

N° d'ordre: 2804

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

par

Raphaël BLOUET

Équipe d'accueil : IRISA/METISS

École Doctorale : Mathématiques, Informatique, Signal, Électronique et
Télécommunications

Composante universitaire : STRUCTURES ET PROPRIÉTÉS DE LA MATIÈRE

Titre de la thèse :

*Approche probabiliste par arbres de décision pour la vérification automatique
du locuteur sur architectures embarquées*

soutenue le 16 12 2002 devant la commission d'examen

M. :	Bernard	DELYON	Président
MM. :	Jean-François	BONASTRE	Rapporteurs
	Paul	DELÉGLISE	
MM. :	Christian	GOIRE	Examineurs
	Jean-Paul	HATON	
	Frédéric	BIMBOT	Directeur de Thèse

Remerciements

Je tiens en premier lieu à remercier Frédéric Bimbot, Chargé de Recherche à l'IRISA, pour avoir encadré mes travaux pendant toute la durée de cette thèse. Son enthousiasme au quotidien, son dynamisme et la qualité de ses choix scientifiques ont énormément contribué à la réalisation de ces travaux.

Je suis tout particulièrement reconnaissant aux rapporteurs de cette thèse, Jean-François Bonastre, Maître de conférence, Habilité à Diriger des Recherches au Laboratoire d'Informatique d'Avignon et Paul Deléglise, Professeur à l'Institut d'Informatique Claude Chappe, pour l'attention qu'ils ont portée à ces travaux. Je remercie Bernard Delyon, Professeur à l'Université de Rennes I qui a bien voulu présider le jury de soutenance de cette thèse ainsi que Christian Goire, Directeur de Recherche à Schlumberger-CP8 et Jean-Paul Haton, Professeur au LORIA, d'avoir bien voulu examiner ce travail.

Je tiens à témoigner de ma reconnaissance envers Mouhamadou Seck, Rémi Gribonval et Gérard Subsol, avec lesquels j'ai successivement eu beaucoup de joie à partager un bureau, pour leur disponibilité, tant scientifique qu'amicale ainsi que pour leur bonne humeur quotidienne. Je remercie tous les membres passés ou présents de l'équipe METISS, Mathieu Ben, Mickaël Bester, Lorcan McDonagh, Fabienne Porée, Guillaume Gravier et Bruno Jacob qui ont tous contribué au cadre agréable de travail tout au long de cette thèse.

Je remercie particulièrement Laurent Benaroya, entre autre, pour toutes nos discussions et son accueil chaleureux lors de mes passages à Rennes.

Je remercie Téva Merlin du LIA pour son accueil et sa disponibilité lors de mes passages à Avignon ainsi que tous ceux et celles qui m'ont fait découvrir et apprécier Rennes, parmi lesquelles Emilie Le Gall et Anna Morali.

Enfin, je remercie Nolwenn Le Gall pour son soutien au quotidien, sa patience et sa compréhension.

Table des matières

Résumé	ix
Glossaire et acronymes	xi
Notations	xiii
Introduction	1
I Présentation de la vérification du locuteur	7
1 La VAL comme technologie d'identification biométrique	11
1.1 L'authentification biométrique	11
1.2 Définition de la biométrie	12
1.3 Discussion sur la caractéristique biométrique et le SAB induit	13
1.4 Propriétés de la caractéristique biométrique	15
1.5 Classification des systèmes biométriques	15
1.6 Avantages / inconvénients de la VAL	17
2 La VAL comme technologie du traitement automatique de la parole	21
2.1 La caractérisation automatique du locuteur	21
2.2 Description des applications de la RAL	22
2.3 Caractéristiques de la vérification automatique du locuteur	26
2.4 Applications visées par la collaboration avec CP8	28
3 Évaluation des systèmes biométriques	29
3.1 Aspects de l'évaluation d'un système d'authentification biométrique	30
3.2 Critères d'évaluation des systèmes biométriques	31
3.3 Corpus d'évaluation	33
3.4 Contraintes et limitations des évaluations	34
3.5 Les évaluations NIST	35

II	Description d'un système de vérification du locuteur	37
4	Formalisme général de la vérification automatique du locuteur	41
5	Modules d'une plateforme de VAL : techniques de l'état-de-l'art	43
5.1	Analyse acoustique	44
5.2	Modélisation des locuteurs	52
5.3	Calcul du score de décision	55
5.4	Normalisation	57
5.5	Le module de décision	60
6	Description du système de référence	61
6.1	Le module de paramétrisation	61
6.2	Le module de modélisation	62
6.3	Le module de normalisation	63
6.4	Calcul du score et décision	63
6.5	En résumé	63
6.6	Orientations de ce travail	63
III	État de l'art, mise en œuvre et améliorations	65
7	Modélisation probabiliste des locuteurs	69
7.1	Les densités de mélange de gaussiennes	70
7.2	Estimation au Maximum de Vraisemblance	72
7.3	Estimation MAP et techniques d'adaptation	73
7.4	Performances de la méthode proposée	80
8	Description d'une nouvelle technique de normalisation : la <i>d-norm</i>	89
8.1	Rappel sur les motivations de la normalisation	89
8.2	Description de la <i>d-norm</i>	90
8.3	Performances	91
9	Limites et premières solutions	93
9.1	Réduction de la complexité : approches existantes	94
IV	Approche probabiliste par arbres de décision	97
10	Caractéristiques du système de VAL à base d'arbres de décision	101
10.1	Estimation locale du score de décision	101
10.2	Formalisme de la mise en œuvre	102
10.3	Partitionnement de l'espace des paramètres	103
10.4	Affectation d'un score à chaque région	109

11 Description de la plateforme à base d'arbres de décision	115
11.1 Vérification du locuteur et arbres de décision	115
11.2 Paramètres de l'analyse	116
11.3 Protocole expérimental	118
12 Résultats principaux	121
12.1 Évaluation du critère d'homogénéité	121
12.2 Évaluation du critère de dispersion	131
12.3 Bilan matériel	133
13 Travaux complémentaires et consolidation de la technique	137
Travaux complémentaires et consolidation de la technique	137
13.1 Motivations	137
13.2 Amélioration de l'estimation du score de décision	138
13.3 Représentation par arbres multiples	139
13.4 Résultats	141
13.5 Bilan matériel	146
Conclusion et perspectives	147
Bibliographie	150

Table des figures

1.1	Description des phases de fonctionnement d'un système d'authentification biométrique	16
1.2	Comparaison des avantages et inconvénients applicatifs de différentes technologies biométriques d'après www.biometricgroup.com	18
1.3	Comparaison des performances de différentes technologies de vérification biométrique d'identité d'après [Mansfield et al., 2001]	20
2.1	Schéma d'un système d'identification du locuteur	25
2.2	Schéma d'un système de détection de locuteur	25
2.3	Schéma d'un système d'indexation par locuteur	25
2.4	Schéma d'un système de suivi de locuteurs	25
2.5	Exemple d'application visée par la collaboration entre METISS et SCHLUMBERGER-CP8	28
3.1	Performances sur des données réelles et sur des données générées d'un système de vérification du locuteur	35
5.1	Description d'un système de vérification du locuteur	45
5.2	Principe de la paramétrisation	51
6.1	Description du système de vérification du locuteur de référence	64
7.1	Principe de l'algorithme Expectation Maximisation	72
7.2	Estimation ML des paramètres des GMM : courbes DET obtenues pour des systèmes de 64, 128, 256 et 512 composantes	74
7.3	$f(n_k)$, contribution du modèle du monde pour l'estimation du modèle client	79
7.4	Comparaison de la contribution relative du modèle du monde dans l'estimation du modèle client pour notre technique d'adaptation et celle proposée dans [Reynolds et al., 2000] avec $r = 15$	83
7.5	Contribution du modèle du monde pour l'estimation du modèle client avec ρ_k calculé selon l'équation 7.22 pour différentes valeurs de α	83
7.6	Estimation MAP des paramètres des GMM : performances d'un système de 128 gaussiennes pour $N_{min} = \{40, 120, 160\}$ et $\rho_{min} = \{0.1, 0.3, 0.5, 0.7\}$	86

7.7	Estimation MAP des paramètres des GMM : performances de systèmes de 256 et 512 gaussiennes pour différentes valeurs de N_{min} et ρ_{min}	87
7.8	Résultat aux évaluations NIST des systèmes de l'IRISA en 1998, 1999, 2000 et 2002	88
8.1	Scores client et imposteur moyens / distances de Kullback	91
8.2	Scores client et imposteur moyens / distances de Kullback d-normalisées	92
8.3	Courbes DET obtenues à partir d'un même système de base, sans normalisation, par <i>z-normalisation</i> et par <i>d-normalisation</i> des scores	92
10.1	Illustration du principe de partition de l'espace des paramètres dans le cas de l'affectation d'un score constant, ici représenté par différents niveaux de gris en chaque région	102
10.2	Architecture des phases d'apprentissage et de test selon la mise en œuvre classique (a) et la mise en œuvre proposée (b)	106
10.3	Partitionnement de l'espace des paramètres	108
10.4	Exemple d'un arbre de classification pour un problème à 2 classes	108
10.5	Classification à deux classes avec des vecteurs de dimension 2	109
11.1	Trois stratégies de représentation de $H_{\bar{X}}$ pour la construction des arbres de décision	120
12.1	Moyenne, écart-type et distance des accès client et imposteur, dans la configuration du système $N_{min} = 10$ et $r = \{1, 2, 3, 4, 5\}$	126
12.2	Moyenne, écart-type et distance des accès client et imposteur, dans la configuration du système $N_{min} = 50$ et $r = \{1, 2, 3, 4, 5\}$	127
12.3	Moyenne, écart-type et distance moyenne des accès client et imposteur, dans la configuration du système $r = 1$ et $N_{min} = \{10, 50, 100\}$	128
12.4	Critère de Gini - Courbes DET associées aux performances des critères d'arrêts $N_{min} = 10$ et $N_{min} = 50$ pour différents r	129
12.5	Critère de Gini - Courbes DET associées aux performances de différents critères d'arrêt du partitionnement de l'espace des paramètres	130
12.6	Critère de dispersion - Critère de Gini : comparaison des meilleures performances	132
13.1	Principe de l'algorithme Discrete AdaBoost	140
13.2	Principe de l'algorithme Real AdaBoost , extrait de Friedman et al. [1998]	141
13.3	Moyenne, écart-type et distance des accès client et imposteur, dans les configurations du système $N_{min} = \{10, 50, 100\}$ et N_c , nombre de composantes de la modélisation additive, $N_c = \{1, 2, 5, 10, 15, 25\}$	144
13.4	Courbes DET obtenues pour des systèmes à base de plusieurs arbres de décision pour différents critères d'arrêt et différents nombres de composantes	145

Liste des tableaux

6.1	Performances selon les critères C_{Det} et $HTER$ de notre système de référence	64
7.1	Estimation ML des paramètres des GMM : performances à l'HTER et au coût de fonctionnement C_{Det} pour différentes tailles de modèles clients	74
7.2	Performances de différentes configurations de l'estimation MAP d'un système à 128 composantes	81
7.3	Performances de différentes configurations de l'estimation MAP d'un système à 256 composantes	81
7.4	Performances de différentes configurations de l'estimation MAP d'un système à 512 composantes	81
7.5	Performances de différentes configurations de l'estimation MAP suivant la configuration [Reynolds et al., 2000] d'un système à 128 composantes	82
7.6	Performances obtenues avec des GMM de 128 composantes en considérant le calcul des ρ_k par la formule de l'équation 7.22	82
12.1	Critère de Gini : C_{Det} et $HTER$ pour différentes configurations de représentation de $H_{\bar{X}}$ avec (a) $s_{R_X}^k$ constant et tel $s_{R_X}^k \in \{-1, 1\}$ et (b) $s_{R_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$	123
12.2	Critère de dispersion : C_{Det} et $HTER$ pour différentes représentations de $H_{\bar{X}}$ avec $s_{R_X}(R_X^k, y)$ tel que défini dans l'équation 10.15	131
12.3	Nombre moyen de partitions (N_f) et nombre moyen de questions (N_Q) pour différentes configurations du système à base d'arbres de décision .	134
12.4	Nombre d'opérations nécessaires au calcul de décision pour les critères d_k et g_k en fonction du nombre N d'observations de test	134
12.5	Comparaison entre les ressources matérielles nécessaires à deux systèmes à base de GMM (128 et 256 composantes) et celles nécessaires à deux systèmes à base d'arbres de décision (Gini, $r = 3$) et (dispersion, $r = 1$) avec N , nombres d'observations de test	135

13.1	Critère de Gini - évaluation de l'algorithme Real Adaboost : C_{Det} et $HTER$ pour différents critères d'arrêt pour le partitionnement de \mathcal{Y} et en considérant 1, 2, 5, 10, 15 et 25 composantes dans la représentation additive. Pour chaque composante $s_{R_X}^k$ (a) constant et tel $s_{R_X}^k \in \{-1, 1\}$ et (b) $s_{R_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$	143
13.2	Nombre moyen de partitions (N_f) et nombre moyen de questions (N_Q) pour différents nombres de composantes de la représentation par arbres multiples des locuteurs	146

Résumé

Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées :

La vérification du locuteur consiste à déterminer automatiquement l'identité d'une personne à partir de sa voix. Actuellement, ses perspectives de mise en œuvre se situent essentiellement dans le domaine de la sécurisation d'accès à des services, des locaux ou de transactions bancaires téléphoniques. Si les outils qu'elle utilise sont souvent issus des recherches et bénéficient des progrès généraux des différentes branches du traitement automatique de la parole, ses caractéristiques techniques et applicatives en font une technologie de l'identification humaine automatique (biométrie) particulièrement intéressante.

Les travaux présentés dans cette thèse ont été effectués dans le cadre d'une convention de recherche entre l'équipe CP8 de SCHLUMBERGER-SEMA (ex BULL-CP8) et le projet METISS de l'IRISA. L'objet de cette collaboration est de mettre en œuvre un système de vérification automatique du locuteur dont tout ou partie des traitements sont effectués sur une carte à microprocesseur. Dans ce but, nous avons tout d'abord cherché à mettre en place et à améliorer un système de vérification automatique du locuteur (VAL) suivant les caractéristiques de l'état-de-l'art, puis nous nous sommes efforcés de réduire autant que possible la quantité de mémoire et la puissance de calcul nécessaires au fonctionnement d'une plateforme de VAL. Dans un premier temps, nos travaux ont donc concerné le développement d'une méthode d'estimation suivant le critère du *Maximum A Posteriori* (MAP) des paramètres des modèles de mélanges de gaussiennes (GMM) ainsi que celui d'une technique originale pour la normalisation du rapport de vraisemblance. Dans un second temps, nous avons développé un système de VAL basé sur les arbres de décision et dont le principe est d'estimer, directement lors de la phase d'apprentissage, le score de décision associé à chaque point de l'espace des paramètres. Une première implémentation de la technique utilise des critères de partition classiquement associés à la construction d'arbres de décision (type critère de Gini et entropie). Une seconde utilise, quant à elle, un critère plus original dans ce domaine et dont la mise en place a permis de fortement augmenter les performances. Enfin, une troisième mise en œuvre de la méthode est basée sur une représentation additive de plusieurs arbres par locuteur.

Les expériences réalisées montrent que si la méthode proposée entraîne une diminution des performances par rapport aux meilleurs systèmes GMM, elle conduit à une réduction considérable des quantités de mémoire et de la puissance de calcul nécessaires à la mise en place d'un système de VAL. Ce travail offre de nombreuses perspectives, qu'il s'agisse de la ré-utilisation de la technique dans d'autres domaines de la biométrie ou des différentes pistes devant permettre son amélioration. Parmi celles-ci, l'incorporation de connaissances a priori pour l'apprentissage des arbres de décision, à l'instar du critère MAP utilisé avec les GMM, semble être la plus prometteuse.

Glossaire et acronymes

AR	:	Auto Régressif
DTW	:	Dynamic Time Warping
CAL	:	Caractérisation Automatique du Locuteur
client	:	Personne connue du système d'authentification biométrique et dont l'identité proclamée est authentique
CMS	:	Cepstral Mean Subtraction
DET	:	Data Error Trade-off
EM	:	Expectation Maximisation
ddp	:	densité de probabilité
GMM	:	Gaussian Mixture Model
HMM	:	Hidden Markov Model
HTER	:	Half Total Error Rate
IAL	:	Identification Automatique du Locuteur
imposteur	:	Locuteur cherchant à tromper le système d'authentification biométrique
ICA	:	Independant Component Analysis
LDA	:	Linear Discriminant Analysis
LPC	:	Linear Predictive Coding

LSP	:	Line Spectrum Pairs
MAP	:	Maximum A Posteriori
ML	:	Maximum Likelihood
non-locuteur	:	Locuteur virtuel, censé représenter l'alternative au client dans l'espace des paramètres
NIST	:	National Institute of Standards and Technology
PCA	:	Principal Component Analysis
PNN	:	Predictive Neural Network
RAL	:	Reconnaissance Automatique du Locuteur
RAP	:	Reconnaissance Automatique de la Parole
ROC	:	Receiver Operating Characteristic
SAB	:	Système d'Authentification Biométrique
TAP	:	Traitement Automatique de la Parole
TCI	:	Transformée en Cosinus Inverse
UBM	:	Universal Background Model, équivalent au modèle de non-locuteur
v.a.	:	variable aléatoire
VAL	:	Vérification Automatique du Locuteur
VQ	:	Vector Quantization

Notations

Représentation du signal de parole

- d : dimension du vecteur de représentation acoustique
- Y : signal de parole reçu à l'entrée du système de VAL
- \mathcal{Y} : espace des paramètres acoustiques
- \mathbf{y}_t : vecteur acoustique extrait de la $t^{\text{ième}}$ trame du signal de parole
- \mathbf{Y} : suite de vecteurs acoustiques extraits de Y , $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^N$

Caractérisation du locuteur

- X : locuteur associé à l'identité proclamée
- \bar{X} : ensemble des locuteurs autres que X
- H_X : hypothèse selon laquelle \mathbf{Y} a été prononcé par le locuteur X
- $H_{\bar{X}}$: hypothèse selon laquelle \mathbf{Y} n'a pas été prononcé par X
- P_X : probabilité *a priori* d'accès client
- $P_{\bar{X}}$: probabilité *a priori* d'accès imposteur
- Γ_X : référence caractéristique du locuteur
- $p_a(y)$: densité de probabilité de H_a dans l'espace des paramètres acoustiques
- λ_a : paramètres de la densité de probabilité associée à H_a

\mathcal{R}_X	:	partition de l'espace des paramètres associée au locuteur X
R_X^k	:	région k de la partition \mathcal{R}_X
$Q_X(y)$:	fonction d'indexation de l'espace des paramètres suivant la partition \mathcal{R}_X
$s_X^k(y)$:	score de décision associée à la région R_X^k

Calcul du score de décision, évaluation et mesure de similarité

Fa	:	fausse alarme, erreur correspondant à l'acceptation à tort d'un imposteur
Fr	:	faux rejet ou non-détection, erreur correspondant au rejet à tort d'un locuteur
P_{fa}	:	probabilité de fausse alarme
P_{fr}	:	probabilité de non-détection
C_{fa}	:	coût associé à l'erreur de fausse acceptation
C_{fr}	:	coût associé à l'erreur de faux rejet
C_{Det}	:	coût de fonctionnement pour l'évaluation des systèmes de VAL : $C_{Det} = C_{fr} \times P_{fr} \times P_X + C_{Fa} \times P_{Fa} \times P_{\bar{X}}$
$S_X(\mathbf{Y})$:	score de décision associé à l'identité proclamée X étant donnée la suite de vecteurs de représentation Y
$D_a(\mathbf{Y})$:	score de l'hypothèse H_a relativement aux observations \mathbf{Y}
$V_a(Y)$:	vraisemblance de la référence associée à l'hypothèse H_a par rapport aux observations Y
θ	:	seuil de décision au dessous duquel l'identité proclamée est rejetée et en dessus duquel elle est acceptée

Introduction

POSITIONNEMENT

Le but de la Vérification Automatique du Locuteur (VAL) est de vérifier l'identité associée à un signal de parole. Celle-ci peut être explicitement proclamée par un individu ou implicite à l'utilisation d'un service. La VAL s'intègre dans deux domaines de recherche de la communication Homme-Machine : le traitement automatique de la parole et l'authentification humaine automatique (ou biométrie).

L'interface Homme-Machine est l'un des éléments qui conditionne pour une large part le succès d'un système informatique. L'interface idéale doit être naturelle, ergonomique, efficace et attrayante. Son utilisation doit être intuitive et conviviale. Une partie importante des travaux de ce domaine tend à la rendre la plus proche possible des moyens usuels de communication humains que sont, par exemple, la parole et l'écriture. On peut ainsi noter la généralisation des commandes vocales dans certains appareils électroniques grand public et l'utilisation fréquente de système de reconnaissance d'écriture manuscrite dans les agendas électroniques. Depuis l'apparition des premiers ordinateurs, les dispositifs matériels et logiciels d'échange d'informations entre l'homme et la machine n'ont cessé de se perfectionner pour des utilisateurs de plus en plus exigeants sur le plan de leur fiabilité et de leur simplicité.

Le Traitement Automatique de la Parole (TAP) regroupe l'ensemble des technologies de la communication Homme-Machine basées sur l'analyse du signal de parole. Si l'avancement des recherches actuelles ne permet pas de créer de systèmes entièrement basés sur le langage naturel, des progrès considérables ont été effectués au cours de ces vingt dernières années. Ainsi, la Reconnaissance Automatique de la Parole (RAP) est maintenant familière à de nombreux utilisateurs de micro-ordinateurs pour qui elle complète le traitement de texte. De même, la synthèse vocale est utilisée dans plusieurs applications grand public (horloge parlante, serveurs téléphoniques, application de type "voicemail" etc.) et plus personne ne s'étonne d'entendre une machine *parler*.

Si les applications actuelles du TAP sont principalement basées sur l'analyse du message linguistique, ce dernier n'est pas la seule information véhiculée par le signal de parole qui nous renseigne aussi sur l'identité et sur certaines caractéristiques (fatigue, stress etc.) de la personne qui l'a généré. C'est à ce type d'informations que

s'intéresse la Caractérisation Automatique du Locuteur (CAL) dont l'une des applications, la vérification automatique du locuteur, constitue le sujet principal de ce document. Les systèmes de VAL cherchent à extraire d'un énoncé toute information permettant de distinguer celui qui l'a prononcé par rapport à l'ensemble des autres locuteurs.

La biométrie rassemble l'ensemble des techniques automatiques d'identification telles que les empreintes digitales, l'identification du visage ou la reconnaissance de signature. Bien qu'on puisse l'inscrire parmi les disciplines de la communication Homme-Machine, elle est généralement rattachée au domaine plus large de l'Identification Humaine Automatique. L'intérêt suscité par la biométrie a considérablement augmenté ces cinq dernières années. Ceci s'explique, entre autres, par la baisse du coût des différents composants (capteurs, systèmes de calcul) nécessaires à la réalisation d'un système d'identification biométrique mais aussi par la généralisation de l'électronique et de l'informatique dans la société.

CONTEXTE

Le travail présenté ici a été effectué à l'IRISA sous la direction de Frédéric BIMBOT et a débuté en janvier 1999. Il a été accompli dans le cadre d'une convention de recherche entre l'équipe CP8 de SCHLUMBERGER (initialement BULL) et le projet METISS (issu du projet SIGMA2) de l'IRISA. Nous nous sommes appuyés sur la plateforme commune du consortium ELISA et sur la participation annuelle aux évaluations NIST (National Institut of Standards and Technology) pour développer puis évaluer nos propres systèmes de VAL.

Le consortium ELISA

Depuis sa création en 1997, le consortium ELISA réunit de manière ouverte et évolutive plusieurs laboratoires de recherche ¹ autour de la caractérisation automatique du locuteur. Il permet d'offrir à tout nouveau membre la possibilité de développer rapidement son propre système de VAL en focalisant son effort de recherche en fonction de son expertise et/ou sur les points qu'il juge les plus intéressants. D'une année sur l'autre, chaque membre du consortium profite des innovations et des progrès réalisés par l'ensemble des participants et a ainsi le moyen de présenter, dans des conditions aussi bonnes que possible, son propre système de VAL aux évaluations NIST. Ce mode de fonctionnement n'a été rendu possible qu'après la mise au point d'une plateforme de référence dont les performances sont les plus proches possibles de celles des systèmes de l'état-de-l'art.

Au début de ce travail, l'utilisation des méthodes probabilistes et plus particulièrement

¹En 2002 on peut compter comme membres actifs du consortium les laboratoires suivants : CLIPS (Grenoble), ENST (Paris), IRISA (Rennes), LIA (Avignon) et DDL (Lyon).

des modèles de mélange de gaussiennes était largement dominante dans l'ensemble des travaux publiés en VAL. Aussi, une bonne partie des efforts de recherche était alors concentrée sur l'apprentissage des paramètres de ces modèles. Si les méthodes d'estimation suivant le critère du *Maximum A Posteriori* (MAP) sont actuellement largement utilisées par les meilleurs systèmes de VAL, elles n'étaient alors qu'émergentes et leur étude ainsi que la réalisation d'un système de VAL utilisant ce critère a fait l'objet de la première partie de ce travail. Notre propre mise en œuvre ainsi que l'estimation de ses performances sont décrites dans la seconde partie de ce document. De plus, nous y présentons des travaux sur la normalisation des scores avec la présentation d'une technique originale pour la normalisation du rapport de vraisemblance : la *d-norm*.

L'implication de l'IRISA dans le consortium ELISA a facilité l'évaluation des performances des systèmes développés dans le cadre de cette thèse. Chacun d'eux a fait l'objet d'un calibrage régulier dans le contexte des évaluations NIST. Celles-ci, auxquelles nous avons consacré un chapitre dans la première partie du document, nous ont permis de tester objectivement et systématiquement les différentes mises en œuvre proposées au cours de nos travaux.

La collaboration avec CP8

La convention de recherche entre METISS et CP8 a démarré avec cette thèse. Son principal objectif vise à étudier différentes stratégies pour l'implémentation de tout ou partie d'un système de VAL sur cartes à microprocesseur (carte à puce). Si le type d'applications visées permet que certains des traitements soient exécutés sur des calculateurs externes à la carte, il requiert qu'au minimum celle-ci prenne en charge le calcul du score de décision. Cette contrainte, liée au niveau de sécurité que l'on souhaite atteindre impose que la référence caractéristique² du locuteur ne soit jamais transmise hors de la carte. Les difficultés de réalisation sont alors induites par les réductions jointes de l'espace de stockage et de la puissance de calcul. Ces dernières interdisent l'utilisation d'un système de VAL suivant les caractéristiques de l'état de l'art et nous ont conduit à proposer l'utilisation d'arbres de décision pour la mise en place d'un système de VAL alternatif. L'avantage principal de cette mise en œuvre est de permettre une forte réduction des besoins en temps de calcul et en espace mémoire par rapport à celle de l'état-de-l'art.

²Les notions de score de décision et de référence caractéristique sont expliquées au cours du premier chapitre de ce document.

PRINCIPALES CONTRIBUTIONS

Les principales contributions du travail présentés dans cette thèse peuvent être résumées ainsi :

1. La mise en œuvre et l'évaluation d'une méthode originale pour l'estimation suivant le critère du *Maximum a Posteriori* (MAP) des paramètres des modèles de mélange de gaussiennes (GMM). L'originalité de notre approche réside dans l'estimation des paramètres de la densité *a priori*.
2. Une méthode de normalisation du rapport de vraisemblance basée sur la distance de Kullback entre la densité associée à l'hypothèse pour l'acceptation de l'identité proclamée et celle associée à son rejet. L'avantage principale de cette technique ³ est de permettre d'obtenir une amélioration des performances comparable à celle de normalisations classiques (du type de la *z-norm*) sans nécessiter d'un ensemble de signaux de parole pour l'estimation de paramètres.
3. L'utilisation d'arbres de décision dans un cadre probabiliste est l'une des originalités des systèmes de VAL que nous avons développés pour répondre aux impératifs matériels de notre collaboration avec CP8. De plus, nous avons lors de leur mise en œuvre défini un critère original pour le partitionnement de l'espace des paramètres avec des arbres de décision. Celui-ci est basé sur la dispersion des observations en chacune des régions et s'est avéré particulièrement efficace pour la tâche que nous avons considérée.
4. L'amélioration d'une des mises en œuvre proposée au chapitre 10 grâce au partitionnement par arbres multiples de l'espace des paramètres et en particulier l'utilisation de l'algorithme `adaboost` pour leur estimation et la mise en place d'un système de vérification du locuteur.

PLAN DU DOCUMENT

Ce document se compose de quatre parties et se termine par un chapitre sur les conclusions et perspectives des travaux présentés.

Dans la première partie, nous présentons la vérification du locuteur d'un point de vue technologique. On la situe d'abord parmi les deux domaines d'applications dans lesquels elle s'inscrit : celui du traitement automatique de la parole et celui de l'authentification biométrique. Nous présentons ensuite les critères d'évaluation des systèmes biométriques avant les protocoles expérimentaux d'apprentissage et de test des évaluations NIST. Enfin, nous donnons une description du type des applications visées par notre collaboration avec CP8.

³mise au point en collaboration avec Mathieu Ben, actuellement doctorant à l'IRISA

La seconde partie permet d'introduire le formalisme utilisé par les systèmes de vérification du locuteur et de présenter leurs fondements théoriques. Nous y décrivons la structure complète d'un système de VAL en donnant différentes solutions de mises en œuvre pour chacun des éléments le constituant. Cette partie permet de comprendre les enjeux scientifiques de la recherche en vérification du locuteur et d'apprécier la diversité des domaines qui lui sont liés.

La troisième partie contient une description détaillée de la méthode la plus utilisée et actuellement la plus efficace en VAL : la modélisation statistique à base de modèles de mélange de gaussiennes. Après un rappel des principales caractéristiques de ces densités, nous exposons nos travaux sur l'estimation suivant le critère MAP des paramètres des GMM ainsi que l'évaluation des performances qu'ils permettent d'obtenir. Puis, nous présentons la technique de normalisation que nous avons mise au point. Enfin, ce chapitre se termine par l'exposé des limites propres à l'utilisation des GMM et par la mise en évidence de l'impossibilité de les utiliser dans le cadre de l'application visée par la convention de recherche entre l'IRISA et CP8.

La quatrième et dernière partie de ce document est au cœur de la collaboration entre CP8 et l'IRISA. Il décrit une approche basée sur les arbres de décision pour la mise au point d'un système de vérification du locuteur sous la contrainte d'une forte réduction des capacités de calcul et de l'espace de stockage. Dans un premier temps, nous y présentons deux familles de critères permettant la construction d'arbres de décision. La première correspond à un critère d'homogénéité dont l'utilisation est fréquente dans la littérature sur les arbres de décision. La seconde correspond à un critère de dispersion permettant d'obtenir un arbre dont chaque feuille regroupe les vecteurs acoustiques les plus proches dans l'espace de représentation acoustique. Dans un second temps, nous y présentons différents développements visant à améliorer la mise en œuvre proposée. La piste explorée concerne l'utilisation du critère d'homogénéité avec plusieurs arbres pour définir la partition de l'espace des paramètres.

Première partie

Présentation de la vérification du locuteur

Cette partie est une présentation de certains aspects applicatifs de la vérification automatique du locuteur. On y discute de sa place dans chacune des deux familles de technologies dans lesquelles elle s'inscrit : celle de l'authentification biométrique et celle du traitement automatique de la parole. De plus, nous y décrivons l'élément indispensable à la conception d'un système de VAL : l'évaluation.

Après avoir exposé dans l'introduction le contexte général et les motivations de ce travail, nous consacrons la première partie de ce document à la présentation précise des différentes familles de technologies auxquelles on rattache habituellement la vérification automatique du locuteur ainsi qu'à la description de certains des éléments liés à l'évaluation des systèmes de VAL. Ce premier point permet de situer la VAL parmi les technologies de l'authentification humaine automatique et de présenter les différents domaines de recherches dont les progrès génèrent souvent des retombées en VAL. Le second permet de décrire les principales caractéristiques de l'évaluation, phase complexe, coûteuse mais incontournable du développement de tout système d'identification biométrique.

Les deux premiers chapitres de cette partie présentent la VAL parmi les technologies de l'authentification biométrique (ou biométrie) puis parmi celles du traitement automatique de la parole. D'autres domaines de recherche comme celui des sciences criminalistiques ou du traitement d'image (par exemple à partir de l'analyse des mouvements des lèvres du locuteur [Zhang et Broun, 2001]) ont contribué au développement de la VAL mais nous nous sommes limités aux deux domaines pré-cités. Dans le cas des sciences criminalistiques, nous renvoyons le lecteur pour plus de précision à [Boë et al., 1999] pour une mise en garde sur l'utilisation de la VAL dans le domaine judiciaire, ainsi qu'à [Kunzel, 1994], [Meuwly et Drygajlo, 2001] et [Nakasone et Beck, 2001] pour une description des besoins et des limitations de la vérification du locuteur dans ce domaine.

Dans le premier chapitre, nous rappelons tout d'abord la définition de la biométrie puis, après une description générale des principes de fonctionnement des systèmes d'authentification biométrique (SAB), nous présentons une classification de ces systèmes ainsi qu'un tour d'horizon des différentes technologies actuellement les plus utilisées. Un des enjeux de cette section étant de situer la VAL parmi l'ensemble de ces technologies, elle se termine par la présentation de deux évaluations, l'une quantitative, l'autre qualitative, de plusieurs technologies de SAB.

Dans le deuxième chapitre, nous situons la VAL dans son contexte le plus usuel *i.e.* celui du traitement automatique de la parole (TAP) et plus particulièrement celui de la reconnaissance automatique du locuteur (RAL). Après une brève présentation des disciplines du TAP, nous décrivons les différentes applications de la RAL. Les caractéristiques de celles ayant de fortes interactions avec la VAL sont précisées. La dernière section de ce chapitre détaille plus avant les caractéristiques de la vérification du locuteur et donne une description quantitative des principes de fonctionnement des

systemes de VAL ainsi que des principales difficultes de realisation associees.

Le troisieme chapitre porte sur l'evaluation des systemes biometriques. Elle met en evidence certaines de leurs difficultes tant du point de vue de la definition du protocole que de celui de l'interpretation des resultats. Nous y presentons nos motivations et justifions notre interet pour les campagnes d'evaluation qui constituent actuellement le seul moyen objectif d'estimation et de comparaison des performances de ces systemes. Cette section se termine par la presentation des corpus ainsi que des protocoles d'apprentissage et de test des evaluations proposees par l'institut americain NIST. Les donnees utilisees lors de l'estimation des performances de tous nos systemes sont extraites de ce corpus.

Chapitre 1

La VAL comme technologie d'identification biométrique

Dans ce chapitre, nous précisons la place et le rôle de la vérification automatique du locuteur parmi les différentes technologies de l'identification biométrique. Deux raisons nous ont incités à situer la VAL dans ce contexte. La première est que ces technologies font directement partie du contexte de la collaboration entre CP8 et l'équipe METISS de l'IRISA, la seconde est liée à l'émergence de la biométrie et à l'intérêt grandissant qu'elle suscite dans de nombreux secteurs¹.

1.1 L'authentification biométrique

L'authentification ou identification biométrique est une des technologies de l'identification humaine automatique. Ces technologies sont essentiellement caractérisées [Liu et Silverman, 2001] par l'élément qui permet l'identification. Celui-ci peut être :

1. quelque chose de connu par l'utilisateur (mot de passe, code etc.),
2. quelque chose de détenu par l'utilisateur (clef, carte etc.),
3. quelques chose de spécifique à l'utilisateur : la caractéristique biométrique.

Les technologies de l'authentification biométrique regroupent un ensemble de procédés dont le but est d'identifier automatiquement une personne à partir de la mesure directe de l'une de ses caractéristiques physiques ou comportementales. Alors que les mots de passe, les clefs ou les cartes sont facilement oubliés, perdus ou volés, l'identification biométrique permet de s'en affranchir et de sécuriser, sans ce type de contrainte, l'accès à un service, des locaux ou des données protégées. Cette caractéristique fait de la biométrie l'une des technologies privilégiées pour sécuriser les applications pour

¹ Ainsi depuis 1995, on peut dénombrer un certain nombre de projets européens ayant trait de près ou de loin à la biométrie (ASPECT, M2VTS, CAVE, BANCA, etc..) et noter la tenue de nombreux WORKSHOP spécialisés sur le domaine. Enfin, on observe depuis quelques années une forte hausse des bénéfices commerciaux des entreprises de ce domaine.

lesquelles le client n'est pas physiquement en contact avec son prestataire comme dans [Jain et al., 1998] où elle permet de sécuriser l'accès à certains sites sur Internet.

Apparues il y a une trentaine d'années au sein de la société Shearson Hamil à Wall Street sous la forme d'un système vérifiant la taille des doigts des employés [Frye, 2001], les techniques automatiques d'authentification biométrique n'ont, dès lors, cessé de se diversifier et de se perfectionner. Si elles n'ont pas connu l'essor annoncé dans le milieu des années 70, le regain d'intérêt qu'elles suscitent actuellement peut s'expliquer d'une part grâce à la baisse du coût de mise en œuvre des SAB mais aussi grâce à l'amélioration des performances que ces systèmes permettent d'obtenir. Ainsi, plusieurs mises en œuvre à grande échelle de systèmes biométriques ont déjà vu le jour, notamment aux Jeux Olympiques de 1996 où un SAB contrôlait l'accès des 65000 spectateurs et à Disney World où l'identité des possesseurs de passe saisonnier est vérifiée grâce à leurs empreintes digitales. Notons aussi le projet du gouvernement jamaïcain d'utiliser les empreintes digitales lors des élections et celui des sécurités sociales espagnole et sud-africaine de mettre en place un SAB pour vérifier l'identité des assurés.

D'une manière générale, l'évolution des modes de consommation de la société avec, par exemple, la forte augmentation des transactions et achats en ligne et la généralisation de l'électronique personnelle, a considérablement accru l'intérêt suscité par la biométrie.

1.2 Définition de la biométrie

Dans le dictionnaire *Le Petit Robert édition 2001*, la biométrie est définie comme suit :

La science qui étudie, à l'aide des mathématiques, les variations biologiques à l'intérieur d'un groupe déterminé.

Cette définition s'applique naturellement aux caractéristiques "physiologiques" et peut être étendue aux caractéristiques "comportementales".

De cette définition générale peuvent être facilement déduits les principes et les moyens de fonctionnement d'un système d'authentification biométrique : il s'agit d'étudier les variations de certaines caractéristiques biologiques au sein d'un groupe, puis de déterminer, parmi celles jugées les plus pertinentes, les techniques permettant de distinguer les individus entre eux sur la base des caractéristiques retenues.

Une définition plus technique de l'authentification biométrique peut se trouver dans [Braghin, 1998], pour qui elle rassemble l'ensemble des procédés automatiques d'identification et de vérification d'identité basés sur des caractéristiques physiologiques et/ou comportementales. Dans la suite de ce document, on emploiera le terme biométrie pour

parler de l'authentification biométrique. Cette définition implique l'utilisation systématique de systèmes automatiques pour prendre une décision. Un SAB doit être capable de mesurer la caractéristique biométrique de l'utilisateur puis de la comparer au modèle qui lui est associé avant de prendre une décision.

1.3 Discussion sur la caractéristique biométrique et le SAB induit

Nous avons vu que la caractéristique biométrique peut être de deux types : physiologique ou comportementale. Celle-ci implique la plupart des propriétés importantes du SAB. Il nous semble donc fondamental de la décrire.

Les caractéristiques physiologiques correspondent à des attributs anthropométriques tels que la morphologie de l'iris ou de la rétine, la forme de la main et les empreintes digitales. Par définition, elles sont difficiles à modifier volontairement ou involontairement par l'utilisateur et n'évoluent que lentement et sensiblement avec le vieillissement de l'individu. Les caractéristiques comportementales correspondent à des attributs acquis par l'individu tels que la parole ou l'écriture. Ils dépendent fortement de son milieu socio-culturel, régional, professionnel, etc. et évoluent selon les cas plus ou moins rapidement avec son vieillissement. En outre, ils dépendent d'éléments comme l'état émotionnel et pathologique et sont généralement facilement modifiables par l'individu lui-même.

Les deux points ci-dessous permettent d'apprécier les différences de principes fondamentales entre les deux familles d'attributs pour la réalisation d'un SAB, la première regroupant les systèmes fonctionnant sur une caractéristique physiologique (système P) et l'autre regroupant ceux fonctionnant sur une caractéristique comportementale (système C) :

1. Alors que, pour le système P, la vérification est basée sur l'observation d'un attribut quasi-déterministe, elle est pour le système C basée sur l'observation d'un processus aléatoire dont les paramètres évoluent avec le temps.
2. Alors que pour le système P, la vérification est basée sur une caractéristique intrinsèquement ² liée à l'utilisateur, elle est pour le système C basée sur un attribut facilement transformable.

Ces deux éléments peuvent selon les points de vue (celui de l'utilisateur ou celui du concepteur) et le contexte (l'environnement socio-culturel de l'application, les utilisateurs visés) être considérés autant comme des avantages que comme des inconvénients.

²Nous entendons par intrinsèquement liée à l'utilisateur une caractéristique qu'il ne peut pas modifier sans porter durablement atteinte à son intégrité physique.

Si les propriétés des attributs physiologiques conduisent généralement à des systèmes biométriques plus robustes et précis que ceux basés sur une caractéristique comportementale, on constate que dans ce premier cas, l'obtention du signal biométrique est souvent plus coûteuse et techniquement difficile.

Contrairement aux caractéristiques comportementales qui peuvent être facilement modifiées par un utilisateur mal-intentionné ou par un client ne souhaitant pas être identifié, le caractère invariant et quasi unique pour chaque individu de certaines caractéristiques physiologiques (ex : les empreintes digitales, la rétine) en font des candidates privilégiées lorsque l'on souhaite réaliser un SAB performant (précis et robuste). Ces avantages sont cependant autant d'inconvénients lorsque l'on considère l'acceptabilité du système par l'utilisateur. En effet, un individu peut facilement considérer comme insupportable qu'autrui puisse disposer d'une mesure caractéristique de son identité, aussi fiable et sur laquelle il n'a aucun contrôle. La possibilité de décider d'être ou non identifié par le système est pour l'utilisateur une garantie supplémentaire de non atteinte à sa liberté personnelle. Un système d'identification à but commercial et utilisant une caractéristique intrinsèquement propre à l'individu conduit à un changement radical du rapport client/fournisseur. En effet, comment considérer une organisation qui dispose d'une mesure si fiable de notre personne, au moment par exemple de résilier un contrat ? De plus, un tel système comporte des risques de dérives, comme la création d'un système monolithique d'information sur chaque individu car il permet de relier différentes sources jusqu'alors (volontairement ou non) tenues distinctes par le client.

Les différences fondamentales entre les deux types de caractéristiques biométriques nous ont conduits à reconsidérer la classification des technologies d'identification biométrique présentée par exemple dans [Wayman, 2000a]. Dans cet article, la classification des SAB est basée non pas sur la caractéristique biométrique supportant le système mais sur la mesure de cette caractéristique (ou signal biométrique). Celle-ci comportant inévitablement certaines imprécisions liées au comportement de l'individu, l'auteur finit par ne considérer qu'une seule famille de caractéristiques dont les éléments ont à la fois des composantes physiologiques et comportementales. Même s'il est en fait extrêmement difficile de séparer l'aspect purement comportemental d'un signal biométrique de celui purement physiologique, nous considérons que la différence fondamentale entre ces deux types d'attributs est telle qu'il peut être trompeur, voire dangereux de la faire disparaître dans la définition des caractéristiques d'un SAB. Aussi, nous préférons classer un système non par rapport au signal biométrique mais par rapport à la caractéristique qu'il cherche à mesurer.

1.4 Propriétés de la caractéristique biométrique

D'une manière générale, les propriétés souhaitées pour la caractéristique biométrique étudiée sont :

- *la robustesse* : la caractéristique doit être la plus stable possible au cours du temps et la plus difficilement altérable par le contexte d'utilisation,
- *la distinctibilité* : la caractéristique doit être la plus fortement dépendante de l'utilisateur,
- *l'accessibilité* : elle doit être facilement et efficacement mesurable par un capteur,
- *l'acceptabilité* : elle ne doit pas être perçue comme intrusive par l'utilisateur. Cette propriété relativement subjective dépend du contexte culturel voir politique dans lequel le SAB est mis en œuvre.
- *la disponibilité* : pour chaque utilisateur, une quantité suffisante de mesure de la caractéristique doit être simplement disponible.

Une classification des SAB se déduit naturellement et simplement des caractéristiques ergonomiques, de fiabilité, de robustesse ainsi que des coûts d'installation et de fonctionnement des applications visées.

1.5 Classification des systèmes biométriques

On distingue deux phases dans l'utilisation d'un SAB (cf. figure 1.1) : la phase d'apprentissage et la phase opérationnelle. La première est la phase durant laquelle une suite de vecteurs de paramètres est extraite du signal biométrique puis envoyée au module de modélisation. Celui-ci attribue à partir de ces observations une référence caractéristique (notée Γ_X dans la suite du document) propre à chaque client (X) dans l'espace de représentation. La seconde est la phase durant laquelle la suite de vecteurs de paramètres, extraite selon le même procédé que lors de la phase d'apprentissage, permet le calcul du score de décision conduisant à l'acceptation ou au rejet de l'identité proclamée.

D'après [Wayman, 2000b], une application d'authentification biométrique peut être décrite grâce aux sept modes de fonctionnement suivants. Ceux-ci permettent de caractériser totalement une application d'identification biométrique :

1. coopératif / non-coopératif :

La distinction entre ces deux modes de fonctionnement se comprend à partir du comportement de l'individu cherchant à tromper le système (imposteur). Dans un mode coopératif il collabore avec le système, par exemple pour usurper une identité, dans un mode non-coopératif il s'oppose au système, par exemple pour ne pas être reconnu.

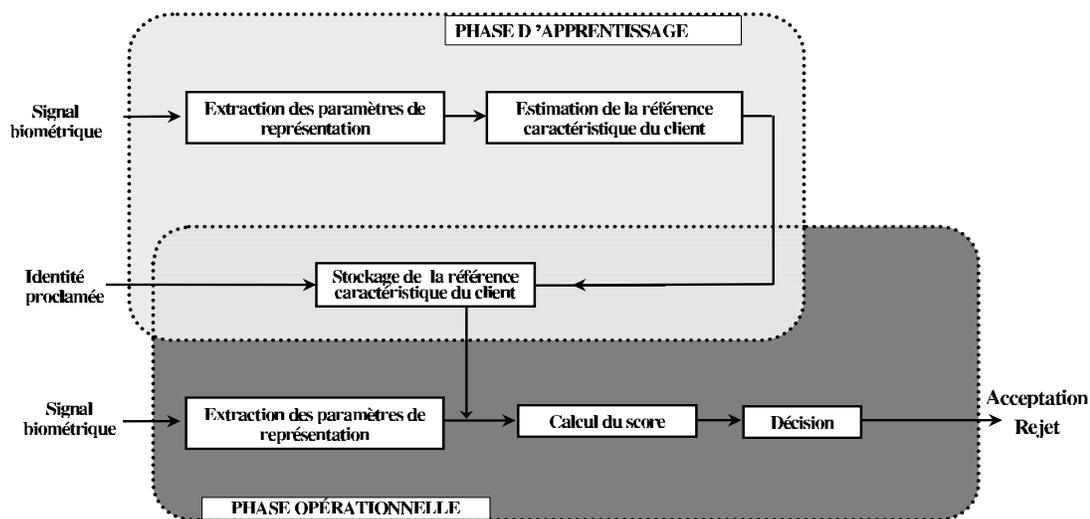


FIG. 1.1 – Description des phases de fonctionnement d'un système d'authentification biométrique

2. manifeste / caché :

Si l'utilisateur sait qu'il est sujet à un test d'authentification biométrique, le mode de fonctionnement est manifeste, sinon il est caché.

3. stable / instable :

Ce mode caractérise le comportement de l'utilisateur en fonction de son niveau de familiarité avec le système. Il permet d'expliquer l'évolution des performances en fonction de l'accoutumance des utilisateurs. Si malgré des accès fréquents et espacés dans le temps les performances restent constantes : le mode de fonctionnement est stable, sinon il est instable.

4. supervisé / non supervisé :

Si l'utilisateur est guidé étape par étape lors de l'utilisation du système, le mode de fonctionnement est supervisé, dans le cas contraire il est non-supervisé. Le mode supervisé concourt à garantir un mode d'utilisation stable.

5. environnement standard / non-standard :

Il s'agit de savoir dans quelle mesure l'environnement de test est homogène à l'environnement d'apprentissage ainsi que de connaître le rapport signal à bruit (RSB) du signal dont on dispose. Dans la suite du document, un environnement

standard est tel que les conditions de test et d'apprentissage sont contrôlées et homogènes avec en plus un signal biométrique peu bruité. Un environnement non-standard correspond à des environnements de test et d'apprentissage dont les caractéristiques sont non contrôlables, donc souvent fortement variables et dans lequel le signal peut être fortement bruité.

6. public / privé :

Ce mode caractérise le type d'utilisateurs. S'il s'agit de clients de la société proposant le système, l'utilisation est dans un mode public. S'il s'agit d'employés d'une entreprise, l'utilisation est dans un mode privé.

7. ouvert / fermé :

Si la référence biométrique caractéristique du client peut être divulguée et utilisée par d'autres systèmes : le système est ouvert, sinon il est fermé.

Chacun de ces modes de fonctionnement conduit à des simplifications ou des complications pour la conception d'un SAB. On ne peut *a priori* prédire les résultats obtenus par un système fonctionnant sous certaines conditions à partir de ceux obtenus sous d'autres contraintes.

1.6 Avantages / inconvénients de la VAL

Les caractéristiques biométriques les plus courantes sont : les empreintes digitales, l'iris, la rétine, la voix, l'écriture, la géométrie de la main et le visage. La figure 1.2 appelée analyse de Zephyr et disponible sur le site www.biometricgroup.com présente une évaluation des avantages et inconvénients applicatifs de différents SAB basés sur différents attributs biométriques courants.

Sur cette figure, les quatre critères d'évaluation des systèmes sont :

- la précision (accuracy) : mesure objective des performances obtenues par le système,
- le coût (cost) : mesure objective du coût de mise en œuvre et de fonctionnement du système,
- l'ergonomie (effort) : mesure subjective du nombre et de la difficulté des démarches à suivre par l'utilisateur lors de son utilisation,
- le caractère intrusif (intrusiveness) : mesure subjective, dépendant fortement de la culture de l'utilisateur et permettant d'évaluer sa perception du système.

Selon les critères d'évaluation choisis, aucune caractéristique biométrique n'est globalement supérieure à une autre.

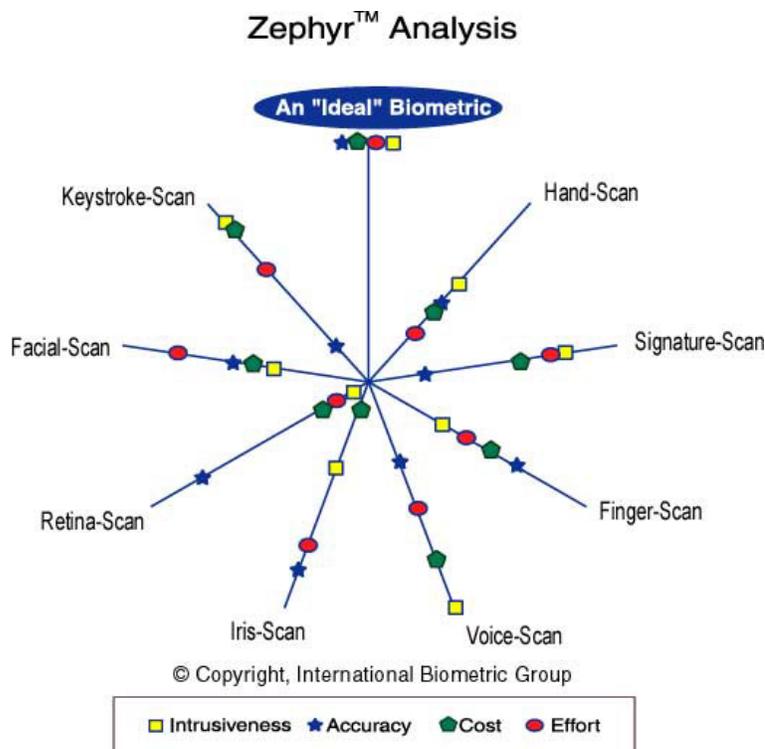


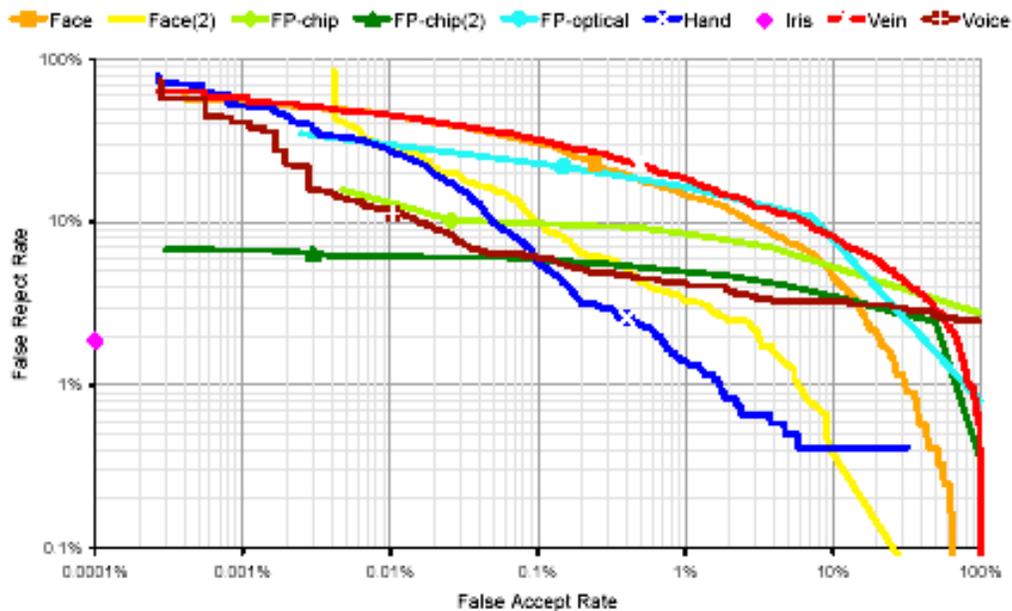
FIG. 1.2 – Comparaison des avantages et inconvénients applicatifs de différentes technologies biométriques d'après www.biometricgroup.com

L'identification vocale est considérée par la plupart des utilisateurs comme l'une des technologies biométriques la plus naturelle : elle n'est pas intrusive, n'exige aucun contact physique avec le système. De plus, elle correspond à une manière usuelle pour chaque individu de reconnaître l'un de ses proches. Un autre avantage de la VAL est qu'elle est souvent la technologie la plus adaptée pour de nombreuses applications telles que la sécurisation de transaction bancaire téléphonique, etc.

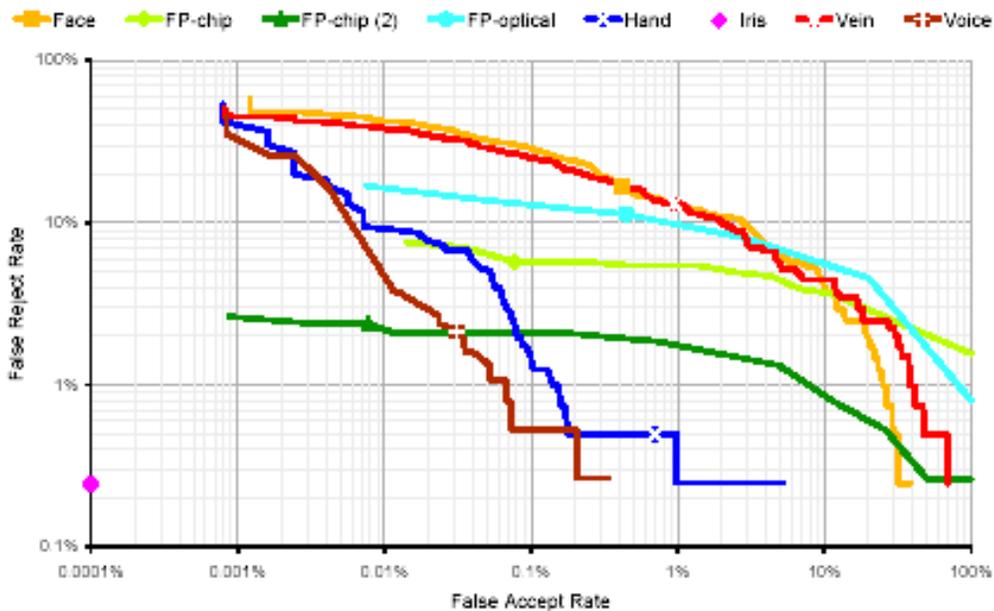
Pour un type d'application donné, les figures (1.3,a) et (1.3,b) extraites de [Mansfield et al., 2001] permettent de comparer objectivement les performances de différentes caractéristiques biométriques. Sur ces figures sont tracés les résultats obtenus par des systèmes basés sur les empreintes digitales (trois systèmes sont présentés : FP-chip, FP-chip(2) et FP-optical), la voix (Voice), le visage (deux systèmes : Face et Face(2)), la main (Hand), l'iris (Iris) et les vaisseaux sanguins (Vein). Ces résultats ont été obtenus avec des données enregistrées et testées selon le protocole défini dans [BWD, 2001]. Ils sont présentés sous forme de courbes sur lesquelles les probabilités de fausse acceptation (acceptation d'un imposteur) sont tracées en fonction des probabilités de faux rejet (rejet d'un vrai client) dans une échelle logarithmique.

La figure (1.3,a) représente les performances obtenues par ces différentes technologies lorsque le système prend une décision après un unique accès alors que la figure (1.3,b) représente les performances obtenues lorsque la décision est basée sur le score le plus favorable parmi trois. Dans les conditions de ce test, la VAL est l'une des technologies obtenant les meilleurs résultats et l'on peut observer que la perte de performances liée à la variabilité des conditions d'accès peut efficacement et simplement être corrigée en multipliant les sessions d'acquisition pour prendre la décision.

Même si la vérification du locuteur n'est actuellement, à notre connaissance, pas utilisée pour des applications grand public, elle apparaît comme une technologie attrayante du point de vue de son ergonomie, de son coût de mise en œuvre et de ses performances. Pour conclure sur sa situation parmi les technologies de l'authentification biométrique, ajoutons qu'actuellement un effort particulier est porté sur le développement de plateformes fusionnant la décision de plusieurs sous-systèmes basés sur différentes caractéristiques biométriques et dans lesquelles la VAL s'intègre de manière quasi systématique [Choudhury et al., 1998], [Duc et al., 1997], [Verlinde et Acheroy, 2000] et [Poh et Korczak, 2001].



(a) Décision basée sur un unique accès.



(b) Décision basée sur le meilleur score calculé parmi trois accès.

FIG. 1.3 – Comparaison des performances de différentes technologies de vérification biométrique d'identité d'après [Mansfield et al., 2001]

Chapitre 2

La VAL comme technologie du traitement automatique de la parole

Dans le chapitre précédent, nous avons inscrit la vérification automatique du locuteur dans le contexte de l'authentification biométrique. Ceci nous a permis de la situer par rapport aux technologies concurrentes pour la conception d'applications d'authentification automatique. Dans ce chapitre, nous présentons la VAL dans le contexte des technologies du traitement automatique de la parole (TAL) et plus particulièrement celui de la caractérisation automatique du locuteur (CAL). La motivation principale pour situer la VAL dans ce contexte est liée aux fortes interactions de réalisation entre les systèmes de VAL et ceux des différentes applications du TAL.

2.1 La caractérisation automatique du locuteur

La caractérisation automatique du locuteur est une branche du traitement automatique de la parole, sous-ensemble des disciplines de l'interaction Homme-Machine regroupant l'ensemble des technologies ayant pour objet l'étude du signal de parole. Les principales disciplines de la TAL sont : la synthèse vocale, le codage de la parole, la reconnaissance de la parole et la caractérisation automatique du locuteur.

Le but de la CAL est d'extraire du signal de parole toutes sortes d'informations relatives à l'individu l'ayant prononcé. La nature des caractéristiques recherchées est très variée (identité, pathologie, origine géographique, etc.) et dépend du type d'applications visées. La reconnaissance automatique du locuteur est une discipline dérivée de la CAL. Elle regroupe l'ensemble des applications dont le but est de déterminer l'identité d'une personne à partir de sa voix.

En RAL, on considère habituellement que le signal de parole est source de trois

informations différentes : la première correspond à l'information linguistique, la seconde à l'information sur le support de transmission du signal, et la troisième à l'information propre à caractériser le locuteur. Toutes trois sont à l'origine de la variabilité du signal de parole et la dernière est celle que les systèmes de RAL vont chercher à isoler et à interpréter pour identifier le locuteur.

Classification des systèmes de RAL

Une classification essentielle des systèmes de RAL est basée sur la dépendance au texte. Pour définir d'autres critères de classification, on pourra par exemple fixer le type d'environnement opérationnel ou le matériel d'acquisition du signal de parole. Cependant, ces caractéristiques de conception sont générales à tout système d'authentification biométrique et nous détaillons ici uniquement les différents types de dépendance au texte des systèmes de RAL.

On distingue les systèmes dépendants du texte des systèmes indépendants du texte. En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur un signal de parole dont le contenu linguistique (mot de passe, phrase, code) est connu du système. Les différentes configurations possibles sont [Eagles, 1995] :

- systèmes à messages fixes : la vérification de l'identité du client est alors précédée d'une étape de reconnaissance de la parole. La personne doit selon les cas prononcer un message qu'elle aura préalablement choisi [Jacob et al., 2000] ou qui sera imposé par le système [Higgins et al., 1991]. Dans le dernier cas, le message peut être différent à chaque nouvel accès. La motivation de cette approche est de se protéger des imposteurs disposant d'un enregistrement de la voix d'un client.
- systèmes à unités segmentales fixes : lors d'un accès au système, le client doit prononcer un signal de parole contenant soit une séquence de mots (ex : chiffres [Melin, 1998]), soit des traits phonétiques connus du système [Lastrucci et al., 1994].
- systèmes indépendant du texte. Le système de reconnaissance n'impose alors aucune contrainte sur le contenu linguistique du signal de parole.

L'information apportée par la connaissance *a priori* du contenu linguistique permet généralement d'augmenter les performances. Cependant cette amélioration est obtenue au prix d'une baisse de l'ergonomie du système : l'utilisateur doit alors soit se souvenir d'un mot de passe, soit être en mesure de lire un message prompté.

2.2 Description des applications de la RAL

Après l'identification automatique et la vérification automatique du locuteur, applications liées à la préoccupation originelle de la RAL (*i.e.* l'authentification automatique humaine), sont apparues de nouveaux objectif plus complexes et liés à l'extraction

et à la gestion d'informations dans des bases de données multimédia. Ainsi, les deux principaux domaines d'applications de la RAL sont actuellement liés d'une part à la sécurisation par authentification du client et d'autre part à l'indexation pour l'aide à la navigation dans les documents audio.

La suite de cette section présente brièvement les principales tâches de la RAL autres que la vérification du locuteur et donne pour chacune d'elle un schéma de leur principe de fonctionnement ainsi que des exemples d'applications.

Identification automatique du locuteur

L'Identification Automatique du Locuteur (IAL) consiste à déterminer, parmi une population de N locuteurs connus, celui qui a prononcé un message donné. Lors d'un accès à un système d'IAL, le signal de parole fourni à l'entrée du système est comparé à la référence caractéristique de chacun des locuteurs connus et l'identité retournée est celle dont la référence est la plus proche du signal de test. Le signal est la seule entrée du système d'IAL. Dans un système d'identification du locuteur sur un ensemble fermé, le locuteur est supposé être l'un des N locuteurs du système. Dans un système d'identification du locuteur sur un ensemble ouvert, le système peut décider qu'aucune des N identités connues n'est celle du locuteur. Il doit pour cela disposer d'un modèle de rejet.

Les performances obtenues par les systèmes d'IAL sont directement liées au nombre N de locuteurs du système. La figure 2.1 représente un schéma illustrant le fonctionnement d'un système d'IAL.

Applications :

En ensemble fermé, les applications d'un système d'IAL sont peu nombreuses. L'identification automatique du locuteur peut cependant être utilisée de manière très efficace pour simplifier l'accès des membres d'une population d'individus à des données ou à des services personnalisés (mise en place automatique de paramètres d'utilisation, etc.). En ensemble ouvert, les applications de l'IAL sont essentiellement liées à des problèmes de sécurisation comme la protection de l'accès à des sites sensibles.

Détection automatique des locuteurs

La détection automatique des locuteurs consiste à déterminer la présence ou non d'un locuteur donné sur un enregistrement audio. Si l'on fait l'hypothèse que le signal sonore est mono-locuteur, cette tâche est équivalente à la vérification automatique du locuteur. Comme dans le cas de la VAL, l'identité recherchée ainsi que le signal de parole constituent les deux entrées des systèmes de détection automatique du locuteur (*c.f.* figure 2.2).

Sur les figures 2.2, 2.3 et 2.4, la présence de différents locuteurs sur le signal d'entrée est symbolisée par différents niveaux de gris.

Applications :

Comme pour la VAL, les applications actuelles de la détection des locuteurs sont principalement liées à la sécurisation de service (authentification de l'interlocuteur dans une communication téléphonique pour la validation de transactions, etc.). Cependant, d'autres applications du domaine de l'indexation de documents multimédia telles que la recherche d'information dans un document audio numérisé ou la navigation dans les données sonores sont actuellement étudiées. Ainsi, les futurs moteurs de recherche permettront sans doute de retrouver des fichiers audio contenant la voix d'un individu donné.

Suivi de locuteurs - Indexation par locuteurs

Le suivi de locuteurs consiste à segmenter un signal de parole pour indiquer les instants et durées de prise de parole d'un locuteur cible. L'identité de ce locuteur ainsi qu'un signal de parole multi-locuteurs sont les entrées du système.

L'indexation par locuteur consiste à déterminer le nombre de locuteurs présents sur un document audio ainsi que leurs intervalles de prise de parole. Les systèmes d'indexation du locuteur fonctionnent sans aucun *a priori* sur l'identité des locuteurs présents sur l'enregistrement sonore.

Les schémas du fonctionnement d'un système d'indexation du locuteur et de suivi de locuteurs sont présentés respectivement sur les figures 2.3 et 2.4.

Applications :

Le domaine d'application de ces deux tâches est principalement le traitement de bases de données audio. Citons par exemple [Delacourt, 2000] : la recherche d'information dans des séquences d'émissions télévisées ou radiophoniques, l'estimation du temps de parole de chaque intervenant lors d'un débat, la recherche des interventions d'une personne dans des archives audio, etc..

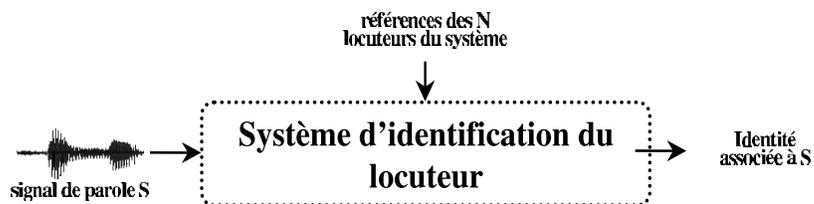


FIG. 2.1 – Schéma d'un système d'identification du locuteur

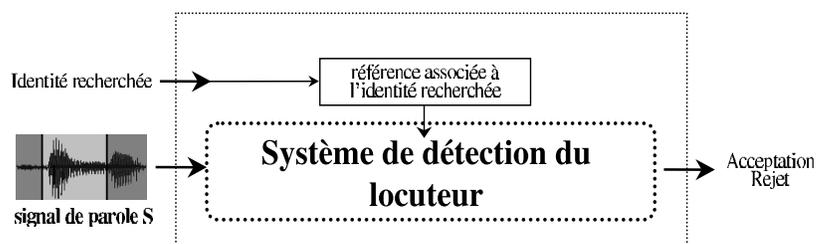


FIG. 2.2 – Schéma d'un système de détection de locuteur

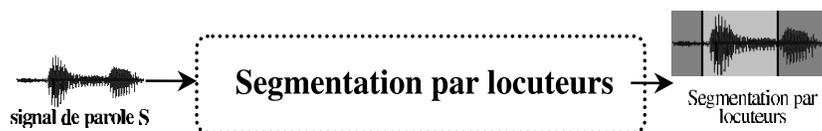


FIG. 2.3 – Schéma d'un système d'indexation par locuteur

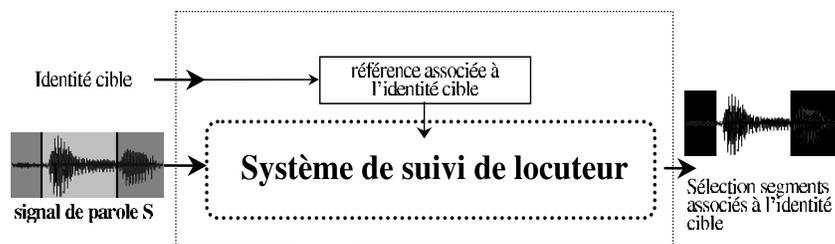


FIG. 2.4 – Schéma d'un système de suivi de locuteurs

2.3 Caractéristiques de la vérification automatique du locuteur

La VAL est une technologie de sécurisation bien adaptée aux applications exigeant de la simplicité et de la transparence vis-à-vis de l'utilisateur. De plus, elle apparaît actuellement comme le moyen le plus adéquat pour sécuriser les transactions ou échanges de données sur le réseau téléphonique et sur Internet. Dans cette section, nous proposons de décrire les principes de fonctionnement ainsi que les difficultés de mise en œuvre propres à la VAL ainsi que de définir la famille d'applications visée par la collaboration avec CP8.

Généralités sur les systèmes de VAL

Le signal de parole est fortement corrélé avec certains attributs physiologiques et comportementaux du locuteur. Leurs influences se retrouvent dans la densité spectrale de puissance du signal à court terme (caractéristique du conduit vocal, de la source glottique et du timbre de la voix) ainsi que dans d'autres éléments supra-segmentaux tels que la prosodie et d'autres facteurs para linguistiques. La vérification du locuteur repose sur tous ces attributs qui définissent la variabilité inter-locuteur, *i.e* l'ensemble des dissemblances observables sur un signal de parole et permettant de caractériser un individu.

Difficultés rencontrées en VAL

La VAL repose sur la variabilité inter-locuteur du signal de parole. Cependant, le signal tel qu'il est transmis au système de vérification contient trois autres sources de variabilité difficilement dissociables de cette première :

1. la variabilité intra-locuteur, liée aux changements et à l'évolution de la voix d'un locuteur,
2. la variabilité introduite par les modifications des caractéristiques du matériel d'acquisition et de transmission du signal de parole,
3. la variabilité due au changement d'environnement et par exemple à la présence de bruit dans le signal de parole.

L'obtention d'une représentation robuste à la variabilité intra-locuteur est cruciale pour la réalisation d'un système de VAL. Les deux autres facteurs de variabilité sont des causes majeures de la dégradation des performances et l'une des principales gageures des systèmes de VAL est de réduire autant que possible leurs influences.

Variabilité intra-locuteur

Deux signaux de parole produits par la même personne prononçant le même énoncé sont toujours différents. Le signal de parole est par nature un processus aléatoire et il est impossible pour un individu de produire plusieurs fois exactement le même signal. De plus, la voix d'un locuteur est fortement influencée par son état physique et émotionnel. Tous les changements observables dans la voix d'une même personne définissent la variabilité intra-locuteur. Ils ont pour origine différents facteurs qui peuvent être :

- occasionnels : l'état pathologique (rhume, mal de dents, etc.), émotionnel (stress, angoisse) ou de fatigue d'une personne modifie temporairement sa voix.
- à moyen terme : le comportement d'un individu se modifie lorsqu'il s'habitue au système, ainsi s'il s'applique et parle distinctement lors des premiers accès, sa parole évolue et devient plus naturelle au cours des accès suivants. L'évolution à moyen terme de la voix des utilisateurs est un des effets du mode de fonctionnement stable/instable des systèmes d'authentification biométrique définis à la section 1.5.
- à long terme : la voix évolue avec l'âge.

Même si les variations à court terme sont très préjudiciables aux systèmes de vérification du locuteur, des travaux ont permis de mettre en évidence une dégradation croissante des performances à mesure de l'augmentation du temps entre la phase d'apprentissage et celle de test. La mise à jour régulière par un apprentissage incrémental des modèles de locuteurs avec de nouvelles données permet d'assurer une amélioration des performances obtenues [Fredouille, 2000].

Variabilité matérielle d'acquisition et du canal de transmission

De nombreux travaux expérimentaux comme par exemple [Vuuren, 1996] et surtout l'étude des performances sous différentes configurations de test des nombreux systèmes présentés aux évaluations NIST [Martin et Przybocki, 2000] et [Reynolds, 1996] ont permis de mettre en évidence l'influence des variations de la chaîne d'acquisition entre la phase d'apprentissage et de celle test. Ces variations sont à l'origine d'une forte dégradation des performances obtenues. Ceci est particulièrement vrai lorsque le signal est fortement perturbé par le canal de transmission, comme c'est le cas avec la parole téléphonique (limitation de la bande utile et distorsions dues au combiné et au canal de transmission). Ainsi, la compensation et l'annulation des effets de cette variabilité est l'un des enjeux fondamentaux de la recherche actuelle, d'autant plus qu'il s'agit maintenant de mettre en œuvre des systèmes de VAL opérationnels sur les téléphones cellulaires utilisant de la parole bas débit, avec en outre de nombreuses pertes de données.

2.4 Applications visées par la collaboration avec CP8

La famille d'applications visée par la collaboration entre METISS et CP8 concerne principalement la sécurisation de transactions ou d'achats via le réseau téléphonique ou Internet. On cherche à utiliser la carte à microprocesseur non seulement pour stocker la référence caractéristique du client, mais aussi pour prendre en charge une partie des traitements nécessaires aux différents processus d'un système de VAL.

La figure 2.5 décrit la mise en œuvre d'un procédé de sécurisation d'achat en ligne. Sur cette figure, la carte à puce est reliée à un micro-ordinateur faisant office de terminal hôte. Celui-ci assure, lors d'une transaction en ligne, l'acquisition du signal de parole ainsi que l'extraction des paramètres de représentation. Ceux-ci sont ensuite transmis à la carte à puce qui calcule par ses propres moyens le score de décision associé. Selon la configuration du système, celui-ci peut être transmis à un serveur distant pour l'étape de décision ou rester sur la carte qui assure elle-même cette fonction. Suivant la décision du système, la transaction est ensuite autorisée ou refusée.

Idéalement et pour atteindre un niveau de sécurité maximale, la référence caractéristique du locuteur ne doit en aucun cas quitter la carte à puce. Il est donc impératif que celle-ci soit non seulement capable de la stocker mais aussi d'effectuer le calcul du score de décision. Cette contrainte a fixé l'un des enjeux principaux de ce travail : réduire au maximum la quantité de mémoire et du MIPS (Million d'Instructions par Seconde) nécessaires au processus de vérification du locuteur. En effet, on peut actuellement disposer de cartes à puces dont les capacités mémoire sont de l'ordre de 256 kilo-octets de ROM et de 8 kilo-octets de RAM et dont la puissance de calcul varie entre 3 et 30 MIPS [Urien, 2001], ceci rendant impossible, comme on le voit au chapitre 9, une implémentation sur ces cartes des systèmes de VAL suivant les caractéristiques directes de l'état-de-l'art.

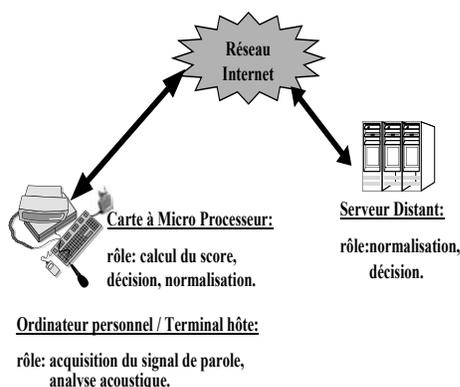


FIG. 2.5 – Exemple d'application visée par la collaboration entre METISS et SCHLUMBERGER-CP8

Chapitre 3

Évaluation des systèmes biométriques

L'évaluation de nos systèmes de vérification du locuteur dans les meilleures conditions possibles a été une préoccupation forte tout au long de ce travail. Nous présentons maintenant quelques éléments généraux relatifs à l'évaluation des systèmes d'authentification biométrique.

Évaluer un SAB est long, coûteux et difficile. Cela requiert une base de données d'échantillons de signaux biométriques soigneusement étiquetée, issus d'une large population d'individus et enregistrés dans un environnement similaire à celui de l'application visée. Seule l'évaluation permet de mettre en évidence certains artefacts et l'origine des dysfonctionnements éventuels d'un SAB. Elle intervient à tous les niveaux de sa réalisation. Lors du développement d'une part, car elle permet d'évaluer de nouveaux algorithmes et/ou paramètres de réglage du système. Lors de la mise en service d'autre part, car elle permet d'estimer dans des conditions opérationnelles réelles le coût de fonctionnement ainsi que d'autres facteurs tels que l'acceptabilité du système.

3.1 Aspects de l'évaluation d'un système d'authentification biométrique

On peut considérer trois aspects différents dans l'évaluation d'un SAB :

1. le premier correspond à l'évaluation de la technologie. Il consiste à tester les algorithmes sur une base de données standardisée qui permet d'estimer les performances du système. Cette étape permet par exemple d'évaluer une nouvelle technique de représentation ou de modélisation du signal de parole.
2. la seconde correspond à l'évaluation d'un scénario. Elle permet de déterminer les performances globales d'un système dans une configuration donnée. Les tests sont alors effectués sur un système complet et dans un environnement qui modélise le monde réel. Cette étape permet par exemple de déterminer certains réglages à l'apprentissage ou lors de l'accès qui pourront permettre d'accroître fortement les performances.
3. la troisième et dernière étape est celle de l'évaluation de la phase opérationnelle proprement dite. Elle permet de déterminer les performances du système dans un contexte applicatif donné et sur une population correspondant à celle de l'application.

Les phases 2 et 3 demandant une logistique plus lourde et plus coûteuse, nous avons limité nos évaluations à la première phase en travaillant sur les données fournies par NIST (National Institute of Standard and Technology) pour leurs campagnes d'évaluation annuelles.

Bien que la plupart des éléments présentés dans la suite de cette section s'appliquent à tous systèmes biométriques, nous nous sommes spécifiquement placés dans le cadre de l'évaluation des systèmes de vérification du locuteur.

Les trois sections suivantes de ce chapitre cherchent respectivement à répondre à ces trois questions relatives à l'évaluation des SAB :

1. Que cherche-t-on ? Où l'on décrit les motivations et les critères objectifs de l'évaluation d'un système SAB. Cette première partie présente les outils les plus fréquemment utilisés par l'estimation des performances d'un SAB.
2. De quoi avons nous besoin ? Où l'on décrit les propriétés indispensables du corpus nécessaire à l'évaluation.
3. Ce que l'on peut conclure ? Où l'on discute des limitations des résultats de l'évaluation.

La quatrième et dernière section porte sur les évaluations NIST. Elle discute de l'intérêt de ces évaluations et décrit la base de données que nous avons utilisée pour valider nos algorithmes.

3.2 Critères d'évaluation des systèmes biométriques

On peut trouver dans [Wayman, 2000c] la liste des différents critères usuels d'évaluation des systèmes d'authentification biométrique. Voici les principaux :

- la probabilité de fausse alarme P_{fa} qui correspond à la probabilité d'accepter à tort l'identité proclamée,
- la probabilité de non-détection ou de faux rejet P_{fr} qui correspond à la probabilité de rejeter à tort l'identité proclamée,
- les échecs à l'apprentissage,
- les échecs lors de l'acquisition des données biométriques lors de la phase de test,
- d'autres critères plus subjectifs tels que la perception et l'acceptabilité du système.

Dans le cadre de ce travail, nous avons uniquement considéré P_{fr} et P_{fa} . Notons que d'une manière générale, les deux types d'échec (à l'apprentissage et lors de l'acquisition) sont particulièrement déterminants dans le cas de systèmes biométriques destinés à des utilisateurs non-coopératifs.

La décision d'un système de VAL, étant donné un signal de parole Y et une identité proclamée X , est basée sur la comparaison d'un score $S_X(\mathbf{Y})$ à un seuil θ fixé *a priori* en fonction des caractéristiques applicatives du système. \mathbf{Y} correspond à la suite de vecteurs extraite du signal de parole Y . La règle de décision pour l'acceptation ou le rejet de l'identité proclamée est telle que décrite ci-dessous :

$$S_X(\mathbf{Y}) = \log D_X(\mathbf{Y}) - \log D_{\bar{X}}(\mathbf{Y}) \begin{matrix} \stackrel{H_X}{\geq} \\ \stackrel{H_{\bar{X}}}{<} \end{matrix} \theta \quad (3.1)$$

avec $D_X(\mathbf{Y})$, score associé par le système à l'acceptation de l'identité proclamée et $D_{\bar{X}}(\mathbf{Y})$ score associé à son rejet.

À partir de tous les accès du corpus d'évaluation, le système fournit une suite de scores client S_X correspondant aux tests pour lesquels l'identité proclamée est bien celle du locuteur et une suite de scores imposteur $S_{\bar{X}}$ pour lesquels l'identité proclamée n'est pas celle du locuteur. Pour chaque valeur de $\theta \in [\min(S_X, S_{\bar{X}}), \max(S_X, S_{\bar{X}})]$ on estime $(P_{fa}(\theta), P_{fr}(\theta))$. L'ensemble des couples obtenus permet de tracer la courbe DET (Data Error Trade-off, [Martin et al., 1997]) variante de la courbe ROC (Receiver Operating Characteristic) qui constitue l'élément de base le plus couramment utilisé pour l'évaluation globale des systèmes de VAL.

Sur un ensemble d'évaluation comportant N_{cl} accès client \mathbf{Y}_X (pour lesquels l'identité proclamée est bien celle du locuteur) et N_i accès imposteur (pour lesquels l'identité

proclamée n'est pas celle du locuteur), les estimations de P_{fr} et P_{fa} sont simplement obtenues comme :

$$P_{fr}(\theta) = \sum_{k=1}^{N_{cl}} \frac{\mathbb{1}(S_X(\mathbf{Y}_X(k)) < \theta)}{N_{cl}}$$

$$P_{fa}(\theta) = \sum_{k=1}^{N_i} \frac{\mathbb{1}(S_X(\mathbf{Y}_{\bar{X}}(k)) \geq \theta)}{N_i}$$

avec

$$\mathbb{1}(a) = \begin{cases} 1 & \text{si } a \text{ est vrai,} \\ 0 & \text{si } a \text{ est faux.} \end{cases}$$

et \mathbf{Y}_X ($\mathbf{Y}_{\bar{X}}$) suite de vecteurs acoustiques extraits d'un signal de parole prononcé par le locuteur X (\bar{X}).

Différentes valeurs du couple $(P_{fa}(\theta), P_{fr}(\theta))$ sont considérées pour évaluer de manière plus fine les performances d'un système de RAL. Habituellement, on en distingue au moins trois :

1. le couple $(P_{fa}(\theta_{Det}), P_{fr}(\theta_{Det}))$ obtenu en minimisant le coût de fonctionnement C_{Det} associé au seuil θ_{Det} minimal à partir des coûts C_{fa} et C_{fr} , respectivement associés à l'erreur de fausse acceptation et à celle de faux rejet. On a :

$$C_{Det} = \min_{\theta_{Det}} \{C_{fa} \cdot P_{\bar{X}} \cdot P_{fa}(\theta_{Det}) + C_{fr} \cdot P_X \cdot P_{fr}(\theta_{Det})\} \quad (3.2)$$

où $P_{\bar{X}}$ désigne la probabilité *a priori* d'accès imposteur et P_X celle des accès client.

Alors que P_X , $P_{\bar{X}}$, C_{fa} et C_{fr} sont fixés *a priori* et dépendent directement des conditions de fonctionnement de l'application visée, P_{fr} et P_{fa} sont liées aux performances intrinsèques du système.

2. le couple $(P_{fa}(\theta_{HTER}), P_{fr}(\theta_{HTER}))$ obtenu en minimisant $HTER$, la moyenne des probabilités des deux erreurs.

$$HTER = \min_{\theta_{HTER}} \left\{ \frac{1}{2} (P_{fa}(\theta_{HTER}) + P_{fr}(\theta_{HTER})) \right\} \quad (3.3)$$

On parle alors du point de fonctionnement à l' $HTER$ (Half Total Error Rate). Ce point de fonctionnement revient à considérer les probabilités $P_{\bar{X}}$ et P_X et les coûts C_{fa} et C_{fr} comme respectivement égaux dans le cas précédent.

3. le couple $(P_{fa}(\theta_{EER}), P_{fr}(\theta_{EER}))$ qui correspond à l'intersection de la courbe DET avec la première bissectrice.

$$EER = P \text{ avec } P = P_{fa} = P_{fr}$$

Ce point n'a pas d'interprétation particulière en terme de coût mais est souvent utilisé comme première évaluation des performances du système car il est directement obtenu à partir de la lecture de la courbe DET et permet d'obtenir une indication grossière des performances du système.

3.3 Corpus d'évaluation

Les différentes sources de variabilité du signal de parole sont propres à l'application visée et les conclusions induites d'une évaluation n'ont de sens que dans un contexte applicatif similaire. Ceci est notamment une limitation des applications criminalistiques où l'on ne connaît pas *a priori* les conditions d'acquisition du signal et où l'on ne peut estimer précisément les performances des systèmes. De plus, tous les locuteurs ne sont pas égaux devant un système biométrique : alors que certains auront un taux remarquable de bonne acceptation et ne seront que difficilement victimes d'une imposture, d'autres, seront plus sensibles et rejetés à tort par le système. On peut trouver dans [Doddington et al., 1996] une classification des locuteurs :

- mouton (sheep) : locuteur pour lequel le taux de bonne acceptation est remarquablement élevé,
- chèvre (goat) : locuteur pour lequel le taux de faux rejet : est exceptionnellement élevé,
- agneau (lamb) : locuteur pour lequel le taux de fausse acceptation : est exceptionnellement élevé,
- loup (wolf) : locuteur performant pour commettre une imposture.

Notons que les chèvres et les agneaux sont en général peu nombreux dans une population de locuteurs mais que leur détection et leur rejet du système permet d'améliorer significativement les performances d'un système de VAL [Doddington et al., 1996].

On peut observer de grandes variations dans les caractéristiques des distributions des scores client et imposteur pour différents locuteurs. Intuitivement, cela implique que le corpus d'évaluation contienne le plus grand nombre possible de locuteurs afin de réduire au maximum les effets de ces variabilités sur l'estimation des performances.

Pour un locuteur donné, les scores client et imposteur ne sont pas stationnaires et leur distribution varie à court et à long terme. Cette non-stationarité implique de disposer d'enregistrements issus de sessions différentes et espacées dans le temps. Ceci afin d'estimer au mieux les effets de la variabilité intra-locuteur sur le score pour chacun des locuteurs du corpus d'évaluation.

L'expérience présentée ci-dessous a pour but d'évaluer l'influence de l'évolution du signal de parole sur les performances d'un système de VAL. Dans cette expérience, on a

tout d'abord estimé pour chacun des 506 locuteurs de sexe féminin de l'évaluation NIST de 2001 un modèle de mélange de gaussiennes de 128 composantes, comme décrit au chapitre 7. Puis, nous avons effectué l'ensemble des tests du plan d'évaluation associés à chacun de ces locuteurs :

1. en utilisant les fichiers de test fournis par NIST,
2. en générant les observations de test par la méthode de Monte-Carlo (MC) à partir du modèle associé à l'identité du locuteur du fichier de test original.

Dans les deux cas, la quantité des données de test est équivalente. Les courbes (1, ■) et (2, ■) de la figure 3.1 représentent les performances associées à chacun de ces deux tests.

Les résultats obtenus dans le cas réel sont très inférieurs à ceux obtenus avec des données artificielles. Ainsi, alors que le taux de fausse acceptation diminue, celui de faux rejet n'augmente que sensiblement sur la courbe 2, alors qu'il croît fortement sur la courbe 1. Le taux relativement bas de faux rejet sur la courbe 2 s'explique entre autre par la neutralisation des variations intra-locuteur du signal de parole grâce à la génération lors d'un accès client de vecteurs acoustiques suivant la loi du modèle du locuteur. Les taux d'erreurs permettant de tracer la courbe 2 peuvent être considérées comme une borne inférieure de ceux atteignables par la représentation des locuteurs utilisée. Trop de facteurs tels que ceux décrits dans la section 2.3 ne sont pas modélisables et seule une évaluation permet d'en estimer l'influence sur les performances.

3.4 Contraintes et limitations des évaluations

La confiance que l'on peut avoir dans l'estimation de P_{fa} et P_{fr} est directement liée à la taille du corpus. L'incertitude sur l'estimation, à partir de N accès, des performances d'un système biométrique est difficilement estimable principalement à cause de la complexité de la densité de probabilité du score de décision. Un principe assez répandu pour fixer la taille d'un corpus est connu sous le nom de *règle de Doddington* :

Un ensemble d'évaluation n'est fiable que si le système se trompe au moins 30 fois.

Cette règle, justifiée dans [Porter, 1997] sous l'hypothèse que les N scores de l'évaluation sont indépendants et qu'ils suivent une loi de Bernoulli, permet d'obtenir des bornes inférieures et supérieures dans l'intervalle de confiance de l'estimation de P_{fa} et P_{fr} : à partir de 30 erreurs, on est sûr à 90% que la vraie probabilité d'erreur est dans un intervalle autour de 30% de la valeur mesurée.

Finalement, il n'existe pas de moyen formel pour déterminer la taille d'un corpus et deux manières de procéder co-existent :

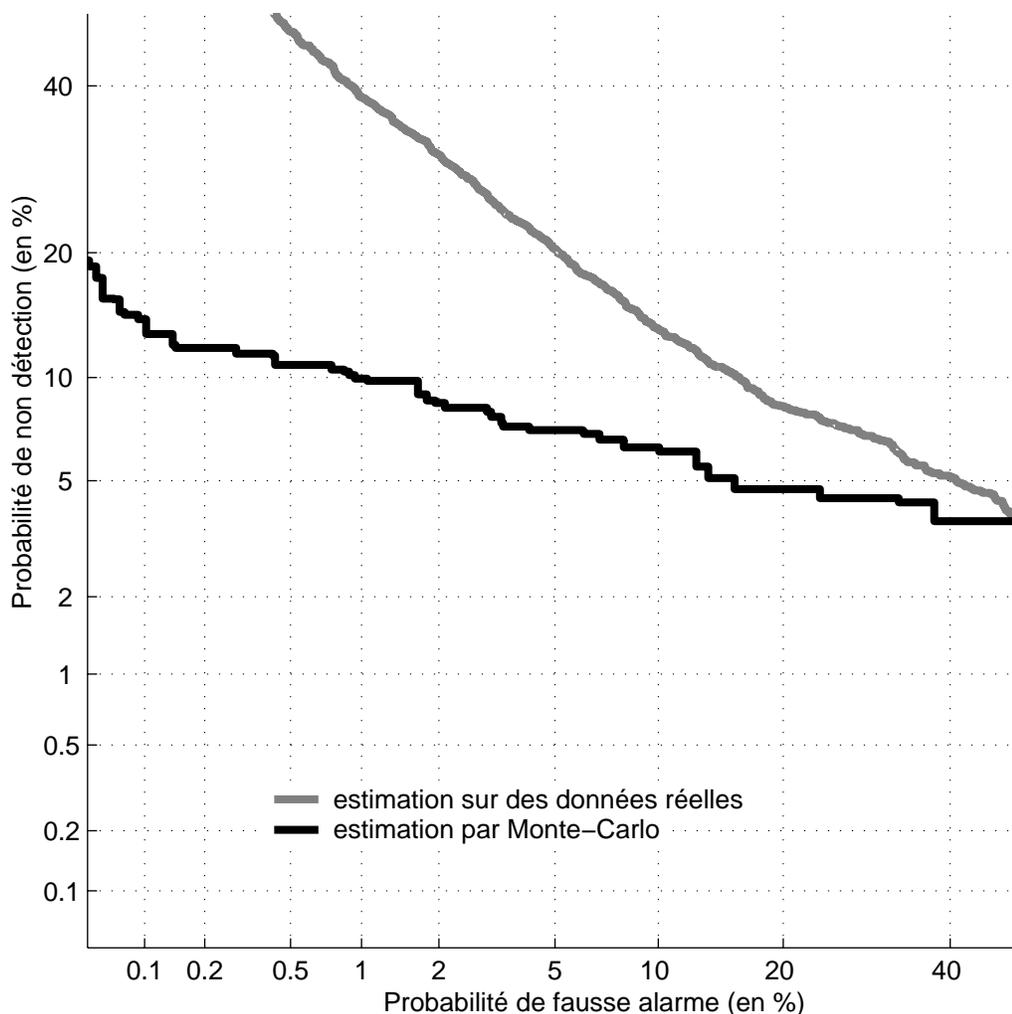


FIG. 3.1 – Performances sur des données réelles et sur des données générées d'un système de vérification du locuteur

- suivre la règle de Doddington dont l'inconvénient est de ne pas permettre de fixer la taille de la base *a priori*;
- plus prosaïquement, rassembler selon le budget disponible le maximum de données.

3.5 Les évaluations NIST

Dans le domaine des technologies de l'identification du locuteur, les évaluations annuelles proposées par le National Institute of Standard and Technology (NIST) sont un élément moteur pour la recherche (nouvelles idées, collaborations, autres.). Depuis la première campagne en 1996, ces évaluations ont permis d'une part de soulever de nou-

velles problématiques dans les technologies de la caractérisation du locuteur et d'autre part de mettre en évidence certaines des difficultés rencontrées par les systèmes. De plus, grâce à notre participation, nous avons pu disposer d'un large corpus pour évaluer nos algorithmes. Enfin, ces évaluations nous permettent de comparer de manière objective les performances de nos systèmes de VAL par rapport à celles de nombreux centres de recherches.

Le protocole expérimental de ces campagnes est précisément décrit dans [Martin et Przybocki, 2000]. Pour l'apprentissage, nous dirons simplement ici que nous disposons d'environ 1000 locuteurs (500 hommes et 500 femmes) et que, pour chacun d'eux, nous disposons d'environ 2 minutes de parole issues d'une seule session. Pour le test, on dispose d'environ 5000 fichiers d'accès dont la taille varie de 3 à 60 secondes. Ceux-ci sont ensuite utilisés de manière totalement indépendante et au total on effectue environ 50000 tests. Les accès client sont obtenus à partir d'enregistrements obtenus sur une durée de plusieurs mois. Tous les segments de parole correspondent à des conversations téléphoniques enregistrées de manière transparente sur une population d'étudiants nord-américains parlant l'anglais américain. Ils sont étiquetés, entre autres, suivant l'identité du locuteur, le type de combiné utilisé lors de l'enregistrement et leur durée.

Les résultats présentés dans ce travail correspondent à l'évaluation des performances des systèmes sur la population de locuteurs féminins suivant la condition primaire [Martin et Przybocki, 2000] des évaluations NIST adaptées à ce sous-ensemble. Sous cette configuration de test, ce corpus comprend environ 31000 tests avec 28000 accès imposteur et 3000 accès client.

Les évaluations NIST nous ont permis de calibrer et de valider les différents systèmes que nous avons proposés. Elles sont un moyen fiable et efficace d'estimer les performances des systèmes de VAL. Leur existence annuelle est un moteur important de la recherche en vérification du locuteur, qui justifie à lui seul le fait que notre équipe ait tenu à y participer depuis 1997.

Deuxième partie

Description d'un système de vérification du locuteur

Cette partie présente le formalisme mathématique sur lequel se base la majorité des systèmes de vérification du locuteur actuels. Il permet d'introduire la plupart des notations qui sont utilisées dans la suite de ce document. Enfin, différentes mises en œuvre pour chacun des modules d'un système de VAL y sont décrites.

Alors que nous avons jusqu'alors abordé la vérification automatique du locuteur d'un point de vue essentiellement applicatif, nous nous intéressons dans cette partie aux descriptions théorique et technique des systèmes de VAL.

Il s'agit ici d'expliquer le principe de fonctionnement et de présenter les principales technologies associées aux modules nécessaires à la mise en œuvre d'une plateforme de VAL. Ceci permet de mettre en évidence certaines des modifications susceptibles, d'une part d'améliorer les performances, et d'autre part d'apporter une baisse significative des ressources nécessaires à leur fonctionnement. À partir de ces différents objectifs, nous avons découpé cette partie en quatre chapitres.

Dans le premier, nous décrivons les éléments de la théorie de la détection statistique dont le formalisme est à la base de la réalisation des systèmes actuels.

Le second concerne les technologies associées aux différents modules d'une plateforme de vérification du locuteur. Nous proposons pour chacun d'eux un aperçu de différentes solutions de mise en œuvre en nous intéressant plus particulièrement à celles de la modélisation et du calcul du score de décision.

Dans le troisième chapitre de cette partie, nous présentons les éléments des systèmes de VAL sur lesquels s'est focalisé notre travail. Nous donnons pour chacun d'eux certaines des motivations qui nous ont conduits à les modifier.

Enfin, dans le quatrième chapitre, nous présentons notre plateforme de VAL de référence dont les modules sujets de notre étude sont tels qu'au début de nos travaux. Celle-ci nous permettra d'évaluer objectivement l'influence sur les performances de nos différentes propositions de mise en œuvre.

Chapitre 4

Formalisme général de la vérification automatique du locuteur

Soit une suite de vecteurs acoustiques \mathbf{Y} extraits d'un signal de parole Y et une identité proclamée ou supposée X , on formalise le problème de la vérification du locuteur en considérant les deux hypothèses H_X et $H_{\bar{X}}$ suivantes :

$$\begin{aligned} H_X & : \text{ l'énoncé a bien été prononcé par } X, \\ H_{\bar{X}} & : \text{ l'énoncé n'a pas été prononcé par } X. \end{aligned}$$

Une autre formulation de $H_{\bar{X}}$ est souvent associée à un locuteur imaginaire \bar{X} censé représenter l'ensemble des alternatives à X dans l'espace des paramètres. On nomme ce locuteur le 'non-client'.

Tout système de décision $\mathcal{D}(\cdot)$ permettant de choisir entre H_X et $H_{\bar{X}}$ partitionne l'espace des paramètres \mathcal{Y} en deux sous-espaces disjoints \mathcal{Y}_X et $\mathcal{Y}_{\bar{X}}$ tels que :

$$\mathcal{D}(\mathbf{Y}) = \begin{cases} H_X & \text{pour } \mathbf{Y} \in \mathcal{Y}_X \\ H_{\bar{X}} & \text{pour } \mathbf{Y} \in \mathcal{Y}_{\bar{X}} \end{cases}$$

et dans lesquels les probabilités de faux rejet (P_{fr}) et de fausse acceptation (P_{fa}) s'écrivent :

$$P_{fr} = \int_{\mathcal{Y}_{\bar{X}}} p_X(y) dy \quad \text{et} \quad P_{fa} = \int_{\mathcal{Y}_X} p_{\bar{X}}(y) dy$$

où p_X et $p_{\bar{X}}$ sont respectivement les densités de probabilité de H_X et $H_{\bar{X}}$ dans l'espace des paramètres acoustiques.

Comme on l'a vu dans le chapitre précédent, la décision d'un système de vérification du locuteur est basée sur la comparaison d'un score $S_X(\mathbf{Y})$ à un seuil θ suivant la règle décrite sur l'équation 3.1 :

$$S_X(\mathbf{Y}) = \log D_X(\mathbf{Y}) - \log D_{\bar{X}}(\mathbf{Y}) \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta$$

Le score de décision $S_X(\mathbf{Y})$, étant donné la suite de vecteurs acoustiques \mathbf{Y} et l'identité proclamée X , est basé sur le calcul de deux mesures de similarité : $D_X(\mathbf{Y})$ et $D_{\bar{X}}(\mathbf{Y})$. Celles-ci correspondent, en général et selon le système de VAL mis en œuvre, au calcul d'une distance, d'une erreur de prédiction ou d'une vraisemblance $D_a(\mathbf{Y})$, où $D_a(\mathbf{Y})$ mesure le degré de concordance des observations \mathbf{Y} avec l'hypothèse H_a .

Dans le cas d'une modélisation probabiliste de H_X et $H_{\bar{X}}$, $S_X(\mathbf{Y})$ correspond à l'estimation du rapport de vraisemblance entre les deux hypothèses H_X et $H_{\bar{X}}$. $D_X(\mathbf{Y})$ et $D_{\bar{X}}(\mathbf{Y})$ sont alors respectivement équivalents à $p_X(\mathbf{Y})$ et $p_{\bar{X}}(\mathbf{Y})$. La règle de décision se réécrit alors :

$$S_X(\mathbf{Y}) = \log p_X(\mathbf{Y}) - \log p_{\bar{X}}(\mathbf{Y}) \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta \quad (4.1)$$

La théorie du test d'hypothèse bayésien permet, dans ce cas, de fixer théoriquement le seuil de décision permettant de minimiser le coût de fonctionnement du système, étant donné les coûts de fausse acceptation et de faux rejet (respectivement C_{fa} et C_{fr}) associés à l'application visée avec :

$$\theta = \log \frac{C_{fa} P_{\bar{X}}}{C_{fr} P_X}$$

on constate que dans la pratique, l'utilisation de ce seuil conduit à une forte dégradation des performances et qu'en outre le θ optimal est dépendant du locuteur. Si bien qu'en pratique, quelque soit la méthode utilisée pour calculer les $D_a(\mathbf{Y})$, le seuil θ est fixé de manière à minimiser la fonction de coût sur un ensemble d'évaluation dont les caractéristiques sont similaires à ceux du corpus de test.

Chapitre 5

Modules d'une plateforme de VAL : techniques de l'état-de-l'art

Nous avons jusque-là considéré les systèmes de VAL comme étant simplement des «boîtes noires» permettant d'associer une décision binaire d'acceptation ou de rejet à partir d'une identité prétendue X et d'un signal de parole Y . Si nous avons vu que cette décision est obtenue par la comparaison à un seuil de la différence entre deux mesures de similarité, l'une associée à l'acceptation, l'autre au rejet de l'identité prétendue, nous n'avons pas considéré les différentes techniques associées à leur calcul, ni l'espace dans lequel elles sont définies. Ces deux points constituent l'objet de ce chapitre. L'espace de définition de la mesure de similarité est défini par le module d'analyse acoustique, leur mode de calcul l'est par le module de modélisation et celui de calcul du score.

Nous présentons ici, pour chacun des modules d'une plateforme de vérification du locuteur, certains des procédés de réalisation publiés dans la littérature. Lors de la conception d'une application nécessitant un système de VAL, le choix de chacun des modules doit être effectué en fonction de contraintes applicatives telles que l'environnement, la puissance de calcul et l'espace de stockage disponibles, etc. Nos propres choix de réalisation ont été déterminés par rapport aux données des différentes évaluations NIST auxquelles nous avons participé : tout d'abord sans contrainte de temps de calcul ni de limitation de l'espace de stockage, puis en y intégrant les contraintes propres à l'application visée par notre collaboration avec CP8.

La figure 5.1 représente une plateforme de vérification du locuteur. On y trouve tous ses éléments essentiels ainsi qu'une schématisation de l'ensemble de leurs interactions avec l'intérieur et l'extérieur du système. Pour chacun d'eux sont indiqués : les ressources logicielles (algorithmiques) avec leurs paramètres, les ressources matérielles (mémoire, CPU, etc.) ainsi que les flux de données entrants et sortants. Nous y retrouvons les modules généraux d'un système d'identification biométrique :

- le module d'extraction des paramètres caractéristiques. Dans le cas de la RAL, il s'agit du module d'analyse acoustique,
- le module de modélisation qui permet de calculer la référence caractéristique du locuteur,
- le module de calcul de la mesure de similarité,
- le module de normalisation, qui n'apparaissait pas dans la figure 1.1 mais qui est couramment utilisé en RAL et dont nous décrivons les buts et principes dans la suite de ce chapitre,
- le module de décision.

5.1 Analyse acoustique

Le module d'analyse acoustique décrit sur le schéma de la figure 5.2 a pour but d'extraire la représentation du signal de parole Y dans l'espace des paramètres acoustiques \mathcal{Y} . Elle permet d'obtenir la suite $\mathbf{Y} = \{y_1, \dots, y_N\}$ de vecteurs acoustiques qui sont, selon le mode de fonctionnement du système (en phase d'apprentissage ou en phase opérationnelle), utilisés pour estimer la référence caractéristique du locuteur ou calculer le score de décision. L'analyse acoustique permet de quantifier sous la forme d'une représentation multidimensionnelle toutes les grandeurs contenues dans Y et susceptibles de nous renseigner sur l'identité du locuteur.

Les performances des systèmes de VAL dépendent en grande partie de la qualité de la représentation acoustique choisie. Les propriétés souhaitées pour ces vecteurs acoustiques sont :

- qu'ils contiennent une information la plus caractéristique possible du locuteur,
- qu'ils soient robustes aux bruits et aux distorsions de canal qui perturbent le signal de parole lors de son acquisition,
- qu'ils soient rapidement calculables.

L'obtention du vecteur de représentation acoustique peut se décomposer en trois étapes.

La première appelée *étape de pré-traitement* a pour rôle principal d'augmenter la robustesse des paramètres qui sont calculés à l'étape suivante.

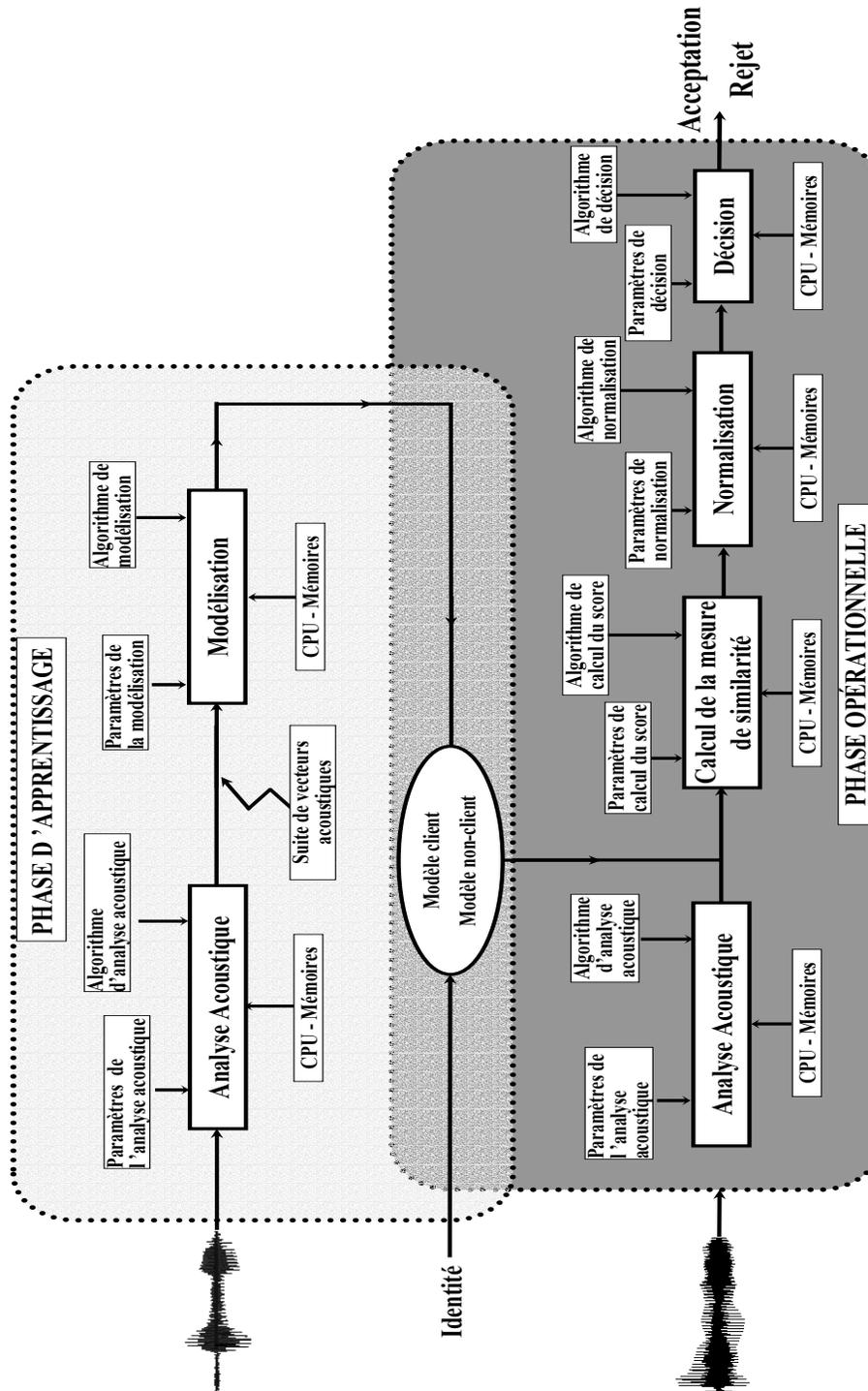


FIG. 5.1 – Description d'un système de vérification du locuteur

La seconde où l'on calcule explicitement les vecteurs acoustiques de la représentation et qui peut être décomposée en trois parties :

1. la première consiste à extraire du signal de parole un vecteur dont les coefficients correspondent souvent aux paramètres d'un modèle de production du signal de parole. Celui-ci est généralement issu d'une modélisation Source-Canal [Calliope, 1989].
2. la seconde consiste à obtenir une représentation plus robuste et ayant de meilleures propriétés statistiques que le vecteur brut (indépendance, propriété discriminante, etc.).
3. la troisième consiste à ajouter une information dynamique au vecteur acoustique de base. Dans ce but, on procède généralement en concaténant à ce dernier un vecteur dont les coefficients sont censés caractériser certains éléments de la dynamique, propre au locuteur, du signal de parole.

À l'issue de cette seconde étape, on dispose des vecteurs acoustiques homogènes à ceux qui sont utilisés par le module de modélisation ou de calcul du score.

La troisième et dernière étape de l'extraction des paramètres acoustiques, que l'on peut appeler *étape de post-traitement*, consiste en un ensemble de traitements effectués dans l'espace des paramètres. Ceux-ci visent par exemple à rendre la représentation plus indépendante du canal de transmission et plus robuste à l'environnement d'acquisition.

Les paragraphes suivants donnent une description détaillée de ces trois étapes ainsi que des méthodes couramment utilisées pour leur mise en œuvre.

Étape de *pré-traitement*

La première étape de la paramétrisation acoustique consiste à découper, à cadence régulière, le signal de parole en fenêtres d'analyse appelées trames. C'est à partir de ces trames que sont extraits les vecteurs de paramètres acoustiques. La taille de la fenêtre est fixée *a priori* selon une approximation de la durée moyenne de stationnarité du signal de parole $\approx 20ms$. À ce niveau, différents traitements tels que le filtrage (de pré-emphase et/ou de limitation de la bande-passante), la sélection de trames et/ou l'amélioration du rapport signal à bruit du signal de parole peuvent être effectués.

Calcul du vecteur de représentation : recherche de l'information utile

Le signal de parole a une structure complexe qui reflète la grande richesse de la production vocale humaine. On modélise habituellement ce signal en considérant une composante *source* associée au signal glottique et une composante *conduit vocal* caractéristique des conduits buccal et nasal du locuteur. L'analyse la plus fréquemment utilisée en RAL vise à extraire du signal de parole ces deux composantes caractérisant

physiquement le système de production du locuteur, *i.e* d'une part le filtre associé à son conduit vocal et d'autre part l'excitation à l'origine de la production sonore.

Il existe de nombreuses techniques permettant de caractériser localement l'enveloppe spectrale du conduit vocal du locuteur, parmi lesquelles : le codage prédictif linéaire LPC [Calliope, 1989] et l'estimation spectrale.

Les paramètres cherchant à caractériser la source du signal de parole sont moins largement utilisés en RAL. Ceci est en partie dû à la difficulté de leur estimation et à leur grande variabilité intra-locuteur. Parmi eux citons : l'énergie, la fréquence fondamentale f_0 et le taux de voisement.

D'autres types d'analyse telle que celle fournie par la décomposition sur une base d'ondelettes ont aussi été expérimentés en RAL [Kadambe, 1994], mais leur utilisation reste peu répandue et leurs performances, pour l'instant décevantes.

Tous ces vecteurs de paramètres sont obtenus par analyse à court terme du signal de parole, mais certains systèmes ont cherché à utiliser des caractéristiques supra-segmentales de la parole telles que :

Les **paramètres prosodiques**, qui correspondent à des grandeurs telles que le débit (vitesse d'élocution), la mélodie (modulation de la fréquence fondamentale) et l'accentuation (variations énergétiques)...

Les **caractéristiques phonétiques** ont été utilisées avec profit par des systèmes de VAL. Comme dans [Sturin et al., 2001]. Certains phonèmes sont plus caractéristiques du locuteur que d'autres et ce type de système cherche à en tirer partie.

Les **caractéristiques linguistiques** ont aussi été utilisées pour identifier le locuteur. Ainsi [Doddington, 2001] présente un système qui identifie puis utilise des groupes de 2 à 3 mots fréquemment prononcés par le locuteur pour le caractériser.

Les paramètres supra-segmentaux et ceux liés à la source du signal de parole (énergie, fréquence fondamentale) étaient jusqu'alors peu utilisés car leur estimation est complexe et qu'ils apparaissent souvent comme difficilement exploitables. En effet, ils sont particulièrement sensibles à de nombreux facteurs de variabilité comme celui de l'environnement d'acquisition. Cependant de récentes expériences ont montré qu'ils pouvaient, dans certaines applications, être extrêmement utiles [Sonmez et al., 1998] , [Carey et al., 1996] notamment lorsque la quantité de données d'apprentissage est importante ou, comme dans [Weber et al., 2001] où est rapportée une expérience utilisant avantageusement les caractéristiques prosodiques.

Calcul du vecteur de représentation : extraction des paramètres de représentation

La seconde étape de l'extraction des vecteurs acoustiques a pour but d'augmenter leur robustesse et leur caractère discriminant. Ses deux fonctions principales sont :

1. Définir une projection des vecteurs acoustiques bruts dans un espace où les propriétés statistiques de leurs coefficients seront meilleures et qui facilitera l'étape de *post-traitement*.
2. Ajouter au vecteur brut une information dynamique.

On compte actuellement de nombreuses transformations du vecteur acoustique brut dont voici les principales :

Le cepstre :

Le calcul du cepstre est la technique la plus fréquemment utilisée. Elle consiste en un calcul successif du logarithme de chacun des coefficients des vecteurs obtenus lors de l'étape précédente puis de la Transformée en Cosinus Inverse (TCI) sur l'ensemble de coefficients. Ces deux transformations permettent de définir l'espace cepstral [Furui, 1981] dont l'utilisation est extrêmement courante dans de nombreux domaines du traitement automatique de la parole. Les principaux atouts des coefficients cepstraux sont leur relative décorrélation et l'efficacité des traitements de compensation de canal qu'ils permettent (cf. l'étape de post-traitement). On peut aussi montrer que, selon le postulat d'un modèle de production source + canal, ils sont bien adaptés pour extraire du signal les caractéristiques du conduit vocal du locuteur. Comme on le voit au paragraphe suivant, cette représentation facilite l'une des fonctions de l'étape de post-traitement : obtenir des paramètres qui soient le plus insensibles possible aux variations du canal de transmission.

Analyse en Composante Principale (PCA) et Analyse Discriminante (LDA) :

Le principe général de ces méthodes est d'estimer à partir de nombreuses observations de vecteurs acoustiques une matrice de filtrage dépendante [Magrin-Chagnollet et Durou, 2000] ou indépendante du locuteur [Jang et al., 1999] puis de l'appliquer aux vecteurs acoustiques bruts. Dans certaines représentations, la matrice de filtrage permet d'ajouter au vecteur acoustique une information dynamique relative à son contexte. Une comparaison des performances obtenues par ces différentes transformations peut se trouver dans [Magrin-Chagnollet et al., 2000].

L'ajout de l'information dynamique consiste essentiellement à concaténer aux vecteurs acoustiques un ou plusieurs vecteurs caractéristiques de son contexte. Le but est d'introduire dans la représentation des éléments caractérisant certaines structures temporelles propres au signal de parole comme par exemple les phénomènes de coarticulation. Les principales stratégies pour l'ajout d'une information dynamique sont :

L'ajout du vecteur d'estimation de la **vitesse instantanée** [Soong et Rosenberg, 1986] (coefficients Δ) et/ou de l'**accélération instantanée** (coefficients $\Delta\Delta$) de chacun des coefficients du vecteur.

La **concaténation** directe de plusieurs trames successives passées et futures de parole.

Le **filtrage vectoriel** du vecteur et de son contexte. Le vecteur y_t est alors remplacé par : $y_t = B(y)$ avec $B(y)$ vecteur de dimension p' avec $p' > p$ [Magrin-Chagnolleau, 1997].

Étape de *post-traitement*

Cette troisième et dernière étape de l'extraction des vecteurs acoustiques consiste en une série de traitements dans l'espace des paramètres. Sa principale fonction est de rendre la représentation acoustique robuste aux variations du canal de transmission et de l'environnement d'acquisition. Ses deux mises en œuvre les plus utilisées sont : l'égalisation de canal et la sélection de trames.

Égalisation de canal :

La **soustraction cepstrale** (CMS pour Cepstral Mean Subtraction) qui permet de compenser les distorsions linéaires introduites lors de l'acquisition du signal [Furui, 1981]. On estime la moyenne cepstrale \bar{y}^* sur une fenêtre glissante, puis on la retranche à chacun des vecteurs. Si l'on considère que le canal de transmission est constant à l'échelle de la fenêtre, sa contribution l'est aussi et la CMS permet de l'annuler. L'hypothèse de cette méthode est que la dégradation des performances causée par la perte de l'information caractéristique du locuteur liée au retrait de \bar{y}^* est inférieure au gain apporté par la neutralisation des distorsions du canal de transmission. Sur les données des évaluations NIST, la CMS permet d'améliorer les performances lorsque le canal de transmission du signal d'apprentissage est différent du signal de test mais dégrade les performances dans le cas contraire.

Le **filtrage RASTA** [Hermansky et Morgan, 1994] (**R**elative **cepST**rAl). Cette technique cherche à supprimer les composantes spectrales dont les variations sont en-dehors, *i.e.* plus rapides ou plus lentes que celles caractéristiques du signal de parole. Diverses expériences décrites dans la

littérature montrent des performances très proches entre la CMS glissante et le filtrage RASTA.

Le **Feature Warping** [Pelecanos, 2001] qui consiste à gaussianiser chaque composante du vecteur d'observation de manière à la considérer comme issue d'un processus gaussien $g(\mathbf{y}_p)$ de moyenne nulle et de variance unité. À partir du contexte de l'observation, on estime pour chaque coefficient y_p du vecteur y un histogramme d'où l'on déduit la probabilité $p(\mathbf{y}_p < y_p)$ que le coefficient p d'un vecteur \mathbf{y} soit supérieur à y_p . Puis on remplace y_p par la valeur de l'observation gaussienne associée à cette probabilité. On cherche y_p^{fw} tel que :

$$p(\mathbf{y}_p < y_p) = \int_{-\infty}^{y_p^{fw}} g(\mathbf{y}_p) d\mathbf{y}_p$$

avec, $g(\cdot)$, loi normale centrée réduite, puis l'on remplace y_p par y_p^{fw} . Le **Feature Warping** permet d'obtenir une amélioration notable des performances. On peut noter qu'une alternative à cette technique consiste à ajouter à la CMS glissante une normalisation de chaque coefficient du vecteur par sa variance. Cette stratégie beaucoup plus efficace d'un point de vue calculatoire permet dans certains cas d'obtenir des performances comparables [Meigner et al., 2002].

La sélection de trames :

Certaines trames, correspondant par exemple à un silence, sont plus dépendantes de l'environnement d'acquisition que du locuteur et causent une nette dégradation des performances. On peut principalement considérer deux classes de méthodes pour sélectionner les trames dont les propriétés de discrimination sont les plus robustes aux perturbations du signal de parole.

Dans les méthodes supervisées, on utilise un *a priori* sur le signal pour sélectionner les trames. La sélection des trames à partir d'une modélisation parole-bruit bi-gaussienne de l'énergie est un exemple de mise en œuvre d'une technique de sélection supervisée.

Dans les méthodes non-supervisées, on ne dispose d'aucun *a priori*. Elles peuvent par exemple consister à estimer directement sur les données la distributions bi-gaussienne de l'énergie. Le retrait se fait ensuite en fonction du taux de trames que l'on souhaite rejeter.

Une méthode de suppression de trame de type non-supervisé a été utilisée avec succès par le consortium ELISA lors des évaluation NIST 2001. On peut trouver sa description dans : [Magrin-Chagnolleau et al., 2001].

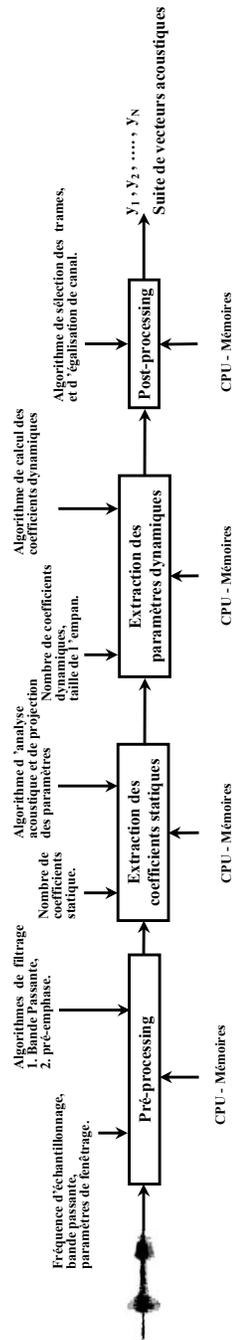


FIG. 5.2 – Principe de la paramétrisation

5.2 Modélisation des locuteurs

La plupart des paramètres utilisés en vérification automatique du locuteur le sont aussi en reconnaissance automatique de la parole où l'on cherche pourtant à exclure toute information extra-linguistique de la représentation acoustique. À partir de la même suite de vecteurs extraits du signal de parole, le module de modélisation doit permettre d'obtenir la référence caractéristique représentant les hypothèses H_X et $H_{\bar{X}}$ dans l'espace des paramètres. Les principales propriétés souhaitées pour ces références sont :

- qu'elles nécessitent le plus faible espace de stockage possible,
- que leur estimation soit la moins complexe possible,
- qu'elles permettent une décision rapide en phase de test,
- qu'elles soient les plus robustes possible aux variations intra-locuteur,
- qu'elles permettent la meilleure séparation des locuteurs entre eux.

La représentation de $H_{\bar{X}}$ pourra être dépendante ou indépendante du locuteur, selon la stratégie utilisée par le module de calcul du score du système.

Nous allons maintenant passer en revue différentes techniques permettant de représenter le locuteur dans l'espace des paramètres acoustiques. L'interprétation du score sur lequel se base la décision des systèmes permet de distinguer deux grandes familles pour la représentation des locuteurs.

La première famille que nous avons considérée correspond aux méthodes basées sur le calcul d'une distance euclidienne entre les vecteurs extraits du signal d'accès et d'autres représentant le locuteur. Le mode d'obtention de ces vecteurs permet de distinguer deux méthodes de ce type.

La seconde famille de méthodes correspond à celles basées sur une représentation statistique du locuteur dans l'espace des paramètres. Le travail effectué durant cette thèse se situe dans ce second cadre et les méthodes concernées sont plus précisément décrites dans la suite de cette section.

Méthodes basées sur une distance euclidienne

Les deux types de méthodes basées sur le calcul d'une distance euclidienne sont :

1. celles pour lesquelles la référence caractéristique du locuteur correspond à une suite de vecteurs de l'espace acoustique. Lors de la phase de test, le score de décision est basé sur le calcul de la distance entre ces vecteurs et ceux extraits du signal d'accès. Parmi les technologies appartenant à cette famille de méthode, citons l'algorithme DTW dont on peut trouver une mise en œuvre dans [Yu et al., 1995] et la quantification vectorielle (VQ) [Soong et al., 1987].

2. celles pour lesquelles la référence caractéristique du locuteur correspond à un modèle de production des vecteurs acoustiques du locuteur. Le score de décision est alors basé sur le calcul de l'erreur entre les vecteurs prédits par le modèle et ceux extraits du signal de test. Ces méthodes cherchent à extraire une dépendance temporelle structurelle de la suite des observations extraites du signal de parole alors que dans la plupart des modélisations (HMM mise à part), toute dépendance temporelle est perdue dans une représentation ergodique des vecteurs acoustiques et n'intervient qu'indirectement au niveau des coefficients Δ et $\Delta\Delta$. On peut distinguer trois méthodes prédictives : la première est basée sur les modèles Auto-Régressifs-Vectoriels [Magrin-Chagnolleau, 1997], la seconde sur les réseaux de neurones prédictifs [Montacie et Le Floch, 1992] et la troisième sur les réseaux de neurones auto-associatifs [Lastrucci et al., 1994], [Kishore et Yegnanarayana, 2000].

Les approches statistiques :

Dans les méthodes statistiques, la phase d'apprentissage consiste à estimer pour chaque locuteur les densités de probabilité p_X et $p_{\bar{X}}$ correspondant aux hypothèses H_X et $H_{\bar{X}}$.

Dans la suite du document, les lois p_{H_X} et $p_{H_{\bar{X}}}$ sont respectivement appelées 'modèle du locuteur' et 'modèle du non-locuteur' (ou encore 'modèle du monde').

L'approche statistique est celle qui semble la mieux adaptée à la modélisation du signal de parole. La solution généralement employée par les systèmes de VAL actuels consiste à utiliser des fonctions de densité paramétriques dont l'avantage majeur est la réduction, dans une certaine mesure, de la taille mémoire nécessaire à la modélisation du locuteur. Il s'agit alors, à partir de l'énoncé d'apprentissage, d'estimer les paramètres λ_X qui permettront la modélisation la plus efficace du locuteur dans l'espace des paramètres. Ce choix introduit deux niveaux d'approximation : le premier porte sur la loi paramétrique choisie et sur sa capacité à être une bonne représentation de la vraie loi des deux hypothèses et le deuxième porte sur l'estimation des paramètres de cette loi.

Trois densités sont principalement utilisées : les Modèles Statistiques du Second Ordre (MSSO), les Modèles de Mélange de Gaussiennes (GMM) et les Modèles de Markov Cachés (HMM).

Les Modèles Statistiques du Second Ordre

Les modèles statistiques du second ordre correspondent à l'utilisation d'une gaussienne $p_X(y)$ pour représenter le locuteur X . Ils ont d'abord été utilisés en identification automatique du locuteur dans [Gish, 1990] et [Bimbot et al., 1995] puis en VAL dans [Zilca, 2001].

Pour un vecteur y_t de dimension d , $p_X(y_t)$ s'écrit :

$$p_X(y_t) = \frac{1}{2\pi^{d/2} |\Sigma_X|^{1/2}} \exp\left(-\frac{1}{2}(y_t - m_X)^T \Sigma_X^{-1} (y_t - m_X)\right)$$

avec $|\cdot|$ déterminant et $(\cdot)^T$ matrice transposée de (\cdot) .

Ainsi, la phase d'entraînement correspond à l'estimation de Σ_X et de m_x pour chacun des locuteurs à partir des N vecteurs acoustiques $\mathbf{Y}_X = \{y\}_{t=1}^N$ extraits de l'énoncé d'apprentissage. À l'issue de cette étape, on dispose de $p_X(y)$ estimation de la vraie loi des observations \mathbf{Y}_X .

L'avantage de cette modélisation est double : d'une part car la référence caractéristique associée est compacte (un vecteur de dimension d et une matrice symétrique comptant $\frac{1}{2} \cdot d \times (d + 1)$ coefficients distincts), d'autre part car le score de décision associé à cette densité nécessite une faible charge CPU. Son inconvénient majeur vient de sa relative simplicité et de la modélisation grossière du locuteur dans l'espace des paramètres. Ainsi, les performances obtenues sont bien inférieures à celles d'autres modélisations comme par exemple celles obtenues par l'utilisation des modèles de mélange de gaussiennes.

Les Modèles de Mélange de Gaussiennes (GMM)

Les modèles de mélange de gaussiennes sont actuellement les densités utilisées par les systèmes de vérification du locuteur les plus performants. La densité de probabilité d'un vecteur y de dimension d suivant une loi de mélange de K gaussiennes s'écrit :

$$p(x) = \sum_{k=1}^K p_k \cdot g_k(x)$$

$$\text{avec } \sum_{k=1}^K p_k = 1 \quad \text{et} \quad g_k(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \cdot |\Sigma|^{1/2}} e^{-\frac{1}{2} \cdot (x - m_k)^T \Sigma^{-1} \cdot (x - m_k)}$$

Les paramètres de ce type de modèle sont : $\lambda = \{(w_k, m_k, \Sigma_k)\}_{1 \leq k \leq K}$ où p_k , m_k et Σ_k sont respectivement le poids, la moyenne et la matrice de covariance de la gaussienne d'indice k .

Le caractère actuellement quasi incontournable des GMM en vérification du locuteur s'explique principalement par le fait que cette famille de densité permet théoriquement d'approcher n'importe quelle loi de probabilité et par l'existence d'un cadre théorique solide permettant d'estimer les paramètres qui lui sont associés. Les *GMM* sont censés fournir une catégorisation implicite des sons présents dans le signal de parole et contenir de manière sous-jacente la modélisation de tous les événements acoustiques productibles par un locuteur. D'un point de vue de la modélisation, une autre de leurs caractéristiques est de ne pas tenir compte de la dépendance temporelle des vecteurs

acoustiques entre eux.

Les deux inconvénients principaux de cette modélisation sont le nombre important d'opérations nécessaires à la décision et la quantité importante de mémoire nécessaire au stockage de la référence caractéristique du locuteur.

Les Modèles de Markov Cachés (HMM)

Les propriétés statistiques des HMM en font l'une des modélisations les plus séduisantes pour représenter un locuteur dans l'espace des paramètres acoustiques. En effet, ils intègrent à la fois les propriétés des distributions de probabilité et celles d'une machine à état et permettent donc de modéliser des processus stochastiques variant dans le temps comme le signal de parole. Comme dans le cas des GMM, la phase d'entraînement consiste à estimer les paramètres (densité de chaque état et transition entre les états). Une description précise de l'utilisation des HMM en VAL peut se trouver dans [Rosenberg et Soong, 1992]. Alors que les HMM fournissent d'excellentes performances en mode dépendant du texte, les GMM sont actuellement la densité la plus utilisée et la plus performante en mode indépendant du texte. La mise en œuvre d'un système à base de GMM est précisément décrite dans le chapitre suivant.

Les modélisations statistiques les plus utilisées en RAL en mode indépendant du texte sont ergodiques, contrairement à celles utilisées en RAP. Elles ne tiennent pas compte de l'enchaînement temporel des trames. Ceci peut s'expliquer par la difficulté d'extraire du signal de parole une information dépendante du locuteur et indépendante du contenu linguistique.

5.3 Calcul du score de décision

Dans un système de vérification automatique du locuteur, le module de calcul du score peut souvent se décomposer en deux sous-modules : celui du calcul de la différence des mesures de similarité entre les hypothèses H_X et $H_{\bar{X}}$ et celui de normalisation.

Étant donné une suite d'observations acoustiques \mathbf{Y} , le score de décision $S_X(\mathbf{Y})$, calculé par un système de vérification automatique du locuteur, est généralement basé sur une estimation de la moyenne sur \mathbf{Y} de la différence trame à trame entre le logarithme des mesures de similarité D_X et $D_{\bar{X}}$.

On a :

$$S_X(\mathbf{Y}) = \frac{1}{N} \sum_{t=1}^N (\log D_X(y_t) - \log D_{\bar{X}}(y_t)) \quad (5.1)$$

Comme on le voit dans la sous-section suivante, les techniques de normalisation courantes sont basées sur une estimation de la moyenne et de la variance de $S_X(y)$ sur un ensemble d'accès imposteur. Pour des raisons de robustesse et pour obtenir une

meilleure estimation de ces paramètres, on préfère baser ce score sur une estimation de $S_X(\mathbf{Y}_L)$ [Fredouille et al., 2000], où \mathbf{Y}_L est une suite de L vecteurs consécutifs extraite de \mathbf{Y} . Le score $S_X(\mathbf{Y})$ utilisé par le module de normalisation s'écrit alors :

$$S_X(\mathbf{Y}) = \frac{1}{P} \sum_{p=1}^P S_X(\mathbf{Y}_{L,p}) \quad \text{avec} \quad S_X(\mathbf{Y}_{L,p}) = \frac{1}{L} \sum_{t=p-L/2}^{p+L/2} (\log D_X(y_t) - \log D_{\bar{X}}(y_t))$$

avec P , nombre de blocs de taille L contenus dans le signal de test et $S_X(\mathbf{Y}_{L,p})$, score associé au bloc centré sur le vecteur y_p .

Dans le cas d'une modélisation probabiliste de H_X et $H_{\bar{X}}$ et sous l'hypothèse de l'indépendance des observations de \mathbf{Y} , $S_X(\mathbf{Y})$ correspond à l'estimation du logarithme du rapport de vraisemblance entre les hypothèses H_X et $H_{\bar{X}}$.

Dans le cas des méthodes statistiques du second ordre, les résultats présentés dans [Zilca, 2001] montrent que le calcul du rapport des mesures de sphéricités présentées dans [Bimbot et al., 1995] remplace avantageusement le calcul du rapport de vraisemblance en terme d'efficacité et de performance. La mesure de sphéricité $S_X(\mathbf{Y})$ est définie comme :

$$S_X(\mathbf{Y}) = -\frac{SM(\Sigma_Y, \Sigma_X)}{SM(\Sigma_Y, \Sigma_{\bar{X}})}$$

avec $SM(\cdot)$ mesure de sphéricité, définie pour deux matrices Σ_1 et Σ_2 comme :

$$SM(\Sigma_1, \Sigma_2) = \frac{1}{2} tr(\Sigma_1 \Sigma_2^{-1}) tr(\Sigma_2 \Sigma_1^{-1})$$

où $tr(\Sigma)$ est la trace de la matrice Σ . Σ_Y est la matrice de covariance des observations. On l'estime sur l'ensemble des vecteurs de test. Dans ce cas, l'estimation de la mesure de similarité a lieu au niveau du segment et n'a pas de sens au niveau de la trame.

Représentation de $H_{\bar{X}}$

Une des difficultés rencontrées lors de la mise en œuvre du module de calcul du score concerne la représentation de l'hypothèse $H_{\bar{X}}$. Nous avons vu que dans le cas d'une modélisation statistique des deux hypothèses H_X et $H_{\bar{X}}$, H_X est caractérisée par une densité $p(\cdot|\lambda_X)$ représentant le locuteur X dans l'espace des paramètres alors que $H_{\bar{X}}$ est caractérisée par une densité $p(\cdot|\lambda_{\bar{X}})$ représentant l'alternative à l'identité proclamée dans l'espace des paramètres. Alors que les paramètres λ_X de la densité représentant le locuteur X sont estimés uniquement à partir d'exemples du signal de parole du locuteur X , les paramètres $\lambda_{\bar{X}}$ de la loi de $H_{\bar{X}}$ sont plus difficiles à obtenir car celle-ci doit théoriquement représenter tous les locuteurs autres que celui associé à λ_X .

Il existe deux stratégies principales pour obtenir $\lambda_{\bar{X}}$. Toutes deux nécessitent un ensemble \mathcal{L} de N_c locuteurs. La première est basée sur une utilisation individuelle de chacun d'eux et associée à chaque client du système une cohorte de locuteurs qui permet

de calculer $D_{\bar{X}}(\mathbf{Y})$. La seconde estime à partir de ces N_c locuteurs un modèle unique λ_Ω , appelé modèle du monde ou UBM (pour Universal Background Model). Dans le cas des méthodes basées sur le calcul d'une distance euclidienne, la représentation de $H_{\bar{X}}$ utilisée est généralement celle de la première proposition présentée ci-dessous. Nous nous situons maintenant uniquement dans le cadre des méthodes probabilistes.

La stratégie basée sur l'utilisation d'une cohorte de locuteurs considère un ensemble $\Lambda = \{\lambda_1 \dots \lambda_N\}$ de locuteurs et la vraisemblance associée à $H_{\bar{X}}$ est une fonction $\mathcal{F}()$ du type max ou moyenne des vraisemblances de tous les locuteurs de la cohorte. Dans ce cas, la vraisemblance de $H_{\bar{X}}$ se calcule :

$$p(\mathbf{Y}|\lambda_{\bar{X}}) = \mathcal{F}(p(\mathbf{Y}|\lambda_1), \dots, p(\mathbf{Y}|\lambda_N))$$

La stratégie basée sur un modèle unique indépendant du locuteur utilise les N_c locuteurs de l'ensemble \mathcal{L} pour estimer les paramètres de la densité $p(\cdot|\lambda_\Omega)$ censée représenter toutes les alternatives à X dans l'espace des paramètres. Cette stratégie est celle qui est généralement utilisée par les systèmes de vérification du locuteur entre autres car elle permet de diminuer la quantité de mémoire nécessaire au stockage des paramètres de la densité associée à $H_{\bar{X}}$ ainsi que le nombre d'opérations nécessaires à la décision. La vraisemblance de l'hypothèse H_X s'écrit alors :

$$p_{H_{\bar{X}}}(\mathbf{Y}) \equiv p(\mathbf{Y}|\lambda_{\bar{X}})$$

5.4 Normalisation

Le module de normalisation est essentiellement utilisé dans le cas d'une mise en œuvre probabiliste des systèmes de VAL. Les approximations liées à l'utilisation des densités p_{H_X} et $p_{H_{\bar{X}}}$, souvent estimées à partir d'un petit nombre d'observations, conduisent à un seuil de décision optimal dépendant du modèle du locuteur. D'autre part, la diversité quantitative et qualitative des observations Y extraites des signaux d'accès entraînent de fortes variations du score $S_X(Y)$ d'un accès à l'autre. Le but du module de normalisation est d'ajuster ce score grâce à une fonction $N_{X,Y}(\cdot)$ dont les paramètres sont calculés à partir de l'étude du comportement de $S_X(Y)$ pour le locuteur X et pour l'accès Y . Il permet de limiter les variations de score dépendantes du locuteur et d'obtenir un score indépendant du locuteur.

La règle de décision (eq 4.1) se ré-écrit alors :

$$S_X^N(Y) = N_{X,Y}(S_X(\mathbf{Y})) \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta^N \quad (5.2)$$

où $S_X^N(Y)$ est la version normalisée du score de décision et θ^N est le seuil optimal associé.

On peut distinguer parmi les différentes techniques de normalisation proposées dans la littérature deux classes de méthode :

1. Les méthodes de normalisation au niveau de la trame ou du bloc :

Il en existe principalement deux :

- la *z-norm* [Kung-Pu et Porter, 1988], [Reynolds et al., 2000] et sa variante particulièrement adaptée aux données des évaluations NIST, la *h-norm* [Reynolds, 1997], [Reynolds et al., 2000],
- la *t-norm* [Auckenthaler et al., 2000].

Le principe de ces deux normalisations est le même. Elles font toutes deux l'hypothèse de la normalité de la distribution des scores imposteurs l'une par rapport aux locuteurs, l'autre par rapport aux signaux de tests. Leur but est de centrer en 0 la moyenne et de normaliser à 1 la variance de ces distributions. Ainsi, pour chaque client X on estime la distribution $\mathcal{N}_X(m_X, \sigma_X)$ et pour chaque signal de test Y on estime $\mathcal{N}_Y(m_Y, \sigma_Y)$.

Les paramètres (m_X, σ_X) et (m_Y, σ_Y) de ces deux densités sont estimés à partir de nombreux segments de parole, prononcés par un groupe de locuteurs censés représenter l'ensemble de la population des imposteurs.

Ainsi, pour le locuteur X , client du système de VAL et à partir de I segments de parole dont l'identité associée n'est pas X , on a pour une mise en œuvre de la *z-norm* sur des blocs de taille L :

$$m_X = \frac{1}{I} \sum_{i=1}^I \frac{1}{P_i} \sum_p^{P_i} S_X(\mathbf{Y}_{L,p}^i)$$

$$\sigma_X^2 = \frac{1}{I} \sum_{i=1}^I \frac{1}{P_i} \sum_p^{P_i} S_X(\mathbf{Y}_{L,p}^i)^2 - m_X^2$$

De même, pour le segment d'accès Y et à partir de J références caractéristiques dont aucune ne correspond à l'identité associée à Y , on a pour une mise en œuvre de la *t-norm* sur des blocs de taille L :

$$m_Y = \frac{1}{J} \sum_{j=1}^J \frac{1}{P_j} \sum_p^{P_j} S_{X_j}(\mathbf{Y}_{L,p}^j)$$

$$\sigma_Y^2 = \frac{1}{J} \sum_{j=1}^J \frac{1}{P_j} \sum_p^{P_j} S_{X_j}(\mathbf{Y}_{L,p}^j)^2 - m_Y^2$$

Ces deux normalisations cherchent à centrer-réduire la distribution des scores imposteur et le score final, toujours dans le cas d'une normalisation par bloc de- vient :

<p>dans le cas de la <i>z-norm</i> :</p> $S_X^z(Y) = \frac{1}{L} \sum_{p=1}^P \frac{S_X(\mathbf{Y}_{L,p}) - m_X}{\sigma_X}$	<p>dans le cas de la <i>t-norm</i> :</p> $S_X^t(Y) = \frac{1}{L} \sum_{p=1}^P \frac{S_X(\mathbf{Y}_{L,p}) - m_Y}{\sigma_Y}$
---	---

La *z-norm* cherche à contrôler pour chaque locuteur la distribution des scores imposteurs, la *t-norm* cherche à maîtriser cette distribution par rapport au signal de test. Au niveau de la mise en œuvre, la différence principale entre ces deux normalisations est qu'alors que les paramètres de la *z-norm* sont estimés une fois pour toutes lors de la phase d'apprentissage, ceux de la *t-norm* doivent être ré-estimés à chaque accès. Cela augmente de façon considérable le temps nécessaire au calcul du score de décision et rend son utilisation dans un contexte opérationnel temps réel beaucoup moins réaliste.

Notons que ces deux normalisations peuvent être utilisées en cascade; on parle alors de *zt-norm* ou de *tz-norm* [Reynolds et al., 2000]. Dans le premier cas, les paramètres (m_Y, σ_Y) sont estimés après application de la *z-norm* pour chacun des J locuteurs de l'ensemble de normalisation. Lors du calcul du score de décision, on *z-normalise* chaque bloc extrait du signal de test avant de les *t-normaliser*. Dans le second cas, les paramètres (m_X, σ_X) sont estimés après application de la *t-norm* pour chacun des I segments d'accès de l'ensemble de normalisation. Lors du calcul du score de décision, on applique à chaque bloc extrait du signal de test la *t-norm* avant la *z-norm*. Ces deux normalisations requièrent un ensemble de segment de parole pour estimer les paramètres de *z-norm* des J références caractéristiques dans le cas de la *zt-norm* ou ceux associés à la *t-normalisation* des I segment d'accès dans le cas de la *tz-norm*.

Lorsque les signaux de parole sont enregistrés sur plusieurs types de micro différents une variante de la *z-norm* : la *h-norm* permet une amélioration importante des performances. Dans le cas de la *h-norm*, les paramètres (m_X, σ_X) sont estimés séparément pour tous les types de micro susceptibles d'être utilisés en phase de test. Les paramètres de normalisation sont ensuite adaptés selon le type la configuration du test.

Une autre normalisation ne nécessitant aucune donnée supplémentaire a été développée au cours de ce travail. Ses principes sont développés au chapitre suivant.

2. Les méthodes de normalisation au niveau du segment

Ces méthodes fonctionnent au niveau du score de décision global. La principale méthode de normalisation de ce type [Fredouille, 2000] consiste à estimer de

manière indépendante du locuteur les densité des scores clients et imposteurs pour un système donné. Le score de décision final est ensuite obtenu selon le critère du *maximum a posteriori*. Pour plus de détails, le lecteur est invité à consulter [Fredouille et al., 2000].

5.5 Le module de décision

La plupart des mises en œuvre actuelles du module de décision consiste simplement en une comparaison du score à une constante. Cependant des travaux récents proposent l'utilisation des Machines à Support de Vecteur (SVM) [Bengio et Mariéthoz, 2001].

Chapitre 6

Description du système de référence

Dans la section précédente, nous avons présenté différentes possibilités de mise en œuvre des modules d'une plateforme de VAL. Ce chapitre présente les caractéristiques des différents paramètres de notre système de référence. Celui-ci correspond à quelques nuances près au système du consortium ELISA dont nous disposions au début de ce travail. Il nous permet d'évaluer objectivement l'influence sur les performances de nos modifications. Comme annoncé dans l'introduction de ce chapitre, les choix de mise en place décrits ici ont été validés sur un ensemble de données extraites de celles des évaluations NIST.

6.1 Le module de paramétrisation

Le choix des coefficients cepstraux de banc de filtres [Furui, 1981] comme vecteurs de paramètres a été fait sur la base des bonnes performances que ceux-ci ont permis d'obtenir dans différentes expériences décrites dans la littérature, mais aussi sur leur relative simplicité d'extraction ainsi que sur l'existence de techniques simples et efficaces de post-traitement, comme par exemple celles d'égalisation de canal. Dans notre plateforme de référence, le vecteur de représentation correspond à 16 coefficients statiques et 16 coefficients dynamiques estimés sur une trame de 20 ms à une fréquence de trame de 10 ms. Le signal de parole du corpus des évaluations NIST étant obtenu via le réseau téléphonique, nous avons restreint la bande de fréquence utile pour calculer ces paramètres à celle du téléphone (300 Hz - 3400 Hz). Nous n'avons pas utilisé l'énergie dont les variations trop sensibles aux conditions d'enregistrement nuisent aux performances du système, ni les coefficients $\Delta\Delta$ dont le gain de performances qu'ils permettent d'obtenir est négligeable comparé au surcoût de calcul que leur utilisation représente.

Pour le module de pré-traitement, un algorithme non-supervisé de suppression de trames est disponible sur la plateforme ELISA. Celui est basé sur une modélisation bi-gaussienne de l'énergie de la trame et sa description se trouve dans [Magrin-Chagnolleau

et al., 2001].

Le choix des paramètres du module de post-traitement est plus délicat car ceux-ci dépendent fortement des conditions expérimentales. Si la soustraction cepstrale (CMS) permet de corriger efficacement certaines distorsions du signal de parole, elle peut aboutir à une dégradation des performances sur un signal pas ou peu perturbé. On peut citer comme exemple l'une des conditions de l'évaluation NIST de 1998 : lorsque l'on contraint le client à utiliser le même combiné lors des phases d'apprentissage et de test avec des accès imposteurs obtenus via des combinés différents. Tous les systèmes alors présentés sous cette condition effectuaient de la CMS excepté un de ceux du consortium ELISA qui a, sous cette condition, obtenu les meilleures performances. La phase de post-traitement dépend donc essentiellement des conditions de fonctionnement de la phase d'exploitation. Dans le cas des évaluations NIST, la configuration opérationnelle sur laquelle se base le classement principal des systèmes correspond à des combinaisons apprentissage-test utilisant le même type de microphone mais des numéros d'abonnement différents. Dans ce cas la CMS glissante avec normalisation des paramètres permet une forte augmentation des performances et nous l'avons intégré à notre plateforme de référence. Nous pouvons tout de même signaler que deux autres techniques de post-traitement ont aussi été implémentées sur la plateforme ELISA. Celles-ci sont : la CMS glissante, le feature warping.

6.2 Le module de modélisation

Un des points communs à tous les meilleurs systèmes de vérification du locuteur présentés aux évaluations NIST de ces cinq dernières années est d'une part l'utilisation systématique des modèles de mélange de gaussiennes comme famille de fonctions de densité de probabilité pour H_X et $H_{\bar{X}}$ et d'autre part d'un unique modèle $p_{H_{\bar{X}}}$ pour caractériser l'hypothèse $H_{\bar{X}}$.

Pour le module de modélisation, l'utilisation des modèles de mélange de gaussiennes s'est donc imposée de manière implicite. Elles apparaissent en effet actuellement comme incontournables en vérification du locuteur. Pour le module de modélisation de notre système de référence, nous avons choisi d'utiliser un GMM composé de 128 gaussiennes avec des matrices de covariances diagonales.

Le critère d'estimation des paramètres des GMM est l'une des caractéristiques fondamentales d'un système de RAL. Même si les techniques d'estimation au *Maximum A Posteriori* de λ_X étaient déjà émergentes lors de la mise en place du système de référence, nous avons utilisé une estimation au Maximum de Vraisemblance afin de mesurer l'apport de différentes techniques MAP.

Une section du chapitre suivant est consacrée aux densités de mélanges de gaus-

siennes et décrit plus particulièrement l'estimation des paramètres qui lui sont associés.

6.3 Le module de normalisation

Au début de nos travaux et lors de la mise en place de notre système de référence, une seule normalisation était principalement utilisée par la plupart des systèmes obtenant les meilleures performances aux évaluations NIST : la *h-norm* [Reynolds, 1997]. C'est donc naturellement que nous l'avons intégrée à notre plateforme de référence.

Une section du chapitre suivant présente en détail les performances obtenues pas différents procédés de normalisation et notamment ceux de la technique développée récemment dans l'équipe METISS de l'IRISA : la *d-norm*.

6.4 Calcul du score et décision

Le score de décision utilisé par notre plateforme de référence est basé sur une estimation du rapport de vraisemblance calculée sur des blocs de 15 observations acoustiques. Le seuil de décision est obtenu par minimisation du critère C_{Det} sur un corpus de développement.

6.5 En résumé

La figure 6.1 représente le système de référence avec toutes ses options. Le tableau 6.1 contient les performances de ce système selon les critères C_{Det} et de l'*HTER*.

6.6 Orientations de ce travail

Du module de modélisation dépendent directement la taille de la référence caractéristique du locuteur et le temps CPU nécessaires au calcul du score de décision. Parmi tous les modules d'un système de VAL, c'est donc celui qui apparaît comme le plus critique dans le contexte des objectifs de notre collaboration avec CP8. C'est donc naturellement que notre travail s'est focalisé sur ce module.

Notre travail sur la modélisation a tout d'abord eu pour but l'approfondissement de notre maîtrise de l'état-de-l'art et a consisté en la mise au point d'une technique permettant l'estimation selon le critère du *Maximum A Posteriori* des paramètres des modèles de mélange de gaussiennes. Dans un second aspect de notre travail, nous avons cherché à modifier plus profondément ce module par la réalisation d'une mise en œuvre originale, n'utilisant pas les GMM et permettant de réduire au maximum les ressources nécessaires au fonctionnement de notre plateforme. Les deux parties suivantes décrivent en détails nos travaux associés à ces deux orientations.

C_{Det}	0.048
$HTER(\%)$	11.2

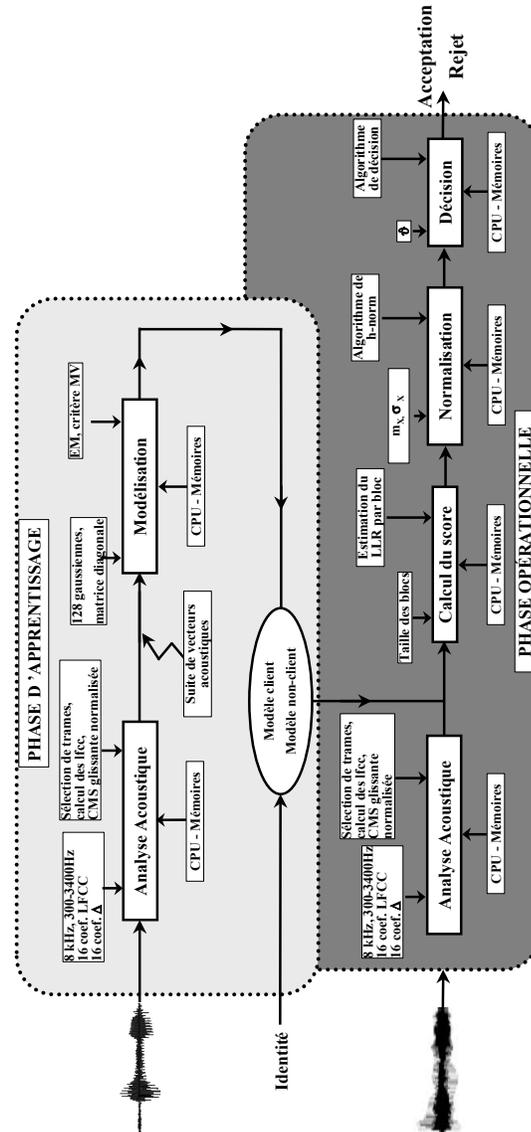
TAB. 6.1 – Performances selon les critères C_{Det} et $HTER$ de notre système de référence

FIG. 6.1 – Description du système de vérification du locuteur de référence

Troisième partie

État-de-l'art, mise en œuvre et améliorations

Cette partie est consacrée à la description de nos travaux visant à améliorer les performances de notre système de référence. Ils concernent l'estimation au Maximum A Posteriori des paramètres des modèles de mélange de gaussiennes ainsi que la description d'une méthode originale pour la normalisation du rapport de vraisemblance.

Nous présentons au cours de cette partie plusieurs modifications du système de référence. Celles-ci s'inscrivent dans la continuité de l'état-de-l'art dont voici, pour mieux situer le cadre de notre étude, certains des éléments clefs : la décision est basée sur un test d'hypothèse statistique, les densités de mélange de gaussiennes (GMM) définissent la référence caractéristique du locuteur et un unique modèle, défini par $p_{\bar{X}}$, permet de modéliser le non-locuteur \bar{X} . Plus précisément, les sujets principaux de ce chapitre portent sur le module de modélisation et sur celui de normalisation.

Nos travaux sur le module de modélisation visent à la mise en place d'une technique d'estimation selon le critère du *Maximum A Posteriori* (MAP) des paramètres des modèles de mélange de gaussiennes alors que le système de référence était basé sur celui du *Maximum de Vraisemblance* (ML pour Maximum Likelihood). Après la description complète de la méthode proposée, nous présentons son évaluation toujours sur la base des locuteurs féminins de l'évaluation NIST de 2001.

Le second aspect des travaux présentés dans ce chapitre concerne une technique de normalisation dont l'atout principal est de ne nécessiter aucune donnée, autre que celles nécessaires à l'estimation de la référence caractéristique du client pour être opérationnelle. Après la description des motivations et des principes de cette méthode, nous proposons la comparaison des performances qu'elle permet d'obtenir par rapport aux techniques usuelles (*z-norm*, *t-norm*).

Enfin, nous présentons les principaux inconvénients de l'approche à base de modèles de mélanges de gaussiennes ainsi que les différentes techniques publiées dans la littérature et dont les motivations sont similaires à celles de notre collaboration avec CP8. Ce chapitre se termine par un bilan des ressources matérielles nécessaires au fonctionnement d'un système de vérification du locuteur.

Chapitre 7

Modélisation probabiliste des locuteurs

L'utilisation des modèles de mélanges de gaussiennes pour la mise en œuvre des systèmes de vérification du locuteur a été proposée pour la première fois en 1992 dans la thèse de D. Reynolds [Reynolds, 1992]. Elle constitue un tournant important pour l'ensemble des technologies de la caractérisation automatique du locuteur et d'une manière plus générale pour les technologies de l'indexation sonore. En effet, les GMM ont largement contribué à l'amélioration des performances des systèmes de RAL observée depuis une dizaine d'années. Dans le cadre de l'approche statistique qui apparaît comme la plus adaptée pour capter les caractéristiques stochastiques du signal de parole, les GMM sont la densité qui permet d'obtenir les meilleures performances en mode indépendant du texte. Ces bons résultats peuvent s'expliquer par certaines de ses propriétés intrinsèques telles que :

- la segmentation implicite des vecteurs acoustiques en K classes de son dans l'espace des paramètres. Chacune d'elle possède sa probabilité d'occurrence *a priori* mais la modélisation ne donne aucune information sur leur dynamique. Cette propriété semble raisonnablement adaptée à la RAL en mode indépendant du texte.
- la possibilité de modéliser un processus stochastique sans être théoriquement limité par sa complexité en augmentant le nombre de composantes du mélange.

L'autre élément qui permet d'expliquer le succès des GMM est l'existence d'un outil puissant pour l'estimation des paramètres qui leur sont associés : l'algorithme Expectation-Maximisation (EM). La mise en œuvre de cet algorithme et plus particulièrement l'estimation selon différents critères des paramètres des GMM constitue l'objet principal de ce chapitre.

7.1 Les densités de mélange de gaussiennes

Comme nous l'avons déjà vu dans la section 5.2, la densité de probabilité d'un vecteur y de dimension d suivant une loi de mélange de K gaussiennes s'écrit :

$$p(y) = \sum_{k=1}^K w_k \cdot g_k(y)$$

avec $g_k(y) = \frac{1}{(2\pi)^{\frac{d}{2}} \cdot |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2} \cdot (y-m_k)^T \Sigma_k^{-1} \cdot (y-m_k)}$ et $\sum_{k=1}^K w_k = 1$.

Les paramètres des modèles multi-gaussiens sont : $\lambda = \{(w_k, m_k, \Sigma_k)\}_{1 \leq k \leq K}$ où w_k , m_k et Σ_k sont respectivement le poids, la moyenne et la matrice de covariance de la gaussienne d'indice k .

Les méta-paramètres associés à cette modélisation sont :

1. Le nombre de composantes du modèle qui permet de fixer la *granularité* de la représentation. Si un nombre de composantes trop faible conduit à une modélisation peu efficace des locuteurs, un nombre trop élevé amène potentiellement à des modèles sur-spécialisés par rapport aux données d'apprentissage et donc peu robustes aux différentes sources de variabilité du signal de parole.
2. La structure des matrices de covariance. En général, on utilise une structure diagonale pour chacune des matrices de covariance du système. Cette structure ne préjuge pas des propriétés de modélisation des GMM. En effet, il suffit d'augmenter le nombre de composantes d'un modèle dont les matrices sont diagonales pour atteindre les mêmes capacités de modélisation qu'un modèle à matrices de covariance pleines. On préfère généralement les matrices diagonales principalement pour des raisons de simplicité d'estimation, d'efficacité des calculs et de réduction de l'encombrement.

L'algorithme utilisé pour obtenir la référence caractéristique des locuteurs est l'algorithme Expectation Maximisation [Dempster et al., 1977]. Cet algorithme itératif permet d'estimer les paramètres d'une densité complexe mal observée. On considère que chaque observation d'apprentissage y_t dont on dispose provient de deux événements distincts :

- celui correspondant au tirage aléatoire de l'état, *i.e.* de la composante du mélange. On lui associe la variable aléatoire (v.a.) l_t , appelée variable cachée,
- celui correspondant au tirage aléatoire de l'observation y_t étant donné la densité associée à l'état l_t .

Dans le cas des GMM, la variable cachée permet d'estimer le poids de chacune des K gaussiennes du mélange. On définit \mathbf{W} , l'ensemble des données complètes tel que $\mathbf{W} = \{(l_1, y_1), \dots, (l_N, y_N)\}$.

Les paramètres associés à l'algorithme EM, sont :

1. Le modèle d'initialisation de l'algorithme qui permet principalement, s'il est bien choisi, d'atteindre plus rapidement la convergence et dont dépend théoriquement l'estimation des paramètres. Cependant, différentes expériences réalisées au cours de ce travail nous ont permis d'observer le peu d'influence du modèle initial sur les performances obtenues.
2. Le nombre d'itérations de l'algorithme.

Dans le cas d'une estimation au maximum de vraisemblance, on cherche à estimer λ^{ML} tel que :

$$\lambda^{ML} = \max_{\lambda} \log p(\mathbf{Y}|\lambda)$$

où \mathbf{Y} constitue l'ensemble des observations d'apprentissage.

Pour une itération i de l'algorithme EM, l'estimation de λ , est basée sur la maximisation de la fonction $Q(\lambda, \tilde{\lambda})$ définie comme la moyenne de la log-vraisemblance des données complètes $\log p(\mathbf{W}|\lambda)$ sur toutes les observations \mathbf{Y} étant donné l'estimation $\tilde{\lambda}$ obtenue à l'itération $i - 1$:

$$Q(\lambda, \tilde{\lambda}) = \mathbb{E}[\log p(\mathbf{W}|\lambda) | \mathbf{Y}, \tilde{\lambda}]$$

Cette fonction a la propriété suivante :

$$\text{si } \lambda \text{ satisfait } Q(\lambda, \tilde{\lambda}) > Q(\tilde{\lambda}, \tilde{\lambda}) \text{ alors } p(\mathbf{Y}|\lambda) > p(\mathbf{Y}|\tilde{\lambda})$$

ce qui permet d'assurer à chaque itération l'augmentation de la vraisemblance $\mathcal{L}(\mathbf{Y}; \lambda)$ des observations \mathbf{Y} par rapport aux paramètres λ et la convergence vers un maximum local. Dans la suite de ce document, les modèles estimés selon le critère ML portent l'indice *. On note : $\lambda^* = \lambda^{ml}$.

Le principe de l'algorithme EM, dans le cas de l'estimation des paramètres d'un GMM est décrit sur la figure 7.1.

Dans le cas d'une estimation selon le critère du *Maximum A Posteriori*, on définit une densité *a priori* $h(\lambda)$ sur les paramètres des GMM. On cherche alors λ^{MAP} tel que :

$$\lambda^{MAP} = \max_{\lambda} \log p(\lambda|\mathbf{Y}) \quad (7.1)$$

En appliquant la formule de Bayes, l'équation 7.1 revient à estimer λ_{MAP} tel que :

$$\lambda^{MAP} = \max_{\lambda} \log p(\mathbf{Y}|\lambda)h(\lambda) \quad (7.2)$$

Entrée : modèle initial λ_1 , suite des observations d'apprentissage \mathbf{Y} , spécification du nombre It d'itérations.
Pour $i = 1, \dots, It$
 1. E-step phase d'Estimation :
 on estime la vraisemblance par rapport à \mathbf{Y}
 de la variable cachée suivant le modèle courant :
 $p(l_k | \mathbf{Y}, \tilde{\lambda})$ pour $k = 1, \dots, K$
 2. M-step phase de Maximisation :
 calcul de λ avec $\lambda = \max_{\lambda_i} \sum_{k=1}^K g_k(\mathbf{Y} | \lambda_i) \cdot p(l_k | \mathbf{Y}, \tilde{\lambda})$
Sortie : $\lambda_X = \lambda$

FIG. 7.1 – Principe de l'algorithme Expectation Maximisation

Une modification simple de l'algorithme tel qu'il est présenté sur la figure 7.1 et proposée dans [Gauvain et Lee, 1994] permet d'obtenir, sous certaines hypothèses, une estimation de λ suivant ce critère. Dans ce cas, la maximisation de la fonction $R(\lambda, \tilde{\lambda})$ assure, à chaque itération, l'augmentation de $p(\mathbf{Y} | \lambda)h(\lambda)$. On a :

$$R(\lambda, \tilde{\lambda}) = Q(\lambda, \tilde{\lambda}) + \log h(\lambda)$$

La phase M-step de la description de l'algorithme EM de la figure 7.1 est alors remplacée par :

$$\lambda = \max_{\lambda_i} \sum_{k=1}^K g_k(\mathbf{Y} | \lambda_i) \cdot p(l_k | \mathbf{Y}, \tilde{\lambda}) + \log h(\lambda_i)$$

À chaque itération, l'estimation de λ est contrainte par la densité *a priori* sur les paramètres. La principale difficulté réside alors dans le choix $h(\lambda)$. Nous en discutons dans la section suivante. Dans la suite de ce document, les modèles estimés selon le critère MAP portent l'indice \dagger : $\lambda^{MAP} = \lambda^\dagger$.

Notons que, quel que soit le critère d'estimation utilisé, on utilise l'UBM comme modèle initial pour la première itération de l'algorithme EM.

7.2 Estimation au Maximum de Vraisemblance

Les estimations des poids w_k , moyennes m_k et matrices de covariance Σ_k du mélange s'écrivent [Reynolds, 1992] pour chacune des composantes k et en fonction des paramètres de l'itération précédente (dont les symboles portent un $\tilde{\cdot}$) :

$$w_k = \frac{1}{N} \sum_{t=1}^N p(l_k | y_t, \tilde{\lambda})$$

$$m_k = \frac{\sum_{t=1}^N p(l_k | y_t, \tilde{\lambda}) y_t}{\sum_{t=1}^N p(l_k | y_t, \tilde{\lambda})}$$

$$\Sigma_k = \frac{\sum_{t=1}^N p(l_k|y_t, \tilde{\lambda}) y_t y_t^T}{\sum_{t=1}^N p(l_k|y_t, \tilde{\lambda})} - m_k m_k^T$$

avec :

$$p(l_k|y_t, \tilde{\lambda}) = \frac{\tilde{w}_k g_k(y_t|\tilde{m}_k, \tilde{\Sigma}_k)}{\sum_{k=1}^K \tilde{w}_k g_k(y_t|\tilde{m}_k, \tilde{\Sigma}_k)}$$

Nous avons évalué ce critère d'estimation pour des modèles de 64, 128, 256 et 512 composantes. Les résultats sont présentés sur le tableau 7.1 et sur la figure 7.2. Dans l'ensemble, les performances de ces différentes configurations sont assez proches, aux alentours de 11% à l'HTER. On observe cependant une faible augmentation de l'HTER lorsque K passe de 128 à 256 puis lorsqu'il passe à 512.

L'estimation ML ne tire pas partie de l'information *a priori*, sur le signal de parole en général, potentiellement contenue dans l'UBM. L'utilisation de ce modèle lors de l'estimation des paramètres des densités associées au client apparaît comme extrêmement attrayante. D'autant plus si ce premier modèle contient énormément de composantes et que l'on parvient, en fonction des données d'apprentissage, à spécialiser uniquement celles qui sont les plus caractéristiques du locuteur. Nous proposons dans la section suivante un moyen de mettre en œuvre une estimation de ce type pour les paramètres des modèles clients.

7.3 Estimation MAP et techniques d'adaptation

Dans le cadre de l'utilisation des GMM en VAL, la volonté d'utiliser un critère différent que celui du *Maximum de Vraisemblance* pour estimer la référence caractéristique des locuteurs trouve ses motivations, entre autres, dans les raisons suivantes :

- l'utilisation d'un modèle unique pour la représentation de $H_{\bar{x}}$ dans le module de calcul du score. Ce modèle doit être capable de caractériser à lui seul l'ensemble des différentes classes de sons contenues dans un signal de parole. Cette caractéristique essentielle de l'UBM amène de manière naturelle à son utilisation, comme *a priori* lors de l'estimation de chaque modèle client. Intuitivement, cette démarche se justifie en considérant le modèle client comme une spécialisation de l'UBM.
- l'augmentation du nombre de composantes du modèle client. Celle-ci, liée à celle de l'UBM lorsqu'on l'utilise comme modèle initial dans la boucle de l'algorithme EM, conduit, si l'on utilise le critère d'estimation ML, à des modèles sur-entraînés, *i.e.* représentant de manière trop spécifique les caractéristiques des données d'apprentissage et conduisant à une dégradation des performances. Pourtant cette

augmentation apparaît comme inévitable si l'on souhaite utiliser un UBM, capable de capter toute les caractéristiques propres au signal de parole.

- la faible quantité de données dont on dispose lors de l'apprentissage des modèles, ainsi que leur faible représentativité, en terme de la variabilité intra-locuteur, dans l'espace des paramètres acoustiques. Cette dernière, spécifique aux évaluations NIST, concerne malgré tout un grand nombre d'applications nécessitant un système de RAL.

K	HTER(%)	C_{Det}
64	11.7	0.05
128	11.3	0.05
256	11.3	0.05
512	12.1	0.05

TAB. 7.1 – Estimation ML des paramètres des GMM : performances à l'HTER et au coût de fonctionnement C_{Det} pour différentes tailles de modèles clients

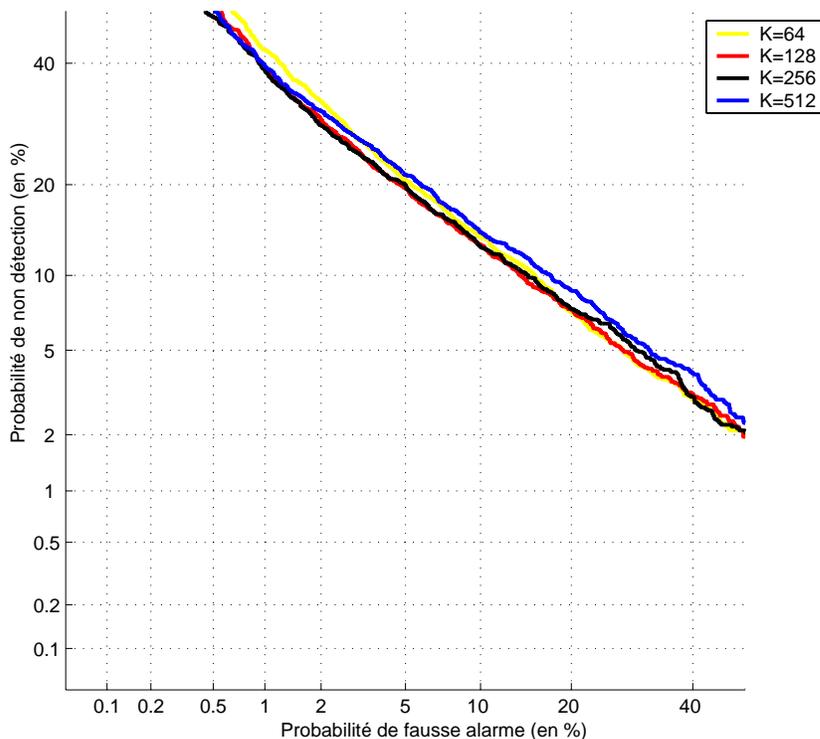


FIG. 7.2 – Estimation ML des paramètres des GMM : courbes DET obtenues pour des systèmes de 64, 128, 256 et 512 composantes

Les techniques d'estimation MAP des paramètres des GMM ont été utilisées et décrites pour la première fois en 1997 dans [Reynolds, 1997]. Les solutions actuellement utilisées se placent toutes dans le cadre de l'utilisation d'un UBM (de paramètres λ_Ω) et se basent sur les développements théoriques présentés dans [Gauvain et Lee, 1994]. Elles conduisent après plusieurs simplifications à considérer chacun des paramètres de λ^\dagger comme une combinaison linéaire entre les paramètres λ^* et λ_Ω . L'apprentissage alors mis en œuvre correspond à une spécialisation pour chaque client des paramètres de l'UBM ce qui justifie le terme souvent employé d'*adaptation au locuteur*. Différentes techniques pour fixer les contributions respectives de ces deux modèles ont été étudiées; elles constituent l'objet de cette section. Nous allons tout d'abord rappeler les principales propriétés de $h(\lambda)$. Puis nous décrivons les techniques d'adaptation que nous avons réalisées avant de proposer une re-formulation de l'apprentissage des paramètres clients.

Estimation au *maximum a posteriori*

Comme le montre l'équation (7.2), l'estimation au maximum de vraisemblance nécessite la définition de la fonction $h(\lambda)$ dont on trouve dans [Gauvain et Lee, 1994] le rappel des caractéristiques principales. Cette fonction peut être représentée par une densité correspondant au produit de deux distributions : la distribution de Dirichlet $\mathbf{d}(w_1, \dots, w_K)$ et la distribution de Wishart $\mathbf{w}(m_k, R_k)$ qui correspondent respectivement aux *a priori* sur les poids, sur la moyenne et sur la matrice de covariance de chaque composante du modèle. On a :

$$h(\lambda) = \mathbf{d}(w_1, \dots, w_K) \prod_{k=1}^K \mathbf{w}(m_k, \Sigma_k) \quad (7.3)$$

où :

$$\mathbf{d}(w_1, \dots, w_K | \nu_1, \dots, \nu_k) \sim \prod_{k=1}^K w_k^{\nu_k - 1} \quad (7.4)$$

et,

$$\mathbf{w}(m_k, \Sigma_k) \sim \frac{1}{|\Sigma_k|^{(\alpha_k - p)/2}} \exp\left[-\frac{\tau_k}{2} (m_k - \mu_k)^t \Sigma_k^{-1} (m_k - \mu_k)\right] \exp\left[-\frac{1}{2} \text{tr}(U_k \Sigma_k^{-1})\right] \quad (7.5)$$

La densité $h(\lambda)$ est caractérisée par les paramètres $\{\nu_k, \tau_k, \alpha_k, \mu_k, U_k\}_{k=1}^K$. Nous essayons maintenant d'expliquer leur rôle ainsi que la façon dont ils interviennent dans l'estimation de λ . L'obtention des formules de ré-estimation MAP (7.6, 7.7 et 7.8) peut se trouver dans [Gauvain et Lee, 1994].

Chaque ν_k représente l'information *a priori* sur le poids de la composante k du mélange. Étant donné w_k^* , estimation ML de w_k , son estimation MAP w_k^\dagger s'écrit :

$$w_k^\dagger = \frac{(\nu_k - 1) + w_k^* N}{\sum_{k=1}^K (\nu_k - 1) + N} \quad (7.6)$$

où N représente le nombre de vecteurs acoustiques extraits du signal d'apprentissage. La quantité ν_k est homogène à un nombre d'observations.

μ_k et τ_k sont relatifs à l'information *a priori* sur la moyenne de la composante k du mélange. μ_k est égal au vecteur moyenne *a priori* de la gaussienne k . τ_k , nombre sans dimension, exprime la confiance que l'on donne à μ_k - plus τ_k est grand, plus m_k^\dagger est proche de μ_k . On peut écrire m_k^\dagger sous la forme :

$$m_k^\dagger = \frac{\tau_k \mu_k + w_k^* N m_k^*}{\tau_k + w_k^* N} \quad (7.7)$$

L'expression 7.7 fait apparaître m_k^\dagger comme la combinaison linéaire des deux vecteurs μ_k et m_k^* chacun étant pondéré par un terme lié à la confiance que l'on lui porte. Ainsi μ_k est pondéré par τ_k et m_k^* par $w_k^* N$, dont la valeur peut s'interpréter comme le nombre de vecteurs acoustiques affectés à la gaussienne k . Plus ce nombre est grand, plus la contribution de l'estimation ML de m_k est importante dans m_k^\dagger .

L'estimation de Σ_k^\dagger fait intervenir U_k , μ_k et τ_k et s'écrit :

$$\Sigma_k^\dagger = \frac{U_k + \tau_k (\mu_k - m_k^\dagger) \cdot (\mu_k - m_k^\dagger)^t + w_k^* N \{ \Sigma_k^* + (m_k^\dagger - m_k^*) \cdot (m_k^\dagger - m_k^*)^t \}}{(\alpha_k - d) + w_k^* N} \quad (7.8)$$

L'expression 7.8, dans laquelle on peut rappeler que d est la dimension du vecteur acoustique, fait apparaître que Σ_k^\dagger est la combinaison linéaire des trois termes :

- Le premier U_k correspond à la matrice de covariance *a priori* de la gaussienne k .
- Le second $\tau_k (\mu_k - m_k^\dagger) \cdot (\mu_k - m_k^\dagger)^t$ est une matrice corrective liée à l'écartement de m_k^\dagger par rapport à la moyenne *a priori* μ_k . τ_k est un scalaire qui permet de pondérer sa contribution en fonction de la confiance exprimée en μ_k .
- Le troisième $w_k^* N \{ \Sigma_k^* + (m_k^\dagger - m_k^*) \cdot (m_k^\dagger - m_k^*)^t \}$ permet d'exprimer l'*a priori* sur Σ_k et contient un terme correctif lié à l'écartement de m_k^\dagger par rapport à la moyenne m_k^* . Il correspond à la somme de deux matrices Σ_k^* et $(m_k^\dagger - m_k^*) \cdot (m_k^\dagger - m_k^*)^t$ pondérées par le nombre d'observations affectées à la gaussienne k .

En résumé, ν_k , τ_k , α_k sont des scalaires homogènes à un nombre de données, μ_k et U_k sont respectivement un vecteur de l'espace des paramètres et une matrice de covariance. Ils correspondent respectivement à la moyenne et à la matrice de covariance *a priori* de la gaussienne k .

Algorithme d'estimation

Nous présentons maintenant la mise en œuvre de notre technique d'estimation des paramètres MAP des GMM. À partir des formules d'estimation 7.6, 7.7 et 7.8, elle vise

à définir ν_k , τ_k , α_k , μ_k et U_k .

Tout d'abord, comme dans [Gauvain et Lee, 1994], nous avons fixé :

$$\nu_k = \tau_k + 1 \quad (7.9)$$

$$\alpha_k = \tau_k + d \quad (7.10)$$

Ensuite, disposant des paramètres λ_Ω associés à l'UBM, c'est naturellement que nous avons utilisé m_Ω et Σ_Ω comme valeur *a priori*, respectivement pour μ_k et U_k .

Il reste maintenant à définir ν_k . Celui-ci doit être homogène à un nombre de données (cf. 7.6) et nous avons choisi de fixer :

$$\nu_k = \delta_k w_k^\Omega N + 1 \quad (7.11)$$

Pour une suite de N vecteurs acoustiques, $w_k^\Omega N$ est le nombre moyen d'observations affectées à la gaussienne d'indice k . δ_k , est un coefficient de proportionnalité qui reste à déterminer. Il est propre à chaque gaussienne et permet de régler la contribution du modèle *a priori* lors de l'estimation de la $k^{\text{ième}}$ composante du modèle client.

En tenant compte de 7.9 et de 7.11 et en remplaçant μ_k et Σ_k par leur valeur, les équations 7.6, 7.7 et 7.8 s'écrivent :

– Adaptation des poids :

$$w_k^\dagger = \frac{\delta_k \cdot w_k^\Omega + w_k^*}{\delta_k + 1} \quad (7.12)$$

– Adaptation des moyennes :

$$m_k^\dagger = \frac{\delta_k \cdot w_k^\Omega m_k^\Omega + w_k^* m_k^*}{\delta_k \cdot w_k^\Omega + w_k^*} \quad (7.13)$$

– Adaptation des matrices de covariance :

$$\Sigma_k^\dagger = \frac{\delta_k \cdot w_k^\Omega (\Sigma_k^\Omega + \Gamma_k) + w_k^* (\Sigma_k^* + \Delta_k)}{\delta_k \cdot w_k^\Omega + w_k^*} \quad (7.14)$$

où,

$$\Gamma_k = (m_k^\dagger - m_k^\Omega) (m_k^\dagger - m_k^\Omega)^t \quad \text{et} \quad \Delta_k = (m_k^\dagger - m_k^*) (m_k^\dagger - m_k^*)^t.$$

Finalement, en posant $\rho_k^w = \frac{\delta_k}{\delta_k + 1}$ et $\rho_k^{m,\Sigma} = \frac{\delta_k \cdot \frac{w_k^\Omega}{w_k^*}}{\delta_k \cdot \frac{w_k^\Omega}{w_k^*} + 1}$, nous obtenons à partir des équations 7.12, 7.13 et 7.14 :

$$w_k^\dagger = \rho_k^w \cdot w_k^\Omega + (1 - \rho_k^w) \cdot w_k^* \quad (7.15)$$

$$m_k^\dagger = \rho_k^{m,\Sigma} \cdot m_k^\Omega + (1 - \rho_k^{m,\Sigma}) \cdot m_k^* \quad (7.16)$$

$$\Sigma_k^\dagger = \rho_k^{m,\Sigma} \cdot (\Sigma_k^\Omega + \Gamma_k) + (1 - \rho_k^{m,\Sigma}) \cdot (\Sigma_k^* + \Delta_k) \quad (7.17)$$

Ces 3 formules introduisent 2 nouveaux coefficients. Le premier ρ_k^w permet de fixer les contributions relatives du modèle du monde et du modèle ML pour l'estimation du poids de chaque gaussienne. Le second $\rho_k^{m,\Sigma}$ permet de fixer les contributions relatives du modèle du monde et du modèle ML pour l'estimation des moyennes et des matrices de covariance du modèle client. Ces coefficients appartiennent tous deux à $[0, 1]$. Par souci de simplification, nous utilisons :

$$\rho_k = \rho_k^w = \rho_k^{m,\Sigma}$$

et l'on a finalement :

$$w_k^\dagger = \rho_k \cdot w_k^\Omega + (1 - \rho_k) \cdot w_k^* \quad (7.18)$$

$$m_k^\dagger = \rho_k \cdot m_k^\Omega + (1 - \rho_k) \cdot m_k^* \quad (7.19)$$

$$\Sigma_k^\dagger = \rho_k \cdot (\Sigma_k^\Omega + \Gamma_k) + (1 - \rho_k) \cdot (\Sigma_k^* + \Delta_k) \quad (7.20)$$

Une fois ces formules d'adaptation posées, il ne nous reste plus qu'à définir ρ_k .

Soit n_k , le taux d'occupation de la $k^{\text{ième}}$ gaussienne du mélange, on a :

$$n_k = \sum_{t=0}^N p(k|y_t) \text{ avec } p(k|y_t) = \frac{w_k g_k(y_t)}{\sum_{k=1}^K w_k g_k(y_t)}$$

n_k peut s'interpréter comme le nombre moyen de vecteurs de paramètres de \mathbf{Y} affectés à la gaussienne k .

Nous proposons ensuite de considérer ρ_k comme fonction de n_k :

$$\rho_k = f(n_k)$$

$f(n_k)$ permet de définir les contributions relatives des paramètres de l'estimation ML et ceux du modèle du monde dans λ^\dagger . Pour chaque composante, elle doit permettre d'augmenter la contribution de λ_Ω , pour de faibles valeurs de n_k ou de la diminuer lorsque n_k augmente.

Dans la première méthode que nous avons développée pour le calcul des ρ_k , nous distinguons trois régions selon la valeur de n_k . Dans chacune d'elle, $f()$ a un comportement différent, caractérisé par les deux paramètres suivants : ρ_{min} et N_{min} avec :

- ρ_{min} : contribution minimale du modèle du monde, on a $0 \leq \rho_{min} < 1$,
- N_{min} : taux d'occupation de la gaussienne ¹ au-delà duquel la contribution du modèle du monde est constante et minimale.

Ainsi :

$$f(n_k) = \begin{cases} 1 & \text{si } n_k \leq 0 \\ an_k + b & \text{si } n_k < N_{min} \\ \rho_{min} & \text{si } n_k > N_{min} \end{cases}$$

avec :

$$a = \frac{N_{min} - 1}{\rho_{min}} \text{ et } b = 1.$$

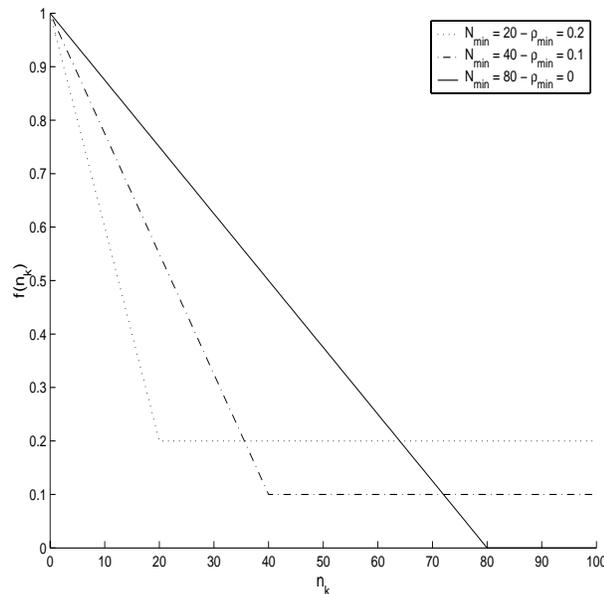


FIG. 7.3 – $f(n_k)$, contribution du modèle du monde pour l'estimation du modèle client

La figure 7.3 montre la forme de $f(n_k)$ ainsi définie pour différentes valeurs de ρ_{min} et de N_{min} . Les valeurs optimales de ρ_{min} et N_{min} sont obtenues de manière empirique

¹i.e. le nombre moyen de trames qui lui sont affectées

par optimisation des performances sur un ensemble de développement. Les principaux résultats obtenus sont présentés à la section suivante.

7.4 Performances de la méthode proposée

Comparaison de différentes configurations du système

Les travaux présentés dans [Reynolds et al., 2000] et vérifiés au cours de ce travail, proposent de limiter l'apprentissage des paramètres des GMM aux seules moyennes. Ceci permet principalement de réduire la taille de la référence caractéristique du locuteur et des calculs nécessaires à l'estimation de λ_X . Cette simplification n'entraîne qu'une baisse négligeable des performances. Le poids de la composante k d'un GMM peut s'interpréter comme la probabilité *a priori* d'avoir un vecteur dont la densité de probabilité est celle de la gaussienne k et donc comme la probabilité *a priori* d'occurrence de l'événement acoustique correspondant. Ne pas estimer les poids revient à faire l'hypothèse que la distribution du contenu acoustico-phonétique des énoncés est indépendante du locuteur ; ne pas estimer les matrices de covariance revient à supposer que l'événement acoustique associé à la gaussienne k a la même variance quelque soit le locuteur.

Dans les expériences présentées ci-dessous, seules les K moyennes des gaussiennes du mélange sont propres à chaque locuteur.

Trois ensembles de résultats sont présentés, chacun étant obtenu sur la totalité des locuteurs féminins de l'évaluation NIST de 2001. Ils ont respectivement été obtenus avec des modèles de mélange de gaussiennes de 128, 256 et 512 composantes.

Les figures (7.6, a-c), (7.7, a-d) correspondent aux courbes DET obtenues avec différentes configurations du couple (N_{min}, ρ_{min}) pour des systèmes dont les modèles de locuteur contiennent 128, 256 ou 512 composantes. Les tableaux 7.2, 7.3 et 7.4 présentent les performances de ces mêmes configurations selon le critère d'évaluation de NIST (C_{Det}) et celui du HTER ($HTER$).

D'une manière générale, on constate que les petites valeurs de ρ_{min} , associées à la diminution du taux de *fausse alarme* (T_{fa}), permettent d'obtenir les taux de *mauvaise détection* (T_{fr}) les plus faibles. Les faibles valeurs de ρ_{min} entraînent, aux fortes valeurs de T_{fa} , les T_{fr} les plus fortes. L'influence de ρ_{min} diminue avec l'augmentation du nombre de composantes du mélange et de N_{min} . Ainsi, c'est à $N_{min} = 40$ que les courbes sont les plus sensibles à ces variations. Pour 512 composantes, l'influence des ρ_{min} sur les performances est quasi nulle pour $N_{min} > 20$.

Nous avons évalué les performances obtenues par la technique d'estimation proposée dans [Reynolds et al., 2000]. Celle-ci consiste à utiliser des formules d'estimation

N_{min}	critère d'évaluation	ρ_{min}				
		0	0.1	0.3	0.5	0.7
20	C_{Det}	0.047	0.046	-	-	-
	$HTER(\%)$	11.0	10.3	-	-	-
40	C_{Det}	0.046	0.048	0.047	0.050	0.053
	$HTER(\%)$	10.0	11.4	10.7	10.3	10.9
120	C_{Det}	-	0.050	0.052	0.055	0.059
	$HTER(\%)$	-	10.8	10.7	11.1	11.7
160	C_{Det}	-	0.054	0.051	-	0.060
	$HTER(\%)$	-	10.9	9.9	-	12.0

TAB. 7.2 – Performances de différentes configurations de l'estimation MAP d'un système à 128 composantes

N_{min}	critère d'évaluation	ρ_{min}				
		0	0.1	0.3	0.5	0.7
20	C_{Det}	-	0.043	0.042	-	-
	$HTER(\%)$	-	9.6	8.8	-	-
40	C_{Det}	-	0.048	0.047	0.050	0.053
	$HTER(\%)$	-	11.4	10.7	10.3	10.9
120	C_{Det}	0.047	0.047	0.048	-	-
	$HTER(\%)$	9.0	9.0	9.1	-	-
160	C_{Det}	0.048	0.048	-	-	-
	$HTER(\%)$	9.1	9.2	-	-	-

TAB. 7.3 – Performances de différentes configurations de l'estimation MAP d'un système à 256 composantes

N_{min}	critère d'évaluation	ρ_{min}			
		0.1	0.3	0.5	0.7
20	C_{Det}	0.043	-	-	0.043
	$HTER(\%)$	9.0	-	-	8.1
40	C_{Det}	0.044	0.045	0.046	-
	$HTER(\%)$	8.6	8.4	8.5	-
120	C_{Det}	0.047	-	0.050	-
	$HTER(\%)$	8.7	-	9.0	-

TAB. 7.4 – Performances de différentes configurations de l'estimation MAP d'un système à 512 composantes

équivalentes à 7.18, mais en considérant un ρ_k tel que :

$$\rho_k = \frac{r}{n_k + r} \quad (7.21)$$

où r est un paramètre fixé *a priori*. Les résultats présentés, suivant cette technique

r	5	10	15	20
C_{Det}	0.044	0.043	0.042	0.042
$HTER(\%)$	9.6	9.2	8.9	8.9

TAB. 7.5 – Performances de différentes configurations de l'estimation MAP suivant la configuration [Reynolds et al., 2000] d'un système à 128 composantes

		$\alpha = 0.8$	$\alpha = 1$	$\alpha = 1.2$
r=10	C_{Det}	0.043	0.042	0.044
	$HTER(\%)$	9.4	9.2	9.8
r=15	C_{Det}	0.043	0.042	0.043
	$HTER(\%)$	9.1	8.9	9.4
r=20	C_{Det}	0.043	0.042	0.043
	$HTER(\%)$	9.1	8.9	9.2

TAB. 7.6 – Performances obtenues avec des GMM de 128 composantes en considérant le calcul des ρ_k par la formule de l'équation 7.22

d'estimation ont été obtenus avec $r = \{5, 10, 15, 20\}$. Les performances obtenues pour une modélisation de 128 composantes se trouvent sur le tableau 7.5. Cette méthode permet d'obtenir de meilleures performances.

La figure 7.4 représente la contribution du modèle du monde pour l'estimation des paramètres suivant la configuration permettant d'obtenir les meilleures performances pour 128 composantes *i.e.* : $N_{min} = 40$ et $\rho_{min} = 0$ et la technique d'estimation proposée dans [Reynolds et al., 2000], considérée avec $r = 15$. Contrairement à la technique de calcul des ρ_k suivant l'équation 7.21, notre stratégie d'estimation implique que le gain de la contribution relative du modèle du monde lorsque n_k passe de 0 à 10 soit le même que lorsqu'il passe de 20 à 40. De plus, le comportement asymptotique de la fonction $f(n_k)$ permet alors d'avoir $\lim_{n_k \rightarrow +\infty} f(n_k) = 0$. L'utilisation d'une fonction hyperbolique semble donc mieux adaptée et il nous a semblé intéressant d'évaluer les performances obtenues en accentuant ou en réduisant les variations du gain relatif en utilisant des fonctions du type :

$$\rho_k = \frac{1}{1 + \left(\frac{n_k}{r}\right)^\alpha} \quad (7.22)$$

La figure 7.5 représente ρ_k pour $\alpha = \{0.8; 1; 1.2\}$ et $r = 15$. Le tableau 7.6 présente les résultats obtenus avec $\alpha = \{0.8; 1; 1.2\}$ et $r = \{10; 15; 20\}$ pour des modèles de 128 composantes. Les meilleures performances sont obtenues avec $\alpha = 1$ correspondant à la formulation du calcul de ρ_k présentée dans [Reynolds et al., 2000].

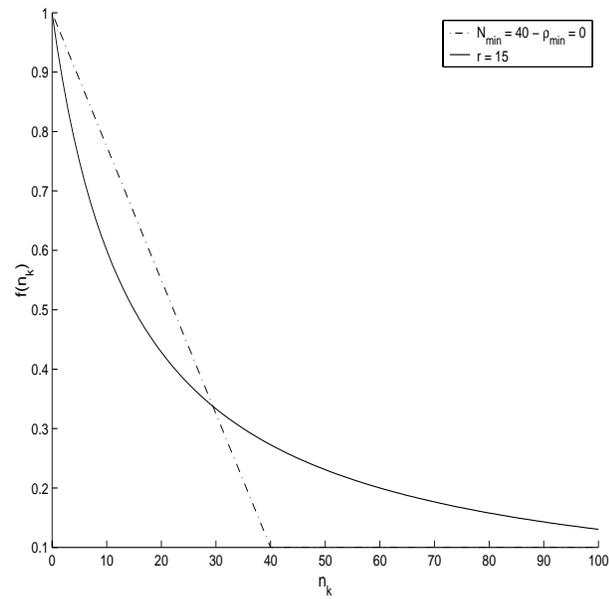


FIG. 7.4 – Comparaison de la contribution relative du modèle du monde dans l’estimation du modèle client pour notre technique d’adaptation et celle proposée dans [Reynolds et al., 2000] avec $r = 15$

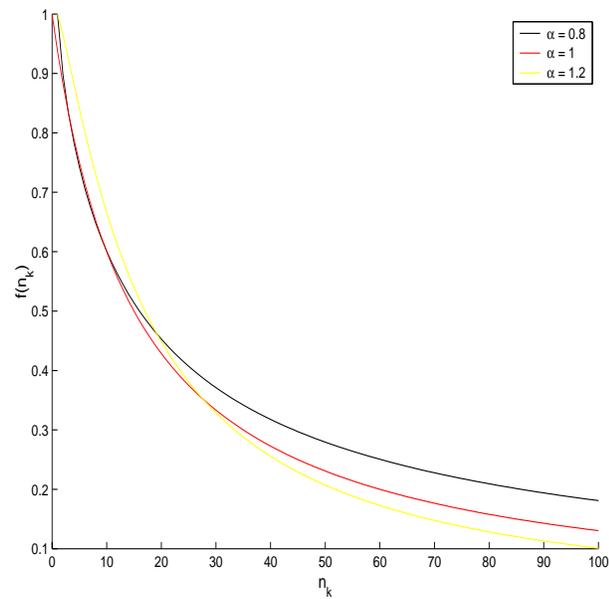


FIG. 7.5 – Contribution du modèle du monde pour l’estimation du modèle client avec ρ_k calculé selon l’équation 7.22 pour différentes valeurs de α

Re-formulation de l'adaptation des modèles clients

Face aux résultats précédemment obtenus, on a cherché à comprendre pourquoi la formule proposée par Reynolds présentait cette "optimalité expérimentale" pour le réglage des paramètres MAP. À cette fin, on a fait l'hypothèse que pour chaque composante du mélange, la moyenne du processus observée m_k^o correspond à la combinaison linéaire de deux moyennes empiriques l'une associée au modèle client m_k^X et obtenue à partir de N_X observations, l'autre associée au modèle non-client m_k^Ω et obtenue à partir de N_Ω observations. N_X et N_Ω correspondent respectivement aux nombres de vecteurs client et non-client associés à la gaussienne k . Pour chaque composante du mélange, cela s'écrit :

$$m_k^o = \gamma_k \cdot * m_k^\Omega + (\mathbb{1} - \gamma_k) \cdot * m_k^X \quad (7.23)$$

avec $m_k^X = \frac{1}{N} \sum_{t=1}^{N_X} y_t^X$, $m_k^\Omega = \frac{1}{N} \sum_{t=1}^{N_\Omega} y_t^\Omega$, γ_k vecteur de dimension d , $\mathbb{1}$ vecteur de dimension d dont toutes les composantes valent 1 et où $a \cdot * b$ désigne le produit terme à terme des vecteurs a et b .

si $y_t^X \in \mathbf{Y}_X$, $w_k g_k(y_t^X) > w_l g_l(y_t^X) \forall l \neq k$
 et si $y_t^\Omega \in \mathbf{Y}_{\bar{X}}$, $w_k g_k(y_t^\Omega) > w_l g_l(y_t^\Omega) \forall l \neq k$.

m_k^o est une variable aléatoire (comme somme des deux variables aléatoire m_k^X et m_k^Ω) dont on va chercher à minimiser la variance en fonction du vecteur γ_k .

On suppose maintenant que les vecteurs y_t^X suivent une loi gaussienne de moyenne m_k^X et de variance Σ_k^X et que les observations y_t^Ω suivent une loi gaussienne de moyenne m_k^Ω et de variance Σ_k^Ω .

On a $y_t^X \sim \mathcal{N}(m_k^X, \Sigma_k^X)$ et $y_t^\Omega \sim \mathcal{N}(m_k^\Omega, \Sigma_k^\Omega)$ donc $\frac{1}{N_X} \sum_{t=1}^{N_X} y_t^X \sim \mathcal{N}(m_k^X, \frac{\Sigma_k^X}{N_X})$
 et $\frac{1}{N_\Omega} \sum_{t=1}^{N_\Omega} y_t^\Omega \sim \mathcal{N}(m_k^\Omega, \frac{\Sigma_k^\Omega}{N_\Omega})$ ainsi m_o suit une loi

$$\mathcal{N} \left((1 - \gamma_k) m_k^X + \gamma_k m_k^\Omega, \frac{1}{N_X} (1 - \gamma_k)^t \Sigma_k^X (1 - \gamma_k) + \frac{1}{N_\Omega} \gamma_k \Sigma_k^\Omega \gamma_k \right)$$

En minimisant la variance de m_o par rapport à γ_k on obtient :

$$Id = \gamma_k \left[Id + \frac{N_X}{N_\Omega} \Sigma_k^\Omega \{ \Sigma_k^X \}^{-1} \right] \quad (7.24)$$

Sous l'hypothèse d'une structure diagonale pour les matrices Σ_k^X et Σ_k^Ω , 7.24 revient, pour chaque composante du vecteur de mélange à :

$$\gamma_k(i) = \frac{1}{1 + \frac{N_X}{N_\Omega} \frac{\sigma_k^\Omega(i)}{\sigma_k^X(i)}} \quad (7.25)$$

avec $\sigma(i)$ composante de la $i^{\text{ème}}$ ligne et de la $i^{\text{ème}}$ colonne de Σ .

Cette formulation de l'estimation des paramètres des GMM permet en posant $r = N_{\Omega} \frac{\sigma_k^{\Omega}}{\sigma_k^X}$ de retrouver une équation équivalente à celle de 7.21. En d'autres termes, la formule proposée par Reynolds revient à faire l'hypothèse que pour chaque composante du vecteur acoustique le produit de N_{Ω} par le rapport des variances est constant et qu'ils sont égaux entre eux.

En outre, on peut proposer comme perspective la modification de l'équation 7.21 en considérant $r' = \frac{N_{\Omega}}{\sigma_k^X}$ constant. Cela s'écrit pour chaque composante i du vecteur γ_k :

$$\gamma_k(i) = \frac{r'}{r' + \frac{N_x}{\sigma(i)_k^X}} \quad (7.26)$$

L'intérêt est d'obtenir pour les K composantes du mélange une contribution relative du modèle du monde différente dans chacune des directions de l'espace des paramètres. Ainsi, si l'on dispose de nombreuses observations de forte variance suivant une direction de l'espace, la contribution des données dans l'estimation de la composante associée sera plus faible qu'une autre suivant laquelle les observations sont plus regroupées (et donc de plus faible variance). Cette mise en œuvre est une perspective intéressante pour l'estimation des paramètres des GMM.

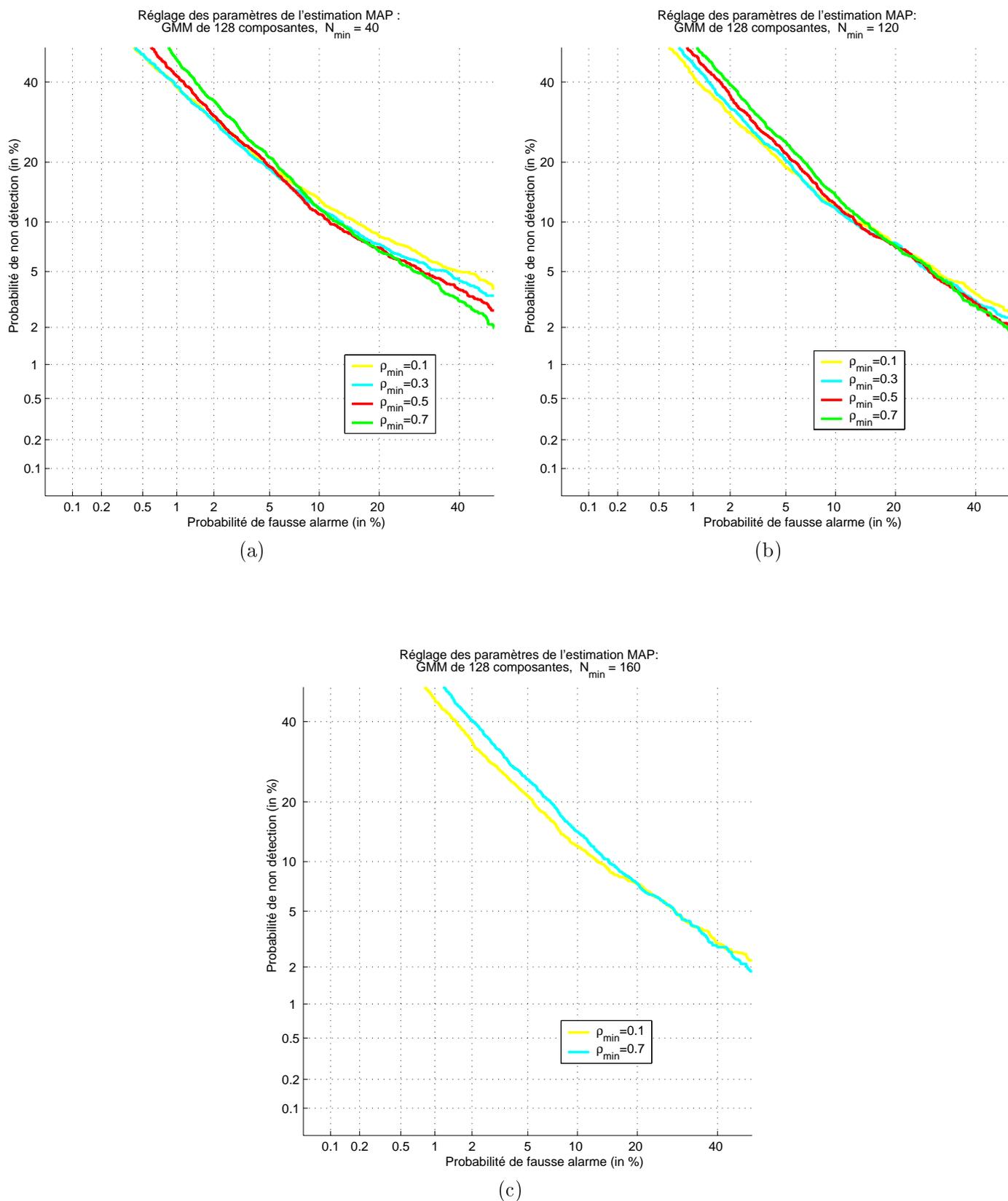
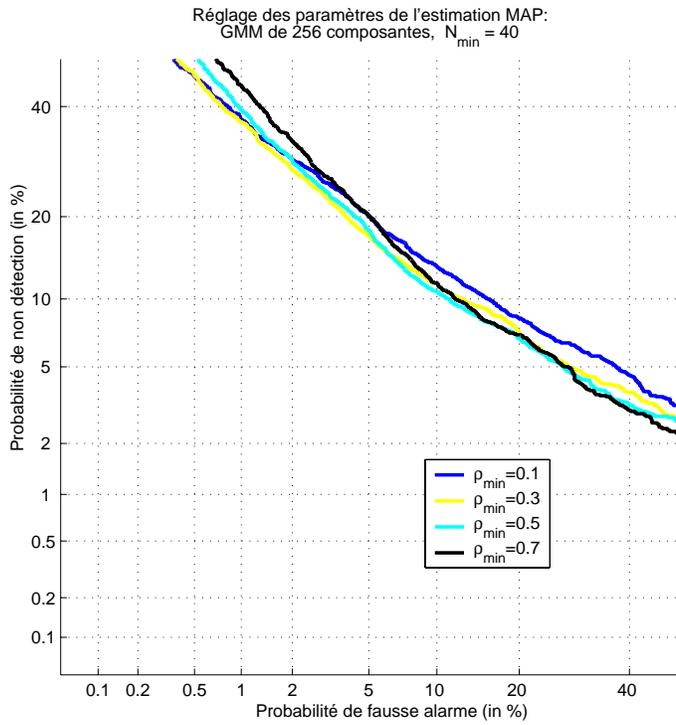
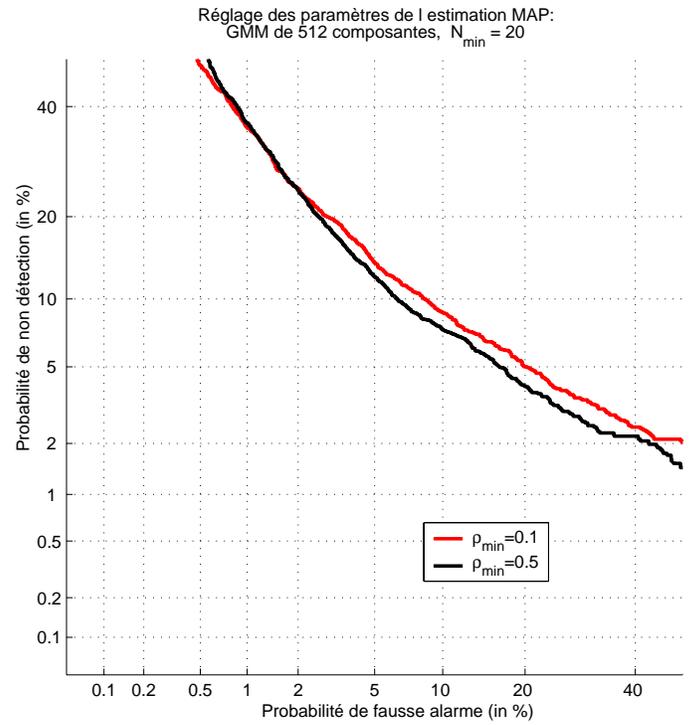


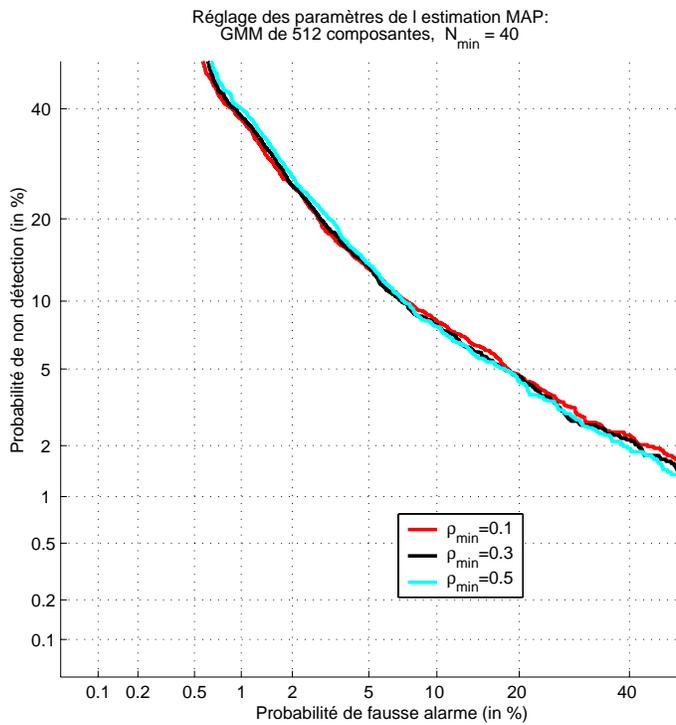
FIG. 7.6 – Estimation MAP des paramètres des GMM : performances d'un système de 128 gaussiennes pour $N_{\min} = \{40, 120, 160\}$ et $\rho_{\min} = \{0.1, 0.3, 0.5, 0.7\}$



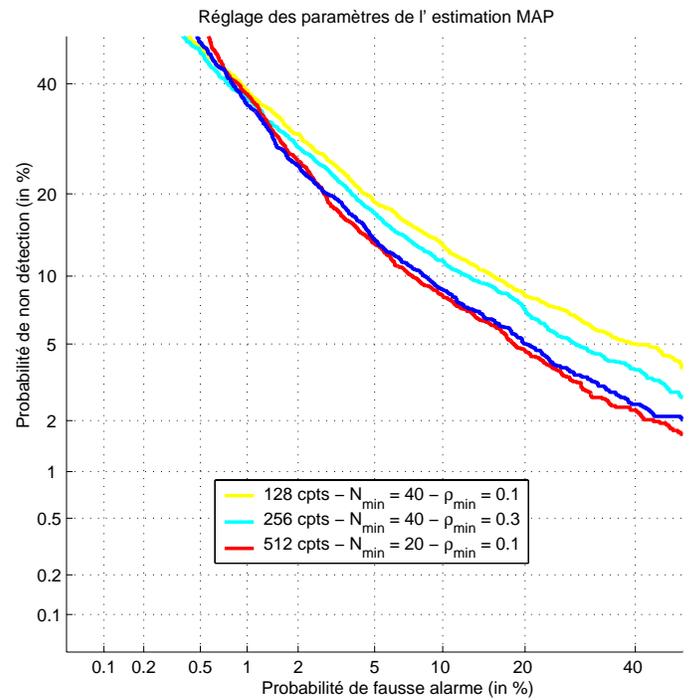
(a.)



(b.)



(c.)



(d.)

FIG. 7.7 – Estimation MAP des paramètres des GMM : performances de systèmes de 256 et 512 gaussiennes pour différentes valeurs de N_{min} et ρ_{min}

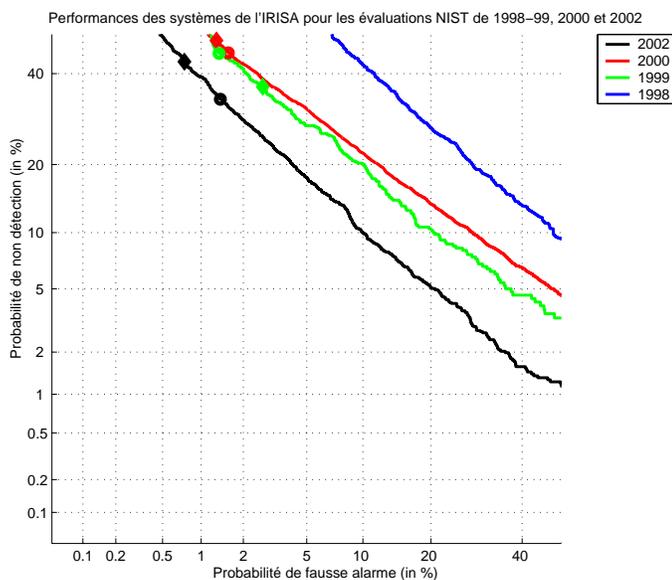


FIG. 7.8 – Résultat aux évaluations NIST des systèmes de l'IRISA en 1998, 1999, 2000 et 2002

Performances aux évaluations NIST

Parallèlement à notre travail sur le module de modélisation, d'autres éléments du système ont été modifiés en collaboration avec tous les membres du consortium ELISA. Les améliorations induites nous ont permis de constamment accroître nos performances au cours des évaluations de 1998 à 2002 malgré l'augmentation de la difficulté du corpus de test (téléphone d'abord fixe puis cellulaire, diminution du nombre de sessions d'apprentissage). Sur la figure 7.8 sont tracées les courbes DET associées pour chaque condition primaire des évaluations de 1998, 1999, 2000 et 2002². En dépit de la relative contre-performance de 2000 qui s'explique par une différence importante des conditions d'apprentissage entre le corpus de développement et celui d'évaluation, on observe une amélioration constante des performances obtenues.

Parmi les travaux autres que ceux portant sur la modélisation effectués dans le cadre du développement de la plateforme commune du consortium ELISA, nous nous sommes particulièrement intéressés au module de normalisation. Nous avons proposé [Ben et al., 2002] une nouvelle technique de normalisation permettant une amélioration des performances comparable à celle offerte par la z -norm avec l'avantage de ne nécessiter aucune donnée supplémentaire. Le chapitre suivant en présente les principales caractéristiques.

²Les performances en 2001 se situaient à un niveau intermédiaire entre celles de 2000 et de 2002, mais les fichiers résultats correspondant ont malencontreusement été perdus.

Chapitre 8

Description d'une nouvelle technique de normalisation : la *d-norm*

La mise en place de l'estimation MAP dans le module de modélisation de la plateforme de VAL du consortium ELISA a contribué aux progrès de ce système dont les performances sont maintenant très proches des meilleurs systèmes actuels. Dans ce chapitre, nous présentons l'autre élément de nos travaux relatifs à l'étude et à l'amélioration de l'état-de-l'art *i.e.* la normalisation du rapport de vraisemblance. Après un bref rappel des motivations ainsi que des caractéristiques des principales mises en œuvre décrites dans la littérature, nous présentons la méthode de normalisation que nous avons développée : la *d-norm*.

8.1 Rappel sur les motivations de la normalisation

La théorie sur les tests d'hypothèses ne permet d'établir un seuil de décision θ optimal indépendant du locuteur que si l'on connaît la vraie distribution des hypothèses H_X et $H_{\bar{X}}$. En pratique ce n'est pas le cas et l'on peut observer une grande variabilité des distributions des scores client $S_X(\mathbf{Y}_X)$ et imposteur $S_X(\mathbf{Y}_{\bar{X}})$. Il n'est alors pas possible de fixer un seuil de décision unique sans obtenir une forte dégradation des performances. Lors d'un accès au système, le score $S_X(\mathbf{Y})$ dépend non seulement de la qualité de l'estimation des paramètres de H_X et $H_{\bar{X}}$ mais aussi de nombreux éléments propres à \mathbf{Y} et souvent indépendants du fait qu'il corresponde à un accès client ou imposteur. Le but de l'étape de normalisation est de définir un ensemble de techniques pour corriger ces sources de variabilité et obtenir un seuil de décision θ le plus indépendant du locuteur possible.

Deux types de normalisation ont été mis au point pour diminuer la variabilité du score et obtenir un seuil de décision unique. Le premier vise à corriger la variabilité du

score liée à la modélisation de H_X ; il s'agit des normalisations de type z -norm [Kung-Pu et Porter, 1988, Reynolds, 1997]. Le second vise à corriger celle liée au signal de test ; il s'agit des normalisations de type t -norm [Auckenthaler et al., 2000]. Elles font toutes les deux l'hypothèse de la gaussianité de la distribution des scores imposteur et visent à les rendre de moyenne nulle et de variance unité.

L'inconvénient principal de la z -norm est de nécessiter un ensemble d'accès imposteur pour estimer les paramètres de normalisation. Malgré le fort gain de performances qu'elle permet d'obtenir [Auckenthaler et al., 2000], le problème majeur de la t -norm est de multiplier le temps nécessaire à la décision par le nombre d'imposteurs nécessaires à l'estimation des paramètres qui lui sont associés.

Nous avons développé la d -norm, une nouvelle technique de normalisation basée sur la distance de Kullback entre les modèles clients et le modèle du monde.

8.2 Description de la d -norm

L'idée de la d -norm est venue de la volonté de mettre en œuvre un outil de diagnostic capable de fournir, en dehors de toute évaluation et pour chaque client X du système, un indice sur la qualité du modèle p_X . Dans ce cadre, nous avons commencé par étudier l'influence de la distance de Kullback $\mathbb{K}(p_X, p_\Omega)$ entre le modèle du locuteur X et le modèle du monde sur la distribution des scores client $S_X(\mathbf{Y}_X)$ et imposteur $S_X(\mathbf{Y}_{\bar{X}})$ pour X . On a :

$$\mathbb{K}(p_X, p_\Omega) = \mathbb{E}_{\mathbf{Y} \sim p_X} \log \frac{p_X(\mathbf{Y})}{p_\Omega(\mathbf{Y})} + \mathbb{E}_{\mathbf{Y} \sim p_\Omega} \log \frac{p_\Omega(\mathbf{Y})}{p_X(\mathbf{Y})};$$

où $\mathbf{Y} \sim p_X$ signifie que les observations \mathbf{Y} sont distribuées suivant p_X .

Ceci nous a conduit à tracer la figure 8.1 sur laquelle sont représentées, pour chaque locuteur du corpus de développement, la moyenne $\bar{S}_X(\mathbf{Y}_{\bar{X}})$ des accès imposteur et celle $\bar{S}_X(\mathbf{Y}_X)$ des accès client en fonction de $\mathbb{K}(p_X, p_\Omega)$. Sur cette figure $\bar{S}_X(\mathbf{Y}_{\bar{X}})$ et $\bar{S}_X(\mathbf{Y}_X)$ sont respectivement représentés par un point et une croix et la modélisation utilisée correspond à un GMM de 128 composantes avec $\rho_{min} = 0.1$ et $N_{min} = 40$. Cette figure fait apparaître une corrélation évidente entre $\bar{S}_X(\mathbf{Y}_{\bar{X}})$ et $\mathbb{K}(p_X, p_\Omega)$ et si l'on suppose qu'il en existe également une entre $\bar{S}_X(\mathbf{Y}_X)$ et $\mathbb{K}(p_X, p_\Omega)$, on peut écrire :

$$\bar{S}_X(\mathbf{Y}_X) \approx a \cdot \mathbb{K}(p_X, p_\Omega) \quad (8.1)$$

$$\bar{S}_X(\mathbf{Y}_{\bar{X}}) \approx b \cdot \mathbb{K}(p_X, p_\Omega) \quad (8.2)$$

L'étape de normalisation consiste à diviser chaque score par $\mathbb{K}(p_X, p_\Omega)$, de manière à pouvoir considérer les scores client et imposteur indépendants de $\mathbb{K}(p_X, p_\Omega)$.

La figure 8.2 représente les scores d -normalisés de la figure 8.1 en fonction de $\mathbb{K}(p_X, p_\Omega)$. L'intérêt principal de cette normalisation est qu'elle ne nécessite aucune

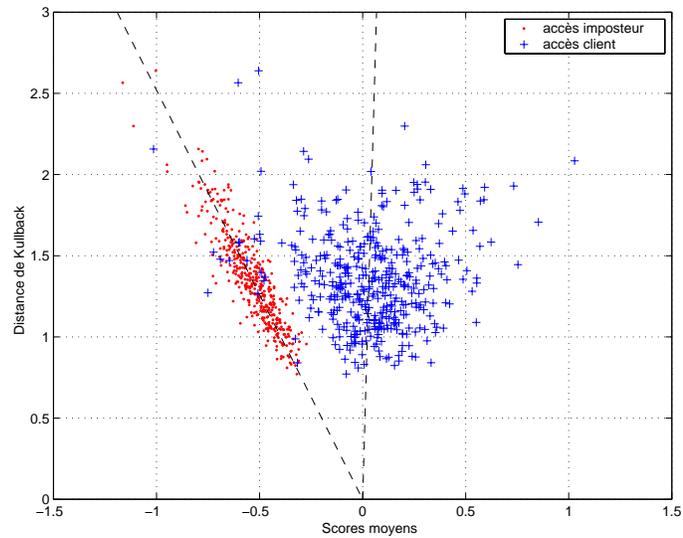


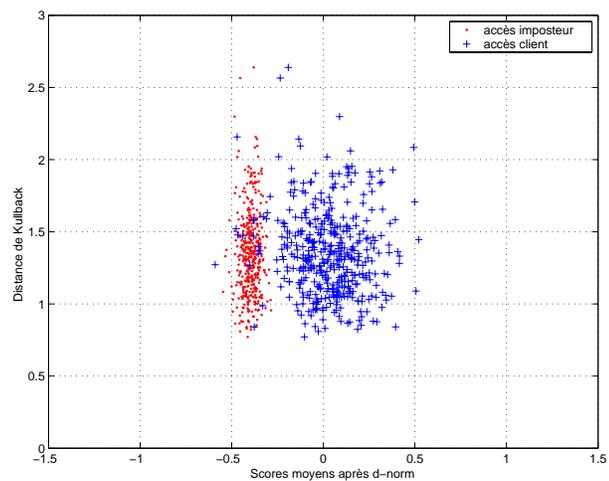
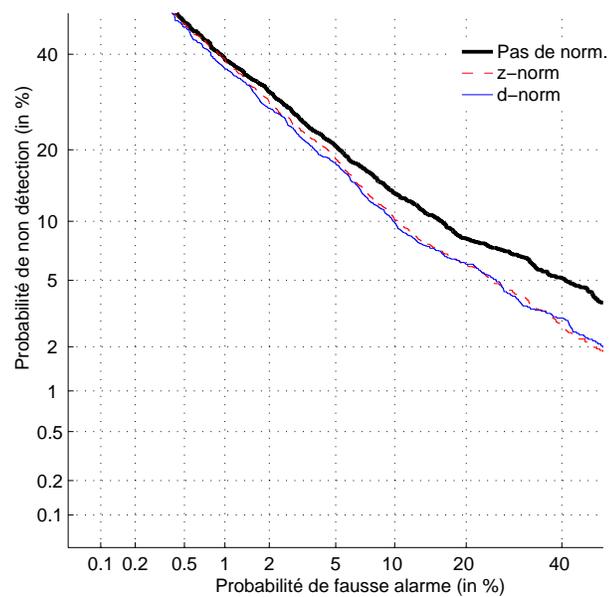
FIG. 8.1 – Scores client et imposteur moyens / distances de Kullback

donnée supplémentaire pour être opérationnelle. En effet, pour chaque locuteur X , $\mathbb{K}(p_X, p_\Omega)$ est estimé à partir de données générées aléatoirement suivant la méthode de Monte-Carlo à partir des modèles p_X et p_Ω .

8.3 Performances

La figure 8.3 présente trois courbes DET obtenues à partir d'un même système de base, sans normalisation, par application de la z -norm et de la d -norm.

Sur cette expérience, la d -norm permet d'obtenir des performances légèrement meilleures que la z -norm, ceci sans nécessiter de données supplémentaires.

FIG. 8.2 – Scores client et imposteur moyens / distances de Kullback *d*-normaliséesFIG. 8.3 – Courbes DET obtenues à partir d'un même système de base, sans normalisation, par *z-normalisation* et par *d-normalisation* des scores

Chapitre 9

Limites et premières solutions

Cette section termine la présentation de nos travaux relatifs à la mise en place et à l'amélioration d'un système de RAL à base de modèles de mélange de gaussiennes. Nous y présentons, dans un premier temps, un bilan des atouts et des limites de cette approche. Puis nous considérons différentes propositions, disponibles dans la littérature, visant à la réduction des ressources matérielles nécessaires à son fonctionnement.

Du point de vue des performances, les GMM sont actuellement la modélisation la plus efficace pour la mise en œuvre d'un système de RAL. En effet, leur forte capacité de modélisation ainsi que la possibilité, grâce à l'estimation MAP, d'obtenir pour chaque locuteur, une adaptation d'un modèle général sont autant d'éléments qui permettent d'expliquer leur succès. On peut cependant relever plusieurs inconvénients dans leur utilisation :

1. d'un point de vue matériel :
 - le calcul du score de décision nécessite un grand nombre d'opérations complexes. Ainsi, pour un signal de test de N trames et une modélisation de K gaussiennes, le système doit calculer $2 \times N \times K$ exponentielles (exp), $2 \times N$ logarithmes (log), $2 \times N \times K \times (2d + 1)$ multiplications (mul) et $2 \times N \times (2d + K)$ additions (add.). Dans ces deux expressions, d correspond à la dimension du vecteur acoustique.
 - la référence caractéristique du locuteur et celle du modèle du monde nécessitent quant à elles une forte capacité de stockage. Pour la modélisation évoquée ci-dessus, celle-ci est de l'ordre de $2 \times 4 \times (2d + 1) \times K$ octets *i.e.* de l'ordre de plusieurs dizaines de kilo-octets pour un système dont les GMM ont 128 composantes.

Ces deux aspects rendent l'utilisation des GMM tels quels rédhibitoire lorsqu'on ne dispose que de quelques centaines d'octets pour stocker la référence caractéristique du locuteur et d'une faible puissance de calcul.

2. d'un point de vue plus conceptuel, le score de décision est obtenu via l'estimation de deux quantités : la vraisemblance du modèle du monde et celle du modèle client. La fonction correspondant au rapport de ces deux quantités est *a priori* de complexité importante mais comme le numérateur est obtenu par adaptation du dénominateur, elle doit probablement pouvoir s'approximer par une fonction plus simple.

Certaines modifications des stratégies de modélisation et de calcul du score, proposées dans la littérature ont permis de réduire la charge de calcul pour l'obtention du score de décision et l'espace mémoire nécessaire au stockage des références caractéristiques des locuteurs. Nous les présentons brièvement dans la sous-section suivante.

9.1 Réduction de la complexité : approches existantes

Nous avons séparé en deux paragraphes les éléments visant à réduire l'espace de stockage de la référence caractéristique du locuteur et ceux visant à réduire la complexité des calculs nécessaire à l'obtention du score de décision.

Réduction de l'espace de stockage

Par espace de stockage, nous entendons la quantité de mémoire nécessaire au stockage de Γ_X , la référence caractéristique du locuteur et celle du non-locuteur.

Une solution fréquemment utilisée et déjà évoquée au début de ce chapitre, consiste à réduire l'ensemble des paramètres des GMM spécifiques au locuteur. Ainsi, on peut contraindre l'apprentissage aux seules moyennes des gaussiennes du mélange, comme dans le cas des expériences reportées à la section 7.4. Si l'architecture du système de RAL est telle que les paramètres de $p_{\bar{X}}$ soient disponibles lors du calcul de $p_X(\mathbf{Y}|X)$, seules les K moyennes spécifiques au locuteur ont besoin d'être stockées. Cette baisse de la taille de la référence caractéristique du locuteur n'entraîne qu'une diminution extrêmement légère des performances du système [Reynolds et al., 2000].

Réduction de la charge de calcul

Les deux principales méthodes pour la réduction de la quantité de calcul nécessaire à l'obtention du score de décision peuvent se trouver dans : [McLaughlin et al., 1999].

La première consiste à ne considérer dans le calcul du score de décision et pour chaque vecteur acoustique que les N_{Best} meilleures gaussiennes, *i.e.* les plus vraisemblables suivant le modèle du monde. Ceci permet de réduire la quantité de calcul d'un facteur proche de $\frac{K+N_{Best}}{2 \cdot K} \approx 0.5$ si $K \gg N_{Best}$. De plus, les différentes expériences que nous avons réalisées montrent que pour $N_{best} \geq 5$, cette réduction du nombre de

composantes effectives dans le calcul du score n'a aucune incidence sur les performances.

La seconde méthode consiste à décimer les vecteurs acoustiques utilisés pour le calcul du score. Ainsi, sur les résultats présentés dans [McLaughlin et al., 1999], l'utilisation d'un quart des observations pour chaque signal de test du corpus ne provoque une augmentation de l'EER que de 10.5% à 10.8%.

Ces solutions, si elles aboutissent effectivement à une réduction de l'ensemble des ressources indispensables au fonctionnement d'un système de RAL, ne permettent toujours pas d'envisager une implémentation efficace sur carte à puce. Nous avons donc réfléchi à la mise au point d'une modélisation dans laquelle on cherche à estimer directement le score en chaque point de l'espace des paramètres. Ceci correspond aux travaux présentés dans les deux chapitres suivants.

Quatrième partie

Approche probabiliste par arbres de décision

Cette dernière partie contient la présentation d'un système de VAL dont l'idée principale est l'estimation directe du score de décision en chaque point de l'espace des paramètres. Dans la mise en œuvre que nous proposons, la phase de modélisation fournit une partition en K régions de l'espace des paramètres et pour chacune d'elle, une fonction simple de calcul du score. L'utilisation d'arbres de décision permet d'obtenir la partition de l'espace des paramètres dans lesquelles nous proposons différentes solutions visant à estimer localement le rapport de vraisemblance entre les hypothèses H_X et $H_{\bar{X}}$.

La collaboration entre l'IRISA et CP8 vise à étudier la possibilité d'intégrer un module de vérification du locuteur sur une carte à microprocesseur. Les contraintes principales d'une telle implémentation concernent la réduction du nombre d'instructions par seconde ainsi que de l'espace de stockage disponible. En effet, le type d'applications visées impose :

- l'utilisation de la carte à puce pour calculer le score de décision,
- l'utilisation de l'espace mémoire disponible sur la carte à puce pour stocker localement la référence caractéristique du locuteur.

Nous avons présenté dans la partie précédente un premier ensemble de travaux. Ceux-ci visaient principalement à rapprocher autant que possible nos performances de celles des meilleurs systèmes présentés aux évaluations NIST. Cette partie propose l'étude et l'évaluation d'une technique de représentation des locuteurs dont les motivations principales sont la réduction des ressources matérielles nécessaires au fonctionnement d'un système de VAL.

Dans ce but, nous avons décidé d'orienter nos travaux vers la modification de l'architecture générale des systèmes de VAL. Celle-ci intervient au niveau des modules de modélisation et de calcul du score. Nous avons cherché à baser la décision non sur l'obtention coûteuse d'un point de vue calculatoire des deux quantités $p_X(\mathbf{Y})$ et $p_{\bar{X}}(\mathbf{Y})$ mais en estimant directement le score de décision en chaque point de l'espace des paramètres \mathcal{Y} . De la rapidité de l'obtention de ce score dans \mathcal{Y} dépend le succès de la méthode. L'approche proposée consiste à partitionner l'espace des paramètres puis à calculer localement le score de décision. Pour un maximum d'efficacité, la partition de \mathcal{Y} est obtenue grâce à l'utilisation d'arbres binaires.

Si, dans un premier temps, les solutions mises en œuvre utilisent les techniques classiques associées aux arbres de décision, nous présentons ensuite différentes solutions originales qui ont permis d'améliorer les performances.

On a fixé trois objectifs à cette partie :

- introduire les concepts et diverses notations de l'approche que nous proposons,

- présenter la solution algorithmique que nous avons retenue pour la réalisation de ce nouveau système de RAL,
- présenter une première évaluation de la technique sur les données des évaluations NIST déjà considérées dans les chapitres précédent.

Aussi, elle se compose de trois chapitres différents.

Dans un premier chapitre, nous donnons l'idée directrice de la méthode ainsi que les motivations qui nous ont conduit à utiliser des arbres binaires et l'algorithme CART (Classification And Regression Tree) [Breiman et al., 1984] pour sa réalisation. Elle se conclut par deux paragraphes portant chacun sur l'une des deux difficultés principales de la méthode : le partitionnement de l'espace des paramètres et l'affectation d'une fonction de score à chacune des partitions.

Le premier chapitre porte sur la réalisation du système de VAL à partir des arbres binaires. Après un premier paragraphe rappelant brièvement certains des travaux similaires publiés dans la littérature, nous présentons les différents éléments sur lequel notre étude s'est focalisée. Enfin, nous décrivons le protocole expérimental utilisé lors des évaluations.

Le second chapitre de cette partie propose une première évaluation des performances obtenues. On y étudie l'influence des différents paramètres de l'algorithme d'apprentissage et d'affectation du score en chacune des régions de l'espace des paramètres. Cette section se termine par un bilan matériel de la méthode proposée ainsi qu'une comparaison avec les GMM.

Enfin, le troisième et dernier chapitre de cette partie propose et évalue une solution de mise en œuvre basée sur l'utilisation d'une représentation par arbres multiples et permettant l'amélioration des performances obtenues précédemment.

Lors de la réalisation et de l'évaluation de ce système de VAL, la partition de l'espace des paramètres et le calcul du score de décision ont été obtenus à partir de l'implémentation de l'algorithme CART fournie par l'Edinburgh Speech Tools Library [Taylor et al., 1999]¹ que nous avons modifiée en fonction des configurations de partitionnement et d'affectation locale de la fonction de score que nous souhaitons obtenir.

¹disponible à l'adresse <http://www.cstr.ed.ac.uk/projects/festival/download.html>

Chapitre 10

Caractéristiques du système de VAL à base d'arbres de décision

Comme on l'a vu au chapitre précédent, les GMM, avec les nombreux paramètres qu'ils nécessitent et la complexité numérique du calcul du score qu'ils impliquent, sont les principaux responsables des besoins matériels des systèmes de VAL. Malgré la croissance constante des capacités de calcul des ressources informatiques, l'utilisation de cette modélisation reste réhhibitoire dans le cadre de l'application visée par notre collaboration avec CP8. En effet, on souhaite d'une part que la référence caractéristique du locuteur soit contenue sur une carte à microprocesseur dont la capacité est de l'ordre de quelques dizaines de kilo-Octets (kO) et d'autre part que la carte à puce prenne en charge le calcul du score de décision. Le but premier du travail présenté ici est donc de réduire conjointement et autant que possible la complexité numérique du calcul du score de décision et la taille mémoire nécessaire au stockage de la référence caractéristique du locuteur.

10.1 Estimation locale du score de décision

La méthode proposée repose sur une représentation du locuteur dont la propriété majeure est de simplifier le plus possible le calcul du score de décision. Dans sa phase d'apprentissage, son principe consiste à partitionner l'espace des paramètres puis à associer, par région, une fonction numériquement simple pour le calcul du score de décision d'un vecteur acoustique. La modélisation induite correspond au partitionnement de l'espace des paramètres et à l'ensemble des fonctions de calcul du score associées.

Plus précisément, partant de l'ensemble des observations acoustiques extraites de l'énoncé d'apprentissage du client et de celles des locuteurs du modèle du monde, on souhaite partitionner l'espace des vecteurs acoustiques en un certain nombre de régions R^k dans lesquelles le score est une fonction numériquement simple de la position du vecteur acoustique dans R^k . En d'autres termes, alors que les techniques classiques

fonctionnent par l'estimation des deux vraisemblances $p_X(\mathbf{Y})$ et $p_{\bar{X}}(\mathbf{Y})$ pour calculer le score de décision, nous souhaitons déduire ce score directement de la position de chacune des observations dans l'espace acoustique.

Pour chaque locuteur X , l'ensemble des régions R_X^k obtenues à l'issue de la phase d'apprentissage définit la partition \mathcal{R}_X de l'espace des paramètres qui lui est associé.

La modélisation directe de la fonction de score doit permettre d'obtenir une forte diminution des ressources matérielles nécessaires à notre application. La figure 10.1 illustre le principe de la méthode dans le cas où un score constant (représenté par différents niveaux de gris) est affecté à chacune des zones obtenues.

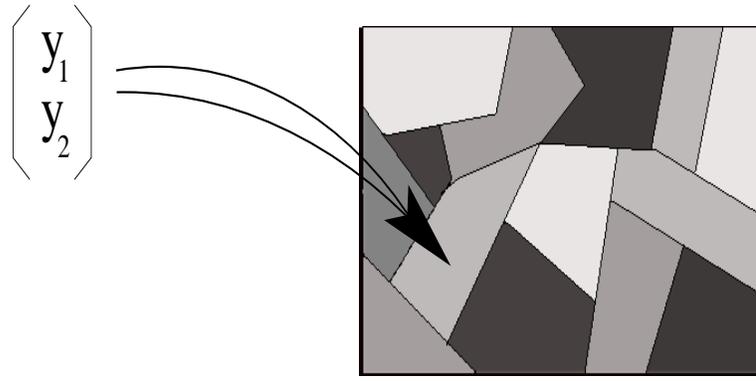


FIG. 10.1 – Illustration du principe de partition de l'espace des paramètres dans le cas de l'affectation d'un score constant, ici représenté par différents niveaux de gris en chaque région

10.2 Formalisme de la mise en œuvre

L'approche proposée comporte deux étapes. La première consiste à obtenir la partition \mathcal{R} de l'espace des paramètres, la seconde à attribuer une fonction de calcul du score de décision à chacune des régions R^k de \mathcal{R} .

Soit $\mathcal{R} = \{R^k\}_{1 \leq k \leq K}$ une partition de l'espace des paramètres \mathcal{Y} :

$$\bigcup R^k = \mathcal{Y} \quad \text{et} \quad R^i \cap R^j = \emptyset \quad \text{pour} \quad i \neq j$$

$Q_X(\cdot)$ est la fonction qui permet d'associer une région de \mathcal{R}_X à chaque point de l'espace des paramètres.

$s_{\mathcal{R}_X}(\cdot)$ est la fonction qui, étant donnée une observation et sa région d'affectation, permet de lui affecter un score de décision.

À l'issue de la phase d'apprentissage, on dispose de la fonction $Q_X(y)$ permettant d'associer une région R_X^k de \mathcal{Y} à toute observation y ainsi que de la fonction $s_{\mathcal{R}_X}(R_X^k, y)$ permettant de leur associer un score de décision. Dans la suite du document, on pourra noter $s_{\mathcal{R}_X}(R_X^k, y) = s_{\mathcal{R}_X}^k(y)$.

La phase de test consiste à affecter à chaque observation y extraite du signal de parole fourni par l'utilisateur à l'une des régions de \mathcal{R}_X puis à considérer $s_{\mathcal{R}_X}^k(y)$ comme contribution associée à y dans le calcul du score de décision.

Les figures (10.2, a) et (10.2, b) représentent respectivement l'architecture d'un système de VAL suivant la mise en œuvre classique et celle que nous proposons. Alors que le formalisme utilisé par la mise en œuvre classique des systèmes de VAL (figure 10.2, a) repose sur l'estimation de deux densités pour affecter un score à un énoncé de test, l'approche décrite sur la figure (10.2, b) est basée sur une estimation directe de ce score en chaque région de la partition \mathcal{R}_X . La référence caractéristique Γ_X du locuteur correspond alors à l'indexation $Q_X(y)$ des régions R_X^k dans l'espace des paramètres et à la fonction $s_{\mathcal{R}_X}^k(y)$ qui leur est respectivement associée : $\Gamma_X = \{Q_X(y), \{s_{\mathcal{R}_X}^k(y)\}_{k=1\dots K}\}$.

10.3 Partitionnement de l'espace des paramètres

On cherche à partitionner l'espace des paramètres acoustiques en régions dans lesquelles le score de décision peut être considéré comme constant ou obtenu par une fonction simple de la position du vecteur dans chacune d'elle. On définit dans ce but c_k et $C(\mathcal{R})$. c_k est une mesure locale à R_X^k d'une certaine propriété de cette région. $C(\mathcal{R})$ est défini tel que :

$$C(\mathcal{R}) = \sum_{k=1}^K p_k c_k$$

avec p_k probabilité *a priori* d'affectation d'une observation à la région R_X^k .

- c_k peut être une mesure de l'homogénéité de la région R_X^k relativement aux deux classes X et \bar{X} . Un exemple est celui de l'entropie :

$$c_k = -p_k(X) \log p_k(X) - p_k(\bar{X}) \log p_k(\bar{X})$$

avec $p_k(X)$ et $p_k(\bar{X})$ probabilités respectives des classes X et \bar{X} conditionnellement à l'affectation de y à R_X^k .

La minimisation de ce critère permet de maximiser l'homogénéité globale de chacune des régions de \mathcal{R}_X . Plus chacun des R_X^k sera homogène à l'une des deux classes, plus c_k et donc $C(\mathcal{R}_X)$ seront petits.

- c_k peut-être une mesure obtenue directement sur les données, *i.e.* indépendamment de leur classe d'appartenance. Ce type de critère peut permettre de regrouper à l'intérieur d'une même région les observations dont les caractéristiques acoustiques sont proches. On pourra par exemple chercher à minimiser la variance des observations à l'intérieur d'une même région, ou à maximiser une distance (comme la distance de Kullback) entre les distributions des données affectées aux différentes régions de \mathcal{Y} .

Dans le cas de la minimisation de la variance des données à l'intérieur de chacune des régions, on a par exemple :

$$c_k = p_k \cdot \prod_{i=1}^d \mathbb{E} (X_k(i) - m_k(i))^2$$

avec $(\cdot)_k(i)$, $i^{\text{ème}}$ coordonnée d'un vecteur $(\cdot)_k$ affecté à R_X^k . Cette approche peut être comparée à la quantification vectorielle dans laquelle chaque locuteur est représenté par un ensemble de K vecteurs, à ceci près que l'on dispose ici d'un moyen d'accès rapide au vecteur *le plus proche* de y , grâce à l'indexation de l'espace des paramètres.

Les figures (10.3, a) et (10.3, b) représentent schématiquement le type de partition que l'on peut espérer obtenir en utilisant un critère du type minimisation de l'entropie (10.3,a) et un critère du type minimisation de la variance (10.3,b). L'un aboutit à une partition des données par rapport à leur classe d'appartenance en minimisant l'hétérogénéité moyenne sur toutes les régions, l'autre à une partition des données minimisant la moyenne sur chaque région d'une mesure de la dispersion des données.

Quel que soit le critère utilisé, la meilleure partition \mathcal{R}_X^* est celle qui minimise $C(\mathcal{R}_X)$:

$$\mathcal{R}_X^* = \arg \min_{\mathcal{R}_X} C(\mathcal{R}_X)$$

Il n'est en pratique pas possible d'obtenir \mathcal{R}_X^* directement et nous avons utilisé un algorithme sous-optimal permettant d'obtenir une partition $\hat{\mathcal{R}}_X^*$, estimation de \mathcal{R}_X^* selon le critère $C(\mathcal{R}_X)$. Cet algorithme de partitionnement de l'espace des paramètres est celui utilisé par la méthode de classification CART proposée dans [Breiman et al., 1984]. Son principe général consiste à diviser récursivement \mathcal{Y} en deux régions (algorithme glouton). À chaque étape la séparation choisie est celle qui maximise la diminution de $C(\mathcal{R}_X)$. La description générale de son fonctionnement est l'objet de la sous-section suivante. Dans la suite du document on note $\hat{\mathcal{R}}_X^* = \mathcal{R}_X$.

Le partage dichotomique et hiérarchique de l'espace des paramètres induit par l'utilisation d'arbres doit permettre de réduire la taille de la référence caractéristique du locuteur. De plus, il assure une affectation rapide d'un vecteur acoustique à chaque région de \mathcal{R}_X .

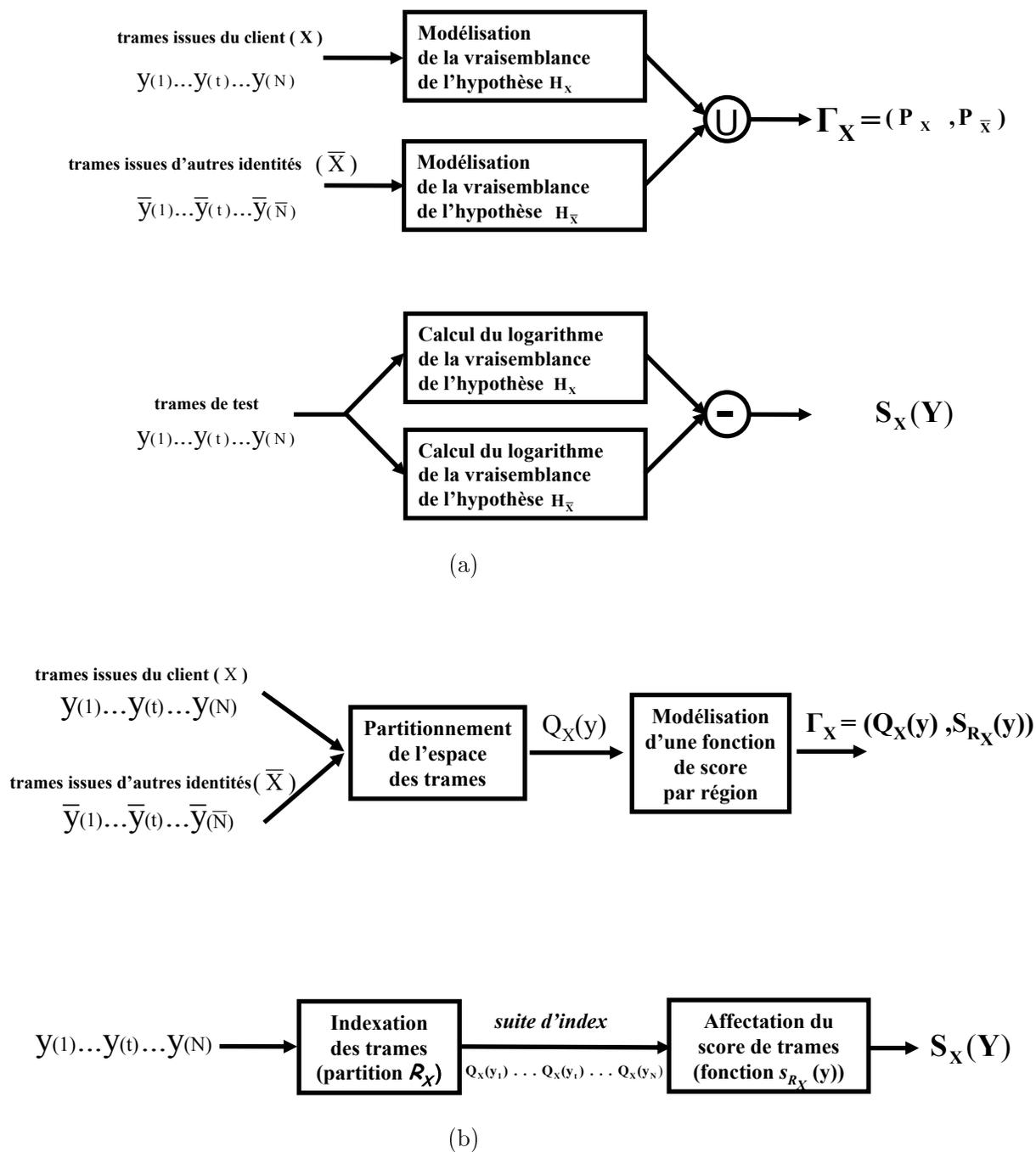


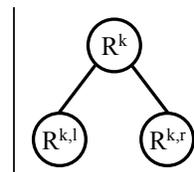
FIG. 10.2 – Architecture des phases d'apprentissage et de test selon la mise en œuvre classique (a) et la mise en œuvre proposée (b)

Description de l'algorithme CART

À chaque étape, la séparation de la région R^k en $R^{k,L}$ et $R^{k,R}$ choisie est celle qui maximise :

$$\Delta c_k = p_k c_k - (p_{k,L} c_{k,L} + p_{k,R} c_{k,R})$$

Sur le dessin de droite, $R^{k,l}$ et $R^{k,r}$ sont appelés les descendants de R^k .



Ce procédé est décrit sur la figure 10.4 (extraite de [Breiman et al., 1984]) dans le cas simple d'un arbre de classification à deux classes. Ainsi, sur cette figure, Y_1 et Y_2 sont disjoints et $Y = Y_1 \cup Y_2$. Les feuilles et les noeuds correspondent à des régions respectivement terminales et non-terminales de la partition. Sur cette figure, les premières sont représentées par des rectangles (Y_4 et Y_5 par exemple), les seconds le sont par des cercles. Chaque séparation est effectuée conditionnellement à l'une des coordonnées de $y_i = (y_1 \dots y_d)$, de manière à maximiser la diminution du critère $C(\mathcal{R}_X)$. Par exemple, sur la figure 10.5 la séparation de Y en Y_1 et Y_2 est du type :

$$Y_1 = \{y; y_1 \leq 0.7\} \text{ et } Y_2 = \{y; y_1 > 0.7\}$$

À chaque ensemble terminal est classiquement associée une classe et dans notre cas une fonction simple de calcul du score. L'ensemble des sous-ensembles terminaux R^k_X définit la partition \mathcal{R}_X de l'espace des paramètres.

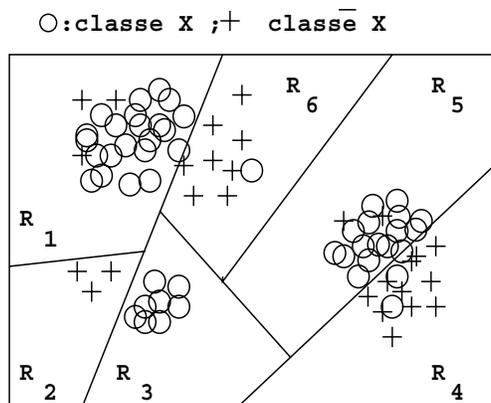
Finalement, l'estimation d'un arbre partitionnant l'espace des paramètres en K régions différentes nécessitent :

1. la définition du critère c_k permettant de définir une propriété de la région R^k_X . À chaque itération, on choisira le partage binaire offrant la plus forte diminution Δc_k .
2. la définition d'un ensemble de questions. Lors du développement de notre méthode, nous avons uniquement considéré des questions portant sur les coefficients des vecteurs acoustiques. Elles sont donc du type :

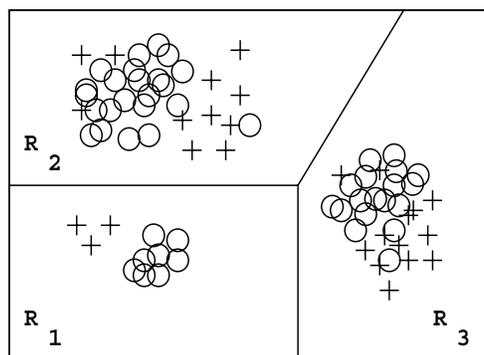
$$y_t(i) < a_i \text{ avec } a_i \in \{\min_Y y(i), \max_Y y(i)\}$$

Cela conduit à un ensemble de régions dont les frontières sont parallèles aux axes canoniques de l'espace de paramètres. On peut espérer que cette simplification ne soit pas en pratique contraignante puisque les coefficients des vecteurs cepstraux sont approximativement indépendants.

3. un critère d'arrêt que nous avons déterminé soit par validation croisée [Breiman et al., 1984] (cross-validation), soit en fixant *a priori* un nombre minimal d'observations par feuille.



(a) : minimisation de l'entropie



(b) : minimisation de la dispersion

FIG. 10.3 – Partitionnement de l'espace des paramètres

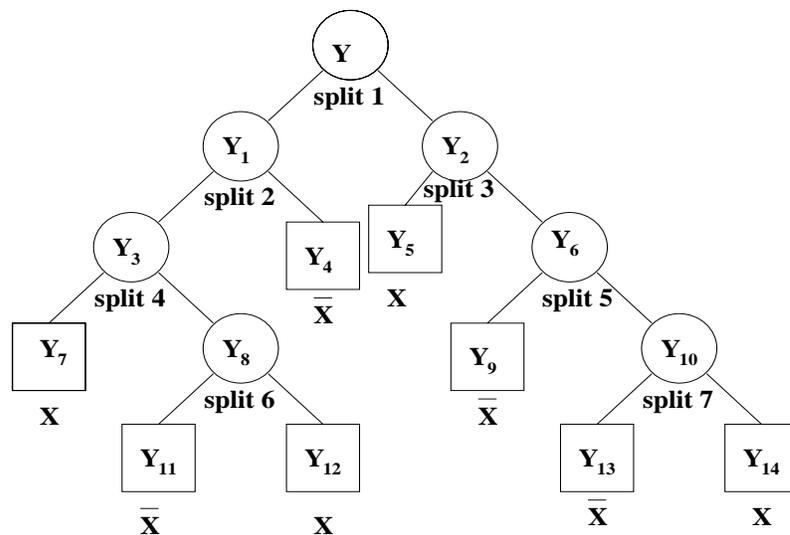


FIG. 10.4 – Exemple d'un arbre de classification pour un problème à 2 classes

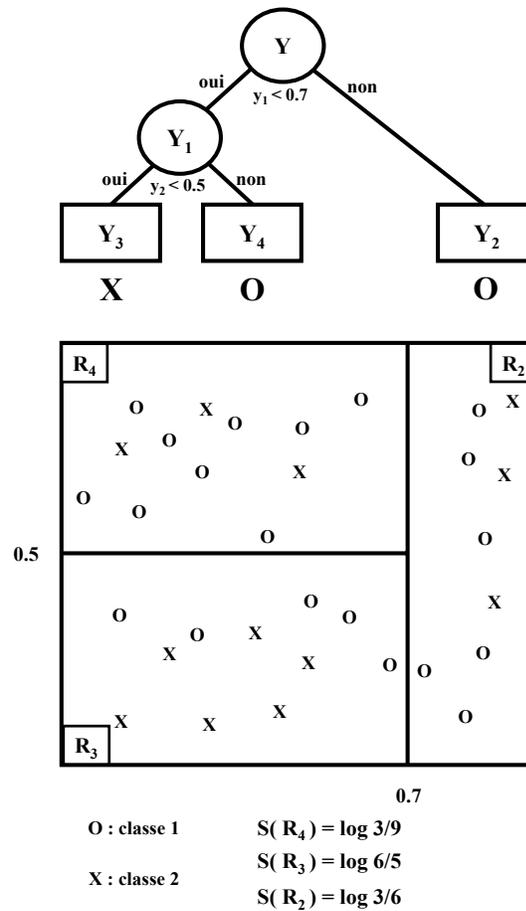


FIG. 10.5 – Classification à deux classes avec des vecteurs de dimension 2

L'estimation de la meilleure partition par partage dichotomique, par l'intermédiaire d'un arbre de décision binaire, permet d'assurer une efficacité maximale en phase de test. En effet, il suffit alors de poser une série de questions simples sur quelques coordonnées du vecteur acoustique pour lui associer un score de décision. De plus elle assure une certaine compacité de la référence caractéristique de chaque locuteur, \mathcal{R}_X étant généralement définie avec peu de partages élémentaires de \mathcal{Y} . L'ensemble des questions nécessaires à l'affectation d'une observation à une région définit $Q_X(y)$.

10.4 Affectation d'un score à chaque région

Le but de cette étape est d'obtenir un score de décision en chacun des R_X^k . Cette fonction notée $s_{\mathcal{R}_X^k}^k(y)$ est définie pour chaque locuteur X et pour chacune des K régions de \mathcal{R}_X . Nous présentons tout d'abord des travaux [Blouet et Bimbot, 2001] dans les-

quels nous avons affecté un score constant à chacune des régions obtenues puis ceux concernant l'affectation d'une fonction de score simple à chaque R_X^k .

Affectation d'un score constant

Dans ce paragraphe, par simplification lors d'une première mise en place de la technique, nous avons considéré les fonctions $s_{\mathcal{R}_X^k}^k(y)$ comme constantes en chacun des R_X^k et l'on note : $s_{\mathcal{R}_X^k}^k = s_{\mathcal{R}_X^k}(R_X^k, y)$. Les différentes techniques d'affectation du score proposées dans ce paragraphe ont été évaluées sur des arbres construits uniquement selon le critère de Gini présenté au chapitre suivant.

Pour chaque R_X^k , les observations de l'ensemble d'apprentissage qui leur sont affectées peuvent provenir du locuteur cible (classe X) ou de tout autre locuteur (classe \bar{X}). Pour chacun d'eux on définit :

$N(k)$: nombre total d'observations de l'ensemble d'apprentissage affectées à R_X^k .

$N_X(k)$ et $N_{\bar{X}}(k)$: nombres d'observations respectives des classes X et \bar{X} de l'ensemble d'apprentissage dans R_X^k . On a $N(k) = N_X(k) + N_{\bar{X}}(k)$.

$P(X|k)$ et $P(\bar{X}|k)$: estimation des probabilités d'avoir une observation de classe X (respectivement \bar{X}) dans la feuille R_X^k .

La partition de \mathcal{Y} et chacun des $s_{\mathcal{R}_X^k}^k$ sont obtenus à partir des N observations d'apprentissage avec :

$$N = \sum_{k=1}^K N_X(k) + \sum_{k=1}^K N_{\bar{X}}(k)$$

À partir de ces différents paramètres nous proposons deux techniques pour l'affectation du score à chacune des partitions de \mathcal{Y} .

La première consiste à attribuer à chacun des R_X^k un score binaire $\{-1, 1\}$ selon la règle suivante :

$$s_{R_X^k}^k = 1 \quad \text{si } \frac{N_X(k)}{N_{\bar{X}}(k)} > 1$$

$$s_{R_X^k}^k = -1 \quad \text{si } \frac{N_X(k)}{N_{\bar{X}}(k)} \leq 1$$

Le score de décision sur l'ensemble des observations s'écrit :

$$S_X(Y) = \frac{1}{N} \sum_{t=1}^N s_{\mathcal{R}_X}(Q_X(y_t), y_t) \quad (10.1)$$

Comme le score est constant en chaque région et que l'on a $s_{\mathcal{R}_X}^k = s_{\mathcal{R}_X}(R_X^k, y)$, l'équation 10.1 peut s'écrire :

$$S_X(Y) = \frac{1}{N} \sum_{t=1}^N \sum_{k=1}^K s_{\mathcal{R}_X}^k \mathbb{1}(Q_X(y_t) = R_X^k) \quad (10.2)$$

Comme $s_{\mathcal{R}_X}^k = \{-1, 1\}$, 10.2 peut s'écrire :

$$S_X(Y) = \frac{1}{N} \sum_{k \in r_X} \sum_{t=1}^N \mathbb{1}(Q_X(y_t) = R_X^k) - \frac{1}{N} \sum_{k \in r_{\bar{X}}} \sum_{t=1}^N \mathbb{1}(Q_X(y_t) = R_X^k) \quad (10.3)$$

avec r_X , ensemble des régions de \mathcal{Y} pour lesquelles le nombre d'observations issues de X est supérieur à celui de celles issues de \bar{X} :

$$r_X = \{R_X^k : N_X(k) > N_{\bar{X}}(k)\}$$

de même, on a $r_{\bar{X}}$ tel que :

$$r_{\bar{X}} = \{R_X^k : N_X(k) \leq N_{\bar{X}}(k)\}$$

Finalement, le score de décision global peut s'écrire :

$$S_X(\mathbf{Y}) = \frac{N_X(\mathbf{Y})}{N} - \frac{N_{\bar{X}}(\mathbf{Y})}{N} = \hat{P}_X(\mathbf{Y}) - \hat{P}_{\bar{X}}(\mathbf{Y}) \quad (10.4)$$

où $N_X(Y)$ correspond au nombre de trames de \mathbf{Y} affectées à l'une des régions de r_X et $N_{\bar{X}}(Y)$ au nombre de celles qui sont affectées à $r_{\bar{X}}$. Si bien que dans ce cas, le score apparaît comme la différence de l'estimation des probabilités des deux classes, suivant les données extraites du signal de test.

Dans la seconde technique que nous proposons pour l'affectation du score à chacune des partitions de \mathcal{Y} , on suppose que les deux densités $p_X(y)$ et $p_{\bar{X}}(y)$ peuvent être approximées par des fonctions constantes en chacun des R_X^k .

$p_X(y)$ et $p_{\bar{X}}(y)$ sont deux densités de probabilité et l'on doit donc avoir :

$$\int_{\mathcal{Y}} p_X(y) = 1 \quad (10.5)$$

$$\text{et, } \int_{\mathcal{Y}} p_{\bar{X}}(y) = 1 \quad (10.6)$$

$p_X(y)$ et $p_{\bar{X}}(y)$ sont constantes dans chaque R_X^k si bien que 10.5 et 10.6 peuvent s'écrire :

$$\sum_k p_X(R_X^k) \cdot \int_{y \in R^k} dy = 1 \quad (10.7)$$

$$\text{et, } \sum_k p_{\bar{X}}(R_X^k) \cdot \int_{y \in R^k} dy = 1 \quad (10.8)$$

De plus, on pose :

$$\int_{y \in R_X^k} dy = A_k \quad (10.9)$$

10.7 et 10.8 s'écrivent :

$$\sum_k A_k p_X(R_X^k) = 1 \text{ et } \sum_k A_k p_{\bar{X}}(R_X^k) = 1 \quad (10.10)$$

Une estimation au maximum de vraisemblance de $p_X(R_X^k)$ et $p_{\bar{X}}(R_X^k)$ est donnée par :

$$p_X(R_X^k) = \frac{N_X(k)}{A_k N_X} \text{ et } p_{\bar{X}}(R_X^k) = \frac{N_{\bar{X}}(k)}{A_k N_{\bar{X}}} \quad (10.11)$$

et le rapport de vraisemblance entre les deux densités s'écrit :

$$\log \frac{N_X(k)}{N_{\bar{X}}(k)} - \log \frac{N_X}{N_{\bar{X}}} \quad (10.12)$$

si bien que le score affecté à chaque région que nous avons utilisé s'écrit :

$$s_{\mathcal{R}_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)} \quad (10.13)$$

Le score de décision global s'écrit :

$$S_X(\mathbf{Y}) = \frac{1}{N} \sum_{t=1}^N s_{\mathcal{R}_X}(Q_X(y), y)$$

soit en procédant de la même manière que lors de l'obtention du score 10.4 :

$$\begin{aligned} S_X(\mathbf{Y}) &= \sum_{k=1}^K \frac{N_k(\mathbf{Y})}{N} s_{\mathcal{R}_X}^k \\ S_X(\mathbf{Y}) &= \sum_{k=1}^K \hat{P}_k s_{\mathcal{R}_X}^k \end{aligned} \quad (10.14)$$

où $N_k(\mathbf{Y})$ correspond au nombre d'observations de \mathbf{Y} affectées à R_X^k .

Au niveau des équations 10.4 et 10.14, on peut espérer améliorer les performances en modifiant intégralement dans les estimations de \hat{P}_X , $\hat{P}_{\bar{X}}$ et \hat{P}_k une information *a priori* sur leur distribution, estimés par exemple sur une population d'imposeurs.

Les processus de partition de l'espace des paramètres et d'affectation d'un score à chacune des partitions obtenues sont schématisés dans un cas simple à deux dimensions sur la figure 10.5.

Simplification pour l'implémentation

Dans le cas d'une fonction de score constante dans chacune des partitions, la désignation de la région R_X^k à laquelle est associée en phase de test chaque observation y s'effectue avantageusement par l'intermédiaire d'un index, noté β_t (avec $\beta_t \in [1, \dots, K]$) correspondant à l'indice de la région R_X^k dans \mathcal{R}_X . Le score individuel $s_{R_X^k}^k$ de chaque vecteur acoustique y_t est alors déduit via le β_t de sa région d'affectation et de son score $s_{R_X^k}^k$. Cette indexation de \mathcal{R}_X trouve son intérêt dans le cas de la réalisation d'un système de VAL distribuée comprenant par exemple un terminal hôte et une carte à puce. En effet, si l'on décide de la répartition des modules telle que celle décrite sur la figure 2.5, elle permet de conserver la référence caractéristique du locuteur sur la carte à puce et de satisfaire les contraintes de taux de transfert des cartes actuelles, pour une application en temps réel. Cette mise en œuvre a fait l'objet d'une demande de dépôt de brevet conjoint entre l'IRISA et CP8.

Affectation d'une fonction de score

Nous proposons ici différentes méthodes pour l'affectation d'une fonction de score simple en chacun des R_X^k . La première technique a été développée et évaluée pour des arbres de décisions construits par minimisation du critère de dispersion d_k présenté au chapitre suivant.

À l'issue de la phase d'apprentissage, on dispose maintenant d'une partition \mathcal{R}_X de l'espace des paramètres ainsi que des estimations m_X^k et Σ_X^k respectivement vecteur moyen et matrice de covariance des observations de \mathbf{Y}_X associées à la région R_X^k lors de la phase d'apprentissage. De même, on dispose pour chaque R_X^k de $m_{\bar{X}}^k$ et $\Sigma_{\bar{X}}^k$, obtenus à partir des observations de $\mathbf{Y}_{\bar{X}}$. \mathbf{Y}_X et $\mathbf{Y}_{\bar{X}}$ sont les suites de vecteurs respectivement extraites d'un signal de parole client et non-client.

La fonction d'affectation du score que nous avons considérée est basée sur le calcul des distances euclidiennes $\mathbf{d}(y, m_X^k)$ entre y et m_X^k et $\mathbf{d}(y, m_{\bar{X}}^k)$ entre y et $m_{\bar{X}}^k$. On a :

$$s_{\mathcal{R}_X}(R_X^k, y) = \frac{1}{2} [\mathbf{d}(y, m_X^k) - \mathbf{d}(y, m_{\bar{X}}^k)] \quad (10.15)$$

avec $\mathbf{d}(u, v) = [\sum_{i=1}^p (u(i) - v(i))^2]^{\frac{1}{2}}$.

Après affectation de la trame acoustique à l'une des régions de \mathcal{Y} , le score qui lui est associé est simplement la différence entre les distances euclidiennes de l'observation avec m_X^k et $m_{\bar{X}}^k$. Cette approche est comparable, dans sa façon de calculer le score de décision, à la quantification vectorielle présentée dans la section 5.2.

Chapitre 11

Description de la plateforme à base d'arbres de décision

Lors du chapitre précédent, nous avons présenté les principes généraux du procédé que nous proposons pour la réalisation d'un système de VAL sous de fortes contraintes matérielles ainsi que le type d'algorithme que nous comptons utiliser pour sa réalisation. Les deux aspects fondamentaux de notre approche sont :

- la représentation de chaque locuteur X par un arbre binaire définissant une partition \mathcal{R}_X de l'espace des paramètres.
- l'affectation d'une fonction simple à chaque région de \mathcal{R}_X pour le calcul local d'un score de décision.

Nous décrivons maintenant de manière plus technique la réalisation de notre plateforme à base d'arbres de décision. Dans ce but nous présentons d'abord les caractéristiques d'autres réalisations de systèmes de RAL utilisant les arbres de décision et publiés dans la littérature. Ensuite, nous exposons notre approche avant de donner le protocole de son évaluation.

11.1 Vérification du locuteur et arbres de décision

On peut trouver dans la littérature plusieurs publications décrivant des réalisations de systèmes de vérification du locuteur utilisant les arbres de décision ou d'une manière plus général d'autres algorithmes de classification binaire. Citons par exemple [Genoud, 1999], [Castellano et Sridharan, 1999] et [Farrell et al., 1994].

Parmi ces différentes publications, nous détaillons [Genoud, 1999] dans laquelle on cherche à réaliser un système de VAL en mode dépendant du texte. La méthode proposée vise à construire, pour chaque mot de la phrase sur laquelle est basée la vérification, N_c arbres de classification de manière à séparer dans l'espace des paramètres les observations du client de celles de chacun des N_c locuteurs (aussi appelé

anti-client) d'une cohorte. Lors de la phase de test, le contrôle du texte est effectué normalement par un système de reconnaissance de la parole qui segmente temporellement la phrase prononcée. Chaque mot m reconnu est alors envoyé à l'ensemble des N_c arbres lui correspondant. Lesquels renvoient un indice entre $[0, 1]$ relatif à la confiance portée à l'hypothèse : le locuteur X a prononcé le mot m . Ensuite, les sorties de chaque classifieur sont fusionnées en un seul score qui est comparé au seuil de décision. Si les locuteurs de la cohorte sont choisis aléatoirement lors de l'apprentissage, plusieurs stratégies de sélection des classifieurs sont proposées lors de la phase de test.

Les quatre principales caractéristiques de cette mise en œuvre sont :

1. Elle fonctionne en mode dépendant du texte.
2. L'hypothèse $H_{\bar{X}}$ est représentée par une cohorte de locuteur qui permet de construire N_c classifieurs binaires. En phase de test, les sorties de ces classifieurs sont fusionnées pour donner le score de décision final.
3. Les arbres de décision construits cherchent à maximiser la *pureté* des données, le critère utilisé est l'entropie. Leur profondeur est fixée par validation croisée.
4. Lors de la fusion des scores, les classifieurs sont sélectionnés pour chaque locuteur en fonction de leurs propriétés discriminantes.

L'inconvénient de cette technique est relatif au nombre d'arbres qu'il faut construire pour chaque locuteur. Ainsi, pour une phrase de N_m mots, il faut construire et stocker $N_c \cdot N_m$ arbres de classifications et calculer autant d'indices de confiance.

11.2 Paramètres de l'analyse

Nous discutons ici des trois principaux paramètres que nous avons considérés pour la construction des arbres. Le premier concernent le critère de partition. Nous en avons utilisé deux dont la description est donnée dans la première des trois sous-sections suivantes. Dans la seconde, nous exposons les différents critères que nous avons évalués pour l'arrêt du partitionnement de \mathcal{Y} . Enfin, la troisième énonce les trois stratégies de représentation de l'hypothèse $H_{\bar{X}}$ que nous avons considérées.

Critère de partition

Nous avons considéré les deux familles de critères de partition présentées au chapitre précédent.

La première porte sur l'homogénéité de la partition R_X^k . Ses deux principaux éléments, largement décrits et utilisés dans la littérature, sont le critère de Gini (\mathcal{G}) et l'entropie (\mathcal{H}). L'utilisation de l'un de ces deux critères est courante pour la construction d'arbres de décision et ils nous ont naturellement servi de point de départ dans nos développements. C'est ce type de critère que nous avons utilisé pour l'affectation d'un

score constant à chacune des régions de \mathcal{Y} .

Le critère de Gini et l'entropie dans la région k de \mathcal{Y} s'écrivent respectivement :

$$\mathcal{G}(k) = p_k(X) \cdot p_k(\bar{X})$$

$$\mathcal{H}(k) = -p_k(X) \log p_k(X) - p_k(\bar{X}) \log p_k(\bar{X})$$

Les deux désavantages principaux de ces critères sont la forte instabilité du processus de partitionnement de \mathcal{Y} et donc celle du score de décision induit [Breiman et al., 1984]. De plus, on peut observer une forte discontinuité du score de décision entre deux régions voisines. Ceci peut entraîner, sous l'effet d'une petite variation du vecteur acoustique, une forte modification du score de décision. Ces deux caractéristiques des critères \mathcal{G} et \mathcal{H} amènent un manque de robustesse lors de la construction de la partition ainsi que lors de l'estimation de score de décision. Notons tout de même que l'utilisation de modules de pre- et de post-processing permet de réduire fortement l'influence de ce défaut.

La seconde famille de critères de partition considérée porte sur la distribution des données à l'intérieur des régions définies par \mathcal{R}_X . Nous avons défini le critère d_k mesurant la dispersion des données à l'intérieur de R_X^k . d_k s'écrit :

$$d_k(y) = \prod_{i=1}^p \mathbb{E} (y(i) - m_k(i))^2$$

c'est-à-dire la variance des données dans la région. À chaque itération, l'algorithme de partitionnement sélectionne les deux régions permettant la plus forte diminution du déterminant de la matrice de covariance, considérée ici comme diagonale. Ce critère permet de regrouper dans une même région les observations proches dans l'espace de représentation acoustique de manière. Les fonctions locales de calcul du score présentées à la section précédente ont été utilisées conjointement à la construction d'arbres selon ce critère.

Critère d'arrêt

Lors de l'utilisation du critère de Gini ou de l'entropie, mis à part la stagnation de la diminution du critère, nous avons considéré les deux stratégies d'arrêt suivantes dans le processus de partitionnement de \mathcal{Y} :

1. la première consiste à fixer *a priori* un nombre minimal d'observations en chacune des régions de \mathcal{R}_X ,
2. la seconde consiste à utiliser la procédure de validation croisée décrite dans [Breiman et al., 1984, chap. 3], pour estimer la taille optimale en terme d'erreur de classification, de la partition de \mathcal{Y} .

Dans le cadre de l'utilisation du critère de dispersion d_k , nous avons uniquement considéré la première stratégie d'arrêt décrite ci-dessus.

Représentation de $H_{\bar{X}}$

La représentation de l'hypothèse $H_{\bar{X}}$ est l'une des difficultés majeures en vérification du locuteur et d'une manière générale de tout système de détection. Nous discutons ici de la représentation de $H_{\bar{X}}$ dans le cadre de notre modélisation basée sur les arbres de décision.

Nous considérons trois stratégies pour la représentation de $H_{\bar{X}}$.

La première est celle, classique, qui consiste à utiliser une cohorte de locuteurs. Cette approche est celle de toutes les mises en œuvre à base d'arbres de décision de système de RAL dont nous avons pu trouver référence dans la littérature. Son inconvénient principal est d'augmenter par autant de locuteurs anti-client la complexité de l'apprentissage et le nombre de tests à effectuer.

La figure (11.1,a) décrit le procédé de construction des arbres binaires selon cette stratégie de représentation de la référence caractéristique du locuteur.

Dans la seconde, nous cherchons à adapter l'approche UBM à l'estimation des arbres de décision. Cette approche a l'avantage de simplifier le calcul du score en permettant de ne considérer qu'un unique modèle de locuteur mais pose le problème de la quantité de données à considérer pour représenter $H_{\bar{X}}$. En effet $\frac{N_{\bar{X}}}{N_X} \ll 1$ peut conduire à fausser l'algorithme d'apprentissage et à fortement dégrader les performances. La méthode que nous avons appliquée consiste en un échantillonnage aléatoire des observations de $H_{\bar{X}}$. À partir de l'ensemble $\mathcal{Y}_{H_{\bar{X}}}$ des observations d'un ensemble de N_c anti-clients nous tirons un ensemble de $N_{\bar{X}}$ observations qui seront utilisées lors de la construction de l'arbre. Lors de l'évaluation nous avons cherché à estimer les performances des systèmes pour différentes valeurs du rapport $r = \frac{N_{\bar{X}}}{N_X}$. La figure (11.1,b) décrit le procédé de construction des arbres binaires selon cette stratégie.

Enfin la troisième stratégie que nous proposons peut être vue comme un mélange des deux premières. En effet, nous considérons alors pour chaque client N_c arbres indépendants et estimés selon la même règle que dans la deuxième. La figure (11.1,c) décrit le procédé de construction des arbres binaires selon une telle représentation de $H_{\bar{X}}$.

11.3 Protocole expérimental

Comme dans le cas de l'évaluation des performances des différentes configurations du système basé sur les GMM, nous avons utilisé le corpus des locuteurs féminins, en-

registrés avec un combiné téléphonique de type électret de l'évaluation NIST 2001. La paramétrisation utilisée est la même que celle utilisée dans la troisième partie du document : 16 coefficients cepstraux, de distribution centrée réduite auxquels sont ajoutés 16 coefficients Δ . L'influence des différents procédés de pre- et de post- processing sur la modélisation que nous proposons a été située en dehors du cadre de ce travail.

L'objectif des évaluations proposées au chapitre suivant a été d'une part d'estimer les performances globales, sur le corpus considéré, d'un système utilisant une modélisation par arbres binaires et d'autre part d'évaluer l'influence des différents paramètres de construction des arbres.

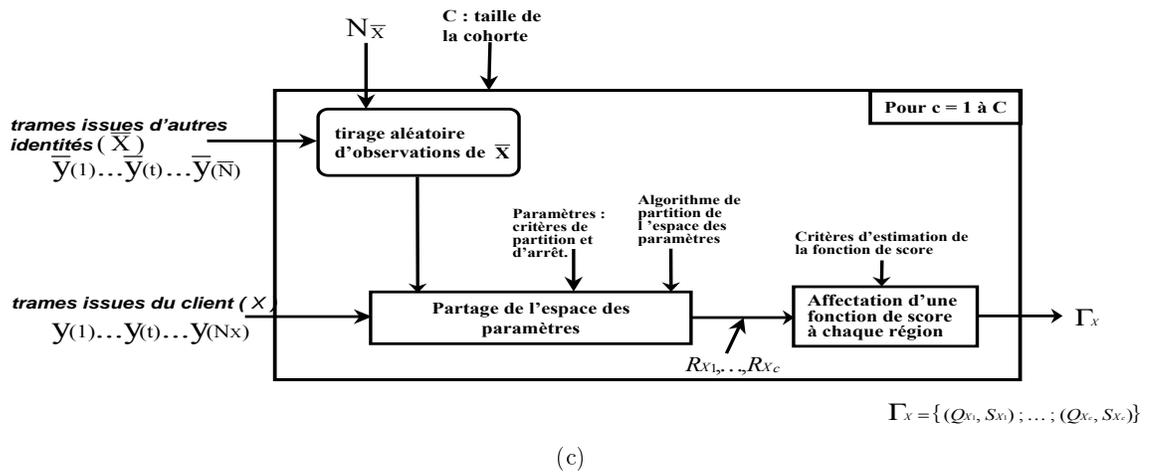
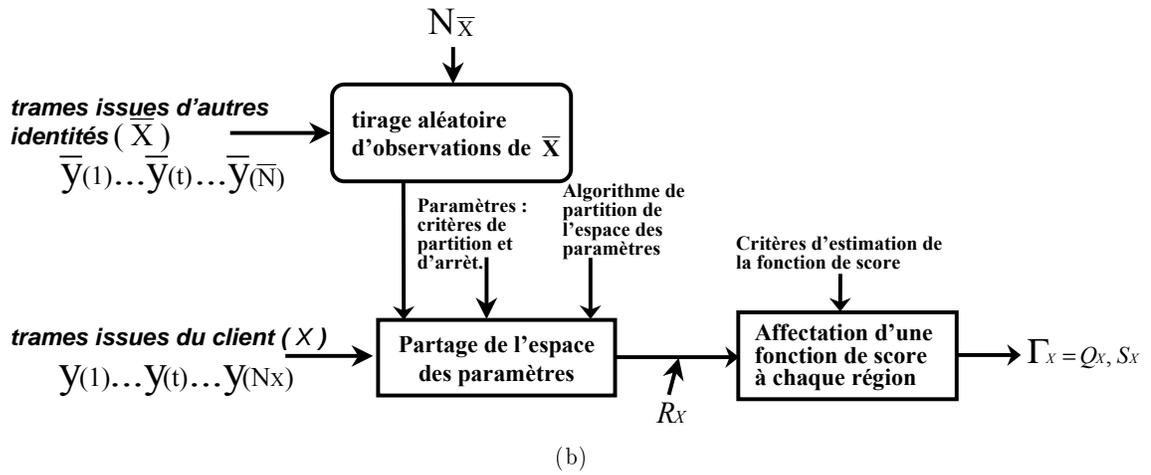
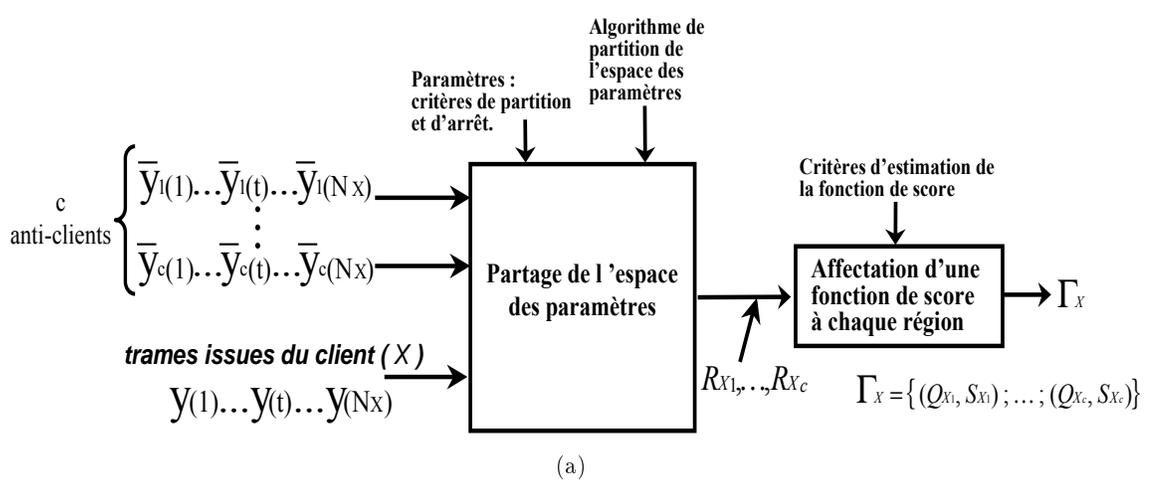


FIG. 11.1 – Trois stratégies de représentation de $H_{\bar{X}}$ pour la construction des arbres de décision

Chapitre 12

Résultats principaux

L'utilisation d'arbres de décision nous permet de représenter des locuteurs par modélisation directe du score de décision dans tout l'espace des paramètres. Le partitionnement de \mathcal{Y} nécessite alors la définition d'un critère qui caractérise certaines propriétés de la partition obtenue. Nous en avons proposé deux types. Le premier, basé sur l'homogénéité des R_X^k , est plus centré sur l'estimation locale des probabilités $P_k(X)$ et $P_k(\bar{X})$ puisqu'il cherche à séparer au maximum dans chaque région les observations de \mathbf{Y}_X de celles de $\mathbf{Y}_{\bar{X}}$. Le second, basé sur la minimisation de la dispersion des observations dans R_X^k , permet de regrouper, par classe de sons, les observations de \mathbf{Y}_X et de $\mathbf{Y}_{\bar{X}}$. Nous avons défini pour chacun de ces critères un moyen d'affecter à chaque région un score de décision. Ceci fait, nous avons ciblé l'autre élément dont nous souhaitons déterminer l'influence : la représentation de $H_{\bar{X}}$.

Cette section concerne l'évaluation de ces deux familles de critères, des fonctions de score qui leur sont associées ainsi que des différents modes de représentation de $H_{\bar{X}}$.

Nous avons séparé en deux sections les résultats obtenus par l'utilisation des deux familles de critères décrits au paragraphe précédent. Dans le premier, nous présentons les résultats obtenus par celui mesurant l'homogénéité de R_X^k alors que dans le second nous présentons celui mesurant la dispersion des observations qui lui sont affectées. Dans la troisième et dernière section de ce chapitre, nous présentons le bilan des besoins matériels de notre approche.

12.1 Évaluation du critère d'homogénéité

Nous proposons dans cette section une série de résultats obtenus pour différentes valeurs de r , rapport entre les nombres $N_{\bar{X}}$ et N_X d'observations extraites de \bar{X} et de X , utilisés lors de la construction des arbres à partir du critère de Gini (\mathcal{G}) ainsi que pour différentes conditions d'arrêt.

On peut trouver dans [Breiman et al., 1984, Chap. 4, section 6] la preuve que la

construction d'arbres suivant le critère de Gini permet de minimiser l'erreur quadratique de l'estimation des probabilités P_X et $P_{\bar{X}}$ en chaque région R_X^k . Nous avons vérifié expérimentalement que des arbres construits suivant \mathcal{G} permettaient d'obtenir systématiquement des coûts de fonctionnement plus faibles que ceux dont la construction est basée sur l'entropie \mathcal{H} . Selon la condition d'arrêt nous avons observé une différence de 0.1% à 0.6% des performances à l'HTER entre \mathcal{G} et \mathcal{H} .

Les tableaux (12.1,a) et (12.1,b) rassemblent les performances obtenues par les différentes stratégies de représentation de $H_{\bar{X}}$, dont nous avons discuté à la section 11.2. Pour chacune d'elle nous avons testé selon les cas un ou plusieurs critères d'arrêt dont les caractéristiques sont précisées dans la première colonne du tableau : il s'agit d'abord de différentes valeurs de N_{min} minorants du nombre d'observations en chaque feuille des arbres puis d'un apprentissage contrôlé par validation croisée. Pour chacun de ces critères, les performances sont estimées par les coûts $HTER$ et C_{Det} tels que décrit dans la section 3.2. Les cinq premières colonnes de ces tableaux correspondent à une représentation de $H_{\bar{X}}$ basée sur un tirage aléatoire de $N_{\bar{X}}$ observations dans l'ensemble des trames du modèle du monde. Cette technique est évaluée pour les valeurs 1, 2, 3, 4 et 5 du rapport $r = \frac{N_{\bar{X}}}{N_X}$.

La colonne dont le premier élément s'intitule *cohorte* présente les résultats obtenus en utilisant un corpus de 10 anti-clients extraits du corpus de NIST 2001.

Enfin, la dernière colonne ($10 \times r = 1$) présente les résultats obtenus en utilisant, pour chaque locuteur, dix arbres construits avec 10 tirages aléatoires et $r = 1$.

Pour les résultats présentés sur le tableau (12.1,a), le score affecté en chaque feuille vaut ± 1 . Il est tel que défini par l'équation 10.1 pour ceux présentés sur le tableau (12.1,b). Pour chacun de ces tableaux, les meilleures performances ont été inscrites en gras.

À l'issue de cette première série d'évaluations, nous pouvons formuler plusieurs constatations :

1. Pour N_{min} , nombre minimal d'observations par feuille, égal à 10 et à 50, l'affectation d'un score binaire permet d'obtenir des résultats meilleurs ou équivalents que lorsque le score est basé sur l'estimation locale des probabilités $P_k(X)$ et $P_k(\bar{X})$, ceci quelle que soit la représentation de $H_{\bar{X}}$. Pour $N_{min} = 100$ et la validation croisée, cette dernière technique d'affectation du score en chaque région permet d'obtenir de meilleures performances.
2. Dans les deux cas, quel que soit le critère d'arrêt pour le partitionnement de \mathcal{Y} et la valeur de r , le tirage aléatoire des données représentant l'hypothèse $H_{\bar{X}}$ permet d'obtenir de meilleurs résultats que ceux fournis par l'utilisation d'une cohorte de locuteurs. Ceci confirme une constatation que nous avons déjà faite dans le cadre

		r=1	r=2	r=3	r=4	r=5	cohorte	$10 \times r = 1$
$N_{min} = 10$	C_{Det}	0.088	0.079	0.078	0.076	0.078	-	0.080
	$HTER(\%)$	20.1	18.3	17.4	16.6	17.0	-	17.9
$N_{min} = 50$	C_{Det}	0.092	0.088	0.089	-	-	0.092	0.086
	$HTER(\%)$	22.5	20.3	23.3	-	-	28.8	19.0
$N_{min} = 100$	C_{Det}	0.096	-	-	-	-	-	-
	$HTER(\%)$	24.0	-	-	-	-	-	-
CV	C_{Det}	0.096	-	-	-	-	-	-
	$HTER(\%)$	24.4	-	-	-	-	-	-

(a)

		r=1	r=2	r=3	r=4	r=5	cohorte	$10 \times r = 1$
$N_{min} = 10$	C_{Det}	0.085	0.081	0.078	0.079	0.079	-	0.080
	$HTER(\%)$	20.2	19.1	17.4	18.2	18.6	-	18.3
$N_{min} = 50$	C_{Det}	0.091	0.088	0.090	-	-	0.094	0.088
	$HTER(\%)$	22.6	21.0	21.9	-	-	29.3	20.0
$N_{min} = 100$	C_{Det}	0.097	-	-	-	-	-	-
	$HTER(\%)$	23.4	-	-	-	-	-	-
CV	C_{Det}	0.090	-	-	-	-	-	-
	$HTER(\%)$	22.8	-	-	-	-	-	-

(b)

TAB. 12.1 – Critère de Gini : C_{Det} et $HTER$ pour différentes configurations de représentation de $H_{\bar{X}}$ avec (a) $s_{R_X}^k$ constant et tel $s_{R_X}^k \in \{-1, 1\}$ et (b) $s_{R_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$

de l'estimation de l'UBM des GMM : la diversité en terme de locuteur des données utilisées pour représenter $H_{\bar{X}}$ est importante que la quantité des données utilisées.

3. Dans les deux cas, quel que soit le critère d'arrêt, la contrainte $N_{min} = 10$ permet d'obtenir les meilleures performances.
4. Dans le cas de l'affectation d'un score dans $\{-1; 1\}$, les performances sont maximales pour $r = 4$, alors qu'elles le sont pour $r = 3$ dans l'autre cas.

Dans le but d'expliquer ces quatre observations, nous avons tracé sur les figures (12.1, a-c), (12.2, a-c) et (12.3, a-c), les moyennes μ_X et $\mu_{\bar{X}}$ des accès client et imposteur (12.1-3, a), leur écart-type (respectivement σ_X et $\sigma_{\bar{X}}$) (12.1-3, b) ainsi que la distance $|\mu_X - \mu_{\bar{X}}|$ (12.1-3, c). Ces grandeurs ont été estimées sur l'ensemble des accès client et imposteur du corpus de développement et pour différentes configurations de construction des arbres.

Sur ces figures sont représentés les accès client et imposteur avec une affectation des scores binaires ± 1 (première stratégie de scoring) ainsi que par l'affectation définie par l'équation 10.1 (deuxième stratégie de scoring).

Les figures (12.1, a-c) et (12.2, a-c) permettent d'analyser l'influence sur les performances de r , rapport du nombre d'observations non-client et client utilisées lors de l'apprentissage de l'arbre. Les figures (12.2, a-c) permettent quant à elles d'expliquer l'influence de N_{min} .

On peut tout d'abord noter que pour $N_{min} = 10$ et $N_{min} = 50$, les comportements de μ_X , $\mu_{\bar{X}}$, σ_X , $\sigma_{\bar{X}}$ et $|S(\mathbf{Y}_X) - S(\mathbf{Y}_{\bar{X}})|$, sont identiques pour les deux stratégies d'affectation des scores lorsque r augmente de 1 à 5. Sur les figures (12.1,a) et (12.2,a), on peut remarquer que les moyennes des scores client et imposteur diminuent lorsque r augmente. Cela s'explique simplement car plus $N_{\bar{X}}$ est grand devant N_X , plus le nombre de régions de \mathcal{Y} favorables à $H_{\bar{X}}$ est important et donc plus la moyenne des scores sur l'ensemble des régions est négative. Cette diminution est nettement plus forte pour les accès dont les scores sont affectés suivant l'équation 10.1. Cette caractéristique, intrinsèque au mode d'affectation des scores dans chacune des régions, n'a cependant pas d'influence sur les performances tant qu'il n'affecte pas la distance moyenne entre les scores client et imposteur. Les figures (12.1,c) et (12.2,c) permettent de visualiser l'évolution de cette distance lorsque r augmente. On constate qu'elle diminue assez fortement pour la première stratégie d'affectation des scores : la baisse est alors d'environ de 50% lorsque r passe de 1 à 5. La seconde stratégie semble moins sensible à l'augmentation de r . Dans ce cas, la baisse de la distance moyenne entre les scores client et imposteur moyen n'est que d'environ 10%.

Quelle que soit la stratégie d'affectation des scores, l'augmentation de r a un effet négatif sur l'estimation des scores client et imposteur et provoque le rapprochement de

ces deux valeurs.

Les figures (12.1,b) et (12.2,b) tracent l'écart-type des deux scores en fonction de r . Elles nous permettent de constater que l'augmentation de r apporte une diminution de l'écart-type des scores clients et impositeurs pour chaque stratégie d'affectation du score. Notons que cette diminution est assez marquée pour la première : environ 50% de baisse lorsque r passe de 1 à 5.

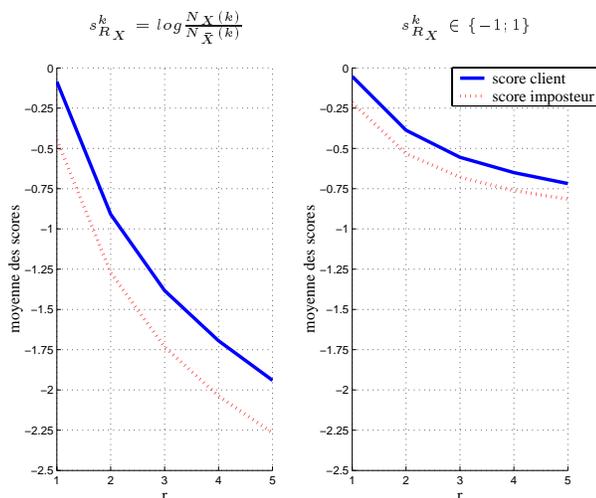
Ces différents éléments permettent de mieux comprendre les résultats présentés dans les premières lignes des tableaux (12.1,a) et (12.1,b). Ainsi, l'augmentation de r provoque une diminution de la distance entre les moyennes des distributions des scores impositeurs et clients ainsi que de leur écart-type. Si la première décroissance a un effet négatif sur les performances, celui de la seconde est positif et, étant plus marquée, permet de les améliorer jusqu'à $r = 4$ (pour $N_{min} = 10$) et $r = 2$ (pour $N_{min} = 50$) pour la première stratégie d'affectation des scores et jusqu'à $r < 2$ (pour $N_{min} = 10$ et $N_{min} = 50$) pour la seconde. De plus, on peut remarquer que les performances sont plus sensibles à l'augmentation de r lorsque $N_{min} = 50$ que lorsque $N_{min} = 10$.

La figure 12.4 correspond aux courbes DET obtenues par un score binaire en chaque feuille pour les critères $N_{min} = \{10; 50\}$ et pour $r = \{1; 2; 3\}$.

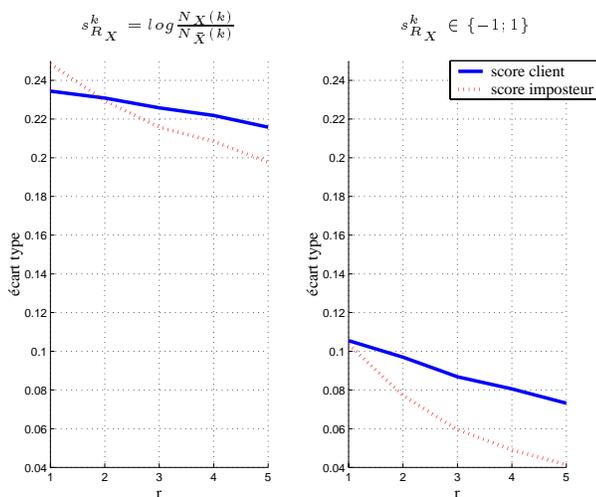
Les figures (12.3,a-c) correspondent aux variations des mêmes grandeurs que sur les figures (12.1-2,a-c) avec des scores obtenus pour $r = 1$ et pour N_{min} prenant les valeurs 10, 50 et 100. Contrairement à ce que l'on a pu constater précédemment avec les variations de r , le comportement des deux stratégies d'affectation du score n'est pas identique lorsque N_{min} varie. Ainsi, si l'on observe dans les deux cas une augmentation de la moyenne des scores associés aux accès client et impositeur lorsque N_{min} augmente, l'écart-type et $|\mu_X - \mu_{\bar{X}}|$ ont, selon les cas, des comportements opposés. Pour $s_{\mathcal{R}_X}^k \in \{-1, 1\}$, ces deux grandeurs diminuent lorsque N_{min} augmente alors qu'elles croissent si l'on considère l'autre technique de calcul des $s_{\mathcal{R}_X}^k$. Malheureusement, dans le cas de la première stratégie de calcul du score, l'augmentation de $|S(\mathbf{Y}_X) - S(\mathbf{Y}_{\bar{X}}|$ étant accompagnée d'une forte augmentation de l'écart-type des scores et dans le cas de la seconde stratégie de scoring la diminution de σ étant accompagnée d'une forte baisse de $|S(\mathbf{Y}_X) - S(\mathbf{Y}_{\bar{X}}|$, on observe dans les deux cas une diminution des performances lorsque N_{min} augmente.

L'une des principales différences entre les représentations obtenues par ces quatre conditions d'arrêt est le nombre de partitions de \mathcal{Y} . De ce point de vue, les arbres construits selon le critère $N_{min} = 100$ et de validation croisée sont très proches. Ceci concourt à expliquer la quasi-similitude des performances obtenues par ces deux approches comme on peut le voir sur le tableau 12.3.

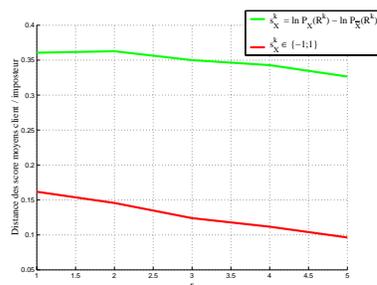
La figure 12.5 correspond aux courbes DET obtenues par un score binaire en chaque feuille pour les critères $N_{min} = \{10, 50, 100\}$ et pour $r = 1$.



(a) Moyenne des accès client et imposteur



(b) Écart-type des accès client et imposteur



(c) Distance entre les scores client et imposteur moyens sur tous les accès

FIG. 12.1 – Moyenne, écart-type et distance des accès client et imposteur, dans la configuration du système $N_{min} = 10$ et $r = \{1, 2, 3, 4, 5\}$

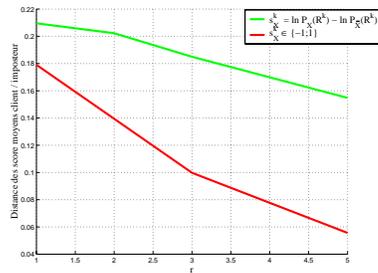
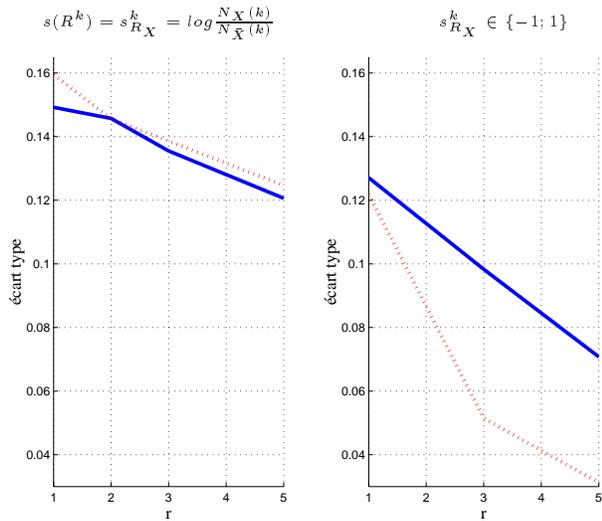
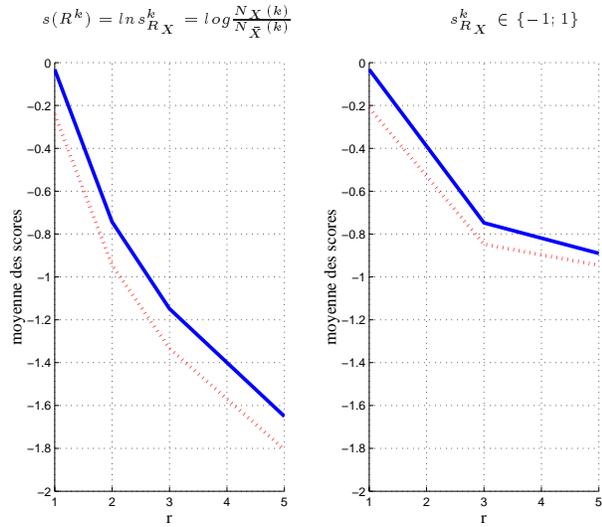


FIG. 12.2 – Moyenne, écart-type et distance des accès client et impoteur, dans la configuration du système $N_{min} = 50$ et $r = \{1, 2, 3, 4, 5\}$

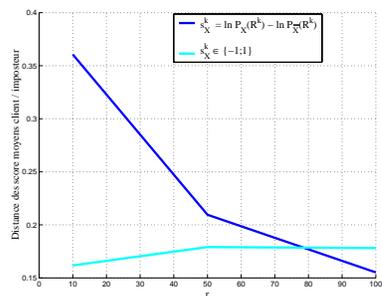
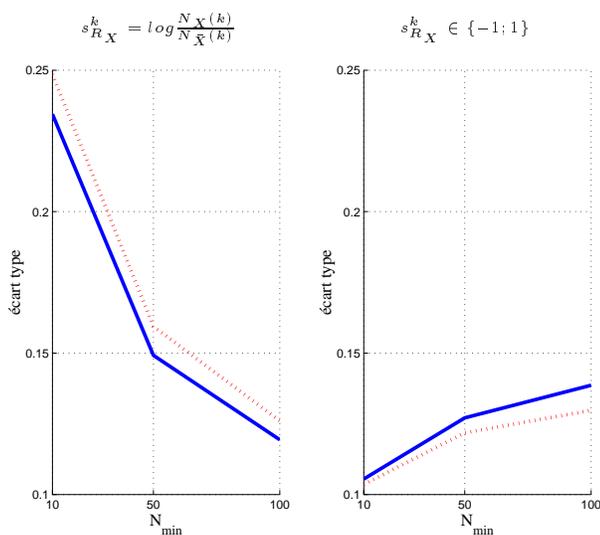
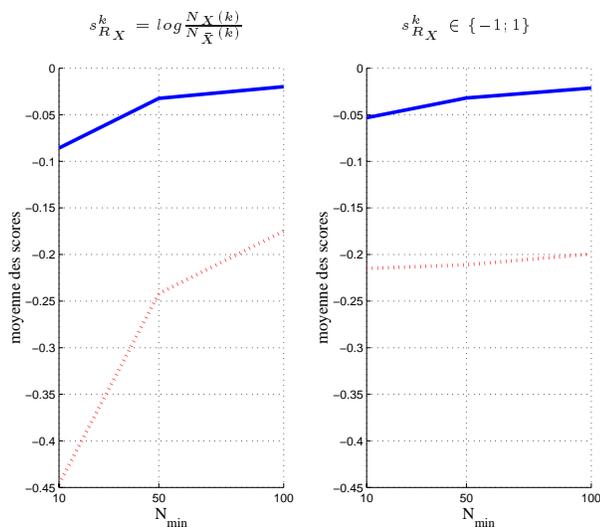


FIG. 12.3 – Moyenne, écart-type et distance moyenne des accès client et imposteur, dans la configuration du système $r = 1$ et $N_{min} = \{10, 50, 100\}$

Même si les performances sont bien inférieures à celles obtenues avec les GMM, on peut remarquer que dans le meilleur des cas, on obtient des résultats équivalents à ceux que nous avons lors de l'évaluation NIST de 1998, ceci nous permet de considérer cette méthode comme prometteuse.

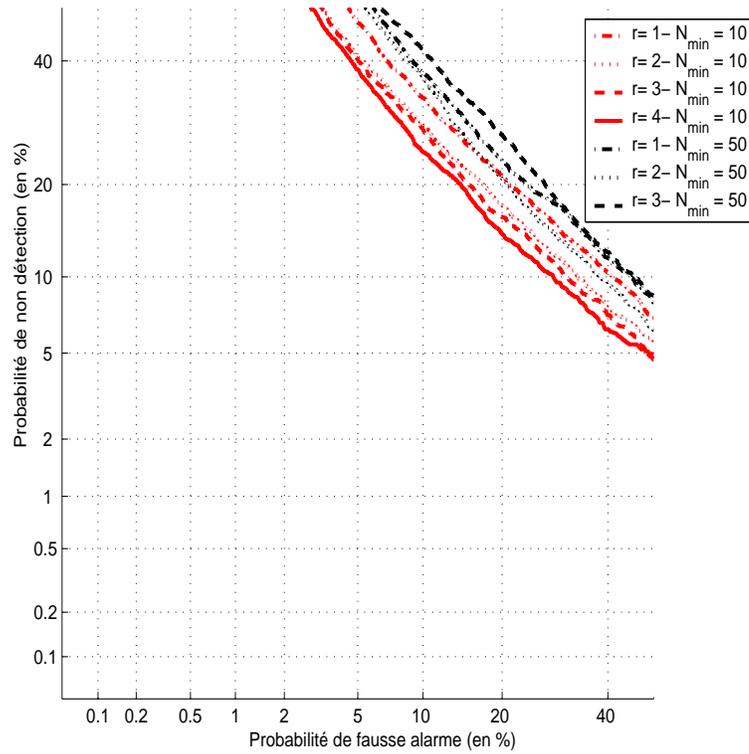


FIG. 12.4 – Critère de Gini - Courbes DET associées aux performances des critères d'arrêts $N_{min} = 10$ et $N_{min} = 50$ pour différents r

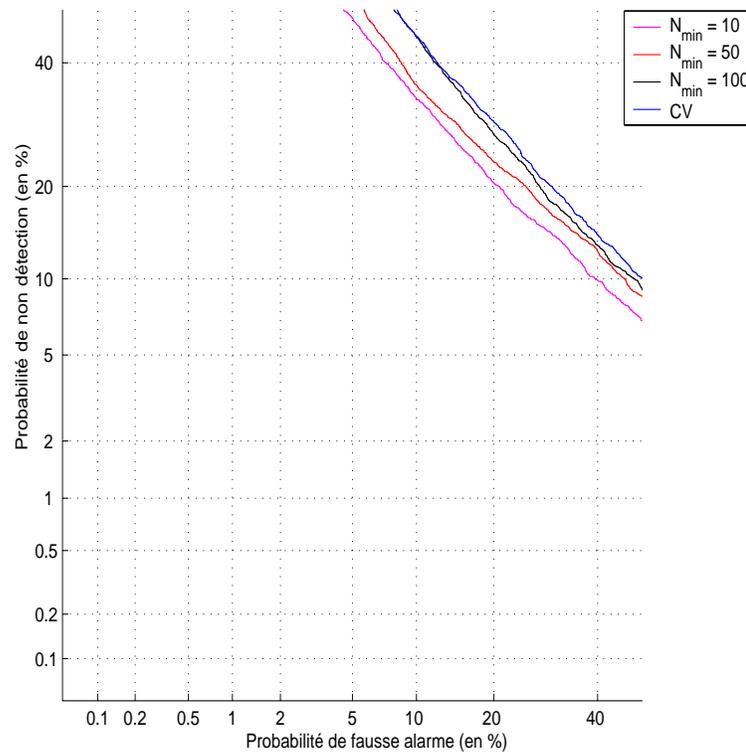


FIG. 12.5 – Critère de Gini - Courbes DET associées aux performances de différents critères d'arrêt du partitionnement de l'espace des paramètres

		r=1	r=2
10	C_{Det}	0.072	-
	$HTER(\%)$	17.8	-
50	C_{Det}	0.069	0.076
	$HTER(\%)$	16.0	18.8
100	C_{Det}	0.070	-
	$HTER(\%)$	16.2	-

TAB. 12.2 – Critère de dispersion : C_{Det} et $HTER$ pour différentes représentations de $H_{\bar{X}}$ avec $s_{\mathcal{R}_X}(R_X^k, y)$ tel que défini dans l'équation 10.15

12.2 Évaluation du critère de dispersion

Cette section présente les résultats obtenus par le critère portant sur la dispersion des observations à l'intérieur de R_X^k .

Le tableau 12.2 présente les résultats obtenus en construisant les partitions de R_X^k par minimisation successive du critère d_k . Lors de la phase de test, le score $S_{\mathcal{R}_X}^k(y)$ affecté à la partition R_X^k correspond au rapport des distances euclidiennes entre \mathbf{Y}^k et $\mathbf{Y}_{\bar{X}}^k$ d'une part et \mathbf{Y}^k et $\mathbf{Y}_{\bar{X}}^k$ d'autre part.

Ce critère permet une forte augmentation des performances et apparaît comme une contribution majeure de nos travaux.

Nous pouvons proposer deux perspectives à cette stratégie d'affectation du score en chaque partition de l'espace des paramètres.

De manière similaire à ce qui a été mis en œuvre pour l'estimation des GMM, on peut envisager, avec l'utilisation du critère sur la dispersion des données, la construction d'une partition de \mathbf{Y} indépendante du locuteur à partir de $\mathbf{Y}_{\bar{X}}$. Lors de l'apprentissage de la référence caractéristique du client, seules les moyennes m_X^k sont estimées et spécifiques du locuteur.

La seconde perspective que nous proposons concerne la phase de normalisation : pour chaque client, nous souhaitons utiliser la distribution $\mathcal{N}_k^S(m_k, \sigma_k)$ des scores imposteurs en chacune des régions.

La figure 12.6 correspond aux meilleures performances obtenues lors de nos évaluations par le critère de dispersion et par le critère de Gini.

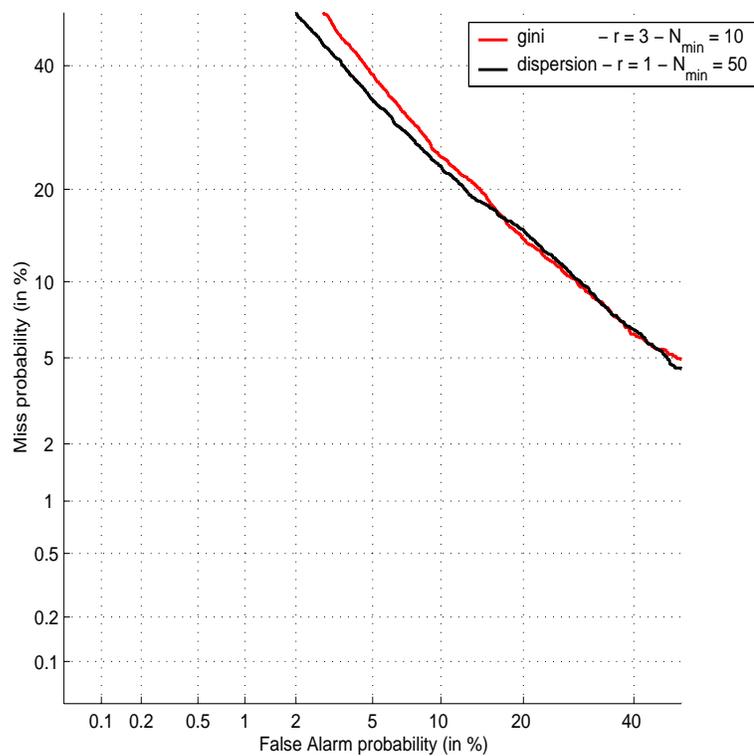


FIG. 12.6 – Critère de dispersion - Critère de Gini : comparaison des meilleures performances

12.3 Bilan matériel

Nous proposons ici l'inventaire des ressources matérielles nécessaires à la phase opérationnelle des différentes configurations de notre système de RAL à base d'arbres de décision. Le bilan concerne la phase opérationnelle du système de vérification du locuteur. Celui-ci comprend une comparaison avec les ressources nécessaires à l'utilisation des GMM. Dans les deux cas, le vecteur de paramètre acoustique est de dimension 32.

Pour chaque configuration présentée à la section précédente sont donnés la taille de la référence caractéristique du locuteur ainsi que le nombre d'opérations élémentaires à effectuer lors d'une phase de reconnaissance du système.

Ressources mémoires et temps de calcul de l'approche à base d'arbres de décision

Sur le tableau 12.3 sont inscrits pour tous les systèmes présentés à la section précédente l'estimation du nombre moyen N_f de feuilles terminales des arbres ainsi que du nombre moyen N_Q de questions nécessaires à l'affectation d'un vecteur acoustique à un R_X^k .

N_f et N_Q sont respectivement estimés sur l'ensemble des 506 clients et sur l'ensemble des tests effectués pour une évaluation.

La taille de la référence caractéristique du locuteur se déduit directement de N_f , puisqu'il permet de définir la partition de \mathcal{Y} . Cette dernière dépend ensuite de l'efficacité du codage de la représentation. Dans le cas des logiciels que nous avons utilisés, elle varie respectivement entre 50 et 300 octets par feuille pour le critère de Gini et le critère de dispersion. Ces deux grandeurs peuvent cependant, sans problème majeur, être réduites de manière à être respectivement de l'ordre de 10 et 140 octets par feuille :

- 1 bit pour décider si la feuille est terminale ou non,
- dans le cas d'une feuille non terminale, on a ensuite 4 octets pour stocker la valeur de la question et 1 octet pour indexer le coefficient du vecteur caractéristique sur laquelle elle porte.
- dans le cas d'une feuille terminale on a alors besoin :
 - pour le critère de Gini de 4 octets pour stocker le score affecté à la feuille,
 - pour le critère de dispersion, en remarquant que le score s_t défini par l'équation 10.15 dans chaque région R^k de \mathcal{Y} peut s'écrire :

$$s_{\mathcal{R}_X}^k = y_t^\dagger \underbrace{[m_{\bar{X}}^k - m_X^k]}_{(1)} - \frac{1}{2} \underbrace{(\alpha_{\bar{X}} - \alpha_X)}_{(2)}$$

avec $\alpha_X = m_X^\dagger \cdot m_X$ et $\alpha_{\bar{X}} = m_{\bar{X}}^\dagger \cdot m_{\bar{X}}$, on a alors besoin de 32×4 octets pour stocker (1) et 4 octets pour stocker (2). Soit finalement 132 octets pour stocker une feuille terminale.

Le nombre d'opérations élémentaires se déduit directement du nombre N_Q de comparaisons pour l'approche utilisant le critère de Gini, car dans ce cas, ces N_Q comparaisons permettent d'obtenir directement le score associé à une observation. Pour les arbres construits en minimisant le critère d_k , il se déduit de N_Q ainsi que des opérations nécessaires au calcul de $s_{\mathcal{R}_X}^k$.

Le tableau 12.4 présente le nombre d'opérations nécessaires au calcul du score de décision pour les critères d_k et g_k en fonction du nombre N d'observations acoustiques extraites du signal de test.

		Gini, r=1	Gini, r=2	Gini, r=3	dispersion
10	N_f	648.3	859.8	963.0	719.9
	N_Q	13.7	14.0	14.3	10.3
50	N_f	133.2	151.4	137.0	132.8
	N_Q	10.2	10.0	9.4	7.2
100	N_f	67.1	-	-	65.3
	N_Q	8.6	-	-	6.0
cross validation	N_f	94.1	-	-	-
	N_Q	8.8	-	-	-

TAB. 12.3 – Nombre moyen de partitions (N_f) et nombre moyen de questions (N_Q) pour différentes configurations du système à base d'arbres de décision

	Comparaisons	Multiplications	Additions	log / exp
g_k	$14.3 \times N$	-	N	-
d_k	$7.2 \times N$	$32 \times N$	$65 \times N$	-

TAB. 12.4 – Nombre d'opérations nécessaires au calcul de décision pour les critères d_k et g_k en fonction du nombre N d'observations de test

Comparaison des ressources matérielles des deux approches

Le tableau 12.5 permet la comparaison concrète entre les besoins matériels de l'approche que nous proposons et l'approche classique. Sur ce tableau, nous avons reporté, pour deux configurations différentes de ces deux approches, la taille de la référence caractéristique du locuteur ainsi que le type et le nombre des différentes opérations nécessaires à la phase de reconnaissance. Ce dernier est reporté en fonction du nombre

	référence caractéristique		opérations			
	mise en œuvre	taille	<	+	×	log / exp
GMM	128	≈ 128 kO	-	$8321 \times N$	$8320 \times N$	$128 \times N$
	256	≈ 256 kO	-	$16642 \times N$	$16640 \times N$	$257 \times N$
arbre de décision	Gini	≈ 10 kO	$14.3 \times N$	N	-	-
	disp.	≈ 20 kO	$7.2 \times N$	$66 \times N$	$32 \times N$	-

TAB. 12.5 – Comparaison entre les ressources matérielles nécessaires à deux systèmes à base de GMM (128 et 256 composantes) et celles nécessaires à deux systèmes à base d’arbres de décision (Gini, $r = 3$) et (dispersion, $r = 1$) avec N , nombres d’observations de test

N de vecteurs acoustiques extraits du signal de test. Les systèmes GMM considérés comportent 128 et 256 composantes dont la structure des matrices de covariance est diagonale. De plus, nous avons considéré que seules les moyennes sont adaptées et propres à chaque locuteur. Les deux systèmes à base d’arbres de décision considérés utilisent, pour le premier le critère de Gini et pour le second celui de la dispersion des données. Ils ont respectivement été construits avec $N_{min} = 10$ et $N_{min} = 500$.

D’un point de vue des ressources matérielles, le gain de notre méthode apparaît important. Le bénéfice en terme de ressources mémoires va d’un facteur 2.5 à 10 alors que celui en terme d’opérations élémentaires (multiplication et addition) varie d’un facteur 500 à 25000. De plus notre méthode ne nécessite l’utilisation d’aucune opération calculatoirement lourde comme les exponentielles ou les logarithmes.

Chapitre 13

Travaux complémentaires et consolidation de la technique

13.1 Motivations

Si du point de vue des ressources matérielles requises, l'approche proposée au chapitre précédent apparaît comme satisfaisante, elle conduit actuellement à une baisse sensible des performances par rapport aux modèles de mélange de gaussiennes. On peut cependant nuancer cette dégradation en remarquant que, comme on peut le voir sur la figure 7.8, nos meilleurs systèmes à base de GMM n'ont été obtenus qu'après plus de trois ans de collaboration entre les différents laboratoires du consortium ELISA, alors que notre mise en œuvre, encore récente, n'a bénéficié que d'à peine plus d'un an de développement.

Dans ce deuxième chapitre consacré à la réalisation d'un système de VAL par modélisation directe du score de décision, nous présentons la mise en œuvre et l'évaluation de quelques propositions visant à améliorer les performances. Les modifications que nous souhaitons mettre en œuvre se basent sur l'analyse des résultats obtenus après les premières évaluations présentées au chapitre précédent. En effet, nous avons constaté que si certaines configurations permettent d'améliorer sensiblement les performances de nos systèmes basés sur les arbres de décision, ces améliorations n'ont lieu qu'au détriment de l'une ou plusieurs des 4 grandeurs dont les valeurs agissent directement sur les performances : les moyennes et variances des scores client et imposteur. Ainsi, dans le cas de $s_X^k = \pm 1$ si l'augmentation du nombre N_{min} d'observations par feuille des arbres permet la réduction de la variance des scores client et imposteur, celle-ci est associée à la baisse de la distance entre les moyennes de ces deux scores et conduit finalement à une baisse des performances. Nous avons donc cherché d'autres stratégies pour la construction des arbres avec, pour objectif, de les réduire simultanément. Cette idée nous a conduit à considérer un des algorithmes fréquemment utilisé dans la littérature des algorithmes de classification : le *boosting*. Cet algorithme permet de construire, à partir du même ensemble de données d'apprentissage un ensemble additif de N_c arbres.

Après la présentation du principe et des propriétés du *boosting*, nous présentons les résultats obtenus par une partition de l'espace des paramètres basée sur un groupe d'arbres obtenu avec cet algorithme. Cette mise en œuvre concerne, dans chaque partition élémentaire, l'affectation d'un score constant et nous l'avons uniquement évaluée pour les arbres construits suivant le critère de Gini.

13.2 Amélioration de l'estimation du score de décision

Le but des travaux présentés dans cette section est de corriger certaines des faiblesses de la première famille de mise en œuvre proposée au chapitre précédent. Celle-ci est basée sur l'utilisation d'un critère d'homogénéité et sur l'affectation d'un score constant en chacune des partitions de l'espace des paramètres.

Nous avons vu que l'une des raisons de la faiblesse de notre technique est liée à la difficulté de réduire simultanément la distance entre les moyennes des scores client et imposteur et leur variance. structurel, nous pouvons distinguer au moins deux causes concourant à l'expliquer. Celles-ci concernent notre stratégie de construction des arbres :

1. l'affectation d'un score constant peut conduire à de fortes discontinuités d'une région à l'autre. Cela amène deux vecteurs acoustiques pourtant proches dans l'espace des paramètres à contribuer de manière opposée au score de décision. La discontinuité des scores à gauche et droite d'une frontière entre deux régions fausse l'estimation locale du score de décision et conduit à l'augmentation de la variance des scores client et imposteur.
2. l'estimation du rapport de vraisemblance en chaque feuille dépend uniquement de l'estimation de \mathcal{R}_X et du nombre d'observations affectées à R_X^k . Or, dans [Breiman et al., 1984, chapitre 5], puis dans [Breiman, 1996, 1997] Breiman met en évidence l'instabilité des partitions produites par les arbres de décision. Il remarque que de petites modifications dans les observations extraites du corpus d'apprentissage peuvent conduire à des arbres radicalement différents. Selon la profondeur de l'arbre, cette instabilité conduit à biaiser plus ou moins fortement l'estimation locale du score de décision en chaque point de l'espace des paramètres.

Ces deux points permettent d'expliquer notre volonté d'utiliser des représentations additives par arbres multiples. D'une part, celles-ci doivent permettre d'améliorer l'estimation du score dans \mathcal{Y} en augmentant la granularité de la représentation. D'autre part, elles doivent permettre de diminuer la variance de ces estimations en lissant la fonction de score (*i.e.* en diminuant les discontinuités du score entre deux régions voisines). L'accumulation additive de différentes partitions doit donc permettre de réduire

simultanément le biais et la variance de l'estimation des densités des classes X et \bar{X} dans l'espace des paramètres.

La section suivante décrit la mise en œuvre d'un système utilisant les arbres multiples.

13.3 Représentation par arbres multiples

Parmi les différentes techniques de construction d'arbres multiples, deux sont principalement décrites dans la littérature : le *bagging* [Breiman, 1996] et le *boosting* [Schapire, 1990, Schapire et Freund, 1997]. Ces deux techniques sont largement sollicitées chaque fois que l'on souhaite accroître les performances d'un classifieur de base (appelé classifieur faible, *weak learner*). Dans notre cas, le classifieur faible correspond à un arbre et chaque composante de la modélisation additive induite correspond à une partition élémentaire de \mathcal{Y} .

Le *bagging* consiste à estimer N_c partitions construites à partir de N_c tirages aléatoires indépendants et avec remise de N observations parmi les N vecteurs d'apprentissage. Il permet d'augmenter la stabilité de la partition globale induite et de réduire les variations du score entre deux régions voisines. Du point de vue de l'estimation du score de décision, sa principale caractéristique est d'en réduire la variance [Pfahringer et Witten, 1997].

Dans le *boosting*, dont le fonctionnement de base est similaire au *bagging*, le $i^{\text{ème}}$ tirage aléatoire des N observations dépend des performances trame à trame du $(i-1)^{\text{ème}}$ arbre binaire. La différence de conception principale entre ces deux procédures réside dans la sélection des données pour la construction d'une des composantes. Alors que les arbres du *bagging* sont construits de manière indépendante après tirage aléatoire suivant une probabilité équiprobable d'occurrence des données de l'ensemble d'apprentissage \mathbf{Y} , on utilise dans le *boosting* les performances des classifieurs antérieurs pour estimer la densité de probabilité d'occurrence des données de \mathbf{Y} . Les modifications, à chaque itération, de la probabilité *a priori* d'apparition de chacune des observations visent à augmenter dans l'ensemble d'apprentissage \mathbf{Y}_i , du $i^{\text{ème}}$ arbre de la représentation, l'influence des observations pour lesquelles le score de décision est le plus éloigné de la valeur désirée.

À chaque itération, on détermine les N vecteurs utilisés pour la construction de la partition par tirage aléatoire suivant la densité de probabilité W d'occurrence des vecteurs acoustiques extraits de Y_X et de $Y_{\bar{X}}$.

W est une densité multinomiale de paramètres $\{\nu_t\}_{t=1}^N$ et l'on a :

$$W(y_1 \dots y_N) = \prod_{t=1}^N y_t^{\nu_t} \text{ avec } \sum_{t=1}^N \nu_t = 1$$

Contrairement au *bagging*, le *boosting* est une procédure purement séquentielle et apparaît [Friedman et al., 1998] comme une technique permettant de réduire conjointement la variance et le biais de l'estimation du score de décision.

La figure 13.1 est une description de la première implémentation du *boosting* basée sur l'algorithme **Discrete AdaBoost** [Schapire, 1990]. À chaque itération i , on construit l'arbre \mathcal{T}_i à partir des observations \mathbf{Y}_i obtenues par tirages aléatoires selon la loi W de vecteurs de $\mathbf{Y} = \mathbf{Y}_X \cup \mathbf{Y}_{\bar{X}}$. On utilise ensuite \mathcal{T}_i pour affecter un score à chacune des observations de \mathbf{Y} . Dans les cas de l'algorithme **Discrete AdaBoost**, les arbres sont des classifieurs binaires et l'on a $\mathcal{T}_i(y_t) : \mathcal{Y} \rightarrow \{-1, 1\}$. L'ensemble des scores obtenus est ensuite directement utilisé pour estimer les paramètres de W nécessaires à l'itération $i + 1$. Dans notre cas, $\{-1, 1\}$ sont respectivement associés à \bar{X} et X . Sur la figure 13.1, $c(y_t) = 1$ si $y_t \in \mathbf{Y}_X$, $c(y_t) = -1$ sinon.

Entrée : N observations d'apprentissage,
 $N = N_{\bar{X}} + N_X$, $\nu_t = 1/N$ pour $t = 1, 2, \dots, N$,
spécification du nombre N_c d'arbre.

Pour $i = 1, \dots, N_c$

- 1.** Construction du $i^{\text{ème}}$ arbre $\mathcal{T}_i(y)$ à partir de l'ensemble \mathbf{Y}_i d'observations tirées aléatoirement parmi celles de \mathbf{Y} , suivant la distribution W des données d'apprentissage,
- 2.** Calcul de $err_i = \mathbb{E}_t[\mathbb{1}_{(c(y_t) \neq \mathcal{T}_i(y_t))}]$, $e_i = \log\left(\frac{1-err_i}{err_i}\right)$
- 3.** Mise à jour des ν_t avec :
 $\nu_t = \nu_t \exp[e_i \cdot \mathbb{1}_{(c(y_t) \neq \mathcal{T}_i(y_t))}]$ avec $t = 1 \dots N$.
Normalisation : $\nu_t = \nu_t / \sum_{t=1}^N \nu_t$

Sortie : calcul du score affecté à l'observation $y_t : \sum_{i=1}^{N_c} e_i \mathcal{T}_i(y_t)$

FIG. 13.1 – Principe de l'algorithme **Discrete AdaBoost**

Dans [Schapire et Freund, 1996, Schapire et Singer, 1998, Friedman et al., 1998] ont été développées différentes variantes de l'algorithme de base du boosting. Dans ces références, l'affectation binaire $\{-1, 1\}$ des scores en chaque région est remplacée par un indice de confiance sur leur prédiction. C'est dans ce dernier article que l'on peut trouver la description des principes et des propriétés de l'algorithme **Real AdaBoost** que nous avons utilisé dans nos propres mises en œuvre.

Dans l'algorithme **Real AdaBoost**, décrit sur la figure 13.2, chaque classifieur faible renvoie à l'itération i du processus d'apprentissage l'estimation des probabilités $P_i(X|y)$ et $P_i(\bar{X}|y)$. Le score de décision final est alors, dans le cas d'un système de classification :

$$\text{sign} \left[\sum_{i=1}^{N_c} \mathcal{T}_i(y) \right]$$

et dans le cas de notre système de vérification du locuteur :

$$s_{\mathcal{R}_X}(y) = \sum_{i=1}^{N_c} \mathcal{T}_i(y)$$

Dans les différentes expériences comparant, le bagging, le boosting suivant les algorithmes **Discrete AdaBoost** et **Real AdaBoost**, décrites dans [Friedman et al., 1998] ce dernier algorithme permet d'obtenir les meilleures performances. On peut trouver dans ce même article une interprétation statistique du *boosting* dont la modélisation induite est équivalente à celle d'une régression logistique. C'est-à-dire qu'il permet d'approximer le rapport $\log(P_X(y)/P_{\bar{X}}(y))$ par un modèle additif $\sum_{i=1}^{N_c} \mathcal{T}_i(y)$ où $P_X(y)$ et $P_{\bar{X}}(y)$ sont les probabilités conditionnelles des classes X et \bar{X} étant donnée l'observation y . Cette caractéristique fait de cet algorithme un outil idéal pour construire la référence caractéristique du locuteur associée à notre représentation, *i.e.* la partition de l'espace des paramètres et la fonction de score associée à chaque région.

Entrée : N observations d'apprentissage,
 $N = N_{\bar{X}} + N_X$, $\nu_t = 1/N$; $t = 1, 2, \dots, N$,
 spécification du nombre N_c d'arbres.

Pour $i = 1, \dots, N_c$

1. Construction du $i^{\text{ème}}$ arbre à partir de \mathcal{Y}_i
 obtenu suivant la distribution W des données d'apprentissage,
 estimation pour $t = 1 \dots N$ de $P_i(X|y_t)$ et de $P_i(\bar{X}|y_t)$

2. On pose alors, $\mathcal{T}_i(y) = \frac{1}{2} \log \frac{P_i(X|y)}{1-P_i(\bar{X}|y)}$

3. Mise à jour de ν_t avec :

$$\nu_i = \nu_i \exp[-c(y_t)\mathcal{T}_i(y_t)]$$

$$\text{Normalisation : } \nu_t = \nu_t / \sum_{i=1}^N \nu_i$$

Sortie : calcul du score affecté à l'observation y_t : $\text{sign} \sum_{i=1}^{N_c} \mathcal{T}_i(y_t)$

FIG. 13.2 – Principe de l'algorithme **Real AdaBoost**, extrait de Friedman et al. [1998]

Nous avons évalué le *boosting*, dans l'implémentation du **Real Adaboost** pour les trois critères d'arrêt ($N_{min} = \{10, 50, 100\}$) présentés au chapitre précédent. Les résultats de la méthode sont présentés à la section suivante.

13.4 Résultats

Le tableau 13.1 présente les résultats obtenus par nos trois critères d'arrêt du partitionnement de l'espace des paramètres $N_{min} = \{10, 50, 100\}$ et en considérant des représentations de 1, 2, 5, 10, 15 et 25 composantes, chacune d'elles étant construite suivant l'algorithme **Real AdaBoost**.

Les meilleures performances à l'HTER et au coût défini par NIST correspondent à des configurations différentes du système. Ainsi pour $s_{\mathcal{R}_X}^k \in \{-1, 1\}$ C_{Det} et le minimum de l'HTER sont atteints pour $N_{min} = 10$, en considérant une vingtaine de composantes. Pour $s_{\mathcal{R}_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$, ils sont respectivement atteints pour $N_{min} = 50$ et $N_{min} = 100$ toujours en considérant une vingtaine de composantes.

Dans certains cas et suivant le critère d'évaluation, on peut noter une dégradation des performances, lorsque N_c augmente par exemple pour $s_{\mathcal{R}_X}^k \in \{-1, 1\}$, $N_{min} = 10$, C_{Det} diminue lorsque N_c passe de 15 à 25. De même, pour $s_{\mathcal{R}_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$, $N_{min} = 50$ HTER diminue lorsque N_c passe de 15 à 25. Cependant celles-ci étant à chaque fois inférieures à 1 %, nous les considérons comme non-significatives.

Pour $s_{\mathcal{R}_X}^k \in \{-1, 1\}$ et $r = 1$, l'utilisation d'*adaboost* permet d'améliorer relativement les performances d'environ 25 % pour une évaluation suivant C_{Det} et d'environ 19 % dans le cas de l'HTER. De plus, on constate que dans ce cas, l'influence de N_{min} s'atténue lorsque N_c , nombre de composantes du modèle augmente. Cela se vérifie sur la colonne de droite des figures (13.3,a-c) où l'on peut observer une convergence des scores client et imposteurs ainsi que de leur écart-type pour $N_{min} = 10, 50$ et 100 lorsque N_c augmente.

Pour $s_{\mathcal{R}_X}^k = \frac{N_X(k)}{N_{\bar{X}}(k)}$ toujours avec $r = 1$, le partitionnement de \mathbf{Y} avec plusieurs arbres permet de diminuer le coût de fonctionnement C_{Det} d'environ 15 % et d'améliorer l'HTER d'environ 22 %. Dans ce cas, on peut observer une disparité des performances selon le critère d'arrêt N_{min} . Celle-ci se vérifie aussi sur les figures (13.3,a-c), colonne de gauche.

D'un point de vue de la distribution des scores client et imposteur, *adaboost* permet de fortement réduire leur variance mais l'on observe une diminution de $|S(\mathbf{Y}_X) - S(\mathbf{Y}_{\bar{X}})|$ lorsque N_c augmente.

Nous avons tracé sur les figures (13.4,a-c) les courbes DET obtenues pour différentes valeurs de N_c et pour $N_{min} = 10, 50$ et 100. Ces figures permettent de visualiser l'amélioration des performances pour chacune des conditions d'arrêt et font apparaître que le gain est maximal pour $N_{min} = 100$, ce que la figure (13.4,d) confirme de manière plus évidente.

L'utilisation d'une représentation des locuteurs par plusieurs arbres améliore considérablement les performances du système de RAL. Nous conjecturons que cette amélioration est principalement due à une réduction de la variance des scores client et imposteur. Cette représentation entraîne une diminution de la distance entre ces scores.

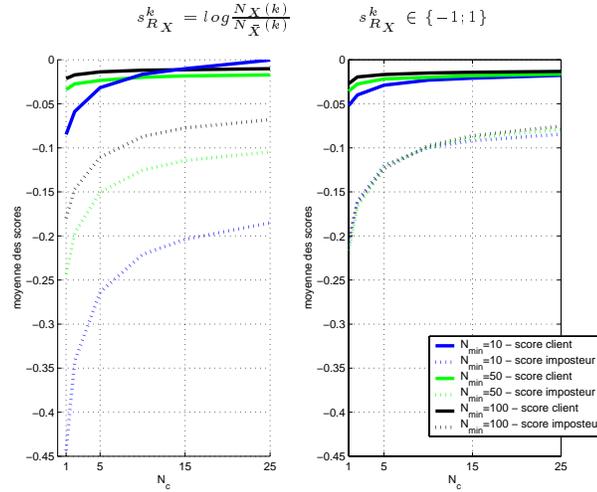
	N_c	1	2	5	10	15	25
$N_{min} = 10$	C_{Det}	0.088	0.080	0.074	0.069	0.068	0.070
	$HTER(\%)$	20.1	18.8	18.0	17.0	16.7	16.4
$N_{min} = 50$	C_{Det}	0.092	0.084	0.075	0.071	0.069	0.066
	$HTER(\%)$	22.5	19.8	18.2	17.4	17.2	16.5
$N_{min} = 100$	C_{Det}	0.096	0.088	0.077	0.075	0.073	0.070
	$HTER(\%)$	24.0	21.6	18.7	17.5	17.3	16.5

(a)

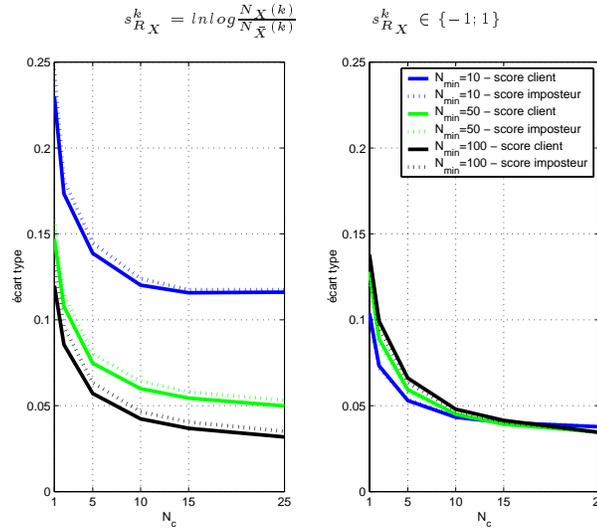
	N_c	1	2	5	10	15	25
$N_{min} = 10$	C_{Det}	0.085	0.079	0.076	0.073	0.077	0.090
	$HTER(\%)$	20.2	19.2	17.9	17.5	17.6	18.5
$N_{min} = 50$	C_{Det}	0.091	0.082	0.074	0.072	0.071	0.074
	$HTER(\%)$	22.6	19.2	17.8	16.7	16.3	16.0
$N_{min} = 100$	C_{Det}	0.097	0.087	0.079	0.074	0.074	0.074
	$HTER(\%)$	23.4	20.3	18.1	16.7	16.5	15.7

(b)

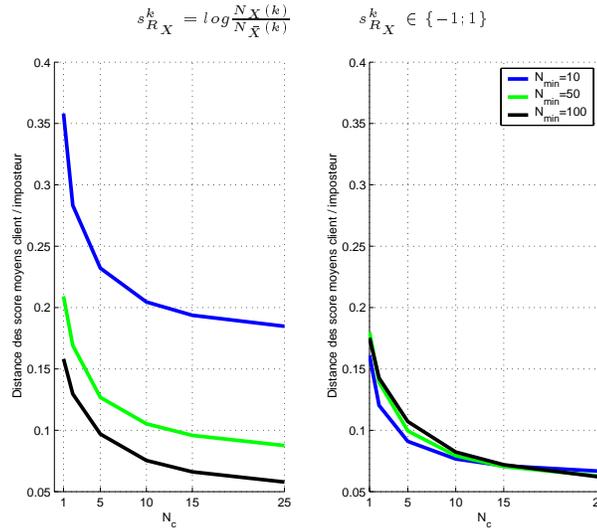
TAB. 13.1 – Critère de Gini - évaluation de l'algorithme **Real Adaboost** : C_{Det} et $HTER$ pour différents critères d'arrêt pour le partitionnement de \mathcal{Y} et en considérant 1, 2, 5, 10, 15 et 25 composantes dans la représentation additive. Pour chaque composante $s_{R_X}^k$ (a) constant et tel $s_{R_X}^k \in \{-1, 1\}$ et (b) $s_{R_X}^k = \log \frac{N_X(k)}{N_{\bar{X}}(k)}$



(a) Moyenne des accès client et imposteur



(b) Écart-type des accès client et imposteur



(c) Distance entre les scores client et imposteur moyens sur tous les accès

FIG. 13.3 – Moyenne, écart-type et distance des accès client et imposteur, dans les configuration du système $N_{min} = \{10, 50, 100\}$ et N_c , nombre de composantes de la modélisation additive, $N_c = \{1, 2, 5, 10, 15, 25\}$

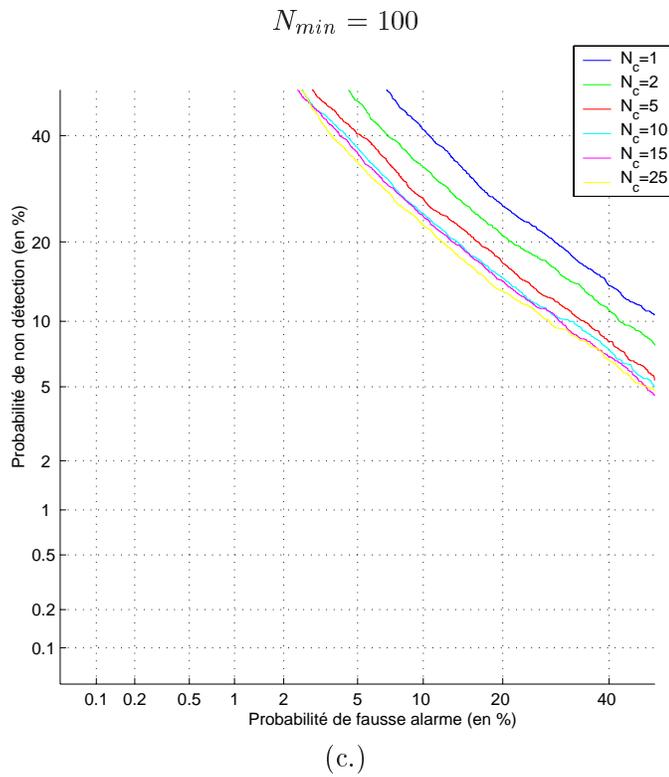
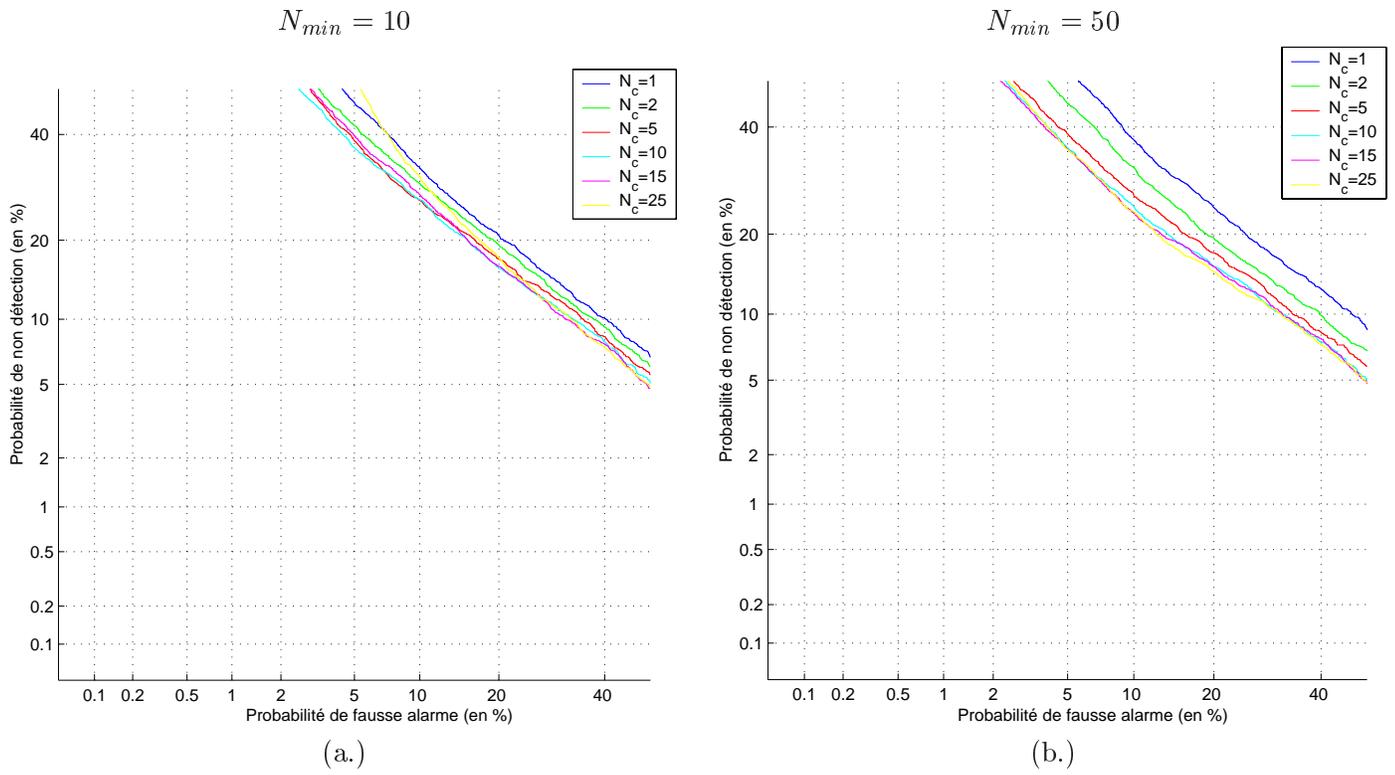


FIG. 13.4 – Courbes DET obtenues pour des systèmes à base de plusieurs arbres de décision pour différents critères d'arrêt et différents nombres de composantes

	$N_{min} = 10$	$N_{min} = 100$		
	$K = 1$	$K = 10$	$K = 15$	$K = 25$
N_f	648.3	436.2	536.8	637.45
N_Q	13.7	55.9	68.8	81.7

TAB. 13.2 – Nombre moyen de partitions (N_f) et nombre moyen de questions (N_Q) pour différents nombres de composantes de la représentation par arbres multiples des locuteurs

13.5 Bilan matériel

Du point de vue des ressources matérielles nécessaires au fonctionnement du système de RAL, avec plusieurs arbres par locuteur, nous pouvons constater que l'utilisation de 25 arbres n'impliquent pas une multiplication par 25 des ressources nécessaires, par rapport à l'utilisation d'un seul arbre.

D'une part, car *adaboost* permet d'obtenir des performances comparables avec $N_{min} = 100$ et avec $N_{min} = 10$ et que dans ce premier cas le nombre de régions de la partition est fortement réduit par rapport au second. Cela implique, comme nous l'avons observé dans la section 12.3, une diminution de l'espace de stockage et du nombre de questions nécessaires à l'affectation du score à une observation.

D'autre part, nous avons observé que la taille des arbres construits suivant *adaboost* nécessite un espace mémoire inférieur, à partir de $N_c = 5$, d'environ 30 % par rapport à celui nécessaire à un unique arbre. Le tableau 13.2 permet de comparer les ressources matérielles nécessaires pour la mise en place de systèmes utilisant 1 arbre avec $N_{min} = 10$ puis 10, 15, et 20 arbres pour $N_{min} = 100$. Sur ce tableau, il apparaît que l'amélioration des performances liée à la représentation par plusieurs arbres des locuteurs s'obtient avec une baisse de la taille mémoire de leur référence caractéristique et un hausse de la quantité de calcul nécessaire à la décision.

Les différentes configurations pour la mise en oeuvre d'un système de VAL à base d'arbres de décision présentées dans les deux chapitres de cette partie, permettent toutes une réduction considérable des besoins matériels nécessaires à leur fonctionnement. Comme dans le cas des GMM, une amélioration importante des performances doit pouvoir être obtenue grâce à l'utilisation d'un a priori sur la référence caractéristique du client. Dans le cas de la méthode proposée, l'information a priori pourra porter sur la partition de l'espace des paramètres et/ou sur la distribution des scores en chaque région. Ces deux points constituent les axes principaux des perspectives de ces travaux.

Conclusion et perspectives

Les travaux présentés dans ce document ont été effectués dans le cadre d'une convention de recherche entre CP8 et le projet METISS de l'IRISA. L'objet de cette collaboration est de mettre en œuvre un système de RAL dont certains des traitements (essentiellement ceux associés au calcul du score de décision) sont effectués sur une carte à microprocesseur.

Au démarrage de cette thèse, nous nous sommes fixés trois objectifs qui nous ont guidés tout au long de nos travaux.

Le premier était l'étude et si possible l'amélioration des systèmes état-de-l'art en vérification du locuteur. Il s'agissait de développer notre propre plateforme de VAL, la plus performante possible. C'est dans ce cadre que nous avons travaillé sur l'estimation suivant le critère du *Maximum A Posteriori* des paramètres des modèles de mélange de gaussiennes et sur la *d-norm*. Ce travail a contribué à la réalisation de la plateforme du consortium ELISA.

Le second objectif était la participation annuelle aux évaluations NIST. Celle-ci, fortement liée à notre implication comme membre du consortium ELISA, nous a permis, entre autre d'évaluer rigoureusement et objectivement les différentes configurations des systèmes de RAL que nous avons mis en œuvre.

Enfin, le troisième objectif était de proposer différentes solutions pour la conception des systèmes de RAL sous de fortes contraintes de réduction de l'espace mémoire et de la puissance de calcul nécessaires à leur fonctionnement.

Étude et amélioration de l'état-de-l'art

L'aspect principal de nos recherches pour l'amélioration des performances de l'état-de-l'art a consisté en l'étude et en la mise en place d'une technique d'estimation MAP des paramètres des GMM. Du point de vue des performances qu'elle permet d'obtenir la méthode que nous avons proposée ne se démarque pas de celles décrites dans la littérature. Cependant, notre reformulation du problème de l'estimation des moyennes du mélange permet d'envisager une amélioration des performances en ajoutant dans le

calcul des contributions un facteur lié à la covariance des observations.

Le deuxième aspect de ce travail a permis la mise en place d'une nouvelle méthode pour la normalisation des scores client et imposteur des systèmes de VAL. Celle-ci permet une amélioration des performances comparables à celle de la *h-norm*, sans nécessiter aucune donnée supplémentaire.

Réduction des ressources matérielles nécessaire à un système de VAL

Dans le cadre de la réduction des ressources matérielles nécessaires à la phase opérationnelle d'un système de VAL, nous avons proposé une stratégie originale pour la mise en œuvre d'un système de RAL. Celle-ci est basée sur l'estimation de fonctions locales pour le calcul direct du score de décision et non sur le calcul de deux vraisemblances, l'une associée à l'acceptation, l'autre au rejet de l'identité proclamée.

L'approche mise en œuvre se compose de deux étapes. La première correspond au partitionnement, grâce à l'algorithme CART, de l'espace des paramètres acoustiques. La seconde consiste à affecter à chacune des régions une fonction de calcul du score de décision. Un des aspects importants de notre étude a été, pour les différents systèmes présentés l'étude systématique de l'influence de la représentation de l'hypothèse $H_{\bar{X}}$ dans la construction de la référence caractéristique du locuteur.

L'un des résultats obtenus pendant ce travail permet de mettre en évidence le gain apporté par le tirage aléatoire des observations de $H_{\bar{X}}$ par rapport à l'utilisation d'une cohorte de locuteurs.

Dans la première configuration du système, nous avons considéré le critère de Gini pour obtenir la partition de l'espace des paramètres. Dans ce cas, le calcul du score de type 'vote majoritaire' sur l'ensemble des observations acoustiques permet, quel que soit le critère d'arrêt et le type de représentation de $H_{\bar{X}}$, d'obtenir des performances supérieures à celles de l'estimation locale du rapport de vraisemblance constant dans chaque région.

Puis nous avons proposé et évalué un autre critère de partition basé sur la distribution des observations acoustiques pour lequel il s'agit de minimiser, à chaque division binaire, la variance des observations affectées à une même région. Pour calculer le score de décision, nous avons alors considéré la distance euclidienne entre les vecteurs de test et les moyennes des observations client et non-client associées à chaque région. Cette approche a permis de fortement améliorer les performances par rapport à celle de la méthode précédemment proposée.

Notre dernière proposition pour la conception d'un système basé sur les arbres de décision repose sur l'algorithme *adaboost*. Celui-ci, évalué uniquement dans le cadre du

critère de Gini, nous a permis de fortement améliorer les performances. L'estimation locale du rapport de vraisemblance constant dans chaque région permet, dans ce cas, d'obtenir des performances supérieures à un score binaire valant $\{-1, 1\}$.

En supprimant complètement du calcul du score de décision les opérations de type multiplication, logarithme et exponentielle, l'approche que nous proposons permet d'obtenir des performances équivalentes à celles que nous obtenions lors de l'évaluation NIST de 1999. Elle permet de réduire conjointement les ressources mémoires et la puissance de calcul nécessaires au fonctionnement d'un système de VAL et d'utiliser la VAL pour des applications où son implémentation n'était pas envisageable.

Perspectives

Une perspective applicative de notre méthode concerne son adaptation à d'autres domaines de la VAL tels que l'indexation ou la recherche d'information dans un document sonore. Une autre est l'intégration des travaux de thèse de Dominique Genoud avec ceux présentés aux chapitres 10 à 13 pour concevoir un système de VAL en mode dépendant du texte.

Les perspectives technologiques que nous proposons concernent principalement notre travail sur la modélisation directe du score de décision dans l'espace des paramètres. Une perspective globale à toutes les configurations à base d'arbres de décision que nous avons proposées consiste en l'utilisation d'un a priori sur la référence caractéristique du client. Dans ce cas, l'information *a priori* pourra porter sur la partition de l'espace des paramètres et/ou sur la distribution des scores en chaque région.

Dans le cadre de l'utilisation du critère de Gini, nous proposons d'utiliser l'algorithme *adaboost* en augmentant le rapport r du nombre d'observations de type non-client et client utilisées dans la phase d'apprentissage. Ces deux points ont séparément permis d'améliorer les performances et leur combinaison semble prometteuse.

Le critère de dispersion a permis d'obtenir une partition de l'espace des paramètres basée sur les propriétés acoustiques des observations. Nous pouvons proposer plusieurs perspectives à l'utilisation conjointe de ce critère et d'une méthode de calcul du score basée sur le rapport de vraisemblance entre les densités mono-gaussienne associées localement au client et au non-client.

La première se déduit directement de la remarque ci-dessus. Si l'utilisation du critère de dispersion permet de disposer d'une partition de l'espace acoustique, on peut construire, uniquement à partir d'observations extraites de \bar{X} , un arbre indépendant du locuteur. Lors de la phase d'apprentissage, les vecteurs acoustiques extraits du signal de parole sont affectés aux différentes régions de \mathbf{Y} puis utilisés pour estimer localement à chacune des régions de la partition les paramètres qui lui sont associés. Chaque

locuteur est alors caractérisé par un ensemble de K moyennes, chacune associée à une région de \mathbf{Y} .

L'autre perspective que nous proposons consiste à approximer localement le score de décision fourni par les GMM, soit en utilisant des arbres de régression, soit en construisant un arbre permettant d'associer à chaque point de l'espace des paramètres les n gaussiennes les plus vraisemblables du modèle du monde. Dans de dernier cas, le score de décision s'obtient alors en considérant les n gaussiennes du modèle client et celle du modèle du monde pour calculer le rapport de vraisemblance.

Les approches présentées dans cette thèse n'apportent certes pas une solution totalement consolidée au problème de l'implémentation de la VAL sur carte à puce. Toutefois, nous pensons que le cadre qu'elles utilisent, *i.e* l'utilisation d'arbres de décision, possède d'intéressantes propriétés d'efficacité qui les rend très attractives dès qu'il s'agit d'utiliser un système de VAL sous de fortes contraintes calculatoires. Les résultats obtenus permettent d'espérer, qu'après quelques années de développement supplémentaire, l'approche par arbre de décision constitue une alternative crédible aux approches par mélange de gaussiennes, pour l'implémentation de techniques biométriques sur cartes à puce.

Bibliographie

- R. Auckenthaler, M. Carey, et L.-T. Harvey. Score Normalisation for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1-3) : 42–55, janvier/avril/juillet 2000.
- M. Ben, R. Blouet, et F. Bimbot. A Monte-Carlo Method For Score Normalisation In Automatic Speaker Verification Using KullbackLeibler Distances. Publié dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Floride, USA, 2002.
- S. Bengio et J. Mariéthoz. Learning the decision function for speaker verification. Publié dans les actes de *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2001. www.idiap.ch/~marietho/publications/orig/rr00-40.ps.gz.
- F. Bimbot, I. Magrin-Chagnolleau, et L. Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17(1-2) : 177–1992, Août 1995.
- R. Blouet et F. Bimbot. A Tree-based Approach for Score Computation in Speaker Verification. Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 223–227, Crète, Grèce, juin 2001.
- L.J. Boë, F. Bimbot, J.F. Bonastre, et P. Dupont. De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Revue Langues*, 2(4) : 270–288, Décembre 1999.
- C. Braghin. Biometric Authentication, University of Helsinki. www.citeseer.nj.nec.com/436492.html, 1998.
- L. Breiman. Bagging predictors. *Machine Learning*, 24(2), 1996.
- L. Breiman. The heuristics of instability in model selection. *Annals of Statistics*, (24) : 2350–2381, 1997.
- L. Breiman, J. Friedman, R. Olshen, et C. Stone. *Classification and Regression Trees*. Wadsworth Int. Group, 1984.

- Biometrics Working Group BWD. Best Practice in Testing and Reporting Performance of Biometric Devices. Rapport de recherche, Communications-Electronics Security Group (CESG), Infosec Assurance and Certification Services, Cheltenham, Gloucestershire, UK, Mars 2001. www.cesg.gov.uk/technology/biometrics.
- Calliope. *La parole et son traitement automatique*. Masson, 1989.
- M. Carey, E. Parris, H. Lloyd-Thomas, et S. Bennett. Robust prosodic features for speaker identification. Publié dans les actes de *International Conference on Spoken and Language Processing (ICSLP)*, volume 3, pages 1800–1803, Philadelphie, Pennsylvanie, USA, 1996.
- P. Castellano et S. Sridharan. Telephone based speaker recognition using multiple binary classifier and gaussian mixture models. Publié dans les actes de *International Conference on Acoustic, Speech and Signal Processing*, volume 2, pages 1075–1078, 1999.
- T. Choudhury, B. Clarkson, T. Jebara, et A. Pentland. Multimodal Person Recognition using Unconstrained Audio and Video. Rapport de recherche TR-472, MIT Media-Lab, 1998. www.citeseer.nj.nec.com/choudhury98multimodal.html.
- P. Delacourt. *La Segmentation et le regroupement pour l'indexation de documents audio*. Thèse de doctorat, Institut Eurécom, Septembre 2000.
- A.P. Dempster, N.M. Laird, et D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal Of Statistical Society of America (JASA)*, 39 : 1–38, 1977.
- G. Doddington. Some Experiments on Idiolectal Differences among Speakers. Publié dans les actes de *NIST 2001 Speaker Recognition Evaluation Workshop*, 2001. www.nist.gov/speech/tests/spk/2001/doc/n-gram_experiments-v06.pdf.
- G. Doddington, W. Liggett, A. Martin, M. Przybocki, et D.A. Reynolds. Comparison of text-independant speaker recognition method on telephone speech with acoustical mismatch. Publié dans les actes de *International Conference on Spoken Language Processing (ICSLP)*, pages 1788–1791, Philadelphie, Pennsylvanie, USA, 1996.
- B. Duc, E. Bigun, J. Bigun, G. Maitre, et S. Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 9 (18) : 835–843, 1997. www.citeseer.nj.nec.com/duc97fusion.html.
- Eagles. Assessment of speaker verification systemes. Rapport de recherche, EAGLES Spoken Language Systems, 1995.
- K. R. Farrell, R. J. Mammone, et K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. Publié dans les actes de *IEEE Transactions On Speech and Audio Processing*, volume 2, pages 194–205, Janvier 1994.

- C. Fredouille. *Approche Statistique pour la Reconnaissance du Locuteur : Informations Dynamiques et Normalisation Bayésienne des Vraisemblances*. Thèse de doctorat, Université d'Avignon et des Pays du Vaucluse, Octobre 2000.
- C. Fredouille, J.-F. Bonastre, et T. Merlin. AMIRAL : a block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing*, 10 (1-3), janvier/avril/juillet 2000.
- J. Friedman, T. Hastie, et R. Tibshirani. Additive logistic regression : a statistical view of boosting. Rapport de recherche, Department of Statistics, Stanford University, 1998. www-stat.stanford.edu/~jhf/ftp/boost.ps.
- M.C. Frye. *The Body As A password : Considerations, Uses, And Concerns of Biometric Technologies*. Thèse de doctorat, Faculty of The Graduate School of Arts and Sciences of Georgetown University, Avril 2001.
- S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, 19(2) : 254–272, Avril 1981.
- J.L. Gauvain et C.-H. Lee. Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) : 291–294, 1994.
- D. Genoud. Reconnaissance et transformation de locuteurs (Quelles identifiées par la parole?). Thèse de doctorat, École Polytechnique Fédérale de Lausanne, janvier 1999.
- H. Gish. Robust discrimination in automatic speaker identification. Publié dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 289–292, 1990.
- H. Hermansky et N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, ASSP-2(4) : 587–589, 1994.
- A. Higgins, L. Bahler, et J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1 : 89–106, 1991.
- B. Jacob, J. Mariéthoz, G. Gravier, et F. Bimbot. Robustesse de la vérification du locuteur par mots de passe personnalisés. Publié dans les actes de *XIIIèmes Journées d'Étude sur la Parole (JEP)*, Aussois, juin 2000.
- A. Jain, A. Ross, et S. Prabhakar. Biometrics-based web access. Rapport de recherche, TR98-33, Michigan State University, 1998. www.citeseer.nj.nec.com/433477.html.
- G.J. Jang, S.J. Yun, et Oh Y.H. Feature vector transformation using independent component analysis and its application to speaker verification. Publié dans les actes

- de *European Conference on Speech Communication and Technology (Eurospeech)*, 1999.
- S. Kadambe. Text independent speaker identification system based on adaptive wavelets. Publié dans les actes de *The International Society for Optical Engineering, Wavelet Applications*, volume 2242, 1994. www.citeseer.nj.nec.com/115365.html.
- S. P. Kishore et B. Yegnanarayana. Online text-independent speaker verification system at iitm. Publié dans les actes de *International Conference on Multimedia Processing and Systems*, pages 178–180, August 2000.
- L. Kung-Pu et J.E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. Publié dans les actes de *ICASSP*, pages 595–597, 1988.
- H. Kunzel. Current Approaches to Forensic Speaker Recognition. Publié dans les actes de *Esca Workshop on Speaker Recognition, Identification, and Verification*, pages 135–141, 1994.
- M. Lastrucci, M. Gori, et G. Soda. Neural autoassociators for phoneme-based speaker verification. Publié dans les actes de *Esca Workshop on Speaker Recognition, Identification, and Verification*, pages 189–192, 1994.
- S. Liu et M. Silverman. A Practical Guide to Biometric Security Technology, publication de *IEEE Computer Society's Web*, 2001. www.computer.org/itpro/homepage/Jan_Feb/security3.htm.
- I. Magrin-Chagnolleau. *Approches statistiques et filtrage vectoriel de trajectoires spectral pour l'identification du locuteur indépendante du texte*. Thèse de doctorat, École Nationale Supérieure des Télécommunications, Janvier 1997.
- I. Magrin-Chagnolleau et G. Durou. Application Of Time-Frequency Principal Component Analysis To speaker Verification. *Digital Signal Processing*, 10(1-3) : 226–237, janvier/avril/juillet 2000.
- I. Magrin-Chagnolleau, G. Gravier, et R. Blouet. Overview of the 2000-2001 ELISA Consortium Research Activities . Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 67–73, Crète, Grèce, juin 2001.
- I. Magrin-Chagnolleau, G. Gravier, M. Seck, O. Boeffard-Dosierre, R. Blouet, et F. Bimbot. Further Investigation in Speech Features for Speaker Characterization. Publié dans les actes de *International Conference on Spoken and Language Processing (ICSLP)*, Beijing, Chine, 2000.
- T. Mansfield, G. Kelly, D. Chandler, et J. Kane. Biometric Product Testing Final Report. Rapport de recherche, Center for Mathematics and Scientific Computing, National Physical Laboratory, Teddington, Middlesex, UK, Mars 2001.

- A. Martin, G. Doddington, T. Kamm, M. Ordowski, et M. Przybocki. The DET Curve in Assessment of Detection Task Performance. Publié dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1895–1898, 1997.
- A. Martin et M. Przybocki. The NIST 1999 Speaker Recognition Evaluation - An Overview. *Digital Signal Processing*, 10(1-3) : 1–19, janvier/avril/juillet 2000.
- J. McLaughlin, D.A Reynolds, et T. Gleason. A Study Of Computation Speed-ups Of The GMM-UBM Speaker Recognition System. Publié dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, septembre 1999.
- S. Meigner, T. Merlin, R. Blouet, et J.-F. Bonastre. LIA System for the NIST 2002 speaker verification evaluation. Publié dans les actes de *NIST 2002 Speaker Recognition Evaluation Workshop*, 2002.
- H. Melin. On Word Boundary Detection in Digit-Based Speaker Verification. Publié dans les actes de *Workshop on Speaker Recognition and its Commercial and Forensic application (RLA2C)*, Avignon, France, Avril 1998. www.speech.kth.se/ctt/publications/papers/rla2c98_46.pdf.
- D. Meuwly et A. Drygajlo. Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM). Publié dans les actes de *Workshop Speaker Odyssey*, pages 145–150, Juin 2001.
- C. Montacie et J. Le Floch. AR-Vector Models for Free-Text Speaker Recognition. Publié dans les actes de *International Conference on Spoken and Language Processing (ICSLP)*, pages 611–614, Banff, Canada, 1992. www.citeseer.nj.nec.com/160881.html.
- H. Nakasone et S.D. Beck. Forensic Automatic Speaker Recognition. Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 139–144, Crète, Grèce, juin 2001.
- J. Pelecanos. Feature Warping for Robust Speaker Verification . Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 213–218, Crète, Grèce, juin 2001.
- B. Pfahringer et I.H. Witten. Improving bagging performance by increasing decision tree diversity. Rapport de recherche TR-97-31, Oesterreichisches Forschungsinstitut für Artificial Intelligence, 1997. www.citeseer.nj.nec.com/friedman98additive.html.
- N. Poh et J. Korczak. Hybrid Biometric Person Authentication Using Face and Voice Features. Publié dans les actes de *3rd International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA'01)*, pages 342–346, Halmstad, Suède, Juin 2001.

- J.E. Porter. On The 30 errors criterion. Rapport de recherche, ITT Industries Defense and Electronics Group, avril 1997. www.engr.sjsu.edu/biometrics/nbtccw.pdf.
- D.A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. Thèse de doctorat, Georgia Institute of Technology, 1992.
- D.A. Reynolds. The effects of handset variability on speaker recognition performance : experiments on the Switchboard corpus. Publié dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, Georgie, USA, 1996.
- D.A. Reynolds. Comparison of Background Normalization Method For Text-Independent Speaker Verification. Publié dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, pages 963–967, Rhôdes, Grèce, septembre 1997.
- D.A. Reynolds, T.F. Quatieri, et R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3) : 19–42, janvier/avril/juillet 2000.
- A. E. Rosenberg et F. K. Soong. Recent research in automatic speaker recognition. *Advances in Speech Signal Processing*, pages 701–738, 1992.
- R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5 : 197–227, 1990.
- R.E. Schapire et Y. Freund. Experiments with a new boosting algorithm. Publié dans les actes de *Machine Learning : Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.
- R.E. Schapire et Y. Freund. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Science*, 55, 1997.
- Robert E. Schapire et Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Computational Learning Theory*, pages 80–91, 1998. www.citeseer.nj.nec.com/schapire99improved.html.
- K. Sonmez, E. Shriberg, L. Heck, et M. Weintraub. Modeling dynamic prosodic variation for speaker verification. Publié dans les actes de *International Conference on Spoken and Language Processing (ICSLP)*, volume 7, pages 3189–3192, Sydney, Australie, 1998.
- F. K. Soong, A. E. Rosenberg, L. R. Rabiner, et B. H. Juang. A Vector Quantization Approach to Speaker Recognition. Rapport de recherche 66, AT&T Bell Laboratories, Mars/Avril 1987.
- F.K. Soong et A.E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. Publié dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 877–880, 1986.

- D. Sturin, B. Dunn, et D. Reynolds. MIT Lincoln Laboratory Site Presentation - Extended Data Task. Publié dans les actes de *NIST 2001 Speaker Recognition Evaluation Workshop*, 2001.
- P. Taylor, R. Caley, A. W. Black, et S. King. Edinburgh speech tools library, system documentation edition 1.2. www.cstr.ed.ac.uk/projects/speech.tools/manual-1.2.0/, 1999.
- P. Urien. Etat de l'art des cartes à puces internet. Schlumberger-Sema, 2001.
- P. Verlinde et M. Acheroy. A Contribution to Multi-Modal Identity Verification using Decision Fusion. Publié dans les actes de *PROMOPTICA*, Bruxelles, Belgique, juin 2000. www.citeseer.nj.nec.com/verlinde99contribution.html.
- S. Vuuren. Comparison of text-independent speaker recognition method on telephone speech with acoustical mismatch. Publié dans les actes de *International Conference on Spoken Language Processing (ICSLP)*, pages 1788–1791, Philadelphie, Pennsylvanie, USA, 1996.
- J.L. Wayman. A Definition of Biometrics. Publié dans les actes de *Collected Works*, pages 21–24. National Biometric Test Center, Août 2000a.
- J.L. Wayman. Fundamentals of Biometric Technologies. Publié dans les actes de *Collected Works*, pages 1–18. National Biometric Test Center, Août 2000b.
- J.L. Wayman. Technicals Testing and Evaluation of Biometric Identification Device. Publié dans les actes de *National Biometric Test Center - Collected Works 1997-2000*, pages 67–89. National Biometric Test Center, Août 2000c.
- F. Weber, L. Manganaro, et B. Peskin. Speaker ID on the Extended Data Set. Publié dans les actes de *NIST 2001 Speaker Recognition Evaluation Workshop*, 2001.
- K. Yu, J. Mason, et J. Oglesby. Speaker recognition using hidden markov models, dynamic time warping and vector quantisation. *IEEE proc. vision, image and signal processing*, pages 313–318, 1995. www.citeseer.nj.nec.com/yu95speaker.html.
- X. Zhang et C. Broun. Using Lip Features for Multimodal Speaker Verification. Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 231–236, Crète, Grèce, juin 2001.
- R.D. Zilca. Using Second Order Statistics For Text Independent Speaker Verification. Publié dans les actes de *A Speaker Odyssey : The Speaker Recognition Workshop*, pages 45–50, Crète, Grèce, juin 2001.

Résumé

La vérification du locuteur consiste à déterminer automatiquement l'identité d'une personne à partir de sa voix. Actuellement, ses perspectives de mise en œuvre se situent essentiellement dans le domaine de la sécurisation d'accès à des services, des locaux ou de transactions bancaires téléphoniques. Si les outils qu'elle utilise sont souvent issus des recherches et bénéficient des progrès généraux des différentes branches du traitement automatique de la parole, ses caractéristiques techniques et applicatives en font une technologie de l'identification humaine automatique (biométrie) particulièrement intéressante.

Les travaux présentés dans cette thèse ont été effectués dans le cadre d'une convention de recherche entre l'équipe CP8 de SCHLUMBERGER-SEMA (ex BULL-CP8) et le projet METISS de l'IRISA. L'objet de cette collaboration est de mettre en œuvre un système de vérification automatique du locuteur dont tout ou partie des traitements sont effectués sur une carte à microprocesseur. Dans ce but, nous avons tout d'abord cherché à mettre en place et à améliorer un système de vérification automatique du locuteur (VAL) suivant les caractéristiques de l'état-de-l'art, puis nous nous sommes efforcés de réduire autant que possible la quantité de mémoire et la puissance de calcul nécessaires au fonctionnement d'une plateforme de VAL. Dans un premier temps, nos travaux ont donc concerné le développement d'une méthode d'estimation suivant le critère du *Maximum A Posteriori* (MAP) des paramètres des modèles de mélanges de gaussiennes (GMM) ainsi que celui d'une technique originale pour la normalisation du rapport de vraisemblance. Dans un second temps, nous avons développé un système de VAL basé sur les arbres de décision et dont le principe est d'estimer, directement lors de la phase d'apprentissage, le score de décision associé à chaque point de l'espace des paramètres. Une première implémentation de la technique utilise des critères de partition classiquement associés à la construction d'arbres de décision (type critère de Gini et entropie). Une seconde utilise, quant à elle, un critère plus original dans ce domaine et dont la mise en place a permis de fortement augmenter les performances. Enfin, une troisième mise en œuvre de la méthode est basée sur une représentation additive de plusieurs arbres par locuteur.

Les expériences réalisées montrent que si la méthode proposée entraîne une diminution des performances par rapport aux meilleurs systèmes GMM, elle conduit à une réduction considérable des quantités de mémoire et de la puissance de calcul nécessaires à la mise en place d'un système de VAL. Ce travail offre de nombreuses perspectives, qu'il s'agisse de la ré-utilisation de la technique dans d'autres domaines de la biométrie ou des différentes pistes devant permettre son amélioration. Parmi celles-ci, l'incorporation de connaissances a priori pour l'apprentissage des arbres de décision, à l'instar du critère MAP utilisé avec les GMM, semble être la plus prometteuse.

Mots clés : *Biométrie, arbres de décision, vérification automatique du locuteur, architecture embarquée, carte à microprocesseur.*