modèles a obtenu des performances légèrement meilleures que celles du système de base avec T-norm. De plus, l'application de la T-norm ne rallonge que de façon négligeable le temps de vérification dans l'espace des modèles.

L'espace Euclidien défini dans cette section, pour la représentation des modèles et le calcul des scores de vérification, nous fait envisager des techniques de normalisation vis-à-vis des variabilités d'environnement matériel, telles que le type de microphone ou le canal de transmission. En effet des méthodes d'analyse en composante principale (ACP) peuvent permettre d'identifier dans cet espace, des directions particulièrement affectées par les variabilités d'environnement, et d'autres plus ou moins indépendantes de ces variabilités (cf. illustration figure 5.18). Il serait alors concevable de supprimer les directions perturbées du calcul du score afin qu'il soit le plus indépendant possible des changements de microphone ou de canal.

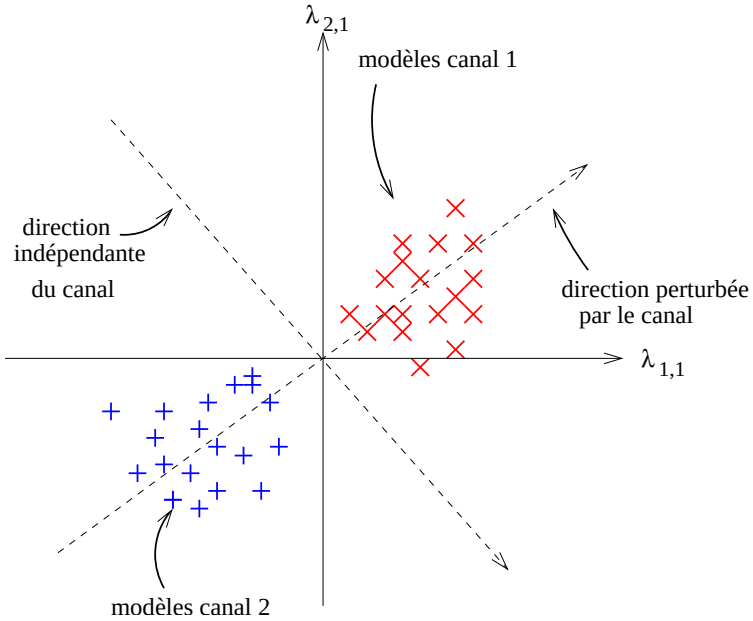


FIG. 5.18 – *Identification d'une direction dépendante du canal dans l'espace des modèles (illustration)*

On peut encore imaginer d'autres techniques de normalisation basées sur des transformations dans l'espace des modèles, par exemple un "model mapping", version du "feature mapping" agissant sur les modèles par projection vers un espace indépendant du canal.

Chapitre 6

Contributions connexes

Nous présentons dans ce chapitre des contributions connexes de nos travaux en VAL. En particulier, les mesures de similarité entre modèles de locuteurs que nous avons définies dans le cadre de la VAL ont été utilisées pour d'autres tâches de reconnaissance et de caractérisation du locuteur.

La section 6.1 présente des travaux effectués lors de la première phase de la campagne d'évaluation ESTER (Évaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) dans des tâches de suivi de locuteur et d'organisation en locuteur de documents audio. Nos travaux en VAL sont directement appliqués pour la tâche de suivi de locuteur que nous traitons comme une tâche de segmentation suivie d'une tâche de détection de locuteur sur les segments. Pour la tâche d'organisation en locuteur, nous utilisons un schéma de regroupement hiérarchique ascendant des segments de parole basé sur une mesure de similarité entre modèles de locuteurs : la distance Euclidienne D_E entre GMMs adaptés (cf. section 5.3).

Dans la section 6.2 nous présentons des applications des mesures de similarité entre GMMs que nous avons utilisées en VAL, pour une tâche de caractérisation des locuteurs. Il s'agit d'une tâche de sélection d'un sous-ensemble restreint de locuteurs représentatifs parmi un ensemble plus important. Les résultats obtenus montrent que ces mesures sont significatives pour la caractérisation du locuteur.

6.1 Contribution en suivi de locuteur et organisation en locuteur

Les travaux en suivi de locuteur et organisation en locuteur présentés dans cette section ont été effectués dans le cadre de la phase 1 de la campagne d'évaluation ESTER. Les documents audio traités sont des journaux d'informations radiophoniques en langue française. Ces travaux ont été réalisés en collaboration avec Guillaume GRAVIER, chargé de recherche CNRS, et Michaël BETSER, ingénieur expert, tous deux de l'équipe METISS de l'IRISA.

6.1.1 Contexte : campagne d'évaluation ESTER

La campagne ESTER [Gravier et al.04], dont la première phase s'est déroulée de juin 2003 à janvier 2004, est dédiée à l'évaluation des systèmes de transcription enrichie et de description automatique d'émissions radiophoniques. Elle est organisée conjointement, dans le cadre du projet EVALDA [1], par les acteurs scientifiques regroupés dans l'AFCP (Association Francophone de la Communication Parlée), le centre d'Arcueil de la DGA (Direction Générale pour l'Armement) et l'organisme ELDA (Évaluations and Language resources Distribution Agency) de collecte et distribution de corpus de parole. Cette campagne est ouverte à tous les acteurs du domaine, académiques ou industriels, sur la base du bénévolat.

On désigne par transcription enrichie le fait d'ajouter à une transcription orthographique de la parole des informations sur le contexte dans lequel elle a été prononcée. Ces informations peuvent être relatives par exemple au locuteur courant, aux tours de parole, à la présence d'événements sonores particuliers ("jingles", musique, ...) voire même aux thèmes abordés. Les émissions radiophoniques visées par l'évaluation ESTER sont essentiellement des journaux d'informations en langue française. Les objectifs de cette campagne sont d'une part de dynamiser la recherche en France sur les systèmes de transcription orthographique et de recherche d'informations dans les documents sonores, et d'autre part de proposer un cadre bien calibré pour mesurer les performances de tels systèmes avec une possibilité de comparaison directe avec d'autres campagnes d'évaluations comme NIST Rich Transcrition (RT) [2] par exemple. En outre, la campagne ESTER permettra à terme de proposer à la communauté scientifique un corpus annoté conséquent adapté aux types de tâches concernées par la campagne.

Le déroulement de la campagne est prévu en deux phases. La première phase qui a déjà eu lieu (juin 2003 à janvier 2004) est un test à blanc destiné à valider, sur un corpus restreint, les protocoles proposés et à analyser les premiers résultats obtenus par les laboratoires participants. La deuxième phase, qui constitue la campagne d'évaluation "officielle", se déroule de mars à décembre 2004 avec un corpus plus conséquent.

1. `http://www.elda.fr/rubrique69.html`
2. `http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf`

Les tâches proposées lors de la deuxième phase s'organisent en trois catégories principales : la transcription (T), la segmentation (S) et l'extraction d'information (E). L'ensemble des tâches est résumé dans le tableau 6.1. L'IRISA a participé à trois de ces tâches (TRS, SES et SVL) lors de la première phase et compte participer également à la tâche SRL lors de la seconde phase.

catégorie	dénomination	intitulé de la tâche
T	TRS	transcription orthographique
T	TTR	transcription temps réel
S	SES	suivi d'événements sonores
S	SRL	segmentation et regroupement en locuteurs
S	SVL	suivi de locuteurs
S	SIL	segmentation en locuteur interactive
E	EN	détection d'entités nommées
E	SD	segmentation thématique de documents
E	ST	suivi de thèmes
E	QR	recherche d'information (question/réponse)

TAB. 6.1 – *Dénomination des tâches de la campagne d'évaluation ESTER*

Dans cette section nous décrivons les travaux relatifs aux tâches de reconnaissance du locuteur uniquement, à savoir les tâches SVL (section 6.1.2) et SRL (section 6.1.3). Pour la première phase de l'évaluation ESTER, 40 heures de données étaient disponibles, divisées en trois corpus distincts :

- un corpus d'apprentissage ("train") comprenant 19 heures 40 minutes de France Inter et 11 heures de RFI;
- un corpus de développement ("dev") comprenant 2 heures 40 minutes de France Inter et 2 heures de RFI;
- un corpus de test ("test") comprenant 2 heures 40 minutes de France Inter et 2 heures de RFI.

6.1.2 Suivi de locuteurs : tâche ESTER-SVL

6.1.2.1 Description de la tâche

La tâche de suivi de locuteur SVL consiste à détecter les zones du document radiophonique où un locuteur donné, connu à l'avance, est présent. Le protocole spécifie une liste de locuteurs à suivre pour lesquels des données d'apprentissage sont disponibles dans le corpus "train". L'ensemble des données disponibles pour chaque locuteur dans

le corpus "train" peut être utilisé pour construire les références caractéristiques des locuteurs.

Pour tous les locuteurs à suivre (33 femmes et 63 hommes sur le corpus "dev") la quantité de données d'apprentissage est au moins égale à 1 minute, mais cette quantité peut être très variable d'un locuteur à un autre (certains locuteurs disposent de plus d'une heure de données d'apprentissage).

Les performances sont mesurées en terme de taux d'erreur de détection, moyenné sur l'ensemble des locuteurs cibles. Pour un locuteur cible X, ce taux d'erreur est défini par :

$$\%Err = 100 \times \frac{T(X|\bar{X}) + T(\bar{X}|X)}{T} \quad , \tag{6.1}$$

où $T(X|\bar{X})$ désigne le temps de fausse alarme (détection à tort) du locuteur X, $T(\bar{X}|X)$ le temps de non-détection de ce locuteur et T le temps total d'analyse.

Cette métrique présente cependant le gros désavantage que les temps de présence $T(X)$ et d'absence $T(\bar{X})$ pour un locuteur X (avec $T(X)+T(\bar{X}) = T$) sur les documents étudiés sont en général très déséquilibrés avec $T(X) << T(\bar{X})$. Par conséquent, un système répondant systématiquement que le locuteur X est absent sur le document, aurait un taux d'erreur faible pour ce locuteur, donné par

$$\%Err = 100 \times \frac{T(X)}{T} \quad ,$$

et un taux d'erreur global sur l'ensemble des locuteurs qui serait faible également.

Cette métrique pour la mesure des performances de la tâche SVL sera probablement remplacée par une autre métrique rendant mieux compte des types d'erreurs lors de la seconde phase de la campagne. Dans le cadre de nos travaux, nous proposons en complément de la métrique $\%Err$ une mesure des performances correspondant à la moyenne des taux de fausse alarme $\mathcal{T}_{FA}$ et de non-détection $\mathcal{T}_{Miss}$ définis de la manière suivante :

$$\mathcal{T}_{FA} = 100 \times \frac{T(X|\bar{X})}{T(\bar{X})} \quad , \tag{6.2}$$

$$\mathcal{T}_{Miss} = 100 \times \frac{T(\bar{X}|X)}{T(X)} \quad . \tag{6.3}$$

La métrique correspondante est l'$HTER$ (Half Total Error Rate) :

$$HTER = \frac{\mathcal{T}_{FA} + \mathcal{T}_{Miss}}{2} \quad . \tag{6.4}$$

Les taux $\mathcal{T}_{FA}$ et $\mathcal{T}_{Miss}$, qui correspondent respectivement à des pourcentages de temps de fausse alarme et de non-détection, peuvent également servir à tracer des courbes DET du système à condition qu'un score de détection soit fourni pour chaque segment et chaque locuteur. Ces courbes DET rendent compte alors de la variation des taux $\mathcal{T}_{FA}$ et $\mathcal{T}_{Miss}$ lorsque l'on fait varier un seuil unique de détection.

6.1.2.2 Mise en oeuvre

La tâche de suivi des locuteurs est traitée séquentiellement par une première tâche de segmentation aveugle des documents, suivie d'une tâche de détection du locuteur sur chaque segment et pour chaque locuteur cible.

Segmentation

Le but de la segmentation aveugle est de détecter des changements abruptes des caractéristiques du signal audio afin d'identifier les frontières des tours de parole et les passages d'un type d'événement sonore à un autre (par exemple parole/musique). Le processus de segmentation aveugle utilisé est basé sur le critère BIC (Bayesian Information Criterion) [Chen et al.98]. Nous rappelons ici les principes de cette approche.

D'une manière générale, étant donné un segment $\mathcal{Y} = \{y_1, ..., y_N\}$ composé de N trames et un modèle de ce segment défini par ses paramètres Λ, la mesure d'information pour le critère BIC est définie par :

$$\mathrm{BIC}_\Lambda = \ln p(\mathcal{Y}|\Lambda) - \lambda \frac{\#\Lambda}{2} \ln N \ , \tag{6.5}$$

où $p(\mathcal{Y}|\Lambda)$ est la vraisemblance du segment étant donné le modèle, $\#\Lambda$ est le nombre de paramètres libres dans le modèle et λ est un paramètre de réglage théoriquement égal à 1.

La détection d'un changement abrupt par l'intermédiaire du critère BIC se fait en calculant la différence entre la mesure BIC sous l'hypothèse qu'il y a un changement dans la séquence et la mesure BIC sous l'hypothèse que la séquence est homogène. En supposant des modèles sous-jacents Gaussiens, un changement est détecté entre deux segments $a = \{y_1 ... y_{n-1}\}$ et $b = \{y_n ... y_N\}$ si :

$$\Delta\mathrm{BIC}(n) = R(n) - \frac{\lambda}{2} \left(d + \frac{d(d+1)}{2} \right) \ln N \tag{6.6}$$

est négatif. d est la dimension des vecteurs acoustiques et $R(n)$ est un log-rapport de vraisemblance donné par :

$$R(n) = N \ln \sqrt{|\Sigma_{ab}|} - N_a \ln \sqrt{|\Sigma_a|} - N_b \ln \sqrt{|\Sigma_b|} \ . \tag{6.7}$$

Dans cette équation, N_x est le nombre de trames dans le segment x, Σ_x est l'estimation au maximum de vraisemblance de la matrice de covariance des données du segment x et ab représente la concaténation du segment a et du segment b.

Le paramètre λ dans l'équation 6.6 permet en pratique de régler la sensibilité de détection des ruptures et d'adopter donc un compromis sur la longueur moyenne des segments détectés. Une sensibilité trop faible conduira à des segments relativement longs mais certains changements dans les tours de parole des différents locuteurs pourront être manqués. Au contraire, une forte sensibilité entraînera une sur-segmentation du flux sonore et conduira à des segments courts, difficilement exploitables ensuite pour les tâches de reconnaissance. La longueur moyenne des segments obtenus par le système

IRISA lors de la phase de segmentation des documents radiophoniques est d'environ 3 secondes.

Détection du locuteur

Après segmentation, les segments sont étiquetés en parole ou musique à l'aide de modèles de chacune de ces deux classes (tâche SES, non décrite ici). Sur chaque segment étiqueté "parole" une tâche de détection du locuteur est effectuée pour chacun des locuteurs cibles, en utilisant le système de VAL de l'IRISA de l'année 2003.

Des modèles GMMs du monde dépendants du genre à 512 Gaussiennes diagonales sont estimés au maximum de vraisemblance à partir de données de parole issues du corpus "train" (environ 3h30 de parole pour les hommes et 1h pour les femmes) en excluant les données correspondant aux locuteurs cibles. Les modèles des locuteurs cibles sont adaptés de ces modèles du monde suivant le schéma d'adaptation Bayésienne décrit à la section 2.3.2. Seules les moyennes des GMMs de locuteurs sont adaptées. Des modèles d'imposteurs sont également estimés à partir du corpus "train" afin d'appliquer une normalisation de score T-norm (75 modèles imposteurs d'hommes et 50 modèles d'imposteurs femmes).

6.1.2.3 Résultats obtenus

Le tableau 6.2 donne les taux minimum d'erreur $\%Err$ et d'$HTER$ obtenus pour la tâche SVL sur le corpus "dev" par le système de suivi de locuteur IRISA 2003. Les valeurs $\min \%Err$ et $\min HTER$ pour différents types de normalisations de score sont donnés, avec également les taux d'erreurs $\mathcal{T}_{FA}$ et $\mathcal{T}_{Miss}$ correspondants . D'une manière générale, ces normalisations de scores apportent peu ou pas d'amélioration par rapport au système de base sans normalisation (No-norm). Cela est probablement dû au fait que les conditions d'acquisition du signal (prise de son en studio pour la majorité du document) sont bien calibrées et relativement constantes d'un document à un autre. De ce fait, il y a peu de disparités entre conditions de test et conditions d'apprentissage.

	$\min \%Err$ $(\mathcal{T}_{FA}$ / $\mathcal{T}_{Miss})$	$\min HTER$ $(\mathcal{T}_{FA}$ / $\mathcal{T}_{Miss})$
No-norm	$0,16$ (0,03 / 20,80)	$7,14$ (0,64 / 13,63)
D-norm	$0,16$ (0,03 / 20,45)	$7,13$ (0,63 / 13,62)
T-norm	$0,15$ (0,02 / 20,59)	$6,47$ (0,74 / 12,20)
DT-norm	$0,14$ (0,03 / 18,86)	$6,30$ (0,69 / 11,90)

TAB. 6.2 – $\min \%Err$ *et* $\min HTER$ *du système de suivi de locuteurs IRISA 2003 avec différentes normalisations de scores (corpus "dev", évaluation ESTER-SVL)*

Les taux d'erreur $\%Err$ obtenus sont très bas (de 0,16% à 0,14%) avec notamment des taux de fausses acceptations $\mathcal{T}_{FA}$ correspondants extrêmement faibles (autour de 0,03%). Les taux d'$HTER$ sont autour de 7% et largement dominés par les taux de

non-détection. Cela est dû au comportement particulier du système pour lequel ce taux de non-détection varie peu lorsque le taux de fausses acceptations augmente, ce qui se traduit par une courbe DET à forte tendance horizontale comme on peut le voir sur la figure 6.1. Sur cette figure ne sont tracées que les courbes DET correspondant au système sans normalisation de scores (No-norm) et celui avec la normalisation de scores donnant les meilleurs résultats (DT-norm). Les courbes DET des différents systèmes sont de toute façon très proches, parfois même quasiment confondues.

Pour conclure sur la tâche préliminaire ESTER-SVL, notons que les taux d'erreur pour cette tâche de suivi de locuteurs rendent compte à la fois des performances de détection du système de VAL mais également de la précision de la segmentation automatique. En effet, si cette segmentation est imprécise, ou si des changements sont purement et simplement manqués, les segments résultants ne sont pas homogènes vis-à-vis des locuteurs y intervenant. Certaines parties d'un segment peuvent avoir été prononcées par un locuteur donné et d'autres parties par un locuteur différent. La décision issue de la phase de détection d'un locuteur étant prise globalement sur l'ensemble d'un segment, la présence successive de plusieurs locuteurs sur un même segment entraîne forcément des temps d'erreur de détection (fausse alarme ou non-détection) pour chacun de ces locuteurs. Ces erreurs sont dues à l'étape de segmentation et ne peuvent être corrigées par l'étape de détection. Il est donc important d'avoir une étape de segmentation aussi précise que possible pour minimiser les taux d'erreur globaux pour la tâche de suivi de locuteurs.

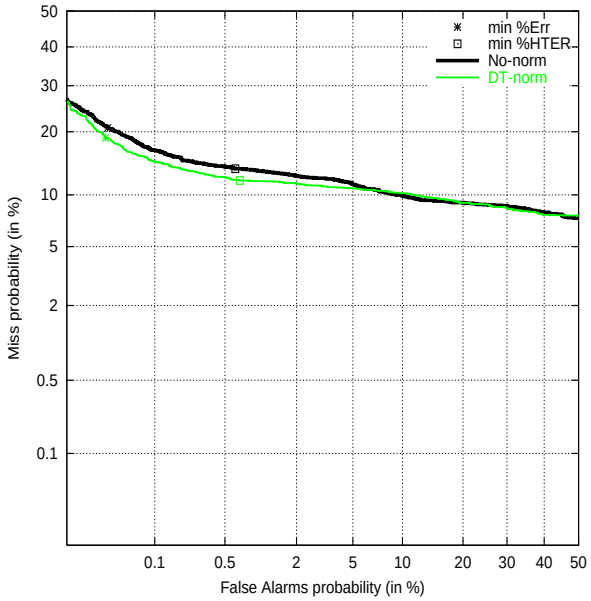


FIG. 6.1 – *Courbes DET du système de suivi de locuteur IRISA 2003 sans normalisation de scores (No-norm) et avec normalisation DT-norm (corpus "dev", évaluation ESTER-SVL)*

6.1.3 Organisation en locuteur : tâche ESTER-SRL

6.1.3.1 Description de la tâche

La tâche ESTER-SRL est une tâche d'organisation en locuteurs des documents radiophoniques. Contrairement à la tâche SVL, on ne dispose pas ici de données d'apprentissage concernant les locuteurs présents sur le document, et on ne connaît pas non plus le nombre de locuteurs y intervenant. L'organisation du document suivant les différents intervenants doit donc se faire entièrement en aveugle. Le système doit retourner une liste descriptive des tours de parole avec un identifiant arbitraire pour chaque locuteur différencié.

La métrique utilisée pour la mesure de performance concernant cette tâche est le taux d'erreur défini comme la somme des taux de parole non-détectée, de fausse détection de parole et de mauvaise détection. Le taux de parole non-détectée correspond aux portions de parole pour lesquelles aucun identifiant de locuteur n'a été étiqueté. Le taux de fausse détection de parole correspond aux portions sans parole ayant été étiquetées avec un identifiant de locuteur. Le taux de mauvaise détection correspond aux erreurs sur les identités (arbitraires) des locuteurs. Au moment de l'évaluation des performances à partir d'une vérité terrain (document étiqueté manuellement en locuteur), une correspondance entre noms de locuteurs et identifiants arbitraires fournis par le système est établie, suivant le meilleur appariement possible (vis-à-vis du taux d'erreur).

Cette façon de mesurer les performances est identique à celle utilisée au cours des campagnes d'évaluation NIST RT, pour les tâches d'organisation en locuteur.

6.1.3.2 Description du système

Nous traitons la tâche d'organisation en locuteur comme une tâche de segmentation aveugle suivie d'une étape de classification hiérarchique ascendante des segments. Ce regroupement des segments doit être contrôlé par un critère d'arrêt qui est censé déterminer le nombre optimal de locuteurs différents sur le document.

Segmentation

L'étape de segmentation est en tous points identique à celle décrite à la section 6.1.2.2 dans le cadre de la tâche SVL. Elle est basée sur l'utilisation du critère BIC avec des modèles de segments Gaussiens.

Regroupement des segments

L'étape de regroupement des segments est basée sur une mesure de similarité entre des modèles de cluster de segments. Au début de l'algorithme, chaque segment constitue un cluster en lui-même et à chaque itération, les clusters les plus proches suivant la mesure de similarité définie sont regroupés pour former un nouveau cluster.

Une approche classique consiste à définir des modèles Gaussiens des clusters et à

utiliser le critère BIC, à la fois comme mesure de similarité entre clusters et comme critère d'arrêt de l'algorithme. Dans ce cas, les deux clusters les plus proches suivant le rapport de vraisemblance[3] défini en 6.7 sont regroupés si la variation du critère BIC 6.6 est négative. Si cette variation est positive, l'algorithme est arrêté.

L'approche que nous proposons consiste à utiliser des GMMs plutôt que des modèles mono-Gaussiens afin de modéliser plus finement la distribution des vecteurs acoustiques, en particulier lorsque les clusters contiennent une grande quantité de parole (vers la fin du regroupement). Cependant, les GMMs ont un nombre plus important de paramètres et leur estimation au maximum de vraisemblance peut être peu fiable quand les clusters contiennent une petite quantité de parole (début du regroupement). Nous utilisons par conséquent l'adaptation Bayésienne pour adapter les GMMs des clusters à partir d'un modèle *a priori*.

Estimation des modèles de clusters

Le modèle *a priori* est estimé au maximum de vraisemblance sur l'ensemble des portions de parole du document traité. Cela permet d'avoir un modèle *a priori* adapté aux conditions générales d'acquisition du signal rencontrées dans le document. Nous appellerons ce modèle *a priori* le DSBM (Document Speech Background Model). Seules les moyennes des GMMs de clusters sont adaptées de celles du DSBM selon la formule donnée à l'équation 2.8.

Mesure de similarité et critère d'arrêt

Pour l'algorithme de regroupement des clusters, nous utilisons la distance Euclidienne D_E définie à la section 5.3 du chapitre précédent. Nous rappelons que cette distance est une mesure de similarité entre des GMMs dont les moyennes ont été adaptées d'un modèle *a priori*. Elle peut être rapidement obtenue à partir des estimations des modèles de clusters puisqu'elle s'exprime comme une fonction simple des paramètres des GMMs. A chaque itération de l'algorithme, les deux clusters ayant une distance inter-cluster minimale sont regroupés et un modèle GMM de ce nouveau cluster est estimé. Nous utilisons également la distance D_E comme critère d'arrêt de notre algorithme de regroupement. Ainsi, si la distance minimale entre clusters est supérieure à un seuil, les clusters correspondants sont supposés appartenir à des locuteurs différents et l'algorithme de regroupement est stoppé.

6.1.3.3 Résultats obtenus

Le nombre de composantes des GMMs (et les facteurs de confiance τ correspondants) ainsi que le seuil de comparaison de la distance minimale D_E pour le critère d'arrêt ont été optimisés sur le corpus "dev". Les taux d'erreur obtenus pour ce corpus sont résumés dans le tableau 6.3, pour chacun des 6 documents du corpus et globalement sur l'ensemble des documents. La partie haute du tableau donne les taux d'erreur lorsque la segmentation est effectuée suivant la référence manuelle et la partie basse du

3. ce rapport de vraisemblance est généralement appelé "distance BIC"

tableau correspond à une segmentation automatique par critère BIC. Les résultats pour la méthode proposée sont donnés en fonction du nombre de Gaussiennes des GMMs de clusters (de 8 à 64) après optimisation globale (sur l'ensemble des documents) du seuil d'arrêt. Les résultats obtenus avec la méthode de regroupement par critère BIC sont également fournis à titre de comparaison.

dev / ref	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	Global
8GMM	8.44	3.87	9.07	**6.97**	5.88	7.76	6.19
16GMM	**5.04**	**2.45**	12.59	9.47	5.72	**7.37**	**6.07**
32GMM	8.26	3.64	8.68	9.47	5.24	10.76	6.94
64GMM	8.65	8.73	7.07	11.04	12.25	18.45	11.84
BIC	6.90	13.94	**6.33**	7.12	**3.85**	19.25	10.27
dev / bic	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	Global
8GMM	10.67	10.71	9.7	10.26	12.53	13.70	11.64
16GMM	9.45	**9.75**	11.46	12.99	10.18	12.44	**10.50**
32GMM	10.49	11.41	13.85	18.19	**8.07**	**10.99**	11.12
64GMM	14.07	10.82	**8.32**	11.65	14.83	20.79	13.58
BIC	**8.77**	16.29	8.76	**8.76**	8.10	26.16	13.69

TAB. 6.3 – *Tâche SRL : taux d'erreur de classification obtenu pour chaque document (colonne 2 à 7) et globalement (colonne 8) sur le corpus "dev" avec segmentation manuelle (haut) et automatique (bas).*

D'après les résultats rapportés dans ce tableau, la méthode de regroupement à base de GMMs que nous proposons apporte un gain de performance significatif sur le corpus "dev" par rapport à la méthode BIC, quelle que soit la segmentation considérée (manuelle ou automatique). Les meilleures performances sont obtenues avec des GMMs à 16 composantes. Cependant, une analyse détaillée des résultats montre que l'amélioration vient essentiellement de deux documents (Inter 2 et RFI 2), tandis que pour les autres documents, les performances sont similaires, parfois à l'avantage de la méthode BIC. De plus, le gain de performance obtenu lorsqu'on utilise la segmentation manuelle par rapport à la segmentation automatique est plus important pour la méthode proposée que pour la méthode BIC. Cela est certainement dû au fait que les segments initiaux obtenus à l'issue de la segmentation automatique sont relativement courts, ce qui avantage la méthode BIC qui utilise des modèles plus simples. Cette remarque suggère que la méthode à base de GMMs est plus sensible à la segmentation que ne l'est la méthode BIC.

Les résultats obtenus sur le corpus "test" avec les réglages optimisés sur "dev" sont donnés dans le tableau 6.4, montrant un léger avantage global pour la méthode BIC.

test / bic	Inter 1	Inter 2	Inter 3	Inter 4	RFI 1	RFI 2	Global
16GMM	14.54	**13.03**	17.76	10.65	16.80	**24.36**	16.4
BIC	**9.77**	16.29	**8.35**	**8.76**	**9.56**	26.16	**15.10**

TAB. 6.4 – *Tâche SRL : taux d'erreur de classification obtenu pour chaque document (colonne 2 à 7) et globalement (colonne 8) sur le corpus "test" avec segmentation automatique.*

Ce résultat peut s'expliquer par deux facteurs. Premièrement, comme mentionné précédemment, l'approche par GMMs semble être plus sensible au comportement de la segmentation automatique. Étant donné que le processus de segmentation a été optimisé sur le corpus "dev", un taux supérieur d'erreurs de segmentation sur le corpus "test" pénalise davantage la méthode par GMM que la méthode BIC. Ceci a été confirmé par des expériences effectuées *a posteriori* en utilisant la segmentation manuelle de référence du corpus "test" avec laquelle l'approche par GMMs donne de meilleurs résultats (9% de taux d'erreur) que l'approche BIC (11% de taux d'erreur).

Deuxièmement, une étude du seuil d'arrêt optimal *a posteriori* sur chacun des documents du corpus "test" a montré que ce seuil optimal varie beaucoup d'un document à un autre dans l'approche par GMMs. C'était également le cas sur le corpus "dev" mais de façon moins marquée. Des expériences complémentaires ont également été faites en utilisant BIC comme critère d'arrêt dans la méthode de regroupement par GMMs, mais ce critère a donné environ 2% de taux d'erreur en plus, sur le corpus "dev" comme sur le corpus "test".

En conclusion, la méthode utilisée pour la tâche d'organisation en locuteur ESTER-SRL, alliant des modèles GMMs adaptés et la mesure de distance D_E entre ces GMMs, a montré un avantage potentiel par rapport à l'approche classique par critère BIC. Cependant, la méthode proposée semble sensible aux erreurs de segmentation, ce qui a mené à des performances légèrement meilleures pour la méthode BIC sur le corpus de test. Les travaux concernant la tâche SRL doivent maintenant se concentrer sur l'amélioration de la segmentation automatique afin que la méthode de regroupement par GMMs adaptés soit plus robuste et compétitive. D'autre part, le critère d'arrêt devra être étudié plus en détail afin d'améliorer sa stabilité.

Ces travaux en organisation en locuteur ont donné lieu à une publication en conférence [Ben et al.04a].

6.2 Contribution en sélection de locuteurs

Les travaux en sélection de locuteurs présentés dans cette section ont été réalisés dans le cadre du projet NEOLOGOS-TECHNOLANGUES. La tâche consiste à sélectionner un sous-ensemble de locuteurs représentatifs parmi un groupe plus important. Les techniques que nous avons utilisées pour cette tâche ont été élaborées en collaboration avec Sacha Krstulovic, ingénieur expert dans l'équipe METISS de l'IRISA, et Frédéric Bimbot, responsable scientifique de l'équipe METISS. L'ensemble des expériences a été mené par Sacha Krstulovic.

6.2.1 Contexte : projet NEOLOGOS

6.2.1.1 Présentation du projet

Le projet NEOLOGOS-TECHNOLANGUES, financé par le Ministère de la Recherche Français, est un projet de création de bases de données de parole en langue Française. Il regroupe des laboratoires académiques (LORIA et IRISA) et des entreprises industrielles (France Telecom, ELDA et TELISMA) chargés de collecter deux types de bases de données :

- une base de données de voix d'enfants (sous-projet PAIDIALOGOS);
- une base de données de locuteurs adultes de référence (sous-projet IDIOLOGOS).

Les travaux présentés dans cette section ne concernent que le sous-projet IDIOLOGOS. Pour plus de précision sur les données récoltées dans le cadre du sous-projet PAIDIALOGOS, on pourra se reporter à [Pinto et al.04].

Le but du sous-projet IDIOLOGOS est de construire une base de données de 200 locuteurs de référence permettant l'utilisation de techniques d'adaptation rapide au locuteur des systèmes de reconnaissance de parole, telles que les techniques "Eigenvoices" par exemple. Les 200 locuteurs de références sélectionnés devront couvrir de façon maximale l'espace des locuteurs et une grande quantité de données de parole (450 phrases) phonétiquement riche sera enregistrée pour chacun de ces locuteurs de référence.

La création de cette base est prévue en deux temps. Dans un premier temps, 200 locuteurs parmi un groupe de 1000 devront être sélectionnés à partir d'un ensemble restreint de phrases prononcées. Dans un deuxième temps, les 200 locuteurs sélectionnés devront enregistrer le corpus complet de 450 phrases chacun. Les 1000 locuteurs de départ couvrent de façon représentative les différents accents régionaux et les différentes tranches d'âge de la population, avec un équilibre homme/femme.

6.2.1.2 Méthodologie de sélection des locuteurs

La tâche à réaliser ici est une tâche de caractérisation du locuteur qui consiste à sélectionner des personnes dont les caractéristiques vocales sont les plus représentatives possible d'un groupe donné. La sélection des locuteurs doit se faire suivant un critère de

"représentativité" de certains locuteurs vis-à-vis de l'ensemble du groupe. Les acteurs du projet ont sélectionné un ensemble de critères relatifs aux domaines de la reconnaissance de la parole et du locuteur. Pour chacun de ces critères, une mesure de qualité des locuteurs sélectionnés est établie et au final, les deux meilleurs critères selon des tests croisés de qualité inter-critères seront sélectionnés. Les locuteurs de référence seront choisis parmi les 200 locuteurs sélectionnés par chacun de ces deux critères.

Pour un critère donné A, les quantités suivantes sont définies :
- $d_A(X_i, X_j)$: mesure de dissimilarité entre les locuteurs X_i et X_j;
- $L^A = \{L_j^A\}_{j=1,\ldots,N}$: ensemble des N locuteurs sélectionnés suivant le critère A (ici $N = 200$);
- $ref_A(X_i|L^B) = \arg\min_{L_j^B} d_A(X_i, L_j^B)$: locuteur de référence associé au locuteur X_i selon le critère A et choisi parmi l'ensemble L^B des locuteurs de références donnés par le critère B.

La sélection d'un ensemble L de locuteurs de référence selon un critère A se fait par optimisation d'une mesure de qualité globale $Q_A(L)$ de l'ensemble L :

$$L^A = \arg\min_L Q_A(L) \ , \tag{6.8}$$

avec :

$$Q_A(L) = \sum_{i=1}^{M} d_A(X_i, ref_A(X_i|L)) \ , \tag{6.9}$$

où M est le nombre total de locuteurs (ici $M = 1000$). Un test croisé de qualité entre un ensemble L^A de locuteurs sélectionnés par le critère A, et mesurée suivant le critère B est alors donné par :

$$Q_B(L^A) = \sum_{i=1}^{M} d_B(X_i, ref_B(X_i|L^A)) \ . \tag{6.10}$$

Parmi les critères étudiés dans le cadre du projet NEOLOGOS on trouve des mesures de similarité classiquement utilisées en reconnaissance de la parole et du locuteur. Par exemple, des critères basés sur des distances entre GMMs de locuteurs, sur des vraisemblances de modèles HMM des phonèmes d'un locuteur, ou encore sur la DTW entre groupes de souffle, ont été testés. De part ses compétences en reconnaissance du locuteur, les travaux de l'équipe METISS de l'IRISA dans NEOLOGOS se sont concentrés sur l'étude de critères basés sur des distances entre GMMs de locuteurs.

6.2.2 Algorithme de sélection des locuteurs

Pour un critère de qualité donné, une recherche exhaustive des 200 locuteurs minimisant ce critère parmi les 1000 locuteurs initiaux n'est pas envisageable. En effet, cela

correspondrait à tester $n = C_{1000}^{200} = \frac{1000!}{200!800!}$ configurations possibles, nombre absolument astronomique (de l'ordre de 10^{215}). Il est donc nécessaire d'employer un algorithme de sélection de ces locuteurs de référence.

L'algorithme utilisé par l'IRISA est basé sur une modélisation des locuteurs par GMMs. Les modèles de locuteurs sont adaptés d'un modèle du monde, lui-même appris au maximum de vraisemblance sur l'ensemble des données des 1000 locuteurs initiaux. Ces modèles sont composés de 512 composantes Gaussiennes à matrice de covariance diagonale dont seules les moyennes sont adaptées du modèle du monde.

Deux mesures de similarité entre GMMs ont été utilisées, en relation avec les travaux déjà effectués en VAL. Ces mesures sont :

1. les distances de Kullback-Leibler $KL2$ entre les GMMs des locuteurs. La mesure de similarité entre les locuteurs X_i et X_j est donc

$$d_{KL2}(X_i, X_j) = KL2(p(y|X_i), p(y|X_j)) \ .$$

   Ces distances KL sont estimées par tirage de Monte-Carlo selon la procédure décrite à la section 4.4.1.

2. les distances Euclidiennes D_E définies à la section 5.3 entre des GMMs dont les moyennes sont adaptées d'un modèle *a priori*. La mesure de similarité entre les locuteurs X_i et X_j est dans ce cas

$$d_{D_E}(X_i, X_j) = D_E(X_i, X_j) \ .$$

   Ces distances D_E sont directement calculables à partir des paramètres des GMMs des locuteurs.

Dans les deux cas, la qualité d'un ensemble de locuteurs de référence sélectionnés correspond à la mesure cumulée des distances entre chaque locuteur de l'ensemble initial et son locuteur de référence associé.

A partir de l'une ou l'autre des mesures définies ci-dessus, la sélection des locuteurs de référence se fait par l'intermédiaire d'un algorithme itératif de regroupement de locuteurs. Cet algorithme consiste à former des clusters de locuteurs et de désigner dans chaque cluster un locuteur centroïde. Son fonctionnement est le suivant.

Initialisation : 200 locuteurs centroïdes initiaux sont tirés au hasard parmi le groupe de 1000 locuteurs.

Après initialisation, les étapes suivantes sont itérées :

1. **Formation des clusters** : chacun des 800 locuteurs restants est rattaché à son locuteur centroïde le plus proche suivant la mesure de distance choisie, formant ainsi 200 clusters de locuteurs.

2. **Détermination des nouvaux locuteurs centroïdes**: pour chaque cluster, un nouveau locuteur centroïde est déterminé. Il correspond au locuteur qui minimise la somme cumulée des distances entre le locuteur considéré et les autres locuteurs du cluster.

Ces deux étapes sont répétées jusqu'à convergence de l'algorithme (plus de changement dans les locuteurs centroïdes).

Le résultat final de cet algorithme dépend fortement des locuteurs tirés au hasard à l'initialisation. Aussi, il est indispensable de répéter l'expérience un nombre important de fois pour étudier le comportement de la méthode et sélectionner l'ensemble de locuteurs de référence donnant la meilleure qualité.

6.2.3 Résultats expérimentaux

Les résultats reportés dans cette section correspondent à l'utilisation de la distance KL entre GMMs de locuteurs.

Afin d'évaluer l'apport de l'algorithme de sélection des locuteurs de référence, les valeurs de qualité obtenues avec cette procédure sont comparées à celles fournies par des locuteurs de référence tirés au hasard. Pour cela, une série d'expériences est effectuée en initialisant l'algorithme de sélection des locuteurs de façon aléatoire et en le laissant converger. D'autre part, une série de sélections aléatoires des locuteurs de référence est également générée.

A chaque ensemble de locuteurs de référence fourni par l'algorithme de sélection, nous associons une grandeur appelée FOM (Figure Of Merit) qui quantifie sa qualité relative vis-à-vis de l'ensemble des qualités obtenues par les sélections aléatoires. La FOM associée à un ensemble de locuteurs de référence est définie comme le pourcentage de sélections aléatoires ayant une qualité inférieure à celle de l'ensemble considéré.

Pour les expériences rapportées dans cette section, l'algorithme de sélection des locuteurs de référence a été lancé 400 000 fois. D'autre part, 400 000 sélections aléatoires ont également été générées. Le tableau 6.5 donne la proportion de sélections obtenues par l'algorithme ayant une FOM supérieure ou égale à 95, 99, 99.8 et 100.

	$FOM \geq 95$	$FOM \geq 99$	$FOM \geq 99,8$	$FOM = 100$
Proportion	45%	20,4%	8,3%	0,15%

TAB. 6.5 – *Évaluation des performances de l'algorithme de sélection de locuteurs de l'IRISA - Proportions des sélections de locuteurs de référence ayant une FOM (Figure Of Merit) supérieure ou égale à* 95, 99, 99,8 *et* 100.

Ce tableau montre que près de la moitié (45%) des ensembles de locuteurs de référence fournis par l'algorithme proposé est meilleure que 95% des sélections aléatoires, suivant la mesure de qualité définie. De plus certains des ensembles sélectionnés (au total 615 sur les 400 000 générés, soit 0,15%) ont une qualité meilleures que n'importe quelle sélection aléatoire.

Dans la suite du projet NEOLOGOS, les résultats obtenus par l'IRISA devront être recoupés avec ceux des autres acteurs du projet. Les ensemble de locuteurs de référence sélectionnés par l'IRISA seront comparés à ceux proposés par les autres laboratoires participants afin de retenir la sélection ayant des valeurs de qualités croisées (en fonction des autres critères étudiés) les plus polyvalentes.

6.3 Synthèse des travaux en reconnaissance et caractérisation du locuteur

Les travaux présentés dans cette section dans diverses tâches de reconnaissance et de caractérisation du locuteur montrent l'utilité de mesures de similarité entre des modèles de locuteur. Les résultats obtenus confirment que les distances que nous avons définies dans le cadre de la VAL dans les chapitres précédents sont des mesures significatives de similarité entre modèles de locuteur.

Les distances KL ont tout d'abord été utilisées dans le cadre de l'évaluation ESTER pour une tâche de suivi de locuteur dans des enregistrements radiophoniques. Cette tâche de suivi de locuteur a été traitée comme une tâche de segmentation aveugle du document suivie d'une tâche de détection de chacun des locuteurs cibles sur chaque segment. Les distances KL ont servi à normaliser les scores de détection de façon similaire à ce qui a été fait pour la VAL (cf la D-norm et la DT-norm section 5.1). L'amélioration des performances apportée par ces normalisations a été peu significative dans ce cas, probablement en raison des faibles disparités dans les conditions d'enregistrement entre les données d'apprentissage et de test.

Les mesures de distance ont ensuite été utilisées pour une tâche d'organisation en locuteur de documents radiophoniques, également dans le cadre de l'évaluation ESTER. Ces distances ont servi de mesures de similarité entre des GMMs adaptés dans un schéma de regroupement ascendant de segments de parole. Elles ont également été utilisées comme critère d'arrêt de l'algorithme de regroupement. La méthode utilisée a montré des performances prometteuses, bien que sa robustesse vis-à-vis des erreurs de segmentation reste encore à améliorer.

Enfin, les mesures de distances entre GMMs de locuteurs ont été appliquées pour une tâche de sélection d'un sous-groupe de locuteurs représentatifs parmi un ensemble plus important d'individus. Dans ce cadre, les distances employées ont permis de sélectionner des locuteurs focaux minimisant un critère de qualité basé précisément sur ces distances entre modèles de locuteurs. La comparaison des qualités relatives des sélections obtenues par l'algorithme proposé avec celle fournies par des sélections aléatoires des locuteurs

a montré l'intérêt d'utiliser ces distances pour identifier un ensemble d'individus ayant des caractéristiques vocales représentatives d'un groupe plus important.

Troisième partie

Adaptation Bayésienne hiérarchique des modèles de locuteurs

Chapitre 7

Connaissance a priori et techniques d'adaptation

Dans ce chapitre nous présentons les différentes techniques d'estimation robustes de modèles acoustiques que l'on trouve dans la littérature. Beaucoup de ces techniques viennent de travaux dédiés à l'adaptation rapide au locuteur des systèmes de reconnaissance de la parole. L'objectif visé par ces techniques est de combler certains manques dans les données d'apprentissage par l'apport d'information *a priori*. En particulier elles ont pour vocation de fournir des estimations robustes des modèles, même lorsque la quantité de données d'apprentissage est réduite. Suivant les cas, l'information *a priori* utilisée peut prendre la forme d'une densité de probabilité *a priori* des paramètres des modèles, ou encore d'une structure sous-jacente de ces paramètres, établissant des liens entre les différentes régions acoustiques occupées par la voix d'un locuteur.

Nous précisons dans un premier temps les motivations principales des techniques d'adaptation robuste. Puis nous présentons trois types de techniques :

- les techniques utilisant un *a priori* probabiliste (méthodes Bayésiennes);
- les techniques utilisant un *a priori* structurel (méthodes d'estimation de modèles par transformation des paramètres);
- les techniques utilisant un *a priori* hiérarchique (méthodes multiéchelles permettant d'adapter la granularité de l'estimation à la quantité de données);

Enfin, nous faisons une synthèse de l'ensemble des techniques évoquées et présentons quelques unes de leurs applications en VAL.

7.1 But recherché : estimation robuste de modèles acoustiques

En reconnaissance de la parole ou du locuteur, l'étape d'estimation de modèles probabilistes est basée sur le calcul de statistiques issues des données acoustiques d'apprentissage. Lorsque les lois d'émission des observations acoustiques sont modélisées par des mélanges de Gaussiennes, les statistiques sur les données sont évaluées localement, pour chaque Gaussienne du mélange. Ces statistiques locales servent à estimer les paramètres des Gaussiennes affectées à la modélisation de la région acoustique correspondante.

Dans de nombreux cas applicatifs cependant, la quantité de données d'apprentissage est faible, en raison de contraintes liées au profil de l'application elle-même. Certaines régions de l'espace acoustique sont alors très peu, ou pas du tout, peuplées par les données d'apprentissage. Les statistiques calculées dans ces régions sont par conséquent peu fiables et l'estimation des paramètres qui en résulte est très imprécise.

Pour ces raisons, de nombreuses techniques ont été développées dans les domaines de la reconnaissance de la parole et du locuteur, visant à améliorer la robustesse de l'estimation des modèles acoustiques face à un manque de données d'apprentissage. En particulier, on trouve dans la littérature d'innombrables travaux sur l'adaptation rapide aux locuteurs de systèmes de reconnaissance de parole indépendants du locuteur. En reconnaissance du locuteur, beaucoup des techniques utilisées pour la reconnaissance de la parole ont été également testées, mais l'une d'entre elle, l'adaptation Bayésienne, prédomine largement actuellement.

Afin de combler un manque de données d'apprentissage, il est possible d'incorporer au schéma d'estimation des informations externes provenant par exemple d'expériences antérieures ou de connaissances expertes. L'estimation se fait alors en contraignant les paramètres du modèle à partir de connaissances *a priori* sur le comportement "possible" ou "moyen" de ces paramètres. Ces contraintes *a priori* trouvent leur fondement d'une part dans des limites physiques inhérentes à l'appareil vocal humain, et d'autre part dans les caractéristiques phonétiques d'une langue. En raison de la morphologie de l'appareil vocal, n'importe quel son ne peut pas être produit par une personne, ce qui limite l'étendue de la distribution des paramètres acoustiques "statiques". De même, l'inertie des articulateurs et le tonus musculaire limitent la vitesse d'enchaînement de différents sons prononçables, contraignant ainsi les paramètres "dynamiques" à certaines plages de valeurs. Enfin, parmi les sons qu'un individu peut émettre, les phonèmes de la langue dans laquelle il s'exprime vont encore contraindre davantage la distribution des paramètres acoustiques à certaines régions caractéristiques de cette langue.

Plusieurs types d'information *a priori* ont été exploités dans les techniques d'adaptation robuste de modèles acoustiques.

Une première catégorie de techniques tire partie d'une analyse statistique de la distribution des paramètres des modèles. Cette analyse est en général effectuée sur une

base de données *a priori* conséquente et sert à estimer une densité de probabilité *a priori* des paramètres. Cette densité de probabilité *a priori* est alors intégrée dans la procédure d'estimation des paramètres du modèle.

Une seconde catégorie de techniques fait l'hypothèse d'une structure sous-jacente des vecteurs acoustiques qui contraint les paramètres des modèles à être liés les uns aux autres, au sein de groupements. Cette hypothèse vient du fait que l'on observe en pratique certains mouvements globaux des vecteurs acoustiques lorsque l'on passe de la voix d'un locuteur à celle d'un autre locuteur, indiquant que les régions acoustiques occupées par la voix d'un locuteur sont inter-dépendantes. Un nouveau modèle de locuteur peut alors être estimé par des transformations groupées des paramètres d'un modèle générique, les paramètres de transformations étant estimés sur l'ensemble d'apprentissage. Une autre façon de procéder consiste à localiser le nouveau modèle à estimer dans un espace de représentation construit à partir d'un ensemble de modèles de référence pré-estimés.

Une troisième catégorie de techniques étend l'hypothèse structurelle mentionnée ci-dessus à une structure hiérarchique des paramètres acoustiques. Ces techniques mettent en jeu des procédés d'adaptation multiéchelles des modèles permettant de s'adapter de façon automatique à la quantité de données d'apprentissage.

Dans les sections suivantes de ce chapitre nous présentons les principes de ces trois types de méthodes. Nous finissons le chapitre en faisant une synthèse des techniques d'adaptation robuste de modèles acoustiques et en présentant quelques-unes de leurs applications en reconnaissance du locuteur.

7.2 A priori probabiliste : méthodes Bayésiennes

7.2.1 Adaptation MAP

La première catégorie de méthodes que nous présentons utilise un *a priori* probabiliste sur la distribution des paramètres des modèles qui sont alors considérés comme des variables aléatoires. Ce type de méthodes, appelé adaptation Bayésienne ou adaptation MAP, passe par la définition d'une densité de probabilité *a priori* $p(\Lambda)$ pour l'ensemble des paramètres Λ du modèle. Les paramètres Λ doivent alors être déterminés de façon à maximiser leur probabilité *a posteriori*, c'est-à-dire conditionnellement à l'observation de l'ensemble d'apprentissage $\mathcal{Y}$. Le critère d'estimation est appelé critère MAP (Maximum A Posteriori) et la règle d'estimation des paramètres s'écrit :

$$\begin{aligned} \Lambda^{(MAP)} &= \arg\max_{\Lambda} p(\Lambda|\mathcal{Y}) \ , \\ &= \arg\max_{\Lambda} p(\mathcal{Y}|\Lambda)p(\Lambda) \ . \end{aligned} \tag{7.1}$$

Le choix des lois *a priori* et l'estimation ou le réglage de leurs paramètres est un enjeu majeur dans ce type de méthode car c'est cela qui va déterminer le comportement du schéma d'adaptation. Une loi *a priori* ayant une très faible variance (fortement

"piquée") va contraindre le paramètre correspondant à rester proche de sa valeur *a priori* (lorsque le volume de données d'apprentissage est limité). Au contraire un *a priori* ayant une variance large (faiblement "piqué") laissera l'estimation du paramètre beaucoup plus "libre". A la limite, un *a priori* non-informatif (i.e. constant) mènera à une estimation au maximum de vraisemblance du paramètre, c'est-à-dire dont le résultat dépend uniquement des données d'apprentissage.

Pour illustrer ce principe, nous rappelons la formule d'estimation au sens du critère MAP de la moyenne m_k de la Gaussienne k d'un GMM via l'algorithme EM [Gauvain et al.94] :

$$\widehat{m}_k = \frac{\gamma_k}{\gamma_k + \tau_k}\overline{y_k} + \frac{\tau_k}{\gamma_k + \tau_k}\mu_k \quad . \tag{7.2}$$

Cette formule est obtenue en supposant que la densité *a priori* de la moyenne m_k est une Gaussienne de moyenne μ_k et de matrice de covariance $\Sigma_k = \frac{1}{\tau_k}S_k$, où S_k est la matrice de covariance de la Gaussienne k du GMM [1]. Étant donnée une matrice de covariance S_k, le paramètre τ_k contrôle alors la covariance de la loi *a priori* et peut être considéré comme un facteur de confiance que l'on attribue à la moyenne *a priori* μ_k.

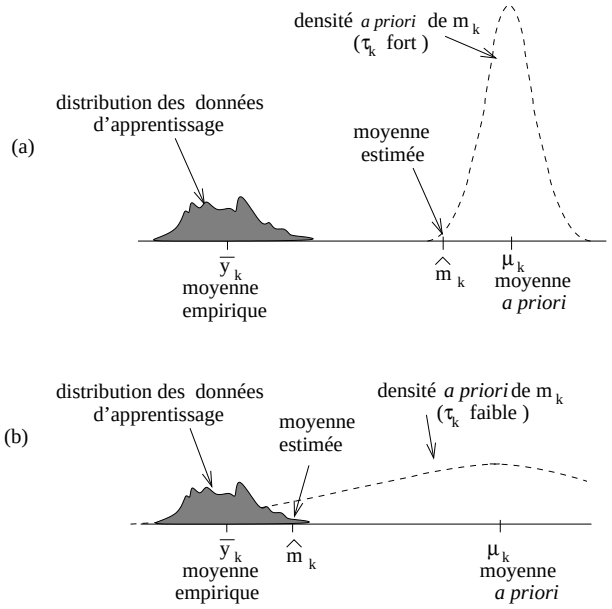


FIG. 7.1 – *Illustration de l'influence du facteur de confiance τ_k sur l'estimation de la moyenne m_k dans l'adaptation MAP d'un GMM - (a) τ_k fort - (b) τ_k faible.*

Une valeur de τ_k grande par rapport au taux d'occupation γ_k contraindra l'estimation $\hat{m}_k$ à rester proche de sa valeur *a priori* μ_k (cf figure 7.1, haut). Au contraire, une valeur de τ_k faible par rapport à γ_k permet à la moyenne empirique $\overline{y_k}$ de dominer l'estimation $\hat{m}_k$ (cf figure 7.1, bas).

1. La densité *a priori* de la moyenne m_k est donc ici définie conditionnellement à la matrice de covariance S_k de la Gaussienne k

Si les facteurs de confiance sont trop grands, l'adaptation ne permettra pas de tirer pleinement partie de l'information contenue dans les données d'apprentissage et le modèle sera peu caractéristique. Des valeurs trop faibles des facteurs de confiance risquent d'entraîner une sur-adaptation du modèle aux données d'apprentissage, le rendant ainsi trop spécifique de cet ensemble (il se généralisera mal à d'autres données provenant du même locuteur).

Les paramètres des lois *a priori*, appelés hyper-paramètres, sont en général déterminés par des méthodes empiriques. Les hyper-paramètres définissant les valeurs *a priori* des paramètres des modèles peuvent être estimés à l'aide d'un ensemble représentatif et important de données *a priori*. On estime sur cet ensemble de données *a priori* un modèle générique qui est alors utilisé comme modèle *a priori*. L'autre partie des hyper-paramètres, en particulier les facteurs de confiance, est optimisée sur une base de données de développement. Très souvent, des contraintes sur ces facteurs de confiance sont fixées afin de diminuer le nombre d'hyper-paramètres à optimiser (voir par exemple [Gauvain et al.94] et [Reynolds et al.00]) .

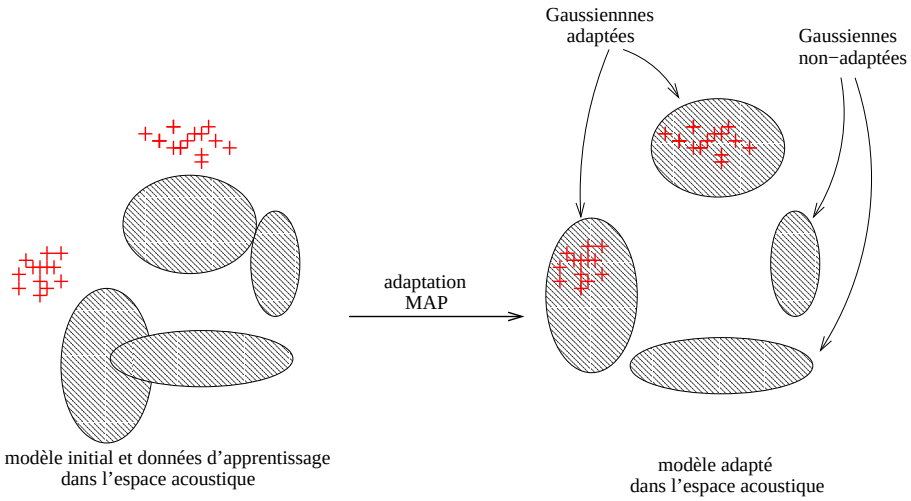


FIG. 7.2 – *Illustration du principe de l'adaptation MAP - Les composantes ne recevant pas de données d'apprentissage ne sont pas adaptées.*

Une propriété intéressante de l'adaptation Bayésienne est qu'elle converge théoriquement vers l'estimation au maximum de vraisemblance lorsque la quantité de données d'apprentissage augmente (cf. section 2.3.2). Cependant, cette convergence est relativement lente car il faut que chaque Gaussienne du modèle ait reçu suffisamment de données dans l'ensemble d'apprentissage pour être modifiée de façon significative. Du fait de l'hypothèse d'indépendance entre les différentes Gaussiennes d'un modèle, une composante ne recevant pas de données ne sera absolument pas modifiée, comme illustré sur la figure 7.2 (cas d'une adaptation des moyennes seulement et de matrices

de covariance diagonales).

En pratique, il n'est pas rare qu'un grand nombre de Gaussiennes ne reçoivent pas suffisamment de données pour être adaptées, ce qui conduit à des modèles peu représentatifs. Pour ces raisons, d'autres techniques d'estimation robuste de modèles acoustiques ont été développées, prenant en compte des informations plus globales dans leur schéma d'estimation.

7.2.2 Extension : utilisation de corrélations entre paramètres

La technique RMP (Regression based Model Prediction) [AS96] consiste à utiliser des corrélations entre les paramètres d'un modèle pour adapter les paramètres des Gaussiennes n'ayant pas reçu suffisamment de données, à partir des paramètres estimés de façon fiable. Un seuil sur les taux d'occupation des Gaussiennes est alors fixé pour définir des paramètres "sources" (ceux qui peuvent être estimés de façon fiable) et des paramètres "cibles" (ceux pour lesquels le taux d'occupation de la Gaussienne correspondante n'est pas suffisant). Les paramètres "sources" sont tout d'abord estimés par adaptation MAP classique et servent ensuite à fournir une prédiction des paramètres "cibles" en utilisant des relations de régression. A titre d'exemple, en considérant une moyenne "cible" m_k, corrélée avec une moyenne source m_j, la prédiction $\hat{m}_k^{(reg)}$ de la moyenne cible est donnée par :

$$\hat{m}_k^{(reg)} = A_{kj}.\hat{m}_j^{(map)} + b_{kj} \quad , \tag{7.3}$$

où $\hat{m}_j^{(map)}$ est l'estimation de la moyenne "source" au sens du critère MAP, et A_{kj} et b_{kj} sont respectivement la matrice et le biais de régression entre m_k et m_j.

L'estimation finale de la moyenne "cible" correspond à une combinaison linéaire de sa prédiction $\hat{m}_k^{(reg)}$ et de son estimation au sens du critère MAP $\hat{m}_k^{(map)}$:

$$\hat{m}_k = \alpha.\hat{m}_k^{(reg)} + (1-\alpha).\hat{m}_k^{(map)} \quad . \tag{7.4}$$

Le coefficient de pondération α tient compte des variances d'estimation de $\hat{m}_k^{(reg)}$ et $\hat{m}_k^{(map)}$ pour effectuer la balance entre ces deux grandeurs. Notons que, pour des raisons de simplicité nous avons considéré ici une régression simple de la moyenne "cible" sur une moyenne "source" unique. Dans la technique RMP, il est possible de prendre en compte des régressions multiples des paramètres "cibles" par rapport à plusieurs paramètres "sources", afin d'améliorer la précision de la prédiction.

L'utilisation des corrélations entre paramètres dans la technique RMP peut être considérée comme un ajout de connaissances probabilistes *a priori* dans le schéma d'adaptation, par rapport à l'adaptation MAP classique. Les matrices et biais de régression entre les différents paramètres sont en effet estimés à partir d'un ensemble de modèles *a priori*, en fonction des corrélations observées sur cet ensemble (seules les régressions correspondant à des coefficients de corrélation suffisamment grands sont retenues). La recherche de ces corrélations sur l'ensemble de données *a priori* est coûteuse

en calcul mais la technique permet d'améliorer la robustesse de l'estimation des paramètres correspondant à des Gaussiennes recevant peu de données d'apprentissage. D'autre part, des composantes ne recevant pas du tout de données peuvent tout de même être adaptées via les relations de régression. Cependant, lorsque la quantité de données est très faible, peu de paramètres peuvent être considérés comme paramètres "sources" et il est alors possible que certains paramètres "cibles" ne puissent pas être prédits par l'intermédiaire des régressions.

7.3 A priori structurel : adaptation par transformation de modèles

La seconde catégorie de techniques que nous présentons fait l'hypothèse d'un *a priori* structurel sur les paramètres d'un modèle acoustique. Cet *a priori* structurel suppose que les paramètres d'un modèle sont obtenus par des transformations appliquées sur les paramètres d'un modèle générique ou sur un ensemble de modèles de locuteurs de référence pré-estimés. Les paramètres du modèle que l'on cherche à estimer sont ici contraints par le type de transformation utilisée.

7.3.1 Techniques MLLR et MAPLR

Un premier type de techniques utilise des transformations linéaires des moyennes d'un modèle GMM générique. La moyenne m_k du modèle estimé est alors déterminée par la formule :

$$\hat{m}_k = A_k.\mu_k + b_k \ , \tag{7.5}$$

où μ_k est la moyenne k dans le modèle générique et A_k et b_k sont respectivement les matrices et biais de régression qui sont des paramètres à estimer. Lorsque les paramètres de régression sont estimés au maximum de vraisemblance, la technique est appelé MLLR (Maximum Likelihood Linear Regression) [Leggetter et al.95b].

En général, les paramètres A_k et b_k ne sont pas estimés individuellement pour chaque Gaussienne. Cela demanderait en effet d'estimer un nombre trop important de paramètres (supérieur au nombre de paramètres dans le modèle lui-même) et mènerait à des estimations peu fiables. Les Gaussiennes sont alors regroupées en clusters, en général en fonction d'une mesure de similarité, et les Gaussiennes d'un cluster donné sont contraintes à partager les mêmes paramètres de régression. Les transformations correspondantes sont alors plus globales et leurs paramètres sont estimés de façon plus robuste car une quantité de données d'apprentissage plus importante peut être utilisée pour chaque transformation. La configuration des regroupements de Gaussiennes est en général optimisée sur des données de développement.

Ce principe d'adaptation est illustré sur la figure 7.3, dans le cas où seules les moyennes sont adaptées et où les matrices de covariances sont diagonales.

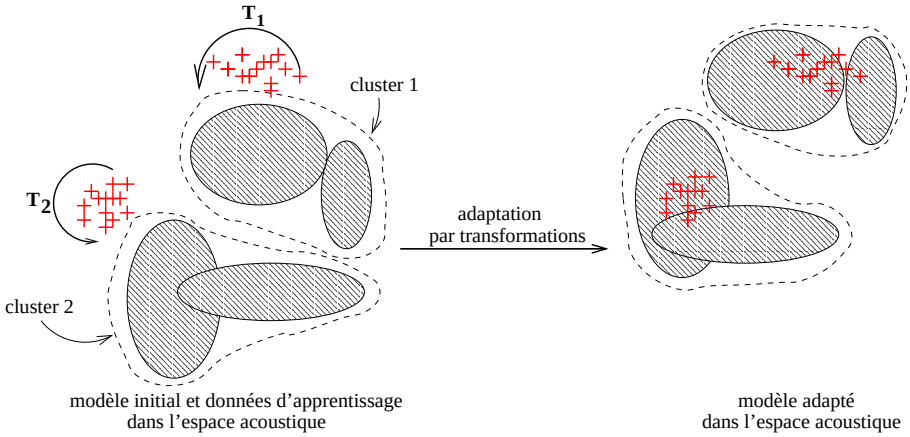


FIG. 7.3 – *Illustration du principe de l'adaptation par transformations - Toutes les composantes sont adaptées, suivant des transformations estimées sur les données d'apprentissage pour chaque cluster de Gaussiennes.*

En pratique, des simplifications permettent encore de diminuer le nombre de paramètres à estimer, en considérant par exemple des matrices de régression diagonales ou en estimant seulement un biais de régression ("bias-only MLLR"). D'autre part, afin d'augmenter la robustesse de l'estimation des paramètres de régression, il est possible d'utiliser des *a priori* probabilistes concernant ces paramètres. Ces lois *a priori* sont introduites dans le schéma d'estimation des paramètres de régression de la même manière que pour l'adaptation Bayésienne des paramètres des modèles, et permettent de régulariser les valeurs estimées. La technique est alors appelée MAPLR (Maximum A Posteriori Linear Regression) [Siohan et al.99].

Les techniques MLLR et MAPLR permettent d'adapter des composantes Gaussiennes ne recevant pas de données dans l'ensemble d'apprentissage car elles utilisent des transformations linéaires globales, communes à un ensemble de composantes. On peut s'interroger sur la validité de ces transformations linéaires globales pour estimer un modèle acoustique d'un locuteur. Ce principe se rattache pourtant à certains phénomènes induits par la configuration physique du conduit vocal lui-même. Notamment, il a été montré dans [Pitz et al.01] que les variations de longueur du conduit vocal d'un individu à un autre se traduisent dans l'espace des coefficients cepstraux par une transformation linéaire globale (surtout pour les voyelles). Cette transformation est similaire à celles utilisées dans MLLR et MAPLR avec certaines contraintes sur la forme des matrices de régression. Ce phénomène a d'ailleurs été utilisé dans [Afify et al.00] pour contraindre les paramètres de transformation dans une technique de type MLLR.

Si les techniques MLLR et MAPLR permettent une adaptation rapide de modèles de locuteur (c'est-à-dire avec peu de données), les estimations qu'elles engendrent ne convergent pas vers les paramètres estimés au maximum de vraisemblance lorsque la

quantité de données augmente, contrairement aux méthodes Bayésiennes. D'autre part, la structure figée des clusters de Gaussiennes ne permet pas d'adapter le schéma d'estimation à la quantité d'apprentissage, en utilisant par exemple des transformations plus globales lorsque cette quantité est faible, et plus locales lorsqu'elle est importante. Pour remédier à ce problème, des méthodes multiéchelles ont été développées permettant de prendre en compte les variations de quantité de données de façon adaptative. Ces méthodes sont présentée à la section 7.4.

7.3.2 Modèles de référence et techniques "Eingenvoices"

Un second type de techniques s'appuyant sur un *a priori* structurel cherche à décrire les paramètres moyennes d'un nouveau modèle dans un espace obtenu à partir d'un ensemble de N modèles de locuteur de référence pré-estimés. Les moyennes $\{m_k\}$ du nouveau modèle s'expriment alors comme une combinaison linéaire des moyennes $\{\mu_{i,k}\}$ des modèles de référence :

$$\hat{m}_k = \sum_{i=1}^{N} w_k.\mu_{i,k} \quad . \tag{7.6}$$

Les paramètres à estimer sont dans ce cas les coefficients $\{w_k\}$ de combinaison linéaire des moyennes des modèles de référence.

Pour diminuer la taille de l'espace de représentation, et donc le nombre de paramètres à estimer, il est possible d'utiliser des techniques de classification de locuteurs afin de sélectionner un sous-ensemble des modèles de référence [Padmanabhan et al.96] ou de construire des modèles de cluster de locuteurs [Pusateri et al.02], servant alors de références. Une autre approche consiste à appliquer des techniques d'analyse en composantes principales (ACP) sur l'espace engendré par les modèles de référence afin d'en diminuer la dimension. Dans ce cas, un certains nombre des premiers vecteurs propres issus de l'ACP sont sélectionnés pour former un nouvel espace de représentation. Les moyennes du modèle à estimer sont alors exprimées par une combinaison linéaire de ces vecteurs propres. La formule d'estimation du super-vecteur $\mathbf{m}$ obtenu par concaténation des vecteur moyennes $\{m_k\}$ est donnée par

$$\hat{\mathbf{m}} = \mathbf{E}.\mathbf{w} \quad . \tag{7.7}$$

où $\mathbf{E}$ est la matrice contenant les vecteurs propres sélectionnés et $\mathbf{w}$ est le vecteur de paramètres à estimer pour l'adaptation du modèle. Ce type de techniques est connu dans la littérature sous le nom d' "Eigenvoices" [Thyes et al.00]. Elle permet en général une adaptation très rapide des modèles car l'espace de représentation est optimisé et le nombre de paramètres représentatifs est donc minimal.

7.4 A priori hiérarchique : méthodes multiéchelles

Pour s'adapter automatiquement à la quantité de données disponibles pour l'estimation d'un modèle acoustique, des méthodes multiéchelles à base d'arbre de classification de Gaussiennes ont été développées. Ces méthodes font l'hypothèse d'une structure hiérarchique des vecteurs acoustiques.

Dans le cadre des techniques MLLR ou MAPLR, un ensemble de transformations sont définies, à différents niveaux de granularité en partant d'une transformation globale de l'ensemble des Gaussiennes (racine de l'arbre) jusqu'à des transformations locales pour chacune des Gaussiennes (feuilles de l'arbre) [Leggetter et al.95a]. Chaque noeud dans l'arbre correspond à un cluster de Gaussiennes, ces clusters contenant de moins en moins de composantes au fur et à mesure que l'on descend dans l'arbre. Au moment de l'adaptation d'un nouveau modèle, un ensemble de noeuds est sélectionné en fonction des données d'apprentissage disponibles dans chaque région de l'espace acoustique (cf figure 7.4). Ces noeuds doivent être les plus bas possible dans l'arbre tout en ayant reçu un nombre suffisant (c.à.d. supérieur à un seuil fixé) de données. Les transformations associées aux noeuds sélectionnés sont alors estimées suivant les techniques MLLR ou MAPLR puis appliquées à l'ensemble des Gaussiennes de chaque cluster.

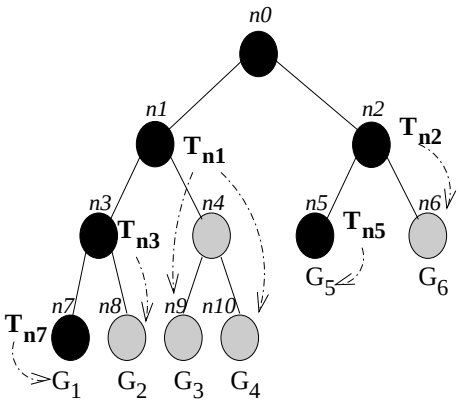


FIG. 7.4 – *Arbre de classification de Gaussiennes - Les noeuds en noir reçoivent suffisamment de données pour qu'une transformation y soit estimée. La transformation est alors appliquée à toutes les Gaussiennes du cluster correspondant.*

Ce type de méthodes permet de garantir qu'une quantité suffisante de données est utilisée pour l'estimation de chaque transformation. D'autre part, elles permettent d'affiner la granularité de l'adaptation lorsque l'on dispose d'une quantité plus importante de données.

L'arbre de classification de Gaussiennes peut être construit de plusieurs manières. Une première possibilité consiste à utiliser des connaissances phonétiques *a priori*. Les Gaussiennes sont alors regroupées de façon hiérarchique selon les classes et macro-classes phonétiques auxquelles elles appartiennent. Cependant, la construction de l'arbre

exige dans ce cas que la base de données acoustiques utilisée à cet effet soit étiquetée phonétiquement. Il n'est pas toujours possible de disposer d'une telle base de données ni d'un système de reconnaissance de la parole permettant d'étiqueter automatiquement une base de données de façon fiable. Dans le cadre de notre travail notamment, les bases de données utilisées proviennent des évaluations NIST et ne sont pas étiquetées phonétiquement.

Une autre manière de procéder pour construire l'arbre est de regrouper les Gaussiennes suivant leur proximité dans l'espace acoustique, de façon hiérarchique. Deux type de procédés peuvent être utilisés dans ce cas :

- un processus divisif descendant permettant de séparer des clusters de Gaussiennes en clusters plus petits.
- un processus agglomératif ascendant basé sur une mesure de similarité entre Gaussiennes, permettant de regrouper des Gaussiennes "proches" au sein de clusters.

Les arbres de classification de Gaussiennes ont été utilisés dans un grand nombre de techniques différentes, principalement destinées à l'adaptation rapide au locuteur de modèles acoustiques dans le cadre de tâches de reconnaissance de la parole. Nous nous limitons ici à mentionner quelques-unes d'entre elles possédant des propriétés intéressantes et certains liens avec les travaux présentés dans la suite de cette thèse.

Dans [Shinoda et al.97] et [Siohan et al.00], des arbres sont utilisés pour définir une structure hiérarchique des lois *a priori* des paramètres de régression. Les techniques correspondantes, respectivement appelées SMAP (Structural MAP) et SMAPLR (Structural MAPLR), permettent de définir les paramètres des lois *a priori* dans un noeud donné de l'arbre à partir des estimations des paramètres de transformation dans le noeud parent. Cela mène à une meilleure définition des lois *a priori* permettant d'obtenir des paramètres de transformation plus pertinents. Un gros avantage de ces techniques est qu'elles convergent vers l'estimation ML des paramètres, contrairement aux techniques MLLR et MAPLR.

Dans [Kannan97] et [Cerisara et al.01], les dépendances entre les paramètres de transformation dans différents noeuds de l'arbre sont modélisées par un processus multiéchelles autorégressif (MAR : Multiscale Autoregressive). La technique SMAP mentionnée précédemment est considérée dans [Kannan97] comme un cas particulier de processus MAR.

Le principe des méthodes SMAP, SMAPLR et MAR est de propager les paramètres estimés de façon robuste dans les niveaux hauts de l'arbre afin de les utiliser pour l'estimation des paramètres dans les niveaux bas, pour lesquels moins de données sont disponibles. La méthode d'adaptation hiérarchique de modèles de locuteur que nous développons dans le chapitre suivant utilise également ce type de procédé.

7.5 Synthèse sur les techniques d'adaptation et application en RAL

Un grand nombre de techniques d'adaptation robuste de modèles acoustiques sont répertoriées dans la littérature. Ces techniques peuvent être réparties en trois catégories principales.

La première catégorie utilise un *a priori* probabiliste sur les paramètres des modèles. Ce sont les méthodes d'adaptation Bayésienne. Ce type de méthodes possède l'avantage de fournir des estimations des paramètres qui convergent vers les estimations obtenues au maximum de vraisemblance lorsque la quantité de données d'apprentissage tend vers l'infini. Par contre, lorsque cette quantité de données est limitée, l'adaptation Bayésienne ne permet d'adapter que les composantes du modèle recevant suffisamment de données, les autres composantes restant inchangées. On risque alors d'obtenir un modèle peu discriminant sur des données de test n'étant pas homogènes aux données d'apprentissage.

La seconde catégorie de techniques fait l'hypothèse d'une structure sous-jacente des vecteurs acoustiques. On trouve parmi elles les techniques MLLR et MAPLR qui adaptent des groupes de Gaussiennes par transformation linéaire des moyennes d'un modèle générique indépendant du locuteur. D'autres techniques utilisant un *a priori* structurel cherchent à représenter les moyennes du modèle à estimer dans un espace formé par un ensemble de modèles réels de locuteur pré-estimés, ou dans un espace fictif de plus petite dimension, optimisé par ACP (techniques "Eigenvoices"). L'ensemble de ces techniques permet d'adapter toutes les Gaussiennes d'un modèle même si la quantité de données est très limitée car elles utilisent des transformations ou des représentations globales. Cependant, les estimations qu'elles délivrent ne convergent vers les estimations au maximum de vraisemblance des paramètres lorsque la quantité de données d'apprentissage augmente, ce qui est conceptuellement génant.

Enfin, une troisième catégorie de techniques utilise des schémas d'estimation multiéchelles à base d'arbre de classification de Gaussiennes, permettant ainsi de régler automatiquement la granularité de l'adaptation à la quantité de données d'apprentissage. Lorsque peu de données sont disponibles, l'adaptation utilise des informations plus globales afin de garantir que les paramètres d'adaptation soient estimés de façon robuste. Lorsque la quantité de données est importante, les informations utilisées sont plus locales, permettant ainsi au schéma d'adaptation d'estimer plus finement le modèle.

On trouve également dans la littérature des techniques associant les méthodes d'adaptation Bayésiennes et d'adaptation par transformation de modèles. Ces techniques ont pour objectif de tirer partie des avantages de chacune des deux méthodes d'adaptation. Mentionnons notamment le cadre unifié pour ces deux types de méthodes proposé dans [Mokbel01].

En reconnaissance du locuteur plus spécifiquement, un grand nombre de techniques d'adaptation ont été expérimentées, afin d'améliorer la robustesse de l'estimation des modèles de locuteur face à un manque de données d'apprentissage.

Une comparaison des performances obtenues en VAL par les techniques MAP, tree-based MLLR et Eigenvoices sur la base de données de l'évaluation NIST 1999 peut être trouvée dans [Mariethoz et al.01], montrant une supériorité des méthodes Bayésiennes dans ce cadre applicatif. Le cadre unifié proposé dans [Mokbel01] a été appliqué en VAL sur les données de l'évaluation NIST 2003 [Blouet et al.04], mais n'a pas permis d'atteindre pour l'instant les performances des systèmes état de l'art utilisant l'adaptation Bayésienne classique. Dans [Charlet04], une technique d'adaptation de GMM par représentation à partir d'un ensemble de locuteurs de référence sélectionnés a été expérimentée pour une tâche d'identification du locuteur en mode indépendant du texte. Les données sont issues d'une base téléphonique de France Télécom R&D, et l'identification se fait à partir d'une unique phrase. La technique d'adaptation proposée n'a pas obtenu de performances meilleures que l'adaptation Bayésienne classique lorsqu'elle est utilisée seule. Elle permet par contre d'observer un gain de performance lorsqu'elle est utilisée comme initialisation de l'adaptation Bayésienne, montrant ainsi que ce type de techniques peut apporter de l'information *a priori* pertinente pour l'adaptation d'un modèle de locuteur.

Malgré l'adaptation moins rapide (en terme de quantité de données) que permettent d'obtenir les méthodes Bayésiennes, celles-ci restent actuellement les plus utilisées pour l'estimation des modèles de locuteur en VAL dans les systèmes état de l'art. Cependant, les évaluations proposées par NIST jusqu'en 2003 n'intégraient pas de tâche avec des quantités de données très réduites. Le développement de techniques robustes d'adaptation dans ce cadre applicatif n'avait donc pas été favorisé jusqu'alors. En 2004, NIST a proposé une tâche de vérification en mode indépendant du texte avec peu de données (10 secondes d'apprentissage et 10 secondes de test) permettant ainsi de confronter les systèmes état de l'art à ce type de conditions.

Dans ce contexte, nous avons développé et expérimenté une nouvelle technique d'adaptation destinée à renforcer la robustesse de l'adaptation Bayésienne lorsque l'on dispose de très peu de données d'apprentissage. Cette technique d'adaptation, qui reste entièrement dans un cadre Bayésien, utilise un arbre hiérarchique de dépendances entre Gaussiennes de GMM multirésolutions. Elle est présentée dans les chapitres suivants de cette thèse.

Chapitre 8

H-MAP : adaptation Bayésienne hiérarchique de GMMs

Ce chapitre est dédié au développement théorique d'une nouvelle technique d'adaptation Bayésienne des GMMs de locuteur utilisant une structure hiérarchique de dépendances entre Gaussiennes : H-MAP. Le but est de pouvoir adapter des composantes n'ayant reçu que très peu de données (ou pas du tout), par le biais des dépendances établies, tout en gardant les avantages que procure l'approche classique d'adaptation Bayésienne.

Nous présentons dans un premier temps nos motivations et le cadre dans lequel s'inscrit ce travail. Ensuite nous décrivons la méthode de modélisation des dépendances entre Gaussiennes que nous avons développée, basée sur une structure hiérarchique en arbre. Puis nous dérivons les formules d'adaptation des moyennes d'un GMM de locuteur par H-MAP, dans le cadre de l'algorithme EM. Les propriétés théoriques de H-MAP sont alors discutées et nous pointons les liens existants entre l'adaptation que nous proposons et les techniques existantes, présentées au chapitre précédent. Enfin nous faisons la synthèse de ce développement théorique en rappelant les avantages potentiels de la méthode et nous identifions certaines difficultés susceptibles d'être rencontrées en pratique .

8.1 Motivations et positionnement

Comme nous l'avons déjà mentionné au chapitre précédent, l'adaptation Bayésienne classique avec l'hypothèse d'indépendance des composantes du modèle ne permet d'adapter que les Gaussiennes d'un GMM qui reçoivent suffisamment de données dans l'ensemble d'apprentissage. Cela peut conduire à des modèles peu représentatifs lorsque la quantité de données est très limitée.

Cependant, de nombreux travaux en adaptation rapide au locuteur des systèmes de reconnaissance automatique de la parole (cf. section précédente) semblent indiquer que les composantes d'un modèle de locuteur ne sont pas indépendantes. Dans le cadre de ces travaux, les Gaussiennes des modèles sont par exemple regroupées par clusters afin d'appliquer un schéma d'adaptation commun à toutes les Gaussiennes d'un même cluster. De plus, des regroupements hiérarchisés de ces Gaussiennes autorisent l'utilisation de schémas d'adaptation multirésolutions qui s'adaptent à la quantité de données d'apprentissage disponible. Cela mène en général à des estimations plus robustes des modèles lorsque les données sont limitées et/ou éparses car les paramètres de l'adaptation sont estimés avec des ensembles d'observations suffisamment importants.

Pour combiner les avantages d'une adaptation Bayésienne directe[1] des paramètres d'un modèle et la robustesse des méthodes multirésolutions, nous étudions la possibilité d'introduire dans le schéma Bayésien, de la connaissance *a priori* supplémentaire sous forme d'une structure de dépendances hiérarchiques entre les Gaussiennes. L'hypothèse sous-jacente est qu'il existe une structure de dépendances multiéchelles entre régions acoustiques occupées par la parole d'un locuteur et que cette structure est partagée par l'ensemble des locuteurs.

Le cadre applicatif visé est celui des évaluations NIST en vérification du locuteur en mode indépendant du texte. Les bases de données de développement dont nous disposons sont également issues de ces évaluations et ne sont pas étiquetées phonétiquement. Nous avons donc opté pour une méthode d'estimation de la structure de dépendance uniquement basée sur les données acoustiques, et n'intégrant pas d'information linguistique.

La définition de cette structure de dépendances est exposée à la section 8.2 et le développement de l'adaptation Bayésienne hiérarchique est donnée à la section 8.3. Nous discutons à la section 8.4 des propriétés de l'algorithme que nous proposons. Les développements expérimentaux et l'évaluation de la méthode sont présentés au chapitre suivant.

1. par opposition à une adaptation indirecte qui utiliserait des transformations pour adapter les paramètres du modèle

8.2 Modélisation structurelle des dépendances entre Gaussiennes

8.2.1 Structure de dépendances

Afin de capturer la structure hiérarchique de la parole, nous définissons des GMMs multirésolutions dont les vecteurs moyennes sont agencés dans un arbre binaire. La structure correspondante est représentée sur la figure 8.1. Dans [Xiang et al.03], une structure similaire en arbre associant des GMMs multirésolutions a été utilisée par Xiang *et al.* où sont définis un modèle du monde structurel (SBM : Structural Background Model) et des modèles de mélange de Gaussiennes structurels (SGMM : Structural Gaussian Mixture Models). Cependant, les auteurs ne définissent pas de dépendances entre niveaux de l'arbre comme nous le faisons dans notre approche.

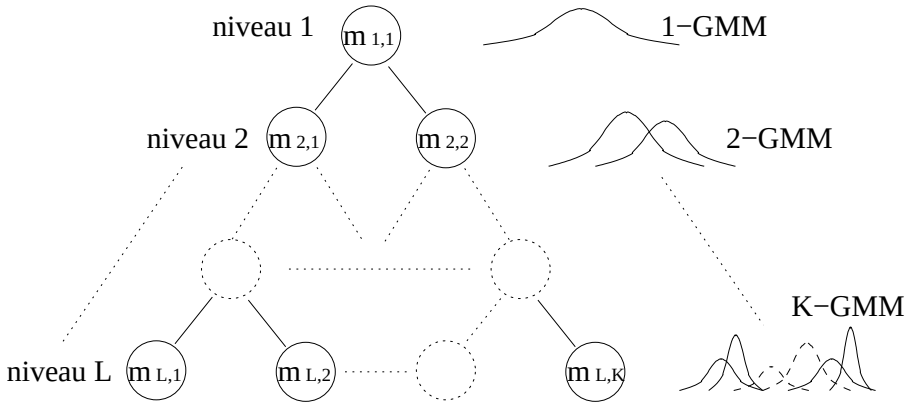


FIG. 8.1 – *Structure en arbre binaire et GMMs correspondants*

A chaque noeud de l'arbre est associé une moyenne (vectorielle) et chaque niveau correspond à un GMM dans une résolution donnée, définie par le nombre de ses composantes. Le plus bas niveau L (feuilles de l'arbre) définit la résolution la plus fine utilisée pour représenter les modèles de locuteur. Une moyenne dans un niveau donné de l'arbre possède une moyenne parent dans le niveau immédiatement supérieur et deux moyennes enfants dans le niveau immédiatement inférieur, à l'exception de la moyenne du premier niveau (racine de l'arbre) qui n'a pas de parent et des moyennes "feuilles" qui n'ont pas d'enfant. La moyenne parent d'une moyenne $m_{\ell,k}$ dans le noeud k du niveau ℓ sera notée $m_{\ell-1,\pi(k)}$.

8.2.2 Relations de dépendances

Un arc dans l'arbre symbolise une dépendance entre une moyenne parent et un de ses enfants, définissant ainsi une densité de probabilité conditionnelle $p(m_{\ell,k}|m_{\ell-1,\pi(k)})$. La

structure de l'arbre, de type réseau Bayésien, suppose de plus que, connaissant la valeur de son parent, une moyenne $m_{\ell,k}$ dans un niveau ℓ est indépendante des autres moyennes de ce niveau et des niveaux supérieurs. En utilisant ces relations d'indépendance conditionnelle, la densité de probabilité jointe de l'ensemble des moyennes d'un niveau ℓ, conditionnellement aux valeurs de toutes les moyennes des niveaux supérieurs peut être écrite sous la forme :

$$p(\mathbf{m}_\ell|\mathbf{m}_1, ..., \mathbf{m}_{\ell-1}) = \prod_{k=1}^{K_\ell} p(m_{\ell,k}|m_{\ell-1,\pi(k)}) \ , \tag{8.1}$$

où $\mathbf{m}_\ell$ dénote l'ensemble des moyennes du niveau ℓ et K_ℓ est le nombre de composantes du GMM correspondant.

Cette structure hiérarchique en arbre binaire introduit donc des dépendances indirectes entre les moyennes d'un GMM d'un niveau donné par l'intermédiaire des relations qui les lient à leur parent et ancêtres dans l'arbre. En se basant sur cette structure, nous développons dans la section suivante un schéma d'adaptation Bayésienne hiérarchique que nous appelons H-MAP (Hierarchical MAP) et qui permet théoriquement d'adapter des composantes ne recevant pas de données à l'apprentissage.

8.3 Adaptation MAP hiérarchique des moyennes d'un GMM

8.3.1 A priori conditionnels Gaussiens

La structure de dépendance décrite à la section précédente définit une densité *a priori* conditionnelle $p(m_{\ell,k}|m_{\ell-1,\pi(k)})$ pour chaque moyenne enfant $m_{\ell,k}$ de l'arbre. Cette densité *a priori* conditionnelle exprime la relation de dépendance statistique de $m_{\ell,k}$ avec sa moyenne parent $m_{\ell-1,\pi(k)}$.

Comme dans le schéma d'adaptation Bayésienne classique exposé à la section 2.3.2, les densités *a priori* marginales des moyennes des GMMs sont supposées être des lois normales multivariées. Sous cette hypothèse, la structure de dépendances utilisée est alors un arbre binaire Gaussien et les relations de dépendances s'expriment comme des régressions linéaires. Les densités *a priori* conditionnelles $p(m_{\ell,k}|m_{\ell-1,\pi(k)})$ sont également des lois Gaussiennes multivariées dont le vecteur moyenne $\mu_{k|\pi(k)}$ et la matrice de covariance $\Sigma_{k|\pi(k)}$ sont donnés par [2] :

$$\mu_{k|\pi(k)} = \mu_k + R_{k|\pi(k)}(\tilde{m}_{\pi(k)} - \mu_{\pi(k)}) \ , \tag{8.2}$$

$$\Sigma_{k|\pi(k)} = \Sigma_k(I - C_{k,\pi(k)}) \ , \tag{8.3}$$

2. dans la suite du document, l'indice de niveau ℓ sera parfois omis lorsqu'il est implicite, afin d'alléger l'écriture des formules

où :
(μ_k, Σ_k) sont les paramètres de la densité *a priori* marginale $p(m_k)$;
$(\mu_{\pi(k)}, \Sigma_{\pi(k)})$ sont les paramètres de la densité *a priori* marginale $p(m_{\pi(k)})$;
$\tilde{m}_{\pi(k)}$ est la valeur observée de la moyenne parent $m_{\pi(k)}$;
I est une matrice identité de même dimension que Σ_k;
$R_{k|\pi(k)}$ est la matrice de régression de m_k par rapport à $m_{\pi(k)}$;
$C_{k,\pi(k)}$ est la matrice de corrélation entre m_k et $m_{\pi(k)}$.

Les matrices $R_{k|\pi(k)}$ et $C_{k,\pi(k)}$ sont définies par les formules suivantes :

$$R_{k|\pi(k)} = \Sigma_{k,\pi(k)}\Sigma_{\pi(k)}^{-1} \ , \tag{8.4}$$

$$C_{k,\pi(k)} = \Sigma_k^{-1}\Sigma_{k,\pi(k)}\Sigma_{\pi(k)}^{-1}\Sigma_{k,\pi(k)}^{*} \ . \tag{8.5}$$

Dans ces équations, $\Sigma_{k,\pi(k)}$ désigne la matrice d'inter-covariance de m_k et $m_{\pi(k)}$. Les coefficients diagonaux $d \times d$ des matrices $R_{k|\pi(k)}$ et $C_{k,\pi(k)}$ seront respectivement notés $r^d_{k|\pi(k)}$ et $(\rho^d_{k,\pi(k)})^2$.

Notons que la matrice de covariance conditionnelle $\Sigma_{k|\pi(k)}$ ne dépend pas de la valeur observée $\tilde{m}_{\pi(k)}$ de la moyenne parent mais seulement des régressions et corrélations qui existent entre cette moyenne parent et la moyenne enfant m_k. La matrice $\Sigma_{k|\pi(k)}$ peut donc être définie avant même d'avoir observé la moyenne parent. D'autre part, la matrice de corrélation $C_{k,\pi(k)}$ étant semi-définie positive, d'après l'équation 8.3 on a :

$$\Sigma_{k|\pi(k)} \leq \Sigma_k \ ,$$

au sens des matrices semi-définies positives. Cela traduit le fait que, au travers de l'observation de la moyenne parent, la densité *a priori* conditionnelle $p(m_k|m_{\pi(k)})$ apporte de la précision [3] sur la distribution de m_k, par rapport à la densité *a priori* marginale $p(m_k)$.

8.3.2 Estimation des paramètres et propagation des dépendances

Dans ce travail, nous ne considérons que l'adaptation des moyennes des GMMs. Les poids et matrices de covariance sont fixés à leurs valeurs *a priori* issues d'un modèle du monde.

Nous décrivons à présent l'algorithme d'estimation hiérarchique des moyennes des GMMs. Celui-ci met en jeu une propagation des dépendances du haut de l'arbre vers le bas : à chaque niveau, l'estimation des moyennes s'effectue en supposant que les moyennes dans les niveaux supérieurs ont été observées.

Étant donné un ensemble d'apprentissage $\mathcal{Y}$ et ayant observé l'ensemble des moyennes des niveaux 1 à $\ell - 1$, la règle d'estimation au sens du critère MAP des moyennes

3. le terme de précision est entendu ici comme l'inverse de la variance

$\mathbf{m}_\ell = \{m_k\}$ du GMM de niveau ℓ s'écrit :

$$\begin{aligned} \widehat{\mathbf{m}}_\ell &= \arg\max_{\mathbf{m}_\ell} p(\mathbf{m}_\ell|\mathcal{Y}, \mathbf{m}_1, ..., \mathbf{m}_{\ell-1}) \ , \\ &= \arg\max_{\mathbf{m}_\ell} p(\mathcal{Y}|\mathbf{m}_\ell, \mathbf{m}_1, ..., \mathbf{m}_{\ell-1}) p(\mathbf{m}_\ell|\mathbf{m}_1, ..., \mathbf{m}_{\ell-1}) \ . \end{aligned} \tag{8.6}$$

En applicant l'équation (8.1), cette expression se développe de la manière suivante :

$$\widehat{\mathbf{m}}_\ell = \arg\max_{\mathbf{m}_\ell} p(\mathcal{Y}|\mathbf{m}_\ell, \mathbf{m}_1, ..., \mathbf{m}_{\ell-1}) \prod_{k=1}^{K_\ell} p(m_k|m_{\pi(k)}) \ . \tag{8.7}$$

Pour pouvoir avancer plus avant dans le développement des formules d'estimation des moyennes, nous faisons à présent l'hypothèse que, connaissant l'ensemble des paramètres de GMMs correspondant à différents niveaux de l'arbre, la vraisemblance des données $\mathcal{Y}$ ne dépend que des paramètres du niveau ayant la résolution la plus fine. Autrement dit, nous considérons que le terme de vraisemblance $p(\mathcal{Y}|\mathbf{m}_\ell, \mathbf{m}_1, ..., \mathbf{m}_{\ell-1})$ de l'équation précédente ne dépend que des moyennes $\mathbf{m}_\ell$ du niveau ℓ et s'exprime sous la forme du GMM correspondant. L'expression finale de la règle d'estimation des moyennes $\mathbf{m}_\ell$ s'écrit alors :

$$\widehat{\mathbf{m}}_\ell = \arg\max_{\widehat{\mathbf{m}}_\ell} \sum_{k=1}^{K_\ell} w_k . \mathcal{N}(\mathcal{Y}|m_k, S_k) \prod_{k=1}^{K_\ell} p(m_k|m_{\pi(k)}) \ . \tag{8.8}$$

De même que pour l'adaptation Bayésienne classique de GMMs présentée dans [Gauvain et al.94], nous utilisons l'algorithme EM pour effectuer cette maximisation de façon itérative. Pour cela, nous considérons la fonction auxiliaire suivante :

$$Q(\mathbf{m}_l, \mathbf{m}_l^{(-)}) = E[\log p(\mathbf{W}|\mathcal{Y}, \mathbf{m}_l^{(-)})] + \sum_{k=1}^{K_l} \log p(m_k|m_{\pi(k)}) \ . \tag{8.9}$$

où $\mathbf{W}$ est l'ensemble des données complètes (observations + variables cachées correspondantes).

Comme dans [Gauvain et al.94], la densité *a priori* marginale de la moyenne m_k, conditionnellement à la matrice de covariance S_k, est supposée être une loi normale de moyenne μ_k et de matrice de covariance $\Sigma_k = \frac{1}{\tau_k} S_k$:

$$m_k \sim \mathcal{N}\left(.; \mu_k, \frac{1}{\tau_k} S_k\right) \ . \tag{8.10}$$

Le coefficient τ_k est le facteur de confiance attribué à la moyenne *a priori* marginale μ_k.

La densité *a priori* de la moyenne m_k conditionnellement à l'observation $\tilde{m}_{\pi(k)}$ de sa moyenne parent est alors une loi normale de moyenne $\mu_{k|\pi(k)} = \mu_k + R_{k|\pi(k)}(\tilde{m}_{\pi(k)} - \mu_{\pi(k)})$ et de matrice de covariance $\Sigma_{k|\pi(k)} = \Sigma_k(I - C_{k,\pi(k)}) = \frac{1}{\tau_k} S_k(I - C_{k,\pi(k)})$:

$$m_k|\tilde{m}_{\pi(k)} \sim \mathcal{N}\left(.; \mu_k + R_{k|\pi(k)}(\tilde{m}_{\pi(k)} - \mu_{\pi(k)}), \frac{1}{\tau_k} S_k(I - C_{k,\pi(k)})\right) \ . \tag{8.11}$$

Nous supposons de plus que les matrices de covariance S_k, de régression $R_{k|\pi(k)}$ et de corrélation $C_{k,\pi(k)}$ sont diagonales, ce qui implique que les dépendances locales inter-dimensions ne sont pas modélisées. Sous ces hypothèses et en dérivant l'expression 8.9 par rapport aux paramètres moyennes, la formule permettant de mettre à jour l'estimation de la composante d de la moyenne m_k, dans l'étape M de l'algorithme EM, est la suivante :

$$\widehat{m}_k^d = \frac{\gamma_k}{\gamma_k + \tau_{k|\pi(k)}^d} \overline{y_k}^d + \frac{\tau_{k|\pi(k)}^d}{\gamma_k + \tau_{k|\pi(k)}^d} \mu_{k|\pi(k)}^d \ , \tag{8.12}$$

avec :

$$\mu_{k|\pi(k)}^d = \mu_k^d + r_{k|\pi(k)}^d (\tilde{m}_{\pi(k)}^d - \mu_{\pi(k)}^d) \ , \tag{8.13}$$

$$\tau_{k|\pi(k)}^d = \frac{\tau_k}{1 - (\rho_{k,\pi(k)}^d)^2} \ . \tag{8.14}$$

Les termes γ_k et $\overline{y_k}^d$ sont les statistiques calculées dans l'étape E de l'algorithme EM selon les équations (2.10) et (2.11), et correspondent respectivement au taux d'occupation de la Gaussienne k et à la moyenne empirique des données calculée selon cette Gaussienne k.

Le coefficient $\tau_{k|\pi(k)}^d$ est un facteur de confiance attribué à la composante d de la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}$. Ce facteur de confiance "conditionnel", dépend du facteur de confiance τ_k attribué à la moyenne *a priori* marginale μ_k et prend également en compte l'intensité de la corrélation existant entre m_k^d et $m_{\pi(k)}^d$, au travers du coefficient de corrélation $\rho_{k,\pi(k)}^d$. Associé aux taux d'occupation γ_k, $\tau_{k|\pi(k)}^d$ permet de faire automatiquement la balance entre la moyenne empirique $\overline{y_k}^d$ et la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}^d$ dans l'équation 8.12.

En considérant les formules ci-dessus pour l'estimation des moyennes du GMM de niveau l par l'algorithme EM, l'adaptation H-MAP opère de la manière suivante :

Initialisation : estimer la moyenne "racine" (premier niveau) par adaptation MAP classique et passer au niveau 2.

Pour tout niveau ℓ tel que $2 \leq \ell \leq L$:

1. **Propagation des dépendances :** considérer les moyennes $\{\hat{m}_{\ell-1,\pi(k)}\}$ estimées précédemment au niveau $\ell-1$ comme leurs valeurs observées $\{\tilde{m}_{\ell-1,\pi(k)}\}$ et mettre à jour les densités *a priori* conditionnelles Gaussiennes des moyennes $\{m_{\ell,k}\}$ selon les formules 8.2 et 8.3.
2. **Estimation des moyennes :** estimer les moyennes $\{m_{\ell,k}\}$ au sens du critère MAP avec les densités *a priori* conditionnelles définies à l'étape 1, en utilisant l'algorithme EM.
   (a) Phase E : calculer les statistiques γ_k et $\overline{y_k}^d$ selon les équations (2.10) et (2.11).
   (b) Phase M : estimer les moyennes $\{m_{\ell,k}\}$ selon l'équation (8.12).

Passer au niveau suivant.

Les étapes 1 et 2 sont répétées jusqu'à atteindre le niveau L de résolution la plus fine.

Le GMM estimé dans le dernier niveau L définit le modèle du locuteur qui sera utilisé dans le processus de vérification. Ce modèle GMM de locuteur est caractérisé uniquement par ses paramètres moyennes, les poids et matrices de covariance gardant leur valeur *a priori* donnée par le modèle du monde correspondant au niveau L.

8.4 Propriétés de H-MAP

8.4.1 Influence de la quantité de données : convergence de H-MAP

L'adaptation H-MAP décrite à la section précédente propage les valeurs estimées des moyennes des niveaux les plus hauts vers les niveaux les plus bas au travers des relations de dépendance parent/enfants. Dans les niveaux hauts, qui ont une résolution grossière, les composantes des GMMs reçoivent plus de données que celles des GMMs des niveaux bas, qui ont une résolution fine. L'estimation des moyennes des GMMs dans les niveaux hauts est donc plus robuste.

Ces moyennes estimées de façon robuste sont propagées vers le bas dans l'arbre afin de mettre à jour les densités *a priori* conditionnelles des moyennes des niveaux inférieurs, permettant ainsi à des composantes qui reçoivent peu ou pas de données d'être tout de même modifiées. En effet, d'après l'équation (8.12), lorsque le taux d'occupation γ_k est nul, c'est-à-dire la Gaussienne k ne reçoit aucune donnée, m_k^d est néanmoins adaptée via l'observation de son parent $\tilde{m}_{\pi(k)}$:

$$\gamma_k = 0 \quad \longrightarrow \quad \widehat{m}_k^d = \mu_{k|\pi(k)}^d = \mu_k^d + r_{k,\pi(k)}^d(\tilde{m}_{\pi(k)}^d - \mu_{\pi(k)}^d) \ . \tag{8.15}$$

Au contraire, lorsque γ_k tend vers l'infini, c'est-à-dire la Gaussienne k reçoit une infinité de données, l'estimation de m_k^d selon l'équation 8.12 fournit sa valeur estimée au maximum de vraisemblance :

$$\gamma_k \to \infty \quad \longrightarrow \quad \widehat{m}_k^d \to \overline{y_k}^d \ . \tag{8.16}$$

Ainsi, l'estimation des moyennes d'un GMM selon H-MAP a la même convergence que celle des critères ML et MAP.

La figure 8.2 résume le comportement de l'adaptation H-MAP en fonction de l'occupation des Gaussiennes. Elle représente l'évolution du coefficient de pondération $\alpha_{k|\pi(k)}^d = \frac{\gamma_k}{\gamma_k + \tau_{k|\pi(k)}^d}$ en fonction du rapport $\frac{\gamma_k}{\tau_{k|\pi(k)}^d}$. Le coefficient $\alpha_{k|\pi(k)}^d$ correspond à la proportion attribuée à la moyenne empirique $\overline{y_k}^d$ dans l'estimation de m_k^d. Lorsque $\gamma_k >> \tau_{k|\pi(k)}^d$ la proportion attribuée à $\overline{y_k}^d$ tend vers 100% (estimation ML). Au contraire lorsque $\gamma_k << \tau_{k|\pi(k)}^d$ cette proportion tend vers 0%, laissant ainsi dominer la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}^d$ dans l'estimation de m_k^d.

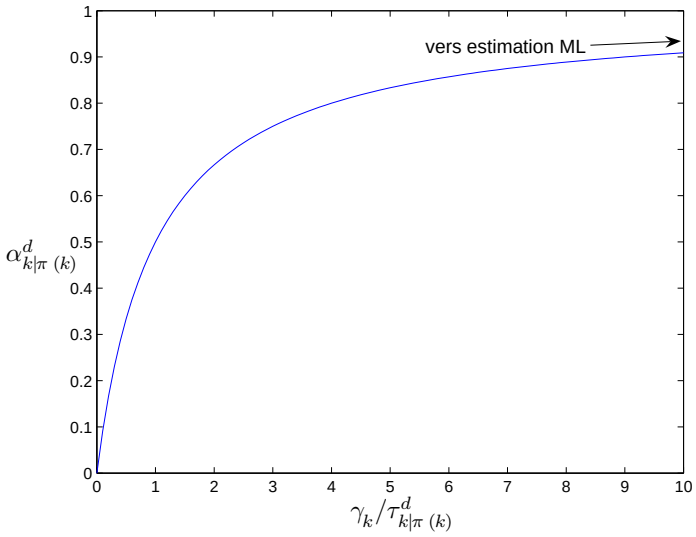


FIG. 8.2 – *Évolution de la proportion attribuée à la moyenne empirique dans l'adaptation H-MAP, en fonction du taux d'occupation de la Gaussienne - Variation du coefficient de pondération* $\alpha^d_{k|\pi(k)} = \frac{\gamma_k}{\gamma_k + \tau^d_{k|\pi(k)}}$ *en fonction du rapport* $\frac{\gamma_k}{\tau^d_{k|\pi(k)}}$.

8.4.2 Influence des dépendances dans H-MAP

Les équations (8.12) et (8.14) définissent une adaptation qui est dépendante de la Gaussienne mais aussi de la dimension considérée. En effet, le taux d'adaptation de la composante d du vecteur moyenne de la Gaussienne k est déterminé non seulement par le taux d'occupation γ_k et le facteur de confiance τ_k de cette Gaussienne, mais également par l'intensité de la corrélation[4] entre les composantes d des moyennes m_k et $m_{\pi(k)}$. Lorsqu'aucune dépendance n'existe entre m^d_k et $m^d_{\pi(k)}$, ou que celles-ci ne sont pas prises en compte (cas de l'adaptation MAP classique), les coefficients de corrélation $\rho^d_{k,\pi(k)}$ et de régression $r^d_{k|\pi(k)}$ sont nuls dans l'équation 2.8. La composante m^d_k est alors estimée selon :

$$\rho^d_{k,\pi(k)} = 0 \quad \text{et} \quad r^d_{k|\pi(k)} = 0 \quad \longrightarrow \quad \widehat{m}^d_k = \frac{\gamma_k}{\gamma_k + \tau_k}\overline{y_k}^d + \frac{\tau_k}{\gamma_k + \tau_k}\mu^d_k \ . \tag{8.17}$$

Cette équation correspond à la formule classique d'adaptation des moyennes d'un GMM au sens du critère MAP (cf. équation 2.8). H-MAP généralise donc l'adaptation MAP classique des moyennes d'un GMM en permettant d'introduire en plus des dépendances entre les moyennes de différentes composantes.

A l'opposé du cas précédent, lorsqu'une relation déterministe lie m^d_k à $m^d_{\pi(k)}$ la valeur absolue du coefficient de corrélation $\rho^d_{k,\pi(k)}$ est égale à 1. Dans ce cas, la valeur

4. c'est-à-dire la valeur absolue du coefficient de corrélation $\rho^d_{k,\pi(k)}$

de m_k^d est entièrement prédictible via l'observation $\tilde{m}_{\pi(k)}^d$ de son parent et les données n'ont plus aucun poids dans l'estimation :

$$|\rho_{k,\pi(k)}^d| = 1 \quad \longrightarrow \quad \widehat{m}_k^d = \mu_k^d + r_{k,\pi(k)}^d(\tilde{m}_{\pi(k)}^d - \mu_{\pi(k)}^d) \ . \tag{8.18}$$

La figure 8.3 illustre l'influence du coefficient de corrélation entre une moyenne m_k^d et son parent $m_{\pi(k)}^d$ sur la valeur du facteur de confiance $\tau_{k|\pi(k)}^d$ attribué à la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}^d$. Elle représente l'évolution du rapport $r_\tau = \frac{\tau_{k|\pi(k)}^d}{\tau_k}$ en fonction du coefficient $(\rho_{k,\pi(k)}^d)^2$. Lorsque le coefficient de corrélation est nul, le rapport r_τ vaut 1 et on se trouve alors dans le cas de l'adaptation MAP classique. Lorsque $\rho_{k,\pi(k)}^d$ est non nul, le rapport r_τ est supérieur à 1 ce qui signifie qu'on attribue un facteur de confiance plus grand à la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}^d$ que celui attribué à la moyenne *a priori* marginale μ_k^q Si $\rho_{k,\pi(k)}^d$ tend vers 1 en valeur abssolue, la relation entre m_k^d et $m_{\pi(k)}^d$ devient déterministe et le facteur de confiance attribué à la moyenne *a priori* conditionnelle $\mu_{k|\pi(k)}^d$ tend vers l'infini (le rapport r_τ également). L'estimation de m_k^d est dans ce cas entièrement déterminée par $\mu_{k|\pi(k)}^d$.

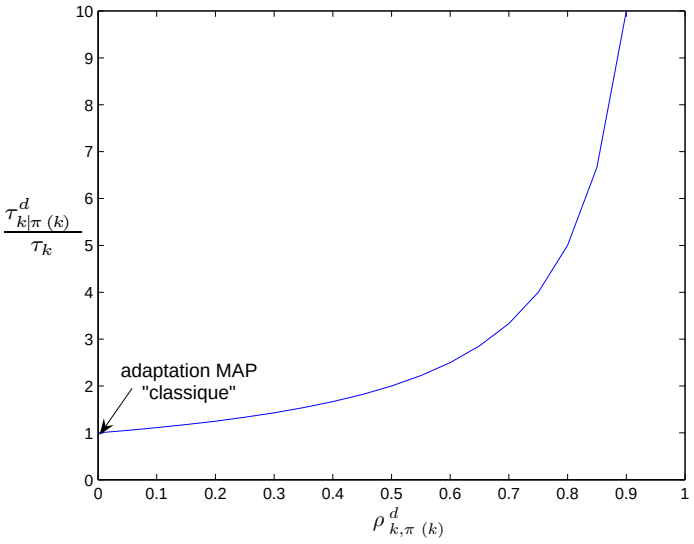


FIG. 8.3 – *Variation du facteur de confiance* $\tau_{k|\pi(k)}^d$ *en fonction du coefficient de corrélation* $\rho_{k,\pi(k)}^d$

8.4.3 H-MAP et autres méthodes d'adaptation

L'adaptation H-MAP a été inspirée de l'adaptation MAP classique développée dans [Gauvain et al.94] et des nombreux schémas hiérarchiques développés dans le cadre de

l'adaptation rapide au locuteur de modèles acoustiques. Dans cette section, nous mettons en évidence des liens existants entre H-MAP et d'autres techniques d'adaptation présentées au chapitre précédent.

H-MAP et méthodes Bayésiennes

Nous avons vu à la section précédente que l'adaptation H-MAP généralise l'adaptation MAP classique en permettant d'introduire, de façon indirecte, des dépendances entre les moyennes de différentes Gaussiennes d'un GMM. Notamment H-MAP conserve la propriété intéressante de convergence vers l'estimation ML.

Notons également une similitude de H-MAP avec la technique RMP (cf. section 7.2.2) utilisant également des relations de régression. Dans le cas de H-MAP, les moyennes parents dans un niveau $\ell-1$ peuvent être considérées comme des paramètres "sources" pour la prédiction des moyennes enfants dans le niveau ℓ, qui sont alors considérées comme les paramètres "cibles". Toutefois, la technique RMP n'utilise pas de structure hiérarchique des dépendances contrairement à H-MAP.

H-MAP et bias-only MLLR/MAPLR

Nous considérons à présent un cas particulier de H-MAP où les relations de dépendance parent/enfant sont déterministes. Cela implique que les coefficients de corrélation $\rho^d_{k,\pi(k)}$ soient tous égaux à 1 en valeur absolue. En considérant de plus des coefficients de régression égaux à 1, une moyenne enfant m_k se prédit directement de l'estimation $\hat{m}_{\pi(k)}$ de son parent selon :

$$\hat{m}_k = \mu_{k|\pi(k)} = \mu_k + (\hat{m}_{\pi(k)} - \mu_{\pi(k)}) \ . \tag{8.19}$$

En notant $\hat{\Delta}_{\pi(k)} = \hat{m}_{\pi(k)} - \mu_{\pi(k)}$ le biais estimé du parent par rapport à sa moyenne *a priori*, cette équation se ré-écrit :

$$\hat{m}_k = \mu_k + \hat{\Delta}_{\pi(k)} \ . \tag{8.20}$$

Cette formule d'adaptation est identique à celle de la technique bias-only MAPLR (ou MLLR si l'estimation de la moyenne parent est faite au maximum de vraisemblance). On pourrait de plus définir le parent $m_{\pi(k)}$ comme correspondant au premier noeud ancêtre de m_k (en remontant dans l'arbre) ayant un taux d'occupation supérieur à un seuil fixé. Cette configuration de H-MAP correspondrait à la technique "tree-based bias-only MAPLR". Notons toutefois que dans H-MAP, le parent d'une Gaussienne dans un niveau donné est également une Gaussienne (de plus basse résolution) dans un niveau supérieur, alors que les techniques MLLR ou MAPLR mettent en jeu des clusters de Gaussiennes pour lesquels des biais globaux sont estimés. La correspondance entre les deux techniques peut être faite directement si l'on approxime la distribution des données pour un cluster de Gaussiennes par une distribution elle-même Gaussienne (de plus basse résolution).

H-MAP et SMAP

On considère maintenant que les coefficients de régression sont égaux à 1 mais que les relations de dépendance parent/enfant ne sont plus déterministes (les coefficients de corrélation sont inférieurs à 1 en valeur absolue). L'équation d'estimation du coefficient d d'une moyenne enfant m_k à partir de l'estimation $\hat{m}^d_{\pi(k)}$ de son parent est alors :

$$\hat{m}^d_k = \frac{\gamma_k}{\gamma_k + \tau^d_{k|\pi(k)}} \overline{y_k}^d + \frac{\tau^d_{k|\pi(k)}}{\gamma_k + \tau^d_{k|\pi(k)}} \left(\mu^d_k + (\hat{m}^d_{\pi(k)} - \mu^d_{\pi(k)}) \right) \quad . \tag{8.21}$$

En notant $\hat{\Delta}^d_k = \hat{m}^d_k - \mu^d_k$ et $\hat{\Delta}^d_{\pi(k)} = \hat{m}^d_{\pi(k)} - \mu^d_{\pi(k)}$ les biais respectifs des coefficients d des moyennes enfant et parent, la relation permettant d'estimer le biais enfants $\hat{\Delta}^d_k$ peut s'écrire :

$$\hat{\Delta}^d_k = \frac{\gamma_k}{\gamma_k + \tau^d_{k|\pi(k)}} \overline{\Delta}^d_k + \frac{\tau^d_{k|\pi(k)}}{\gamma_k + \tau^d_{k|\pi(k)}} \hat{\Delta}^d_{\pi(k)} \quad , \tag{8.22}$$

où $\overline{\Delta}^d_k = \overline{y_k}^d - \mu^d_k$ est le biais enfant estimé au maximum de vraisemblance. Cette dernière relation est similaire à la formule d'estimation des biais de régression dans le schéma d'adaptation SMAP [Shinoda et al.97]. Dans ce schéma, le biais "parent" $\hat{\Delta}^d_{\pi(k)}$ estimé dans le niveau $\ell - 1$ sert de valeur *a priori* pour le biais "enfant" Δ^d_k. L'estimation $\hat{\Delta}^d_k$ du biais "enfant" est alors une combinaison linéaire du biais *a priori* $\hat{\Delta}^d_{\pi(k)}$ et du biais estimé au maximum de vraisemblance $\overline{\Delta}^d_k$, la balance entre les deux se faisant en fonction du taux d'occupation du noeud "enfant".

8.5 Synthèse sur le développement théorique de H-MAP

Le schéma d'adaptation H-MAP est fondé à la base sur l'adaptation Bayésienne classique de modèles GMM mais intègre en plus des informations *a priori* de dépendances entre les Gaussiennes d'un GMM. Ces dépendances interviennent entre les moyennes des Gaussiennes et sont modélisées de façon indirecte par une structure hiérarchique en arbre binaire. Des GMMs multirésolutions sont définis (un GMM par niveau de l'arbre) et des dépendances entre les moyennes de niveaux adjacents sont modélisées par des relations de régression. L'adaptation d'un modèle GMM de locuteur se fait alors de façon hiérarchique en commençant par estimer les moyennes des GMMs de basse résolution (haut de l'arbre) et en propageant ces estimations vers le bas de l'arbre, afin de les utiliser pour l'estimation des moyennes des GMMs de résolution fine. Plus précisément, le procédé consiste à mettre à jour la densité de probabilité *a priori* d'une moyenne "enfant" en fonction de l'estimation de sa moyenne "parent" dans l'arbre. Cela permet d'adapter des composantes n'ayant pas reçu de données dans le niveau le plus bas de l'arbre (GMM de résolution la plus fine) en fonction de paramètres ayant été estimés avec davantage de données, dans les niveaux hauts de l'arbre. Cela peut permettre à l'adaptation H-MAP d'être plus robuste que l'adaptation MAP classique face à de faibles quantités de données d'apprentissage.

L'analyse théorique des propriétés de H-MAP a montré que les estimations fournies par ce schéma d'adaptation convergent vers les estimations au maximum de vraisemblance lorsque la quantité de données augmente. En outre, les formules d'adaptation développées prennent en compte de façon efficace et automatique les relations de dépendance pouvant exister entre les moyennes "enfants" et les moyennes "parents", en tenant compte notamment de l'intensité de la corrélation qui existe entre elles. Cependant, il faut pouvoir estimer de façon fiable ces relations de dépendances pour tirer pleinement partie de la méthode. Cet aspect pratique est discuté dans le chapitre suivant qui traite du développement expérimental de H-MAP.

Signalons enfin que des similitudes entre certaines configurations particulières de H-MAP et d'autres techniques hiérarchiques d'adaptation de modèles acoustiques telles que MLLR, MAPLR ou SMAP ont pu être mises en évidence. L'adaptation H-MAP semble donc pouvoir généraliser un certain nombre de schémas d'adaptation robuste déjà existants.

Chapitre 9

Développement expérimental et évaluation de H-MAP

Dans ce chapitre nous présentons le développement expérimental de la technique d'adaptation Bayésienne hiérarchique H-MAP, et les performances obtenues dans le cadre des évaluations NIST 2003 et NIST 2004.

Nous décrivons la technique de construction de l'arbre de dépendances qui utilise un algorithme hiérarchique de dédoublement de Gaussienne suivi d'une étape d'estimation de modèle. La méthode employée pour estimer les relations de régression est également exposée. Ces régressions sont estimées au maximum de vraisemblance à partir de l'ensemble de données *a priori*, par l'intermédiaire de l'algorihthme EM. Les données *a priori* utilisées pour estimer la structure de dépendances et les matrices de régression proviennent des évaluations NIST 2001 et 2002. Nous présentons ensuite une série d'expériences éffectuées sur NIST 2003 afin de régler le système. Nous l'évaluons également sur NIST2004 pour une tâche de VAL très restrictive en terme de données (seulement 10 secondes de parole à l'apprentissage et 10 secondes au test). Les performances obtenues atteignent le niveau de l'approche classique d'adaptation Bayésienne des modèles de locuteurs, mais H-MAP n'apporte pas d'amélioration concluante même lorsque la quantité de données est très réduite. Nous discutons alors des facteurs applicatifs qui ont pu limiter les bénéfices potentiels de la technique H-MAP, dans le cadre des expériences que nous avons menées, puis nous concluons sur les résultats obtenus.

9.1 Protocole et système de VAL

La technique d'adaptation H-MAP a été développée sur l'évaluation NIST 2003. Elle a ensuite été évaluée sur une tâche à quantité de données très réduite de l'évaluation NIST 2004. Le système de VAL utilisé est dérivé du système IRISA/ELISA 2003 [Ben et al.03b].

Les vecteurs acoustiques délivrés par le module de paramétrisation de ce système sont de dimension 33. Ils comprennent les 16 premiers coefficients cepstraux LFCC, concaténés avec les 16 coefficients Δ correspondants et la Δ-log énergie. Le processus de sélection des trames utiles est basé sur une modélisation bi-gaussienne de l'énergie de chaque énoncé avec une classification parole/silence des trames au maximum de vraisemblance. Les vecteurs acoustiques restants sont ensuite normalisés globalement par centrage/réduction.

Le système de base utilise des GMMs de locuteur à 256 composantes Gaussiennes dont les matrices de covariances sont diagonales. Seules les moyennes des Gaussiennes sont estimées par adaptation MAP des moyennes d'un modèle du monde. Les poids et matrices de covariance ne sont pas adaptés. Ils sont fixés aux valeurs des paramètres correspondants dans le modèle du monde.

Le score brut d'un énoncé correspond à la moyenne des log-rapports de vraisemblances par trame et les normalisations de score D-norm, T-norm et DT-norm sont implémentées.

9.2 Développement sur NIST 2003

9.2.1 Estimation de la structure de dépendance

L'adaptation H-MAP nécessite l'estimation de modèles *a priori* de locuteurs à différents niveaux de granularité. Ces modèles *a priori* sont estimés en entraînant des modèles du monde (UBMs) multirésolutions sur un ensemble de données *a priori*. A partir de ces données *a priori*, il faut également construire une structure hiérarchique représentant les dépendances parent/enfant entre les moyennes des GMMs de niveaux adjacents, et estimer les matrices de régression correspondantes.

Les données *a priori* utilisées pour les expériences de développement sur NIST 2003 sont issues des évaluations NIST 2001 et NIST 2002 (données cellulaires). Elles représentent environ 70 locuteurs par genre, prononçant un énoncé d'environ 2 minutes chacun (mono-session). Dans toutes les expériences effectuées, l'estimation de la structure de dépendance et des UBMs multirésolutions est menée de façon dépendante du genre.

9.2.1.1 Estimation des modèles *a priori*

La structure de dépendance utilisée pour les expériences présentées dans ce chapitre est un arbre binaire balancé à 9 niveaux, de telle sorte que le GMM corres-

pondant au dernier niveau soit composé de 256 Gaussiennes. Ce type de structure a été choisi pour des raisons pratiques liées à la technique de construction de l'arbre. Pour construire cet arbre binaire et estimer les UBMs multirésolutions, un processus hiérarchique de dédoublement de Gaussiennes ("Gaussian splitting") et d'estimation de modèle a été utilisé. Lorsqu'un UBM vient d'être estimé dans un niveau ℓ, chacune de ses composantes est divisée en deux Gaussiennes "enfants" en perturbant légèrement la moyenne "parent" dans deux directions opposées. Le modèle résultant sert d'initialisation à l'UBM du niveau $\ell + 1$. Les matrices de covariance de l'UBM enfant sont initialisées à celles de l'UBM parent et le poids de chaque Gaussienne enfant est initialisée à la moitié du poids de sa Gaussienne parent. L'UBM enfant est ensuite estimé au maximum de vraisemblance via l'algorithme EM.

9.2.1.2 Estimation des matrices de régression et de corrélation

Les relations de régression parent/enfant permettent d'obtenir une prédiction $\hat{m}_k^{(reg)}$ d'une moyenne enfant à partir de l'estimation $\hat{m}_{\pi(k)}$ de sa moyenne parent. Cette prédiction, qui correspond à la moyenne de la loi *a priori* de m_k conditionnellement à $\hat{m}_{\pi(k)}$, est obtenue par la formule:

$$\hat{m}_k^{(reg)} = \mu_{k|\pi(k)} = \mu_k + R_{k|\pi(k)}(\hat{m}_{\pi(k)} - \mu_{\pi(k)}) \ . \tag{9.1}$$

Pour estimer les matrices de régression $R_{k|\pi(k)}$, l'ensemble des données *a priori* $\mathcal{Y}$ est divisé en S sous-ensembles $\{\mathcal{Y}_i\}_{i=1,\ldots,S}$, correspondants aux données de S locuteurs distincts $\{X_i\}_{i=1,\ldots,S}$. A partir de l'estimation au maximum de vraisemblance des moyennes parents (niveau $\ell - 1$) pour chacun des locuteurs X_i, on définit un ensemble de modèles GMMs $\{p(y|X_i)\}_{i=1,\ldots,S}$ pour le niveau ℓ en utilisant la formule de prédiction des moyennes donnée à l'équation 9.1. Les poids et matrices de covariance de ces modèles sont fixés aux valeurs données par l'UBM de niveau ℓ. Un modèle $p(y|X_i)$ est alors défini par l'ensemble des paramètres $\Lambda_{X_i} = (\{w_k\}, \{\hat{m}_{i,k}^{(reg)}\}, \{S_k\})$.

On estime alors les matrices de régression $R_{k|\pi(k)}$ pour le niveau ℓ de façon à maximiser la vraisemblace des données $\{\mathcal{Y}_i\}_{i=1,\ldots,S}$ vis-à-vis des modèles $\{p(y|X_i)\}_{i=1,\ldots,S}$:

$$\{\hat{R}_{k|\pi(k)}\} = \arg \max_{\{R_{k|\pi(k)}\}} \prod_{i=1}^{S} p(\mathcal{Y}_i|\{\{w_k\}, \{\hat{m}_{i,k}^{(reg)}\}, \{S_k\}) \ . \tag{9.2}$$

Nous avons mené cette maximisation via l'algorithme EM. En supposant les matrices de covariance et de régression diagonales, la formule d'estimation de l'élément $r_{k|\pi(k)}^d$ de la matrice $R_{k|\pi(k)}$ est :

$$\hat{r}_{k|\pi(k)}^d = \frac{\sum_{i=1}^{S} \gamma_{i,k}(\overline{y}_{i,k}^d - \mu_k^d)(\hat{m}_{i,\pi(k)}^d - \mu_{\pi(k)}^d)}{\sum_{i=1}^{S} \gamma_{i,k}(\hat{m}_{i,\pi(k)}^d - \mu_{\pi(k)}^d)^2} \ , \tag{9.3}$$

où :
$\gamma_{i,k}$ est le taux d'occupation par les données $\mathcal{Y}_i$ de la Gaussienne k du modèle $p(y|X_i)$;

$\overline{y}_{i,k}^d$ est le coefficient d de la moyenne empirique des données $\mathcal{Y}_i$, calculée selon la Gaussienne k du modèle $p(y|X_i)$.

Le coefficient de corrélation correspondant est donné par :

$$\hat{\rho}_{k,\pi(k)}^d = \frac{\sum_{i=1}^S \gamma_{i,k}(\overline{y}_{i,k}^d - \mu_k^d)(\hat{m}_{i,\pi(k)}^d - \mu_{\pi(k)}^d)}{\sqrt{\sum_{i=1}^S \gamma_{i,k}(\overline{y}_{i,k}^d - \mu_k^d)^2 \sum_{i=1}^S \gamma_{i,k}(\hat{m}_{i,\pi(k)}^d - \mu_{\pi(k)}^d)^2}} \quad . \tag{9.4}$$

Les formules dérivées ci-dessus permettent d'estimer les régressions et corrélations définies par la structure de dépendances en arbre binaire, à partir d'un ensemble de données *a priori*. La qualité des estimations des coefficients de régression va dépendre de la quantité de données *a priori* disponible. D'autre part, les corrélations correspondantes seront également liées à cette quantité de données mais également à la capacité qu'a la structure de dépendances à capturer des liens effectifs entre les moyennes de niveaux adjacents.

Les figures 9.1 et 9.2 montrent respectivement les distributions des coefficients de régression $r_{k|\pi(k)}^d$et de corrélation $\rho_{k,\pi(k)}^d$ obtenus dans l'ensemble de l'arbre, toutes dimensions confondues (courbes continues), pour les coefficients acoustiques statiques seuls (courbes en pointillés/points) et pour les coefficients acoustiques dynamiques seuls (courbes en pointillés). Les données utilisées correspondent à un sous-ensemble des locuteurs femmes (74 individus) des évaluations NIST'01 et NIST'02 (téléphonie cellulaire).

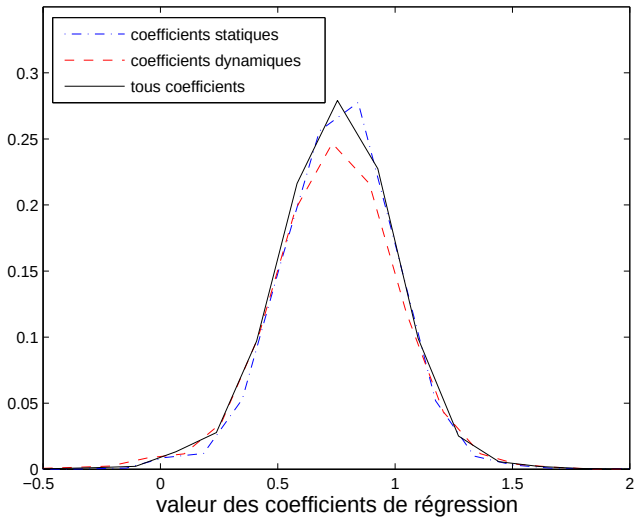


FIG. 9.1 – *Distribution des coefficients de régression estimés sur les données cellulaires NIST'01 et NIST'02.*

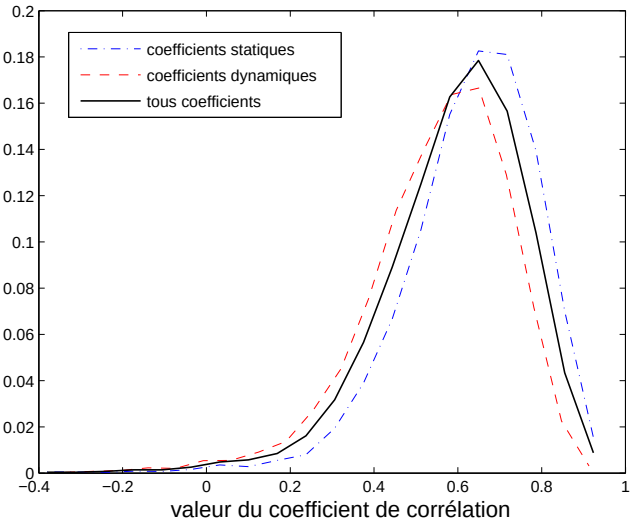


FIG. 9.2 – *Distribution des coefficients de corrélation estimés sur les données cellulaires NIST'01 et NIST'02.*

Les coefficients de régression sont en très grande majorité positifs ce qui indique que les moyennes enfants ont tendance à suivre le "mouvement" de leur moyenne parent dans l'espace acoustique. Cela est cohérent avec le fait qu'une Gaussienne parent et ses deux Gaussiennes enfants modélisent une même région de l'espace acoustique. En outre, l'observation de cette dépendance intuitive parent/enfant indique que l'algorithme de construction de l'arbre permet de capturer des dépendances effectives entre les moyennes de niveaux adjacents. Cependant, les corrélations obtenues sont assez faibles dans l'ensemble, comme le montre la figure 9.2. Les coefficients de corrélation ont une valeur absolue moyenne de $0,6$. Tous niveaux confondus seulement 30% d'entre eux ont une valeur absolue supérieure à $0,7$, ce qui veut dire que seulement une régression sur trois environ est réellement fiable. Cela peut indiquer que la quantité de données *a priori* utilisée pour estimer les régressions n'est pas suffisante ou que l'algorithme de construction de la structure de dépendance ne permet pas de capturer systématiquement des relations fortes entre les moyennes de niveaux adjacents. D'autre part, les dépendances entre coefficients acoustiques statiques semblent légèrement mieux prises en compte par la structure utilisée que celles existant entre les coefficients acoustiques dynamiques. En effet, d'après la figure 9.2 les corrélations obtenues pour les premiers sont un peu meilleures que celles obtenues pour les seconds. Par contre, il n'y a pas de différence notable entre les distributions des coefficients de régression obtenus pour les coefficients acoustiques statiques et dynamiques (cf figure 9.1).

Malgré la faible proportion de corrélations fortes observées entre moyennes parents et enfants, l'utilisation des régressions obtenues peut servir à mieux définir les lois *a*

priori dans le schéma d'adaptation des modèles de locuteur par H-MAP. Cela pourrait néanmoins mener à de meilleures estimations des paramètres des Gaussiennes recevant peu de données d'apprentissage.

9.2.2 Adaptation des modèles de locuteur

L'adaptation des moyennes $\{m_k\}$ des modèles de locuteur pour un niveau ℓ de l'arbre est effectuée via l'algorithme EM par les formules dérivées à la section 8.3.2 du chapitre précédent. Nous rappelons ici la formule permettant d'estimer le coefficient d de la moyenne m_k :

$$\widehat{m}_k^d = \frac{\gamma_k}{\gamma_k + \tau_{k|\pi(k)}^d} \overline{y_k}^d + \frac{\tau_{k|\pi(k)}^d}{\gamma_k + \tau_{k|\pi(k)}^d} \mu_{k|\pi(k)}^d \ , \tag{9.5}$$

avec

$$\tau_{k|\pi(k)}^d = \frac{\tau_k}{1 - (\rho_{k,\pi(k)}^d)^2} \ , \tag{9.6}$$

$$\mu_{k|\pi(k)}^d = \mu_k^d + r_{k|\pi(k)}^d (\hat{m}_{\pi(k)}^d - \mu_{\pi(k)}^d) \ . \tag{9.7}$$

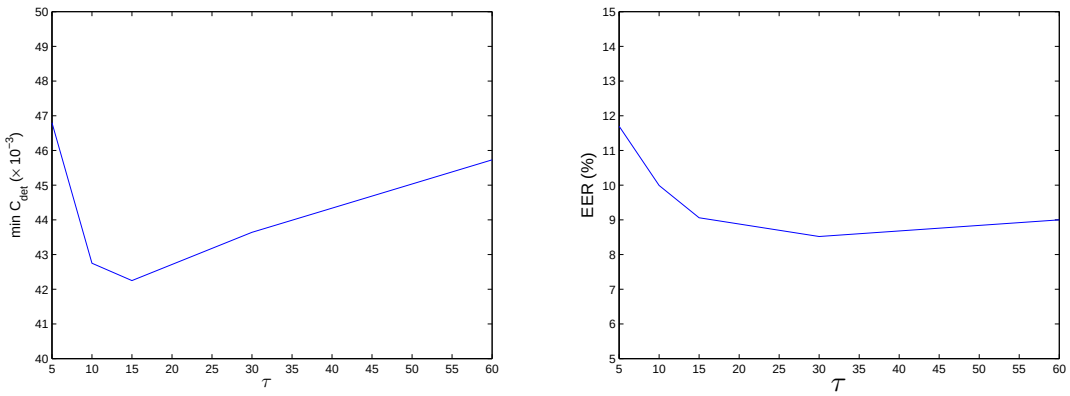
La définition des différents éléments intervenant dans ces équations est donnée à la section 8.3.2. Cette formule d'estimation fait notamment intervenir le facteur de confiance τ_k attribué à la Gaussienne k dans le niveau ℓ. Comme dans [Reynolds et al.00], nous considérons des facteurs de confiance indépendants de la Gaussienne. D'autre part, si l'on considère que les facteurs de confiance sont homogènes à un nombre de trames *a priori* attribuées à chaque Gaussienne, il est logique de les faire varier dans l'arbre proportionnellement au nombre de Gaussiennes de chaque niveau. Le facteur $\tau_k^{(\ell)}$ pour le niveau ℓ est alors déterminé selon l'équation :

$$\tau_k^{(\ell)} = \tau \frac{K_L}{K_\ell} \ , \tag{9.8}$$

où K_L et K_ℓ sont respectivement le nombre de Gaussiennes dans le dernier niveau L de l'arbre, et dans le niveau ℓ. Cette façon de déterminer les facteurs de confiance dans chaque niveau permet de n'avoir qu'un seul coefficient τ à optimiser, correspondant au facteur de confiance affecté au dernier niveau L de l'arbre (on a en effet $\tau_k^{(L)} = \tau$). Ce facteur τ est optimisé sur la base de données de développement.

9.2.3 Performances sur NIST 2003

Pour la tâche de vérification en elle-même, les GMMs de locuteurs estimés dans le dernier niveau de l'arbre sont utilisés pour représenter l'hypothèse "client" H_X et le modèle du monde correspondant à ce niveau sert de modèle de non-locuteur commun pour représenter l'hypothèse "imposteur" $H_{\bar{X}}$.



(a) (b)

FIG. 9.3 – *Évolutions des points* $\min C_{det}$ *(a) et EER (b) en fonction du facteur de confiance* τ *avec l'adaptation H-MAP (corpus femmes NIST'03).*

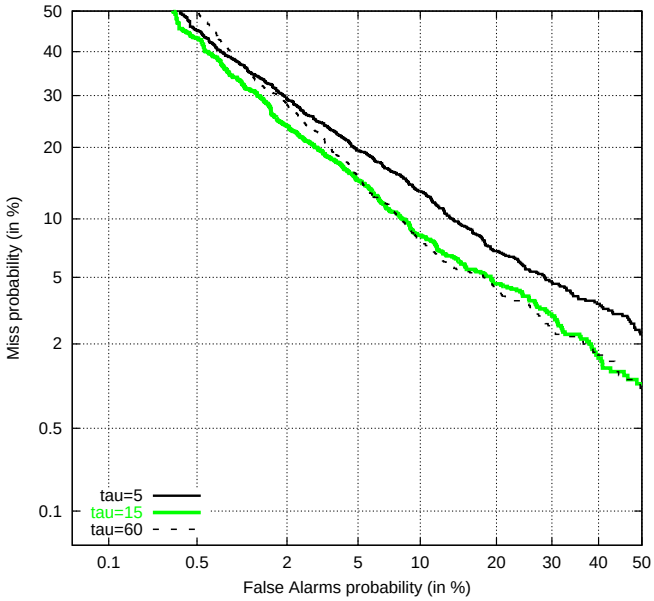


FIG. 9.4 – *Courbes DET obtenues avec l'adaptation H-MAP sans normalisation (corpus femmes NIST'03) : influence du facteur de confiance.*

Réglage du facteur de confiance τ

La figure 9.3 donne l'évolution des points $\min C_{det}$ (a) et EER (b) en fonction du facteur de confiance τ, pour le système utilisant l'adaptation H-MAP sans normalisation de score. D'après ces graphiques, les performances sont optimisées au point $\min C_{det}$ pour la valeur $\tau = 15$ et à l'EER pour $\tau = 30$. Cependant, les différences observées au point $\min C_{det}$ ne sont pas statistiquement significatives[1] (avec un taux de confiance de 90%) pour des valeurs de τ comprises entre 10 et 50, et à l'EER, les différences ne sont pas significatives (avec un taux de confiance de 90%) pour τ variant de 10 à 60.

La figure 9.4 permet d'observer globalement l'influence du facteur τ sur le comportement du système. Elle représente les courbes DET obtenues avec l'adaptation H-MAP sans normalisation de score et avec des facteurs de confiance $\tau = 5$, $\tau = 15$ et $\tau = 60$. D'après ces courbes, une valeur de τ trop faible (illustrée par la courbes DET pour $\tau = 5$) entraîne principalement une perte de performance dans la région à faibles taux de faux rejets. Une valeur trop forte de τ (illustrée par la courbes DET pour $\tau = 60$) entraîne une perte de performance dans la région à faibles taux de fausses acceptations. Une valeur intermédiaire (illustrée ici par $\tau = 15$) permet d'avoir de bonnes performances sur l'ensemble de la courbe DET.

Performances avec les normalisations de score

Le tableau 9.1 résume les valeurs obtenues pour les points $\min C_{det}$ et EER par l'adaptation classique MAP et l'adaptation H-MAP, avec les différentes normalisations de scores implantées (corpus femmes NIST'03).

	MAP		**H-MAP**	
	$\min C_{det}$ ($\times 10^{-3}$)	EER (%)	$\min C_{det}$ ($\times 10^{-3}$)	EER (%)
D-norm	38,6	8,3	40,2	8,7
T-norm	32,6	**7,9**	**31,8**	8,6
DT-norm	32,2	8,1	33,4	8,2

TAB. 9.1 – *Points $\min C_{det}$ et EER obtenus avec l'adaptation MAP classique et l'adaptation H-MAP, avec différentes normalisations de scores (corpus femmes NIST'03)*

Les meilleures performances au point $\min C_{det}$ sont obtenues par l'adaptation H-MAP avec normalisation T-norm. A l'EER c'est l'adaptation MAP classique qui obtient systématiquement les meilleures performances, quelle que soit la normalisation. Il faut noter cependant que pour chaque normalisation considérée dans ce tableau, aucune des différences observées entre l'adaptation MAP classique et l'adaptation H-MAP n'est significative (avec un taux de confiance de 90%), ni à l'EER ni au point $\min C_{det}$.

1. Les tests de confiance utilisés sont ceux proposés dans [Bengio et al.04] et sont basés sur le classique "z-test" entre des proportions

Sur la figure 9.5 sont tracées les courbes DET des systèmes MAP et H-MAP avec les normalisations de score donnant les meilleurs résultats au point $\min C_{det}$ (DT-norm pour MAP et T-norm pour H-MAP). L'adaptation H-MAP donne des performances légèrement meilleures pour les taux de fausses acceptations faibles ($< 1\%$) sinon c'est l'adaptation MAP classique qui permet d'obtenir des résultats un peu meilleurs. Cependant, les écarts sont minimes et pourraient être simplement dus à des différences dans l'optimisation des réglages des deux systèmes.

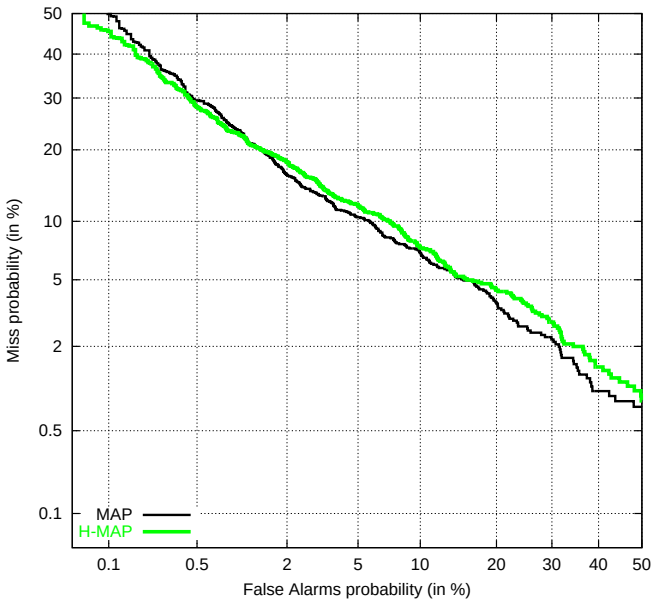


FIG. 9.5 – *Courbes DET du système IRISA 2003 avec adaptation MAP classique et DT-norm et avec adaptation H-MAP et T-norm (corpus femmes NIST'03)*

Sur l'évaluation NIST'03, l'adaptation hiérarchique H-MAP semble donc ne pas apporter de bénéfice clair par rapport à une adaptation MAP classique. Notons cependant que pour cette évaluation, la quantité de données d'apprentissage (environ 2 minutes de parole), bien que limitée, n'est pas extrêmement critique. On peut penser que cette quantité de données est suffisante pour estimer les modèles de locuteurs de façon relativement correcte avec l'adaptation MAP classique et que, par conséquent, l'adaptation H-MAP n'apporte pas grand chose dans ce cas.

Dans la section suivante, nous analysons les performances de l'adaptation H-MAP sur une tâche de VAL proposée à l'évaluation NIST de 2004, beaucoup plus critique du point de vue de la quantité de données disponibles.

9.3 Évaluation sur NIST 2004

En 2004, de nombreuses configurations différentes ont été proposées par NIST lors de ses évaluations, avec notamment une tâche où la quantité de données disponibles à l'apprentissage et au test est limitée à 10 secondes de parole.

L'adaptation H-MAP a été testée sur cette tâche afin d'évaluer l'apport éventuel de cette adaptation hiérarchique lorsque l'on dispose de très peu de données d'apprentissage et de test. Notamment, cette forte limitation du volume de données peut entraîner une grande disparité de contenu linguistique entre les énoncés d'apprentissage et de test. La prise en compte d'informations acoustiques plus globales dans le schéma d'adaptation hiérarchique H-MAP, par rapport à l'adaptation MAP classique, pourrait permettre de diminuer en partie les effets de ces fortes disparités de contenus linguistiques.

9.3.1 Description des données

9.3.1.1 Données d'évaluation

Les données utilisées pour l'ensemble des tâches de l'évaluation NIST'04 sont issues de la base de données MIXER [Ldc]. Elle contient des conversations téléphoniques avec une large variété de conditions d'acquisition. Notamment, la transmission téléphonique peut être filaire ou mobile avec différents formats de transmission. La langue utilisée est en majeure partie de l'anglais américain, mais on peut également rencontrer quelques conversations dans d'autres langues. Notons que les conditions rencontrées dans la base MIXER sont assez différentes de celles de la base SWITCHBOARD qui a servi d'ensemble de développement pour l'évaluation NIST'04. De plus, le traitement opéré par NIST sur les données d'évaluation délivrées aux participants cette année n'est pas le même que celui qui était appliqué les années précédentes (il n'y a pas eu de retrait automatique des silences sur les données MIXER). Ces changements de conditions entre les données de développement et celles d'évaluation expliquent en partie la baisse générale des performances des systèmes observée sur NIST'04. Pour plus de détail on pourra se référer au plan d'évaluation de la campagne NIST 2004 [Przybocki et al.04].

9.3.1.2 Données de développement

Pour estimer l'arbre de dépendance et les UBMs multirésolutions, un sous-ensemble des données de téléphonie cellulaire des évaluations NIST'01 et NIST'02 (74 femmes et 64 hommes prononçant un énoncé d'environ 2 minutes) et un sous-ensemble des données de téléphonie filaire de l'évaluation NIST'01 (75 femmes et 75 hommes prononçant un énoncé d'environ 2 minutes) ont été utilisés. D'autre part, afin d'appliquer la normalisation de score T-norm, un ensemble de modèles imposteurs a été estimé sur une partie des données cellulaires et filaires restantes (50 hommes et 50 femmes en cellulaire, 50 hommes et 50 femmes en filaire).

Les données de l'évaluation NIST'03 ont servi à régler le système.

Il faut rappeler que les données de développement citées ci-dessus et issues des bases SWITCHBOARD et SWITCHBOARD Cellular, ne sont pas homogènes aux données de l'évaluation NIST'04 qui sont issues de la base MIXER. On rencontre dans MIXER plus de variabilité dans les conditions d'utilisation et certains énoncés peuvent être dans une langue différente de l'anglais américain. Rappelons également le pré-traitement différent appliqué par NIST sur les données MIXER par rapport aux données SWITCHBOARD (cf. paragraphe précédent). L'arbre de dépendances et les modèles du monde multirésolutions ont donc été appris sur des données dont les conditions diffèrent légèrement de celles rencontrées dans l'évaluation.

9.3.2 Résultats

Le système IRISA avec l'adaptation H-MAP a été utilisé pour la tâche "10sec-train/10sec-test" proposée par NIST (10 secondes de parole sont disponible, à la fois pour l'apprentissage et pour le test). Nous comparons les performances obtenues par ce système avec celles données par le système IRISA de base utilisant une adaptation MAP classique.

La figure 9.6 montre les courbes DET obtenues avec les deux types d'adaptation, avec normalisation T-norm. Les plages de variation des taux d'erreurs représentés sur les axes ont été largement élargies car la tâche considérée engendre des points de fonctionnement $minC_{det}$ avec un taux de faux rejets qui est au-delà des 80%.

Sur une large plage de ces courbes, aucune différence significative n'est observée entre les adaptations H-MAP et MAP classique. Les points $\min C_{det}$ et EER correspondants sont donnés dans le tableau 9.2. Les différences observées sur les EER ne sont pas statistiquement significatives. La différence entre les points $\min C_{det}$ est significative avec un taux de confiance de 90% mais ne l'est pas pour un taux de confiance de 95%. Bien que les résultats obtenus soient encourageants, l'amélioration apportée par H-MAP sur cette tâche restrictive en données n'est pas complètement concluante pour le moment.

	$\min C_{det}$ ($\times 10^{-3}$)	EER (%)
MAP	95,8 (1,34%FA/82,6%FR)	30,0
H-MAP	92,6 (0,87%FA/84,0%FR)	30,7

TAB. 9.2 – *Points* $\min C_{det}$ *et EER obtenus avec l'adaptation MAP classique et l'adaptation H-MAP pour la tâche 10sec-train/10sec-test de l'évaluation NIST'04*

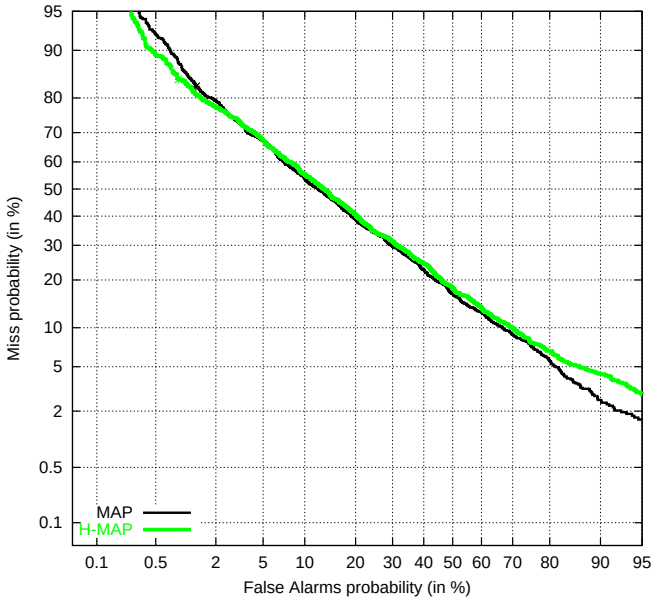


FIG. 9.6 – *Courbes DET du système IRISA avec adaptation MAP classique et adaptation H-MAP, et normalisation T-norm, pour la tâche 10sec-train/10sec-test de l'évaluation NIST'04.*

9.3.3 Discussion sur les résultats obtenus

Pour cette tâche très restrictive du point de vue de la quantité de données disponibles, le schéma d'adaptation hiérarchique utilisé dans H-MAP n'a pas apporté d'amélioration notable par rapport à une adaptation MAP classique. Les corrélations obtenues entre moyennes parents et enfants lors de la construction de l'arbre de dépendance pour l'évaluation NIST'04 sont similaires à celles observées lors du développement sur NIST'03 : la valeur absolue moyenne des coefficients de corrélation est de 0,56 et seulement 27% sont supérieurs à 0,7. Ainsi, malgré l'ajout de données *a priori* supplémentaires (75 hommes et 75 femmes en plus, en téléphonie filaire) les corrélations obtenues ne sont pas meilleures et même légèrement moins bonnes. En résumé, les conditions dans lesquelles a été mise en oeuvre l'adaptation H-MAP sur les bases de données NIST'03 et NIST'04 n'ont pas permis de capturer un nombre important de dépendances fortes entre les moyennes parents et les moyennes enfants. Ce problème peut avoir deux causes principales :

1. la quantité de données utilisée pour estimer les dépendances n'est pas suffisante;
2. l'algorithme de construction de la structure de dépendance ne permet pas de capturer les dépendances les plus fortes pouvant exister entre les Gaussiennes des GMMs de niveaux adjacents.

Dans les deux cas, nous sommes limités par les bases de données dont nous disposons, qui ne sont pas suffisamment conséquentes et qui ne sont pas étiquetées phonétiquement

(ce qui pourrait aider à structurer plus efficacement les relations de dépendances).

9.4 Conclusion sur la mise en application de H-MAP

Malgré l'avantage théorique de H-MAP par rapport à une adaptation MAP classique, ce nouveau schéma d'adaptation hiérarchique n'a pas permis d'obtenir d'amélioration décisive des performances pour l'instant, même lorsque la quantité de données disponible est très réduite. Par contre, H-MAP ne dégrade pas non plus les résultats par rapport à l'adaptation Bayésienne classique. Il peut alors être considéré comme une extension de l'adaptation MAP (cf. section 8.3) dont les conditions où il présente un avantage n'ont pas pu être mises clairement en évidence pour le moment. Malgré tout, les résultats obtenus sont encourageants et offrent de nombreuses perspectives.

Les principales améliorations doivent être apportées à l'estimation de la structure de dépendances et des relations de dépendance elles-mêmes. Dans le cadre de notre travail, l'algorithme utilisé pour construire la structure de dépendance n'a pas permis d'obtenir des corrélations fortes entre les moyennes parents et enfants. Les relations de régression parent/enfant obtenues étaient par conséquent peu fiables.

Cette difficulté à capturer des relations fortes entre Gaussiennes de niveaux adjacents a pu provenir d'un manque de données ou de la façon même de construire l'arbre de dépendance. L'algorithme que nous avons utilisé établit implicitement des liens de dépendance entre Gaussiennes parents et enfants modélisant une même région acoustique. L'observation des régressions obtenues a montré qu'il existe effectivement des dépendances "intuitives" entre des Gaussiennes parents et enfants qui sont proches dans l'espace acoustique, mais les corrélations correspondantes étaient trop faibles en moyenne pour que H-MAP puisse apporté une amélioration significative.

Une stratégie différente de recherche des dépendances entre les Gaussiennes d'un GMM pourrait permettre de capturer des corrélations plus fortes. Notamment, il serait envisageable d'utiliser des informations linguistiques pour construire la structure de dépendance. Ce type d'informations a déjà montré son intérêt en vérification du locuteur pour aider à mieux structurer le modèle du monde et les modèles de locuteurs ("phonetically-structured GMM" [Faltlhauser et al.01]). Il est probable qu'elles peuvent également servir à mieux définir les relations de dépendance hiérarchiques entre les Gaussiennes dans les différents niveaux de l'arbre. Cette approche nécessite cependant d'avoir une base de données étiquetée phonétiquement pour affecter les Gaussiennes aux différentes classes et macro-classes phonétiques dans l'arbre de dépendances.

Quoiqu'il en soit, les développements futurs sur H-MAP devront immanquablement passer par l'utilisation d'une base de données plus importante et fournissant si possible des informations supplémentaires, en particulier linguistiques, afin de pouvoir explorer de nouvelles stratégies pour la structuration et l'estimation des dépendances entre les Gaussiennes des modèles de locuteur. Le développement théorique de l'adaptation H-

MAP offre un cadre puissant pour intégrer ces dépendances. De plus, H-MAP conserve les avantages et les propriétés de l'approche classique d'adaptation MAP des GMMs, qui constitue actuellement l'état de l'art des systèmes de VAL par approche probabiliste.

La technique d'adaptation H-MAP a donné lieu à une publication en workshop [Ben et al.04c].

Conclusion générale

Depuis plus de dix ans, l'approche probabiliste a dominé l'état de l'art des méthodes utilisées en vérification automatique du locuteur (VAL). La majorité des systèmes de VAL actuels utilisent cette approche, parfois associée maintenant à d'autres techniques de classification comme les SVM (Support Vector Machines). Depuis quelques années, un des enjeux majeurs de la recherche dans le domaine a été d'améliorer la robustesse des systèmes de VAL lorsqu'ils sont utilisés en situation réelle et dans des environnements perturbés, pour les applications téléphoniques notamment. Cette robustesse est en effet un point incontournable pour permettre le déploiement de tels systèmes au niveau du grand public. Parmi les techniques développées à cet effet, les normalisations et l'adaptation Bayésienne sont sans aucun doute celles qui ont amené les améliorations les plus remarquables, comme cela a pu être mis en évidence lors des évaluations NIST en reconnaissance du locuteur. Cependant les techniques actuelles ont des faiblesses qui les rendent contraignantes ou inadaptées à certaines situations applicatives. En particulier, les normalisations courantes demandent souvent une mise en oeuvre lourde. L'adaptation Bayésienne des modèles de locuteurs atteint quant à elle ses limites lorsque la quantité de données d'apprentissage est très limitée car elle ne permet pas d'obtenir des modèles suffisamment caractérisants dans ce cas. Dans cette thèse, nous avons développé des techniques destinées à remédier à certaines de ces limitations.

Dans une première partie des travaux, nous avons élaboré et mis en oeuvre des techniques de normalisation permettant d'alléger la procédure de vérification par rapport aux approches courantes. Les normalisations proposées utilisent des distances entre modèles de locuteurs. Une deuxième partie des travaux a été consacrée à la conception et au développement d'un schéma d'adaptation Bayésienne hiérarchique dont l'objectif est d'améliorer la robustesse de l'estimation des modèles de locuteurs dans le cas de données d'apprentissage très limitées. Ce schéma permet de généraliser l'approche d'adaptation Bayésienne des modèles classiquement utilisée en VAL, en offrant de plus la possibilité d'intégrer des dépendances entre régions acoustiques.

Nous rappelons tout d'abord les principales contributions de cette thèse et tirons les enseignements des travaux effectués. Nous terminons la conclusion en présentant les perspectives que nous envisageons en termes de développements supplémentaires de nos travaux.

Contributions

Techniques de normalisation par distance de Kullback-Leibler

Les approches classiques de normalisation de scores utilisées en VAL, notamment la Z-norm et la T-norm, nécessitent une mise en oeuvre contraignante : un ensemble de données externes doit être disponible et une série de tests réels doit être effectuée pour estimer les paramètres de normalisation. Afin d'alléger la procédure de normalisation, nous avons exploré des techniques basées sur des distances entre modèles. Les normalisations proposées ne nécessitent pas l'utilisation de données réelles et permettent donc une mise en application plus facile.

Nous avons tout d'abord mis en évidence le lien théorique existant entre les divergences de Kullback-Leibler (KL) calculées entre les modèles de locuteurs et le modèle du monde, et les log-rapports de vraisemblances (LLRs) classiquement utilisés comme scores de vérification en VAL. En pratique, ce lien s'est traduit par l'observation d'une forte corrélation entre les distances KL et la moyenne des scores imposteurs pour chaque locuteur. C'est à partir de cette relation que nous avons élaboré de nouvelles techniques de normalisation basées sur les distances KL.

La première technique, appelée D-norm, agit au niveau des scores. Elle permet d'homogénéiser la distribution des scores imposteurs par simple division du score brut par la distance KL associée au modèle de l'identité proclamée. Évaluée sur les données NIST, la D-norm a fourni des performances équivalentes à une technique courante de normalisation de scores, la Z-norm, elle aussi destinée à compenser les biais liés aux modèles. D'un point de vue applicatif, la D-norm peut avantageusement remplacer la Z-norm car elle permet une mise en oeuvre moins lourde et un calcul plus rapide des paramètres de normalisation.

La seconde technique de normalisation que nous avons développée agit quant à elle au niveau des modèles. Cette technique, appelée D-MAP (Distance-constrained MAP), permet d'homogénéiser les modèles de locuteurs vis-à-vis de leur distance KL par rapport au modèle du monde. Elle s'appuie sur un schéma d'adaptation Bayésienne contraint qui détermine un facteur d'adaptation propre à chaque locuteur afin de "placer" tous les modèles de locuteurs à la même distance KL du modèle du monde. L'évaluation expérimentale de D-MAP sur les données NIST a montré que ce schéma d'adaptation particulier permet d'obtenir des performances meilleures qu'un schéma d'adaptation classique réglé de façon optimale, lorsqu'on n'utilise pas de normalisation de scores. En outre, les expériences ont mis en évidence que les normalisations de scores destinées à compenser les biais liés aux modèles, comme la Z-norm et la D-norm, ne sont plus nécessaires quand on utilise l'adaptation D-MAP. En effet, ces biais semblent implicitement compensés par la normalisation appliquée au niveau des modèles.

Enfin, nous avons formalisé un nouveau cadre pour la vérification du locuteur dans un espace des modèles offrant de nombreuses possibilités de normalisations. Dans cet espace, chaque modèle de locuteur est représenté par un point, le point origine correspondant au modèle du monde. La vérification du locuteur s'effectue alors par un calcul de score qui fait intervenir des distances Euclidiennes entre le modèle du locuteur pro-

clamé, le modèle du monde et un modèle appris sur les données de test. La distance Euclidienne utilisée définit une mesure de similarité entre modèles qui s'exprime comme une fonction simple des paramètres des modèles. L'espace engendré autorise l'utilisation de la géométrie Euclidienne pour manipuler les paramètres des modèles, ce qui nous a permis d'élaborer une normalisation très simple des modèles inspirée des principes du D-MAP. Elle permet d'homogénéiser l'ensemble des modèles de locuteurs et de test vis-à-vis de leur distance Euclidienne par rapport au modèle du monde. Évaluée sur les données NIST 2004, cette approche de vérification du locuteur dans un espace des modèles a montré des résultats très intéressants. Notamment, lorsqu'aucune normalisation de scores n'est appliquée, les performances obtenues sont meilleures que celles fournies par l'approche conventionnelle basée sur les scores LLRs. Cela est probablement dû à la normalisation de modèles utilisée qui peut être appliquée aussi bien sur les modèles de locuteurs que sur les modèles de tests, permettant ainsi de compenser à la fois des biais liés aux données d'apprentissage et aux données de test. Lorsque la T-norm est appliquée au niveau des scores, l'approche proposée fournit des performances légèrement meilleures que l'approche classique. D'un point de vue applicatif, un avantage de la vérification du locuteur dans l'espace des modèles est que le score de vérification en lui-même est très simplement et rapidement calculable une fois que le modèle de test a été estimé. En particulier, le calcul des paramètres de normalisation nécessaires à la T-norm est beaucoup plus rapide dans l'espace des modèles que dans l'espace des LLRs, ce qui permet d'appliquer cette normalisation en ne rallongeant que de façon négligeable le temps de test.

En connexion avec les travaux précédents, nous avons également utilisé les distances entre GMMs pour des tâches autres que la VAL :

- dans le cadre de la tâche SRL de l'évaluation ESTER, la distance Euclidienne a été utilisée avec succès en organisation en locuteurs de documents radiophoniques. La méthode que nous avons proposée a permis d'obtenir des résultats à un niveau équivalent à ceux des systèmes état de l'art.
- dans le cadre du projet TECHNOLANGUE-NEOLOGOS, nous avons appliqué les distances KL dans une procédure de sélection de locuteurs, le but étant de trouver parmi un groupe de 1000 locuteurs, les 200 locuteurs qui représentent le mieux ce groupe, selon des critères de qualités variés.

Un certain nombre d'enseignements peuvent être tirés de l'ensemble de ces travaux sur l'utilisation de distances entre les modèles de locuteurs et le modèle du monde. Les conclusions auxquelles nous sommes arrivés permettent de mieux comprendre certains liens existants entre les caractéristiques particulières d'un modèle et les scores qu'il engendre. Les travaux sur la D-norm ont montré qu'une partie de l'information utile à la normalisation des scores peut être trouvée directement au niveau des modèles. En effet, la distance KL entre un modèle de locuteur et le modèle du monde semble déterminer le comportement moyen des scores imposteurs engendrés par ce modèle de locuteur. Le biais qui apparaît dans les scores et qui dépend du modèle considéré peut alors être compensé en mesurant directement la distance KL correspondante. Pour

compenser ce type de biais, l'approche classiquement utilisée (normalisation Z-norm) consiste à étudier la distribution des scores imposteurs pour un modèle donné, à partir d'un ensemble d'accès réels. La D-norm a mis en évidence que ces données réelles ne sont pas forcément nécessaires pour normaliser les scores vis-à-vis des différents modèles de locuteurs.

D'autre part, le schéma d'adaptation D-MAP, qui réalise une normalisation des modèles vis-à-vis de leur distance KL, a révélé qu'il est possible d'obtenir un gain de performances similaire à celui que procure une normalisation de scores de type Z-norm ou D-norm, en agissant seulement au niveau de la procédure d'estimation des modèles. On peut interpréter la distance KL associée à un modèle de locuteur comme une mesure de discrimination de ce modèle par rapport au modèle du monde. La normalisation de modèles effectuée par D-MAP semble donc indiquer que le seuil optimal de décision est plus stable lorsque tous les modèles de locuteurs ont le même pouvoir discriminant vis-à-vis du modèle du monde.

Enfin, le nouveau cadre que nous avons développé, pour la vérification du locuteur dans un espace des modèles, a montré qu'il est possible de considérer les ensembles d'apprentissage et de test de façon tout à fait symétrique dans la procédure de vérification, en utilisant un modèle de chacun d'entre eux. Notamment, cette façon de procéder permet d'appliquer le même type de normalisation sur les modèles de locuteurs et sur les modèles de tests. Cette normalisation des modèles de test a conduit à une compensation de certains biais liés aux conditions des données de test, sans avoir recours à un ensemble de modèles imposteurs comme dans le cas de la normalisation de scores T-norm. D'autre part, la vérification du locuteur dans un espace des modèles a montré qu'il est possible d'utiliser un score de vérification basé sur des distances simples entre modèles, comme une alternative au classique score LLR.

En conclusion, ces travaux pourraient déboucher sur un allègement considérable des procédures de normalisation de scores utilisées dans les systèmes de VAL, en levant notamment la contrainte forte de devoir disposer de données réelles additionnelles.

Adaptation Bayésienne hiérarchique des modèles de locuteurs

Dans cette deuxième partie du travail nous avons élaboré et développé une méthode d'adaptation appelée H-MAP, et destinée à rendre plus robuste l'estimation des modèles de locuteur lorsque la quantité de données d'apprentissage est très limitée.

Contrairement à l'adaptation classique au sens du critère MAP, H-MAP permet théoriquement d'adapter des composantes d'un GMM qui ne reçoivent pas suffisamment de données dans l'ensemble d'apprentissage. Cette adaptation est théoriquement rendue possible grâce à l'apport d'informations *a priori* supplémentaires sous forme d'une structure de dépendances entre les moyennes des Gaussiennes. Les formules d'adaptation proposées généralisent l'adaptation MAP classique en levant l'hypothèse d'indépendance des moyennes d'un GMM. H-MAP possède de plus les mêmes propriétés de convergence que l'adaptation MAP classique ou l'estimation ML. D'autre part, de fortes similitudes existent entre H-MAP et d'autres techniques d'adaptation hiérarchique basées sur des transformations de modèles, telles que MLLR/MAPLR ou SMAP. La structure de dépendances que nous avons choisie est une structure

hiérarchique inspirée des nombreux travaux similaires consacrés à l'adaptation rapide au locuteur des systèmes de reconnaissance de la parole. Ainsi, H-MAP a pour but d'associer les capacités théoriques des méthodes Bayésiennes à la robustesse des techniques d'adaptation hiérarchique.

L'adaptation H-MAP a été développée sur la base de données NIST2003 et également évaluée sur une tâche de l'évaluation NIST2004 à quantité de données très restreinte. Malgré l'avantage théorique de H-MAP, cette méthode n'a pas amené pour l'instant d'amélioration décisive des performances par rapport à l'adaptation Bayésienne classique. Ceci peut être imputé à une mauvaise estimation des dépendances ou à un problème de structuration des liens de dépendance. Pour des raisons pratiques liées aux bases de données dont nous disposions, cette estimation des dépendances a été effectuée avec une quantité restreinte de locuteurs. Ce manque de représentativité des locuteurs est une première source de difficulté pour estimer de façon fiable les relations de dépendance. D'autre part, les bases de données utilisées n'étant pas étiquetées phonétiquement, la méthode de construction de l'arbre de dépendances n'intègre pas d'information linguistique. Elle se base uniquement sur les données acoustiques en établissant implicitement des liens de dépendance entre Gaussiennes proches. L'utilisation d'informations linguistiques pourrait aider à identifier des dépendances plus fortes entre des Gaussiennes appartenant à des mêmes classes ou macro-classes phonétiques. Malgré tout, les premiers résultats obtenus avec H-MAP sont encourageants et la technique offre un cadre théorique attrayant qui permet d'intégrer des informations de dépendances entre moyennes d'un GMM, tout en généralisant un certains nombre d'approches état de l'art.

Ces travaux sur H-MAP nous enseignent qu'il est difficile d'intégrer de nouvelles informations *a priori* fiables dans le schéma d'adaptation Bayésienne. Même si le cadre théorique développé montre un avantage potentiel de H-MAP par rapport à l'adaptation Bayésienne classique, nous n'avons pas réussi à mettre en évidence cet avantage pour l'instant, au travers des expériences que nous avons menées. Il est vrai que le nombre de paramètres *a priori* est plus important dans le cadre de H-MAP et que les paramètres supplémentaires, qui représentent des relations de dépendance entre les Gaussiennes des GMMs de locuteurs, sont difficiles à estimer. De plus, on ne sait pas qu'elle est la meilleure façon de capturer ces dépendances entre Gaussiennes. Cette difficulté à estimer des dépendances fortes entre Gaussiennes laisse supposer qu'il faudrait utiliser davantage de connaissances *a priori* pour cette estimation, afin de mieux cerner les relations de dépendance.

Perspectives

Les perspectives que nous envisageons concernant les techniques de normalisation, sont principalement dédiées au nouveau cadre de vérification du locuteur dans un espace des modèles. En effet l'espace Euclidien défini dans cette approche, pour la représentation des modèles, permet d'envisager des techniques de normalisation de

modèles simples et efficaces, basées sur des transformations géométriques. Notamment des normalisations liées au canal de transmission ou au microphone devraient être étudiées dans cet espace. Nous envisageons plusieurs approches pour cela.

Une première approche peut consister à identifier dans cet espace des directions qui sont particulièrement affectées par les changements de canaux ou de microphones. Cette recherche de directions perturbées pourrait être menée grâce à des techniques d'analyse de données telles que l'ACP (Analyse en Composantes Principale). Une solution pour minimiser alors l'influence du canal et/ou du microphone serait de supprimer du calcul du score les directions très dépendantes de ces éléments, en ne conservant ainsi que des informations principalement liées aux locuteurs.

Une autre approche envisageable pourrait s'inspirer d'une technique de normalisation de canal mise au point au niveau des paramètres acoustiques : le "feature mapping". Cette technique consiste à apprendre des transformations permettant de projeter des paramètres acoustiques dépendant d'un canal dans un espace indépendant du canal. Dans l'espace des modèles, l'application d'un tel procédé serait presque immédiate. Il suffirait d'estimer dans un premier temps des modèles du monde de chaque canal et un modèle du monde indépendant du canal, ce dernier étant utilisé comme origine de l'espace des modèles. Les translations permettant de passer des modèles de canal jusqu'au point origine de l'espace pourraient alors servir à normaliser les modèles de locuteurs et de tests vis-à-vis du canal. Cette technique, qui pourrait être dénommée "model mapping", semble simple à mettre en oeuvre et pourrait s'avérer très efficace pour compenser les disparités de canal ou de microphone, au même titre que la technique "feature mapping".

Enfin, la simplicité du calcul de score dans l'espace des modèles laisse envisager la possibilité d'implanter cette technique sur des architectures matérielles à capacité calculatoire limitée telles que des cartes à puce. Il faut cependant pouvoir estimer de façon rapide un modèle de l'énoncé de test mais ce travail peut être délégué à une architecture plus puissante dans le cadre d'une association de la carte à puce à un terminal hôte. En effet, pour des raison de sécurité, seul le score de vérification doit impérativement être calculé au niveau de la carte à puce afin que les informations relatives au locuteur ne transitent en aucun cas en dehors de cette carte. Il serait intéressant d'étudier si les ressources disponibles sur des cartes à puce courantes suffiraient à implémenter la méthode de calcul de score dans l'espace des modèles. Si c'est le cas, cette technologie pourrait permettre de disposer d'applications de VAL fonctionnant sur des systèmes embarqués, avec des performances au niveau de l'état de l'art.

Concernant notre travail sur la technique d'adaptation hiérarchique H-MAP, les perspectives que nous envisageons visent principalement à améliorer l'estimation des dépendances entre Gaussiennes d'un GMM et à mieux définir la structure de dépendances. Les expérimentations effectuées jusqu'ici ont souffert de la faible couverture de la population de locuteurs dans les données dont nous disposions, ne permettant pas d'estimer de façon fiable les relations de dépendance. D'autre part, la technique utilisée pour construire la structure de dépendances, basée uniquement sur les données acoustiques, pourrait ne pas être appropriée pour capturer des relations fortes entre Gaussiennes.

Nous pensons que l'utilisation d'informations linguistiques devrrait permettre de mieux structurer ces dépendances. D'autre part, l'estimation des relations de régression entre moyennes des Gaussiennes devra être envisagée sur une base de données plus importante afin que les dépendances définies soient fiables.

Le développement théorique de la technique H-MAP permet d'englober de nombreux schémas d'adaptation existants. On peut donc penser que H-MAP pourrait tirer partie des avantages de ces différentes approches, à conditions que les informations *a priori* sur lesquelles s'appuie la technique puissent être estimées de façon fiable.

Annexe A

Le consortium ELISA

Le consortium ELISA [1] fut créé en 1998 par plusieurs laboratoires français du domaine du traitement automatique de la parole (ENST, EPFL, IDIAP, IRISA et LIA) dans le but de développer une plateforme commune de reconnaissance du locuteur, et de la maintenir au niveau de l'état de l'art. Un des principaux objectifs du consortium est de participer chaque année aux évaluations des systèmes de reconnaissance du locuteur organisées par l'institut américain NIST (National Institute of Standards and Technology). La participation à ces évaluations demande une quantité de travail importante et le consortium ELISA permet de mutualiser les efforts des différents membres.

La composition du consortium a évolué au cours des années. On compte aujourd'hui parmi les membres actifs le DDL, l'ENST, le CLIPS, l'IRISA, le LIA et l'Université de Fribourg. Depuis 1998, un ou plusieurs systèmes ELISA sont présentés chaque année aux évaluations NIST. On trouvera le descriptif des différents systèmes ELISA et IRISA utilisés lors de l'évaluation de 1999 dans [Elisa00] et [Seck et al.00] . Une présentation des activités du consortium en 2000-2001 est donnée dans [MC et al.01], avec le descriptif du système correspondant et un récapitulatif de l'évolution des performances des meilleurs systèmes ELISA de 1998 à 2001. L'annexe B présente en détail les systèmes ELISA et IRISA de 2001 à 2004.

Notons enfin que le projet TECHNOLANGUE AGILE-ALIZE [2], financé par le Ministère de l'Education Nationale, de la Recherche et de la Technologie, est une émanation du consortium ELISA. Ce projet a permis de développer, de valider et de distribuer librement une plateforme logicielle modulable pour la reconnaissance du locuteur [3]. On pourra télécharger cette plateforme à partir du site internet suivant :

`http://www.lia.univ-avignon.fr/heberges/ALIZE/`

Quelques outils spécifiquement développés pour la reconnaissance du locuteur par le LIA peuvent également être téléchargés à l'adresse suivante :

`http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL/index.html`.

1. `http://www.lia.univ-avignon.fr/heberges/ALIZE/ELISA/index.html/`
2. `http://www.technolangue.net/article76.html`
3. l'intégralité du développement de cette plateforme a été effectué au LIA

Annexe B

Description des systèmes ELISA/IRISA de 2001 à 2004

Les systèmes de reconnaissance du locuteur utilisés au début de ma thèse (jusqu'à la fin 2002) sont basés sur la plateforme commune ELISA, avec des variantes spécifiques à l'IRISA expliquées ci-après. Ils utilisent le logiciel Spro [1], développé par Guillaume GRAVIER pour la paramétrisation du signal, et le logiciel AMIRAL [2] [Fredouille et al.00a], développé au LIA, pour la modélisation et le calcul de score.

A partir de 2003, le système IRISA n'est plus basé directement sur la plateforme ELISA. Un nouveau logiciel, "Audioseg" [3], développé à l'IRISA par Guillaume GRAVIER, Michaël BETSER et Grégoire COLBERT, est utilisé pour la modélisation et le calcul de scores.

B.1 Systèmes ELISA et IRISA 2001

B.1.1 Système de base ELISA 2001

En 2001, la plateforme ELISA est basée sur les logiciels Spro 3.0, pour la paramétrisation, et AMIRAL, pour la modélisation et le calcul de scores.

B.1.1.1 Paramétrisation

Découpage temporel du signal: trames de 20 ms, décalées de 10 ms, fenêtrage de Hamming.

1. `http://www.irisa.fr/metiss/guig/spro/`
2. `http://www.lia.univ-avignon.fr/equipes/TALNO/index.html`
3. le logiciel Audioseg sera prochainement distribué par l'IRISA sous licence GPL

Analyse acoustique :

- 16 MFCC
- + 16 Δ

Retrait de trames :

- modélisation bi-gaussienne de l'énergie des trames
- seuil de rejet fixé par rapport aux paramètres de la gaussienne de haute énergie

Normalisation des vecteurs acoustiques :

- CMS

B.1.1.2 Modélisation

Modèles du monde :

- 2 modèles du monde dépendants du genre
- GMM à 128 composantes diagonales
- estimation EM/ML, 20 itérations (initialisation par tirage aléatoire de trames)
- variance flooring = 0,5 x variance globale

Modèles de locuteur :

- dérivés des modèles du mondes par adaptation MAP sur les moyennes uniquement
- utilisation du MAP version LIA (option MAP 3 du logiciel AMIRAL)
- facteur d'adaptation $\alpha = 0,25$
- estimation EM, 10 itérations

B.1.1.3 Calcul de score et décision

Score brut :

- moyenne des log-rapports de vraisemblances par bloc de 0,3 secondes

Normalisations de score :

- H-norm, T-norm

B.1.2 Variante IRISA 2001

B.1.2.1 Paramétrisation

- identique au système ELISA 2001

B.1.2.2 Modélisation

- identique au système ELISA 2001

B.1.2.3 Calcul de score et décision

Score brut :
- identique au système ELISA 2001

Normalisations de score :
- implémentation de la D-norm

B.2 Systèmes ELISA et IRISA 2002

B.2.1 Système de base ELISA 2002

En 2002, le système de base ELISA a été développé et présenté aux évaluations NIST par le LIA (système primaire LIA 2002).

B.2.1.1 Paramétrisation

Analyse acoustique :
- identique au système ELISA 2001

Retrait de trames :
- identique au système ELISA 2001

Normalisation des vecteurs acoustiques :
- centrage+réduction après retrait de trames

B.2.1.2 Modélisation

Modèles du monde :
- 1 modèle du monde universel obtenu par concaténation de 2 modèles dépendants du genre
- GMM à 256 composantes diagonales
- estimation EM/ML, 20 itérations (initialisation par tirage aléatoire de trames)
- variance flooring = 0,5 x variance globale

Modèles de locuteur :
- identique au système ELISA 2001

B.2.1.3 Calcul de score et décision

- identique au système ELISA 2001

B.2.2 Variante IRISA 2002

B.2.2.1 Paramétrisation

- identique au système ELISA 2002

B.2.2.2 Modélisation

- identique au système ELISA 2001 (2 modèles du monde dépendants du genre : GMM à 128 composantes diagonales)

B.2.2.3 Calcul de score et décision

Score brut :

- identique au système ELISA 2002

Normalisations de score :

- implémentation de la DT-norm

B.3 Système IRISA 2003

En 2003, le système IRISA est basé sur les logiciels Spro 4.0, pour la paramétrisation, et Audioseg, pour la modélisation et le calcul de scores.

B.3.0.4 Paramétrisation

Découpage temporel du signal : trames de 20 ms, décalées de 10 ms, fenêtrage de Hamming.

Analyse acoustique :

- 16 LFCC
- + 16 Δ
- + Δ-log énergie

Retrait de trames :

- modélisation bi-gaussienne de l'énergie des trames
- classification parole/silence des trames au maximum de vraisemblance

Normalisation des vecteurs acoustiques :

- centrage+réduction après retrait de trames

B.3.0.5 Modélisation

Modèles du monde :

- 2 modèles du monde dépendants du genre
- GMM à 256 composantes diagonales
- estimation EM, 20 itérations (initialisation par quantification vectorielle hiérarchique)
- variance flooring = 0,1 x variance globale

Modèles de locuteur :

- dérivés des modèles du monde par adaptation MAP sur les moyennes uniquement

- utilisation du MAP version Reynolds
- facteur de confiance pour l'adaptation $\tau = 30$
- estimation EM, 10 itérations

B.3.0.6 Calcul de score et décision

Score brut :

- moyenne des log-rapports de vraisemblances par trame

Normalisations de score :

- DT-norm

B.4 Système IRISA 2004

En 2004, le système IRISA est basé sur les logiciels Spro 4.0, pour la paramétrisation, et Audioseg, pour la modélisation et le calcul de scores.

B.4.0.7 Paramétrisation

B.4.0.8 Variante 1

- identique au système IRISA 2003

B.4.0.9 Variante 2

Analyse acoustique :

- identique au système IRISA 2003

Retrait de trames :

- identique au système IRISA 2003

Normalisation des vecteurs acoustiques :

- feature warping

B.4.0.10 Modélisation

B.4.0.11 Variante 1

(utilisée dans le système primaire IRISA pour l'évaluation NIST 2004, tâche 1side/1side)

- identique au système IRISA 2003

B.4.0.12 Variante 2

(utilisée dans le système IRISA pour la tâche 10sec/10sec de l'évaluation NIST 2004)
Modèles du monde :

- 2 modèles du monde structurels (SBM) dépendants du genre (arbre de dépendances + GMM multi-résolution)
- GMM à 256 composantes diagonales dans la résolution la plus fine (dernier niveau de l'arbre)
- estimation EM hiérarchique par processus de "gaussian splitting", 20 itérations

Modèles de locuteur :

- dérivés des modèles du mondes structurel par adaptation H-MAP sur les moyennes uniquement
- facteur de confiance dans le dernier niveau : $\tau = 15$
- estimation EM, 10 itérations

B.4.0.13 Calcul de score et décision

Le système primaire IRISA pour l'évaluation NIST 2004 utilise une fusion des scores T-normalisés (moyenne arithmétique) des systèmes utilisant les variantes 1 et 2 de paramétrisation.

Annexe C

Algorithme pour le tirage de Monte Carlo à partir d'un GMM

Soit $p(y|\Lambda)$ un modèle de mélange de Gaussiennes (GMM) à K composantes, et $\Lambda = (\{w_k\}, \{m_k\}, \{S_k\})$ les paramètres poids, vecteurs moyennes et matrices de covariance de ce modèle :

$$p(y|\Lambda) = \sum_{k=1}^{K} w_k . \mathcal{N}(y; m_k, S_k) \quad . \tag{C.1}$$

$\mathcal{N}(\cdot; m_k, S_k)$ est une fonction Gaussienne de moyenne m_k et de matrice de covariance S_k supposée diagonale. Pour générer des données artificielles $\tilde{y}_n$ suivant ce GMM, nous procédons en deux étapes :

1. tirage aléatoire d'une des composantes parmi les K Gaussiennes du mélange, en fonction de leur poids respectif.
2. génération d'une donnée Gaussienne suivant la moyenne et la matrice de covariance de la composante tirée au sort à l'étape 1.

Ces deux étapes sont répétées autant de fois que l'on désire d'échantillons synthétiques. D'autre part, les matrices de covariance étant supposées diagonales, chaque coefficient $\tilde{y}_n^d$ du vecteur $\tilde{y}_n$ peut être généré indépendamment des autres à l'étape 2.

L'étape 1 revient à générer une variable aléatoire discrète $k \in \{1, .., K\}$ représentant l'index des composantes du mélange et qui suit une loi multinomiale de paramètres $\{w_k\}$. Pour cela, on forme des intervalles $\{\mathcal{I}_k\}_{k=1,...,k}$ contigus entre 0 et 1 et dont les longueurs respectives correspondent aux poids $\{w_k\}$ (cf. figure C.1).

On génère ensuite une réalisation $\tilde{U}$ d'une variable aléatoire $\tilde{u}$ de loi uniforme entre 0 et 1. L'indice k de l'intervalle dans lequel "tombe" la valeur générée $\tilde{U}$ donne l'indice de la Gaussienne qui est tirée au sort, en respectant leur probabilité *a priori*. On a en effet :

$$P(\tilde{u} = \tilde{U} \in \mathcal{I}_k) = w_k \quad .$$

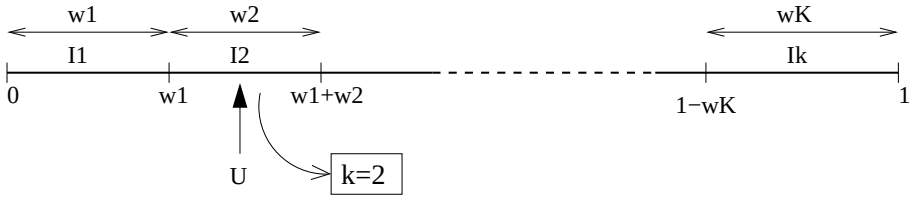


FIG. C.1 – *Principes du tirage aléatoire des Gaussiennes.*

La recherche de cet intervalle peut se faire par comparaisons successives de $\tilde{U}$ avec les bornes supérieures des intervalles $\{\mathcal{I}_k\}_{k=1,\ldots,k}$, ou par une recherche dichotomique (en générale plus rapide).

Une fois la Gaussienne tirée au sort, nous utilisons l'algorithme de Box-Müller [Box et al.58] pour générer une réalisation $\tilde{G}_{(0,1)}$ d'une variable aléatoire (scalaire) $\tilde{g}_{(0,1)}$ Gaussienne, centrée et réduite. Ensuite, cette valeur $\tilde{G}_{(0,1)}$ est transformée en utilisant les paramètres moyenne et covariance de la Gaussienne tirée au sort, afin de former une réalisation $\tilde{Y}_n^d$ du coefficient d d'un vecteur acoustique artificiel $\tilde{y}_n$:

$$\tilde{Y}_n^d = \sigma_k^d.\tilde{G}_{(0,1)} + m_k^d \ , \tag{C.2}$$

où m_k^d est le coefficient d de la moyenne de la Gaussienne k tirée au sort, et σ_k^d est le coefficient $d \times d$ de la matrice de covariance de cette Gaussienne. L'opération est répétée pour chacune des dimensions du vecteur acoustique artificiel.

On montre alors facilement que la variable aléatoire vectorielle $\tilde{y}$ issue de ce tirage suit la densité de probabilité définie par le GMM considéré :

$$p(\tilde{y}) = \sum_{k=1}^{K} p(\tilde{u} \in \mathcal{I}_k).p(\tilde{y}|\tilde{u} \in \mathcal{I}_k) \ , \tag{C.3}$$

$$= \sum_{k=1}^{K} w_k.\mathcal{N}(\tilde{y}; m_k, S_k) \ . \tag{C.4}$$

Bibliographie

[Adami et al.03] Adami (A.), Mihaescu (R.), Reynolds (D.) et Godfrey (J.). – Modeling prosodic dynamics for speaker recognition. *Proceedings of : ICASSP*. – 2003.

[Afify et al.00] Afify (M.) et Siohan (O.). – Constrained maximum likelihood linear regression for speaker adaptation. *Proceedings of : ICSLP*. – 2000.

[Andrews et al.02] Andrews (W.), Kohler (M.), Campbell (J.), Godfrey (J.) et Hernandez-Cordero (J.). – Gender-dependent phonetic refraction for speaker recognition. *Proceedings of : ICASSP*. – 2002.

[AS96] Ahadi-Sarkani (S. M.). – *Bayesian and Predictive Techniques for Speaker Adaptation*. – Thèse de PhD, University of Cambridge, Janvier 1996.

[Atal74] Atal (B.S.). – Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification. *Journal of Statistical Society of America (JASA)*, vol. 55, 1974, pp. 1304–1312.

[Auckenthaler et al.00] Auckenthaler (R.), Carey (M.) et Lloyd-Thomas (H.). – Score normalization for text-independent speaker verification systems. *Digital Signal Processing Vol 10,num 1-3*, 2000.

[Barras et al.03] Barras (C.) et Gauvain (J-L.). – Feature and score normalization for speaker verification of cellular data. *Proceedings of : ICASSP*. – 2003.

[Barras et al.04] Barras (C.), Meignier (S.) et Gauvain (J.L.). – Unsupervised online adaptation for speaker verification over the telephone. *Proceedings of : Odyssey, The Speaker and Language Recognition Workshop*. – 2004.

[BB et al.03] Bailly-Baillière (E.), Bengio (S.), Bimbot (F.), Hamouz (M.), Kittler (J.), Mariéthoz (J.), Matas (J.), Messer (K.), Popovici (V.), Porée (F.), Ruiz (B.) et Thiran (J.-P.). – The BANCA database and evaluation protocol. *Proceedings of : 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, éd. par Springer-Verlag. – 2003.

[Beaufrays et al.97] Beaufrays (F.) et Weintraub (M.). – Model transformation for robust speaker recognition from telephone data. *Proceedings of: ICASSP.* – 1997.

[Ben et al.02a] Ben (M.), Blouet (R.) et Bimbot (F.). – ELISA/IRISA 2002 system description. *Proceedings of: NIST 2002 Speaker Recognition Evaluation Workshop.* – 2002.

[Ben et al.02b] Ben (M.), Blouet (R.) et Bimbot (F.). – A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances. *Proceedings of: ICASSP.* – 2002.

[Ben et al.03a] Ben (M.) et Bimbot (F.). – D-MAP : a distance-normalized MAP estimation of speaker models for automatic speaker verification. *Proceedings of: ICASSP.* – 2003.

[Ben et al.03b] Ben (M.), Gravier (G.), Ozerov (A.) et Bimbot (F.). – IRISA 2003 speaker recognition system. *Proceedings of: NIST 2003 Speaker Recognition Evaluation Worshop.* – 2003.

[Ben et al.04a] Ben (M.), Betser (M.), Bimbot (F.) et Gravier (G.). – Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. *Proceedings of: ICSLP.* – 2004.

[Ben et al.04b] Ben (M.) et Bimbot (F.). – The IRISA 2004 speaker recognition system. *Proceedings of: NIST 2004 Speaker Recognition Evaluation Worshop.* – 2004.

[Ben et al.04c] Ben (M.), Bimbot (F.) et Gravier (G.). – Enhancing the robustness of Bayesian methods for text-independent automatic speaker verification. *Proceedings of: Odyssey, The Speaker and Language Recognition Workshop.* – 2004.

[Bengio et al.01] Bengio (S.) et Mariéthoz (J.). – Learning the decision function for speaker verification. *Proceedings of: ICASSP.* – 2001.

[Bengio et al.04] Bengio (S.) et Mariéthoz (J.). – A statistical significance test for person authentication. *Proceedings of: Odyssey 2004, The Speaker and Language Recognition Workshop.* – 2004.

[Bimbot et al.95] Bimbot (F.), Magrin-Chagnolleau (I.) et Mathan (L.). – Second-order statistical measures for text-independent speaker identification. *Speech Communication*, vol. 17, n° 1-2, Août 1995, pp. 177–192.

[Bimbot et al.00] Bimbot (F.), Blomberg (M.), Boves (L.), Genoud (D.), Hutter (H.-P.), Jaboulet (C.), Koolwaaij (J.), Lindberg (J.) et Pierrot (J.-B.). – An overview of the CAVE project research activities in speaker verification. *Speech Communication*, vol. 31, n° 2-3, 2000, pp. 155–180.

[Blouet et al.04] Blouet (R.), Mokbel (C.), Mokbel (H.), Sánchez-Soto (E.), Chollet (G.) et Greige (H.). – BECARS: A free software for

speaker verification. *Proceedings of: Odyssey, The Speaker and Language Recognition Workshop.* – 2004.

[Blouet02] Blouet (R.). – *Approche Probabiliste par Arbre de Décision pour la Vérification Automatique du Locuteur sur Architecture Embarquée.* – Thèse de PhD, Université de Rennes, Décembre 2002.

[Boe et al.99] Boe (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. *Revue Langues*, vol. 2, n° 4, Décembre 1999, pp. 270–288.

[Bogert et al.63] Bogert (B. P.), Healy (M. J.) et Tukey (J.W.). – The quefrency analysis of time series for echoes: cepstrum, pseudoautocovariance, cross-cepstrum, and shape cracking. *Times Series Analysis*, éd. par Rosenblatt (M.), pp. 209–243. – J. Wiley, New York, 1963.

[Bonastre et al.04a] Bonastre (J-F.), Bimbot (F.), Boë (L.J.), Campbell (J.P.), Reynolds (D.A.) et Magrin-Chagnolleau (I.). – Authentification des personnes par leur voix : un nécessaire devoir de précaution. *Actes des Journées d'Etude sur la Parole (JEP).* – 2004.

[Bonastre et al.04b] Bonastre (J-F.), Scheffer (N.), Fredouille (C.) et Matrouf (D.). – NIST'04 speaker recognition evaluation campaign : New LIA speaker recognition platform based on ALIZE toolkit. *Proceedings of: NIST 2004 Speaker Recognition Evaluation Workshop.* – 2004.

[Box et al.58] Box (G. E. P.) et Muller (M. E.). – A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, vol. 29, 1958, pp. 610–611.

[Campbell et al.02] Campbell (W.M.), Assaleh (K.T.) et Broun (C.C.). – Speaker recognition with polynomial classifiers. *IEEE Transactions on Speech and Audio Processing*, vol. 10, n° 4, Mai 2002, pp. 205–212.

[Campbell et al.03] Campbell (J. P.), Reynolds (D. A.) et Dunn (R. B.). – Fusing high- and low-level features for speaker recognition. *Proceedings of: EUROSPEECH.* – 2003.

[Campbell et al.04] Campbell (W. M.), Reynolds (D. A.) et Campbell (J. P.). – Fusing discriminative and generative methods for speaker recognition: Experiments on switchboard and NFI/TNO field data. *Proceedings of: Odyssey 2004, The Speaker and Language Recognition Workshop.* – 2004.

[Campbell03] Campbell (W.M.). – A SVM/HMM system for speaker recognition. *Proceedings of: ICASSP.* – 2003.

[Carey et al.91] Carey (M.), Parris (E.), J. et Bridle. – A speaker verification system using alphanets. *Proceedings of : ICASSP.* – 1991.

[Cerisara et al.01] Cerisara (C.) et Daoudi (K.). – Modeling dependency between regression classes in MLLR using multiscale autoregressive models. *Proceedings of : ISCA Workshop on Adaptation methods for speech recognition.* – 2001.

[Charlet04] Charlet (D.). – Neighborhood-adapted GMM for speaker recognition. *Proceedings of : Odyssey, The Speaker and Language Recognition Workshop.* – 2004.

[Chen et al.98] Chen (S.) et Gopalakrishnan (P. S.). – Speaker, environment and channel change detection and clustering via the Bayesian information criterion. *Proceedings of : Broadcast News Transcription and Understanding Workshop.* – 1998.

[Davis et al.80] Davis (S. B.) et Merlmelstein (P.). – Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Speech and Audio Processing*, vol. 28, 1980, pp. 357–366.

[Delacourt00] Delacourt (P.). – *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio.* – Thèse de PhD, Institut Eurécom, Sophia Antipolis, Septembre 2000.

[Dempster et al.77] Dempster (A.P.), Laird (N.M.) et Rubin (D.B.). – Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Statistical Society of America (JASA)*, vol. 39, 1977, pp. 1–38.

[Do03] Do (M. N.). – Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, vol. 10, Avril 2003, pp. 115–118.

[Doddington01] Doddington (G.). – Speaker recognition based on idiolectal differences between speakers. *Proceedings of : EUROSPEECH.* – 2001.

[Elisa00] ELISA. – The ELISA systems for the NIST 99 evaluation in speaker detection and tracking. *Digital Signal Processing*, vol. 10, n° 1–3, 2000.

[Faltlhauser et al.01] Faltlhauser (R.) et Ruske (G.). – Improving speaker recognition performance using phonetically structured Gaussian mixture models. *Proceedings of : EUROSPEECH.* – 2001.

[Fredouille et al.00a] Fredouille (C.), Bonastre (J.F.) et Merlin (T.). – AMIRAL : A block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing*, vol. 10, n° 1–3, 2000.

[Fredouille et al.00b] Fredouille (C.), Mariethoz (J.), Jaboulet (C.), Hennebert (J.), Bonnastre (J.-F.), Mokbel (C.) et Bimbot (F.). – Behaviour

of a Bayesian adaptation method for incremental enrollment in speaker verification. *Proceedings of: ICASSP.* – 2000.

[Fredouille00] Fredouille (C.). – *Approche Statistique pour la Reconnaissance Automatique du Locuteur: Informations Dynamiques et Normalisation Bayesienne des Vraisemblances.* – Thèse de PhD, Université d'Avignon, Octobre 2000.

[Furui81] Furui (S.). – Cepstral analysis technique for automatic speaker verification. *IEEE Transaction on Acoustic, Speech and Signal Processing (ASSP)*, vol. 19, n° 2, Avril 1981, pp. 254–272.

[Ganchev et al.03] Ganchev (T.), Tasoulis (D.K.), Vrahatis (M.N.) et Fakotakis (N.). – Locally recurrent probabilistic neural network for text-independent speaker verification. *Proceedings of: EUROSPEECH.* – 2003.

[Gauvain et al.94] Gauvain (J. L.) et Lee (C. H.). – Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Proc.*, vol. 2, n° 2, April 1994.

[Geiger et al.98] Geiger (D.) et Heckerman (D.). – *Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability.* – Rapport technique, Microsoft Research, Advanced Technology Division, Octobre 1998.

[Gravier et al.04] Gravier (G.), Bonastre (J.-F.), Geoffrois (E.), Galliano (S.), McTait (K.) et Choukri (K.). – ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. *Actes des Journées d'Etude sur la Parole (JEP).* – 2004.

[Heck et al.97] Heck (L. P.) et Weintraub (M.). – Handset-dependent background models for robust text-independent speaker recognition. *Proceedings of: ICASSP.* – 1997.

[Hermansky et al.94] Hermansky (H.) et Morgan (N.). – Rasta processing of speech. *IEEE Trans. on Speech and Audio Proc.*, vol. 2, n° 4, 1994.

[Kannan97] Kannan (A.). – *Adaptation of spectral trajectory models for large vocabulary speech recognition.* – Thèse de PhD, Boston University, 1997.

[Klusacek et al.03] Klusacek (D.), Navratil (J.), Reynolds (D.) et Campbell (J.). – Conditional pronunciation modeling in speaker detection. *Proceedings of: ICASSP.* – 2003.

[Ldc] LDC. – The linguistic data consortium. http://www.ldc.upenn.edu/.

[Leggetter et al.95a] Leggetter (C. J.) et Woodland (P. C.). – Flexible speaker adaptation using maximum likelihood linear regression. *Proceedings of: Workshop on Spoken Language Systems Technology.* – 1995.

[Leggetter et al.95b] Leggetter (C.J.) et Woodland (P.C.). – Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 1995.

[Li et al.88] Li (K-P.) et Porter (J. E.). – Normalizations and selection of speech segments for speaker recognition scoring. *Proceedings of: ICASSP.* – 1988.

[Mak et al.02] Mak (M.-W.) et Kung (S.-K.). – Combining stochastic feature transformation and handset identification for telephone-based speaker verification. *Proceedings of: ICASSP.* – 2002.

[Mami03] Mami (Y.). – *Reconnaissance du locuteur par héritage et voix propres.* – Thèse de PhD, France Télécom R&D, Lannion, Octobre 2003.

[Mariethoz et al.01] Mariethoz (J.) et Bengio (S.). – *A Comparison of Adaptation Methods for Speaker Verification.* – Rapport technique n° 01-34, IDIAP, Suisse, 2001.

[Martin et al.97] Martin (A.), Doddington (G.), Kamm (T.), Ordowski (M.) et Przybocki (M.). – The DET curve in assessment of detection task performance. *Proceedings of: Eurospeech.* – 1997.

[MC et al.96] Magrin-Chagnolleau (I.), Wilke (J.) et Bimbot (Frédéric). – A further investigation on ar-vector models for text-independent speaker identification. *Proceedings of: ICASSP.* – 1996.

[MC et al.00] Magrin-Chagnolleau (I.) et Durou (G.). – Application of time-frequency principal component analysis to speaker verification. *Digital Signal Processing*, vol. 10, n° 1-3, January/April/July 2000, pp. 226–236.

[MC et al.01] Magrin-Chagnolleau (I.), Gravier (G.) et Blouet (R.). – Overview of the 2000-2001 ELISA consortium research activities. *Proceedings of 2001 : A Speaker Odyssey-The Speaker Recognition Workshop.* – 2001.

[MC97] Magrin-Chagnolleau (I.). – *Approches Statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte.* – Thèse de PhD, ENST, Janvier 1997.

[Meignier02] Meignier (S.). – *Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection.* – Thèse de PhD, Université d'Avignon, Novembre 2002.

[Mokbel01] Mokbel (C.). – Online adaptation of hmms to real-life conditions: A unified framework. *IEEE Trans. on Speech and Audio Proc.*, vol. 9, n° 4, Mai 2001.

[Navratil et al.03] Navratil (J.) et Ramaswamy (G. N.). – The awe and mystery of t-norm. *Proceedings of: EUROSPEECH.* – 2003.

[Nist] NIST. – The NIST speaker recognition evaluation. http://www.nist.gov/speech/tests/spk/.

[Padmanabhan et al.96] Padmanabhan (M.), Bahl (L. R.), Nahamoo (D.) et Picheny (M. A.). – Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. *Proceedings of : ICASSP*. – 1996.

[Pelecanos et al.01] Pelecanos (J.) et Sridharan (S.). – Feature warping for robust speaker verification. *Proceedings of : Odyssey, The Speaker and Language Recognition Workshop*. – 2001.

[Perronnin et al.02] Perronnin (F.) et Dugelay (J.L.). – Intoduction à la biométrie : Authentification des individus par traitement audio-vidéo. *Revue Traitement du Signal*, vol. 19, n° 4, 2002.

[Pinto et al.04] Pinto (E.), Charlet (D.), François (H.), Mostefa (D.), Boëffard (O.), Fohr (D.), Mella (O.), Bimbot (F.), Choukri (K.), Philip (Y.) et Charpentier (F.). – Development of new telephone speech databases for french: the neologos project. in international conference on language resources and evaluation. *Proceedings of : International Conference on Language Resources and Evaluation - LREC'04*. – 2004.

[Pitz et al.01] Pitz (M.), Schlüter (S. Molau R.) et Ney (H.). – Vocal tract normalization equals linear transformation in cepstral space. *Proceedings of : EUROSPEECH*. – 2001.

[Porter97] Porter (J.E.). – *On The 30 Errors Criterion*. – Rapport technique, ITT Industries Defense, Avril 1997.

[Przybocki et al.04] Przybocki (M.) et Martin (A.). – The NIST year 2004 speaker recognition evaluation plan, 2004. http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf.

[Pusateri et al.02] Pusateri (E. J.) et Hazen (T. J.). – Rapid speaker adaptation using speaker clustering. *Proceedings of : ICSLP*. – 2002.

[Reynolds et al.00] Reynolds (A.), Quatieri (T.F.) et Dunn (R.B.). – Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, vol. 10, n° 1–3, 2000.

[Reynolds et al.03] 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. *Proceedings of : ICASSP*. – 2003.

[Reynolds92] Reynolds (D. A.). – *A Gaussian Mixture Modeling Aproach for Text-Independent Speaker Identification*. – Thèse de PhD, Georgia Institue of Technology, Août 1992.

[Reynolds94] Reynolds (D.A.). – Experimental evaluation of features for robust speaker identification. *IEEE Trans. on Speech and Audio Proc.*, vol. 2, 1994.

[Reynolds03] Reynolds (Douglas A.). – Channel robust speaker verification via feature mapping. *Proceedings of: ICASSP.* – 2003.

[Roberts et al.99] Roberts (C.P.) et Casella (G.). – *Monte Carlo Statistical Methods*, chap. 2, p. 46. – Springer-Verlag New York, Inc, 1999.

[Rosenberg et al.92] Rosenberg (A. E.), DeLong (J.), Lee (C-H.), Juang (B-H.) et Soong (K.). – The use of cohort normalized scores for speaker verification. *Proceedings of: ICSLP.* – 1992.

[Rosenberg et al.96] Rosenberg (A. E.) et Parthasarathy (S.). – Speaker background models for connected digit password speaker verification. *Proceedings of: ICASSP.* – 1996.

[Seck et al.00] Seck (M.), Blouet (R.) et Bimbot (F.). – The IRISA/ELISA speaker detection and tracking systems for the NIST 99 evaluation. *Digital Signal Processing*, vol. 10, n° 1–3, 2000.

[Seck01] Seck (M.). – *Détection de ruptures et suivi de classe de sons pour l'indexation sonore.* – Thèse de PhD, Université de Rennes, Janvier 2001.

[Shinoda et al.97] Shinoda (K.) et Less (C.-H.). – Structural MAP speaker adaptation using hierarchical priors. *Proceedings of: IEEE Workshop on Automatic Speech Recognition and Understanding.* – 1997.

[Siohan et al.99] Siohan (O.), Chesta (C.) et Lee (C.-H.). – Hidden Markov model adaptation using maximum a posteriori linear regression. *Proceedings of: Workshop on Robust Methods for Speech Recognition in Adverse Conditions.* – 1999.

[Siohan et al.00] Siohan (O.) et Less (T.-A. Myrvolland C.-H.). – Structural maximum a posteriori linear regression for fast HMM adaptation. *Proceedings of: ISCA Workshop on Automatic Speech Recognition: Challenges for the new Millenium.* – 2000.

[Solomonoff et al.04] Solomonoff (A.), Quillen (C.) et Campbell (W. M.). – Channel compensation for SVM speaker recognition. *Proceedings of: Odyssey 2004, The Speaker and Language Recognition Workshop.* – 2004.

[Teunen et al.00] Teunen (R.), Shahshahani (B.) et Heck (L.). – A model-based transformational approach to robust speaker recognition. *Proceedings of ICSLP2000.* – 2000.

[Thyes et al.00] Thyes (O.), Kuhn (R.), Nguyen (P.) et Junqua (J.-C.). – Speaker identification and verification using Eigenvoices. *Proceedings of: ICSLP.* – 2000.

[Viikki et al.98] Viikki (O.) et Laurila (K.). – Cepstral domain segmental feature vector normalization for nois robust speech recognition. *Speech Communication*, vol. 25, 1998, pp. 133–147.

[vL et al.04] van Leeuwen (D.A.) et Bouten (J.S.). – Results of the 2003 NFI-TNO forensic speaker recognition evaluation. *Proceedings of: Odyssey, The Speaker and Language Recognition Workshop.* – 2004.

[Wan03] WAN (V.). – *Speaker Verification using Support Vector Machines.* – Thèse de PhD, University of Sheffield, United Kingdom, Juin 2003.

[Xiang et al.02] Xiang (B.), Chaudhari (U.), Navr´atil (J.), Ramaswamy (G.) et Gopinath (R.). – Short-time gaussianization for robust speaker verification. *Proceedings of: ICASSP.* – 2002.

[Xiang et al.03] Xiang (B.) et Berger (Toby). – Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. *IEEE Trans. Speech Audio Process*, vol. 11, n° 5, 2003.

Résumé

Cette thèse est consacrée à l'élaboration et l'évaluation de techniques visant à renforcer la robustesse des systèmes de vérification automatique du locuteur. La vérification automatique du locuteur (VAL) consiste à authentifier l'identité d'une personne en analysant les caractéristiques de sa voix. Ses applications vont du contrôle d'accès à l'authentification d'enregistrements sonores, en passant par des tâches d'étiquetage automatique de documents audio en fonction des intervenants. Utilisés en situation réelle et dans des environnements perturbés, comme les applications téléphoniques notamment, les systèmes de VAL peuvent être confrontés à de fortes variations de conditions d'utilisation, entraînant une augmentation importante des erreurs de reconnaissance. Pour diminuer ce type d'erreurs, les systèmes actuels doivent intégrer des techniques de compensation dont l'objectif est d'atténuer les effets des disparités entre les données d'apprentissage et celles de test. Les approches courantes de compensation ont cependant des faiblesses qui les rendent contraignantes ou inadaptées à certaines situations applicatives. Dans cette thèse, nous développons des techniques destinées à remédier à certaines de ces limitations. Nos travaux s'inscrivent dans l'approche probabiliste pour la VAL, pour laquelle les locuteurs sont modélisés par des modèles de mélange de Gaussiennes et l'étape de décision est basée sur un test d'hypothèses Bayésien. Nous élaborons dans un premier temps de nouvelles techniques de normalisation dont le but est de renforcer la robustesse du processus de décision face aux variabilités rencontrées. Ces normalisations, basées sur l'utilisation de distances de Kullback-Leibler entre modèles de locuteurs, sont appliquées au niveau des scores de vérification et au niveau des modèles eux mêmes. Elles se distinguent des approches conventionnelles par le fait qu'elles ne font appel à aucun corpus de données réelles pour l'estimation des paramètres de normalisation. Nous formalisons également un nouveau cadre pour la vérification du locuteur dans un espace des modèles. Ce cadre conduit à un calcul simplifié des scores de vérification et autorise une manipulation efficace des modèles de locuteurs, offrant ainsi de nombreuses possibilités de normalisation. Les résultats expérimentaux montrent que les techniques proposées peuvent avantageusement remplacer certaines approches courantes, en allégeant considérablement la procédure de vérification. Nous concevons dans un deuxième temps un schéma d'adaptation Bayésienne hiérarchique qui a pour but d'améliorer l'estimation des modèles de locuteurs lorsque la quantité de données d'apprentissage est faible. La technique proposée généralise l'approche classiquement utilisée en VAL, en offrant de plus la possibilité d'intégrer des dépendances entre différentes régions acoustiques occupées par la voix d'un locuteur. La mise en pratique de cette technique est cependant délicate et les conditions dans lesquelles elle a été utilisée n'ont pas permis pour l'instant de mettre en évidence un apport décisif de la méthode. Néanmoins, le cadre théorique introduit est attrayant et offre de multiples perspectives. L'ensemble des techniques étudiées a été évalué sur des bases de données téléphoniques en parole naturelle, dans le cadre des évaluations NIST en reconnaissance du locuteur.