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

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 166 I2S « Mathématiques et Informatique »
Laboratoire d'Informatique d'Avignon (EA 4128)

Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur

par
Alexandre PRETI

Thèse soutenue publiquement le 10 décembre 2008 devant un jury composé de :

M.	Patrick VERLINDE	Professeur, Royal Military Academy, Bruxelles	Président du jury
M.	Frédéric BIMBOT	DR/CNRS, IRISA/INRIA, Rennes	Rapporteur
M.	Sébastien MARCEL	Senior Researcher, IDIAP, Martigny, Suisse	Rapporteur
M.	John MASON	Professeur, UWS, Swansea, Royaume Uni	Examinateur
M.	Claude BARRAS	Maître de Conférences, LIMSI, Orsay	Examinateur
M.	Jean-François BONASTRE	Professeur, LIA, Avignon	Directeur de thèse
M.	François CAPMAN	Ingénieur, THALES, Colombes	Co-Encadrant

Laboratoire d'Informatique d'Avignon

Résumé

Ce travail de thèse s'intéresse à la reconnaissance automatique du locuteur (RAL) dans les réseaux professionnels de communication (*Private Mobile Radio networks : PMR*). Plus précisément, nous nous intéressons à la surveillance des utilisateurs en cours de communication pour détecter un changement de locuteur, issu du vol ou du prêt d'un terminal de communication. Les systèmes « état de l'art » de RAL présentent aujourd'hui de très bonnes performances sur des signaux de conversations téléphoniques. Néanmoins, l'application envisagée entraine différentes contraintes liées au fonctionnement du réseau PMR et à l'ergonomie particulière d'une telle application. En effet, la RAL doit être effectuée en continue et les réseaux PMR offrent une qualité du signal de parole plus faible que les réseaux de téléphonie classique. Dans ce travail, nous évaluons l'impact de ces contraintes applicatives sur les performances d'un système de RAL et nous proposons des solutions pour pallier les différents problèmes énoncés. Plus particulièrement, nous nous intéressons à la phase de paramétrisation qui doit être réalisée en ligne et dans l'environnement des réseaux PMR, ainsi qu'à l'adaptation non supervisée des modèles de locuteurs. Cette technique permet d'utiliser des données de test pour améliorer les modèles de locuteur ; elle répond au problème des durées courtes d'apprentissage et permet de mieux modéliser les variabilités intra-locuteur et inter-session.

Mots-clé : reconnaissance du locuteur, adaptation non supervisée, paramétrisation, milieux bruités, réseaux professionnels de communication.

Abstract

This thesis work deals with automatic speaker recognition for professional telecommunication networks (PMR). More precisely, the targeted application is the online monitoring of communications on this kind of networks. State of the art speaker recognition systems show good performance on telephonic data. Therefore, the targeted application introduces specific constaints. We evaluate the impact of these constraints on a baseline speaker recognition system and propose solutions to limit their influence on recognition error rates. Firstly, we propose an optimised speech parameterization. Some technics are introduced to compensate the effects of noisy environments, low bitrate voice coding and channel transmission variations. Moreover, this parameterization is compliant with the online recognition processing needed by the targeted application. Then, we introduce a new approach for unsupervised speaker model adaptation to reduce the issue of the poor quantity of learning data. Unsupervised adaptation is also a way to reduce the impact of the intra-speaker and inter-session variabilities. We propose a continuous progressive speaker model adaptation able to take into account all the test data withdrawing threshold based data selection.

Keywords : automatic speaker recognition, unsupervised speaker model adaptation, parameterization, noisy environments, professional telecommunication networks.

Remerciements

Ces remerciements s'adressent tout d'abord à mon directeur de thèse, le professeur Jean-François Bonastre. J'ai bénéficié tout au long de ces trois années de sa grande connaissance du domaine et de ses conseils avisés pour proposer les contributions de ce travail. J'ai été très heureux de travailler sous sa direction. Un grand merci à François Capman, mon encadrant au sein de l'entreprise Thales Communications, qui lui aussi m'a guidé dans mes travaux de recherche. Merci à Bruno Sourdillat, directeur du laboratoire MMP de Thales Communications, qui a toujours encouragé mon travail.

Merci à tous les membres du jury pour avoir évalué mon travail et m'avoir offert cette belle journée qu'à été celle de ma soutenance de thèse, quel beau souvenir !

Ce travail est aussi le fruit de collaborations, partagées avec les membres du LIA et du laboratoire MMP de Thales Communications. Merci à mes amis Nico et Driss pour leur patience et leur pédagogie et à Bertrand pour son dynamisme. Je n'oublie pas les pros de la programmation et amis qui ont toujours su me dépanner, c'est facile il s'agit de **Ben**(oit) (le grand), de **Ben**(jamin) (le petit) et de **Ben**(jamin) (le moyen). Il y a aussi Fred W., je l'ai beaucoup agacé au début, merci pour sa patience.

Et puis pour faire du bon travail il faut une bonne ambiance, alors merci à tout ceux qui font régner la joie et la bonne humeur (et ils sont nombreux !), merci à Will, Corinne, JP, Georges, Pascal, Cyril (même avec son humour...), Rachid, Gwen (merci pour les relectures), les Christophe, Nathalie,...une page ne suffirait pas, j'arrête. Merci à tous.

Je garde le meilleur pour la fin, je tiens à remercier et à dédier ce document à mes parents et à ma bonne étoile. Merci pour ce merveilleux soutien et pour avoir toujours fait en sorte que tout se passe dans les meilleures conditions possibles.
Merci enfin à celle qui m'a soutenu (supporté ?), ma belle Camille.

Table des matières

Avant propos

De nos jours, les services sécurisés sont omniprésents dans notre quotidien. Différentes techniques d'identification basées sur la possession d'un secret, mot de passe ou badge d'accès, sont aujourd'hui employées. Cependant, la recherche s'attache à trouver des moyens plus fiables et moins contraignants pour identifier l'utilisateur, notamment par la biométrie. Cette dernière est basée sur la reconnaissance de caractéristiques biologiques propres à l'individu. Contrairement aux systèmes basés sur la possession d'un secret, un mot de passe par exemple, la biométrie utilise des caractéristiques corporelles de l'individu qui ne peuvent être perdues ou volées. La recherche est notamment très active dans le domaine du traitement de la parole. Ce média, naturellement employé par l'homme, présente de nombreux avantages. Il véhicule notamment beaucoup d'informations sur le locuteur. Les techniques de reconnaissance automatique du locuteur ont très fortement progressé ces dernières années et présentent aujourd'hui des performances permettant la mise en oeuvre d'un système de surveillance de réseaux de communication. Bien que la biométrie voix ne soit pas assez fiable pour remplacer des biométries très sûres, comme la rétine ou les empreintes digitales, elle est souvent le seul élément disponible pour authentifier un utilisateur sur les réseaux de communication.

Ce travail de thèse s'intéresse à la mise en place de la surveillance de réseaux professionnels de communication par la reconnaissance du locuteur. Les terminaux de communications utilisés sur ces réseaux sont sécurisés par un identifiant et un mot de passe. Ceci permet de gérer leur autorisation d'accès au réseau. Une fois le terminal initialisé, aucun moyen de surveillance n'est disponible. Le vol ou le prêt d'un terminal ne sont pas détectés. L'authentification de l'utilisateur par sa voix, tout au long de la communication, constitue un niveau de sécurité supplémentaire. Nous présentons dans ce travail de thèse la mise en oeuvre d'une telle solution de reconnaissance du locuteur en nous intéressant particulièrement à quelques facteurs : le codage à bas débit de la parole, l'influence des bruits ambiants, le traitement de RAL en ligne et les durées courtes d'apprentissage.

Chapitre 1

Introduction

Sommaire

Dans de multiples domaines, il s'avère nécessaire d'authentifier les utilisateurs afin, par exemple, de sécuriser les transactions bancaire, ou de gérer les accès à des ressources protégées. En règle générale, l'authentification des utilisateurs est réalisée par un identifiant associé à un mot de passe secret. Les identifiants et mots de passe, clés ou badges d'accès, sont très employés. Pourtant, ceux-ci sont vulnérables aux techniques de falsification et au vol. De plus, il est difficile pour un utilisateur de ne pas égarer son badge d'accès ou de se remémorer son ou ses mots de passe. Une enquête [1] a été publiée en septembre 2005 sur les problèmes rencontrés par l'employé dans la gestion de ses mots de passe et sur les risques potentiels pour la sécurité de l'entreprise. Il ressort de cette étude que neuf personnes sur dix s'estiment agacées par la gestion de plusieurs mots de passe. En résultent des comportements à risque, comme l'utilisation de mots de passe issus de l'environnement familier, faciles à détourner. Sachant que le secret peut être divulgué à une tierce personne, ou volé, il se peut que l'utilisateur soit alors un imposteur. Dans le but de proposer une authentification plus sûre, de nouvelles techniques d'authentification, basées sur la biométrie, ont émergé. En effet, l'authentification biométrique répond à cette problématique car elle utilise le fait incontournable que chaque individu a ses propres caractéristiques physiques, qui sont immuables et donc ne peuvent être perdues, volées ou prêtées, comme un simple mot de passe. L'authentification biométrique peut être ponctuelle, pour un accès protégé,

[1] Enquête publiée par RSA Security, société spécialisée dans l'authentification et le chiffrement. http://www.rsasecurity.com

comme une serrure, ou périodique, *i.e.* répétée dans le temps, comme la surveillance tout au long d'une communication.

1.1 La biométrie

L'authentification biométrique des utilisateurs est basée sur des éléments liés à l'individu, à son corps. Ces éléments ne varient pas ou très peu dans le temps et ne peuvent être modifiés par un individu. Nous pouvons par exemple citer les méthodes d'authentification basées sur les empreintes digitales, sur les empreintes dentaires ou encore sur l'iris. Les méthodes d'authentification biométrique reposent sur deux étapes :

1. la collecte d'un échantillon de référence, appartenant à l'utilisateur,
2. la comparaison d'un échantillon de test à l'échantillon de référence.

La mesure de la correspondance entre les deux échantillons permet de déterminer une décision d'authentification. Deux modes de fonctionnement existent en authentification biométrique :

- l'identification : il s'agit de déterminer l'identité de l'utilisateur à partir d'une base de données d'échantillons de référence. L'échantillon de test est comparé à tous les échantillons de la base. L'identité du locuteur de la base reconnu est retournée ;
- la vérification d'identité : il s'agit de valider l'identité proclamée de l'échantillon test, à partir d'un échantillon de référence. Dans ce cas, le signal de test n'est comparé qu'à un seul échantillon de référence. Le locuteur a prononcé le signal de test si la mesure de correspondance dépasse un seuil prédéterminé.

Pourquoi la voix pour authentifier un utilisateur ?

Dans le domaine de la sécurité militaire ou civile, les techniques d'authentification biométrique majoritairement employées sont la reconnaissance d'empreintes digitales ou de l'iris. Ces dernières autorisent un niveau de sécurité maximal et des taux d'erreurs minimaux. Néanmoins ces méthodes sont dites « intrusives » et restent lourdes à mettre en oeuvre. L'intrusivité d'une méthode biométrique est définie comme le niveau d'acceptation de la méthode par l'utilisateur. Les empreintes digitales sont ainsi peu acceptées du public pour des raisons socio-culturelles car elles ont une connotation criminalistique. L'authentification par empreinte rétinienne, réputée pour être la plus fiable des technologies, est également très peu acceptée par les utilisateurs car elle nécessite un faisceau lumineux pour éclairer le fond de l'œil en vue de déterminer les positions des veines de la rétine [Bolle et Pankanti, 1998; Jain et al., 2001]. Cette méthodologie constitue un frein psychologique fort qui limite l'utilisation de ce type de biométrie pour l'authentification. Il faut néanmoins tempérer cette observation car, dans des contextes bien particuliers comme l'accès à des ressources militaires, l'acceptabilité de l'utilisateur ne prime pas sur la robustesse du système. La figure 1.1, issue de travaux conduits par *l'International Biometric Group*, propose une comparaison des modalités biométriques les plus répandues (comme l'empreinte digitale, l'iris ou la rétine)

en fonction de quatre critères : l'intrusivité, la fiabilité, le coût de mise en place et l'effort de mise en oeuvre. La biométrie idéale, représentée par les quatre indicateurs les

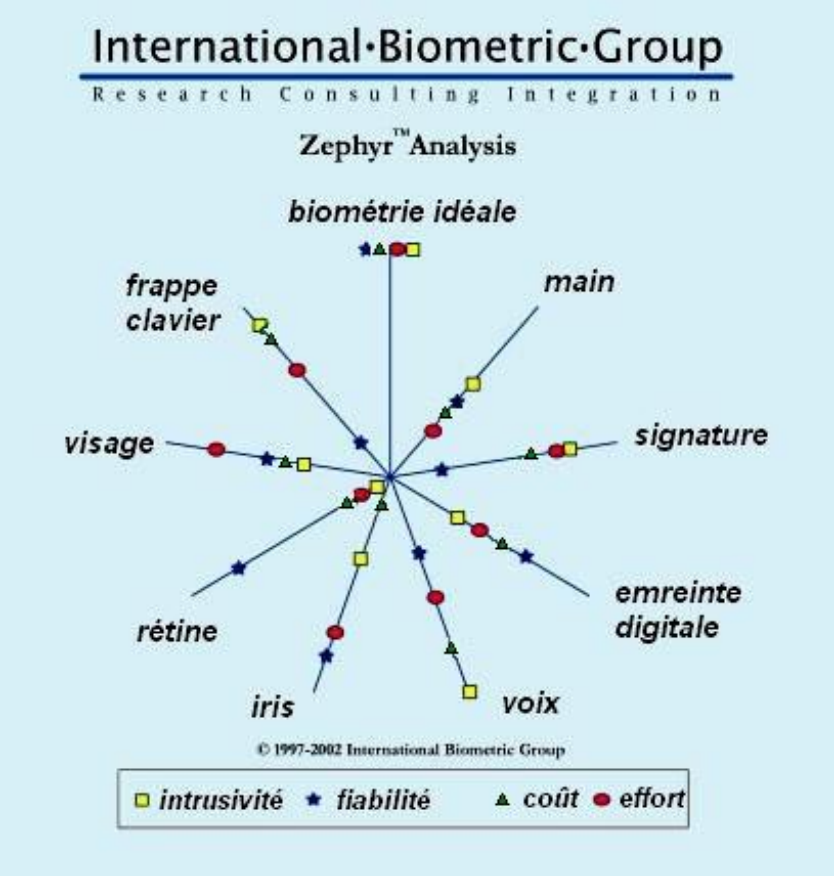


FIG. 1.1 – *Classement des différentes biométries selon quatre critères : le coût, la fiabilité, l'effort de mise en place et l'intrusivité. La biométrie idéale présente les caractéristiques les plus éloignées du centre. (Traduit de l'anglais).*

plus éloignés du centre, présenterait alors une fiabilité élevée, une faible intrusivité, une grande facilité d'utilisation et un faible coût. La rétine est classée comme ayant la plus grande fiabilité mais présente également les plus forts coûts, intrusivité et d'effort de mise en oeuvre. Ce classement démontre aussi que la reconnaissance du locuteur offre l'avantage d'être bien acceptée par l'utilisateur et d'être simple à mettre en oeuvre. Basée sur un échantillon de voix du locuteur, elle n'implique que la prise de son à travers un microphone. De plus, c'est souvent le seul média disponible. Les systèmes de Reconnaissance Automatique du Locuteur (RAL) s'appuient sur les caractéristiques de la parole permettant de reconnaître les individus. La reconnaissance du locuteur en tant que technique d'authentification présente les avantages suivants :

- l'acquisition du signal audio est très simple à mettre en oeuvre,
- l'enregistrement du signal audio n'est pas considéré comme intrusif (mais peut cependant présenter des difficultés au niveau législatif),
- le signal audio est naturellement véhiculé dans la majorité des réseaux de com-

munication,
- les techniques de stockage et de compression du signal audio sont très efficaces,
- dans de nombreuses applications (serveurs vocaux), l'utilisateur emploie déjà la parole pour communiquer avec la machine. Le coût supplémentaire de la RAL, en coût de mise en oeuvre comme en contraintes ergonomiques, est faible.

La recherche dans le domaine de la RAL est très active depuis quelques dizaines d'années. Cependant, les taux d'erreurs des techniques de reconnaissance du locuteur ne permettent pas d'égaler les performances des tecniques biométriques les plus robustes, comme la reconnaissance des empreintes digitales ou de la rétine. De ce fait, la voix est le plus souvent utilisée en complément d'une autre modalité (voix + mot de passe, voix + image) ou des techniques basées sur la possession d'un secret (mot de passe + voix).

1.2 Application visée

La surveillance de réseaux de communication

Ce travail de thèse présente l'utilisation de la RAL pour surveiller les réseaux professionnels de communications, ou réseau *PMR (Private Mobile Radio)*. Pour ce type de réseau, la sécurité représente un enjeu majeur. Elle se limite cependant à une vérification d'identité à la mise en route du terminal de communication. Lors de la mise en route du terminal, un identifiant et un mot de passe permettent d'authentifier l'utilisateur sur le réseau. Une fois la phase d'authentification du terminal effectuée, plus aucun moyen de contrôle n'est mis en oeuvre. En effet, aucun moyen de contrôle ne permet de détecter le vol d'un terminal, ou un changement de locuteur durant la communication.

Pour surveiller les communications, *i.e.* vérifier périodiquement l'identité du locuteur en cours de communication, la RAL représente une solution intéressante. Elle permet de déclencher une alerte lors du vol d'un terminal ou de relever un fonctionnement anormal, comme le prêt du terminal à un autre utilisateur, avec peu de contraintes pour l'utilisateur. Un opérateur distant peut recevoir une telle alerte et décider de couper la communication, ou demander une authentification ponctuelle à l'utilisateur.
Dans les réseaux professionnels de communication, la majorité de la bande passante disponible est dédiée au transport de la voix. Un canal dédié au transfert de données est souvent présent, mais il ne bénéficie pas d'une bande passante suffisante pour transporter des média pouvant servir à l'authentification des utilisateurs (images ou vidéo). Dans ce contexte applicatif, la reconnaissance du locuteur présente de multiples avantages :
- le signal audio est disponible en de multiples points de l'architecture maillée de ces réseaux de communication, permettant la mise en place d'un système de RAL distribué ;
- la surveillance peut s'effectuer en continu lors des communications, l'utilisateur ne doit pas interrompre son activité pour s'authentifier ;
- la réactivité du système à un fonctionnement anormal est quasiment instantanée.

1.3 Problématique

Les techniques de reconnaissance du locuteur souffrent des grandes variabilités du signal vocal et des conditions d'enregistrements. De nombreux facteurs affectent la qualité du signal de parole :

- les conditions de bruits d'environnement,
- le contenu linguistique,
- le canal de transmission utilisé (constitué du canal acoustique entre l'appareil phonatoire et le terminal ainsi que du canal de transmission vers le récepteur),
- les facteurs physiologiques du locuteur (santé, état de stress),

Ces facteurs rendent difficile la comparaison entre deux enregistrements. Le défi majeur en RAL réside dans la compensation des variabilités introduites précédemment. La mise en place d'un système de surveillance de réseaux professionnels de communications implique la prise en compte de contraintes plus spécifiques :

- le codage à bas ou très bas débit de la parole ;
- des durées d'enregistrements très courtes (application militaire par exemple) ;
- des conditions d'enregistrements particulièrement difficiles (bruits environnants, stress des utilisateurs, fatigue/effort),
- le type de fonctionnement souhaité : traitement en ligne, temps réel ou sur fichier.

Les techniques de RAL développées par la communauté scientifique ne s'intéressent pas toujours à ces contraintes applicatives. En effet, la majeure partie des efforts de recherche est orientée vers la reconnaissance du locuteur sur des enregistrements de communications téléphoniques. La durée de ces communications est de plusieurs minutes et les enregistrements (sous forme de fichiers) constituent une base de données qui est traitée globalement. Il n'existe alors pas de notion de traitement en cours de communication.

1.4 Contributions

Dans ce travail de thèse, nous proposons une solution de RAL pour l'application de surveillance des réseaux PMR. Nous utilisons le système de RAL état de l'art, du Laboratoire d'Informatique d'Avignon, pour proposer un système optimisé pour la surveillance des réseaux professionnels de communication. Les thèmes majeurs développés dans ce travail sont :

1. la robustesse à la variabilité de l'environnement d'acquisition,
2. les durées courtes d'apprentissage.

L'environnement d'acquisition constitue l'ensemble des facteurs qui altèrent le signal de parole entre son émission par l'utilisateur et la réception au niveau du moteur de reconnaissance. Nous nous focalisons sur les dégradations dues à la prise de son (bruits ambiants et variations du canal acoustique), au codage à bas débit de la parole, et à la transmission. L'acceptabilité des utilisateurs constitue une spécificité importante de l'application que nous envisageons. Pour ne pas contraindre les utilisateurs à une phase d'initialisation fastidieuse du système, les durées d'apprentissage doivent être courtes.

Ceci pose un problème pour modéliser la variabilité intra-locuteur et adapter correctement le modèle du locuteur à l'environnement. Cette problématique nous amènent à définir une méthode d'adaptation non supervisée des modèles de locuteurs qui permet d'utiliser les signaux de test pour mieux modéliser le locuteur. Enfin, nous abordons les traitements spécifiques de la RAL en ligne qui permettent de vérifier, périodiquement, l'identité du locuteur en cours de communication.

1.5 Cadre de travail de la thèse

Ces travaux ont été réalisés dans le cadre d'une collaboration entre le Laboratoire d'Informatique d'Avignon (LIA) et l'entreprise Thales Communications France (TCF), sous la forme d'une thèse CIFRE. Le laboratoire d'accueil a été le LIA pour une période de 18 mois, puis le laboratoire *Multimedia Processing*, de l'entité *Embedded Digital Systems* de TCF pour les 18 mois suivants. Le système de reconnaissance du locuteur utilisé pour ces travaux est le système ALIZE/SpkDet, développé au LIA. Thales Communications et le LIA ont conjointement participé aux campagnes d'évaluation de reconnaissance du locuteur NIST 2005, 2006 et 2008. Les différentes contributions proposées dans cette thèse sont évaluées sur les bases de données de ces campagnes d'évaluations. Le LIA et TCF sont également impliqués dans le projet MISTRAL, subventionné par l'Agence Nationale de la Recherche (ANR). Un démonstrateur de reconnaissance du locuteur sur PC a été réalisé dans le cadre de ce projet. Il est aujourd'hui intégré sur la plate-forme TETRA Digicom 25, réseau professionnel de communication commercialisé par Thales Communications. Ce démonstrateur est présenté dans l'annexe D.

1.6 Organisation du document

Ce document se divise en deux grandes parties. La première partie est consacrée aux principes généraux de RAL et la seconde aux contributions et conclusions. Dans la première partie, le chapitre 2 décrit les spécificités et variabilités du signal de parole ainsi que son utilisation comme biométrie. Nous détaillons les différentes tâches de RAL qui permettent de mettre en place des scénarios applicatifs distincts. Les approches majoritairement employées, ainsi que les méthodes d'évaluations des performances des systèmes de RAL terminent ce chapitre. Le chapitre 3 revient sur l'approche statistique majoritaire pour la vérification automatique du locuteur (VAL). Différentes méthodes sont employées dans cette approche pour passer du signal de parole à la décision de vérification d'identité. Nous détaillons les techniques de représentation du signal de parole ainsi que la modélisation statistique qui permet de créer des références de locuteurs et, finalement, le test, qui compare les références aux échantillons de test.

La seconde partie de ce document concerne la problématique centrale de cette thèse. Le chapitre 4 présente le système de RAL ALIZE/SpkDet qui a servi de base à nos travaux. Le chapitre 5 introduit les contraintes de mise en oeuvre d'une solution de RAL

sur les réseaux professionnels de communication. Nous verrons que l'architecture de ces réseaux ainsi que les spécificités de l'environnement d'utilisation introduisent des contraintes fortes sur les performances et la réalisation d'une solution de RAL. Les chapitres 6, 7 et 8 sont consacrés à la proposition d'un système de RAL pour la surveillance des réseaux PMR. Le chapitre 6 propose des traitements spécifiques pour un fonctionnement de RAL adapté à l'architecture et l'utilisation de ce type de réseaux. Les chapitre 7 et 8 décrivent une nouvelle méthode pour la compensation de la variabilité de l'environnement et des faibles durées d'apprentissage. Nous concluons finalement en résumant nos contributions et nos principaux résultats ainsi qu'en ouvrant des perspectives pour des futurs travaux de recherche.

Première partie

Principes généraux de la reconnaissance du locuteur

Chapitre 2

Reconnaissance automatique du locuteur

Sommaire

Ce chapitre est consacré aux principes de la reconnaissance automatique du locuteur. Il présente tout d'abord les mécanismes de production de la parole et les principales sources de variabilités pour comprendre comment un individu peut être reconnu par sa voix. Il souligne également les difficultés majeures associées à la RAL. Nous exposons ensuite les traitements numériques appliqués au signal audio dans un système de reconnaissance du locuteur. Dans ce chapitre, sont présentées les différentes tâches liées à la RAL, telles que l'Identification et la Vérification Automatique du Locuteur, ou encore l'Indexation par Locuteur de flux audio. Nous exposons quelques approches utilisées pour les systèmes de RAL et présentons, enfin, les méthodes d'évaluation des performances des systèmes de RAL.

2.1 La parole

2.1.1 La production de la parole

La production de la parole fait intervenir différents organes. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales. Il va ensuite traverser le conduit vocal (cavité nasale et buccale) et les articulateurs tels que les lèvres et la langue (cf. figure 2.1). Cet ensemble agit comme un filtre, considéré comme linéaire, dont la réponse

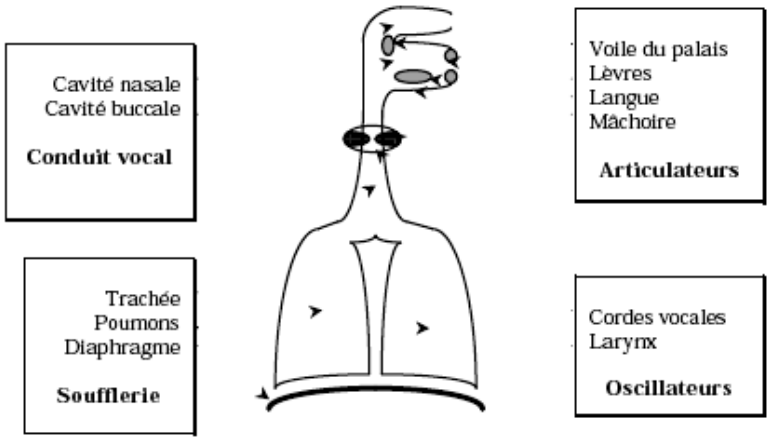


FIG. 2.1 – *Modèle physiologique de la production de la parole.*

impulsionnelle comporte des fréquences de résonance caractérisées par des pics, appelés formants, dans le spectre du signal de sortie. Le signal résultant est globalement non stationnaire mais peut être considéré comme stationnaire sur de très courtes périodes, de l'ordre de 20ms (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes :

1. voisée lorsqu'il y a vibration des cordes vocales, le signal est alors quasi-périodique,
2. non voisée dans le cas d'un simple soufflement, le signal est alors considéré comme aléatoire.

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodique, de fréquence dite de voisement F0, qui correspond à la fréquence de vibration des cordes vocales, la fréquence fondamentale ou pitch ; dans le second cas, la source est modélisée par un bruit blanc. Cette représentation binaire de la production de la parole a été introduite par [Fant, 1960]. Elle est reprise sur la figure 2.2.

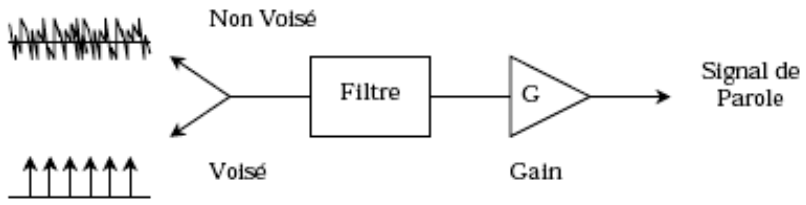


FIG. 2.2 – *Modèle de production de la parole.*

2.1.2 Les variabilités du signal de parole

Les informations transmises par le signal de parole sont multiples [Doddington, 1985]. La variabilité du signal de parole entre locuteurs est majoritairement utilisée en reconnaissance du locuteur pour reconnaître les individus. Il s'agit de la variabilité inter-locuteur. La capacité des systèmes de RAL à authentifier une personne repose essentiellement sur la capacité à discriminer les individus grâce à cette variabilité. Mais d'autres facteurs de variations altèrent la parole. Le signal de parole est par exemple considéré comme non reproductible par son locuteur. Il existe une variabilité propre au locuteur dépendante de son état physique mais aussi psychologique. Il s'agit de la variabilité intra-locuteur. De plus les conditions d'environnement influent sur l'onde de parole ; les bruits additifs ambiants ou les bruits de convolution engendrés par la prise de son modifient le signal de parole.

2.1.2.1 La variabilité inter locuteur

Le signal de parole permet la communication entre les individus. Il véhicule un message linguistique mais aussi quantités d'informations extra linguistiques et des informations liées au locuteur. La personnalité, au sens large, du locuteur influence la production du signal de parole. Ceci permet notamment de discriminer les locuteurs. Le signal de parole est un signal très complexe où se mêlent différents types d'informations classées par leur « niveau » de représentation. Les informations dites « bas-niveau » sont facilement utilisables à partir de l'analyse numérique du signal de parole. Elles regroupent des informations liées principalement à des traits physiques de l'individu (facteurs morphologiques et physiologiques). Les informations de « haut niveau », comme la linguistique ou l'état émotionnel du locuteur sont beaucoup plus complexes à caractériser. Ces informations sont relatives aux facteurs socio-culturels de l'individu. [Ben, 2004] propose une hiérarchie composée sur 6 niveaux d'informations différents :

1. le niveau acoustique : les paramètres sont liés à l'analyse de l'enveloppe spectrale du signal ;
2. le niveau prosodique qui désigne la « mélodie » de l'énoncé de parole ;
3. le niveau phonétique : la distinction des différents sons identifiables d'une langue ;
4. le niveau idiolectal qui se rapporte aux particularités langagières propres à un individu ;

5. le niveau dialogal qui définit la façon de communiquer d'un individu, comme ses temps de parole dans une conversation ;
6. enfin, le niveau sémantique qui caractérise la signification du discours.

Le niveau acoustique est le plus utilisé en RAL où l'influence de l'anatomie du locuteur sur l'émission du signal de parole est retenue [Furui, 1986; Rosenberg et Sambur, 1975]. Ces approches se basent sur une représentation numérique de l'enveloppe du signal, définie comme une suite de paramètres, les cepstres. Ces informations, présentes au niveau de l'enveloppe du signal, sont facilement extraites. Les niveaux prosodique et phonétique sont basés sur la représentation du signal de parole à un niveau supérieur. La prosodie caractérise le style d'élocution du locuteur. Les méthodes pour analyser cette information sont plus complexes bien que souvent basées sur l'analyse numérique du signal. Le niveau phonétique implique l'utilisation d'une segmentation en phonèmes, le plus souvent réalisée par un reconnaisseur de parole. Ces approches sont de plus en plus employées en RAL, en combinaison avec les approches acoustiques. Enfin les niveaux idiolectal, dialogal et sémantique ne sont, à notre connaissance, pas utilisés en RAL.

Il est à noter qu'aucune « empreinte vocale » n'est aujourd'hui définie, *i.e.* que le signal de parole d'un individu acquis à un instant donné sous certaines conditions d'enregistrements ne peut en aucun cas être considéré comme unique, [Bonastre et al., 2003]. Aussi des associations comme l'AFCP (l'Association Française de la Communication Parlée) et la communauté scientifique rappellent que le résultat d'un système de RAL ne peut être considéré comme preuve d'authentification d'un individu. La justice doit tenir compte de ce fait et par conséquent ne pas baser ses conclusions sur ce type d'authentification.

2.1.2.2 La variabilité intra-locuteur

Il est impossible pour un même individu de reproduire exactement le même signal de parole. Les facteurs de variabilités pour un même individu sont multiples. Ils peuvent être liés à la nature physiologique de l'individu. Dans ce cas cette variabilité intra-locuteur est induite par l'évolution naturelle (volontaire ou non) de la voix d'une personne. L'état pathologique est un exemple de variation de la voix involontaire d'une personne. De plus, une altération de la voix due à l'âge est présente chez tous les individus. Cette variabilité est une difficulté majeure en RAL.

2.1.2.3 Les facteurs « extérieurs »

La variabilité inter-session (entre sessions d'enregistrements) fait apparaître l'influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l'onde de parole est considérée comme idéale, car aucune déformation/distorsion de l'environnement extérieur ne l'a modifiée. L'environnement sonore lors de l'enregistrement, le matériel d'acquisition ou le canal de transmission utilisé vont ensuite déformer l'onde sonore originelle. Le canal de transmission, par exemple, agit comme

un filtre en fréquence sur l'onde sonore. Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d'un même individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances [Vuuren, 1996]. Par exemple, l'acquisition d'un signal de parole sur le réseau GSM introduit les dégradations suivantes sur le signal de parole :

- l'ajout du bruit de l'environnement,
- le sous-échantillonnage à 8kHz du signal,
- le filtrage sur la bande de fréquence [300-3400]Hz,
- le codage à bas débit de la parole,
- l'ajout du bruit de quantification des paramètres émis,
- la transmission sur un lien sans-fil avec pertes.

2.1.3 Analyse numérique du signal de parole

Les techniques d'analyse du signal de parole décrites dans ce chapitre sont désormais éprouvées et offrent une base solide aux techniques d'authentification. Ainsi la reconnaissance du locuteur a bénéficié des nombreux efforts de recherche en traitement du signal, originellement destinés au codage, en synthèse ou en reconnaissance de la parole. Les mécanismes de production ainsi que les paramètres caractéristiques du signal présentés ici en sont directement issus.

2.1.3.1 Du signal analogique à la représentation numérique

Les traitements effectués sur la parole sont aujourd'hui réalisés dans le domaine numérique. Au-dessus de 8kHz l'information vocale est négligeable, la bande de fréquence généralement utilisée est [0-8000Hz]. Un échantillonnage du signal de parole à 16kHz convient pour conserver la quasi-totalité de l'information (théorème d'échantillonnage de Nyquist/Shannon). L'amplitude est alors quantifiée généralement sur 16bits afin d'obtenir une bonne qualité. Pour un codage bas-débit l'échantillonnage est réalisé à 8kHz, ce qui permet de conserver la bande téléphonique (300-3400Hz). Le signal est représenté dans le domaine fréquentiel par l'utilisation des transformées de Fourier, ou encore sous une forme pouvant regrouper les informations de temps et de fréquence : le spectrogramme (cf. figure 2.3).

Du fait de sa quasi-stationnarité sur de courtes périodes, le signal de parole est généralement analysé sur des trames découpées par une fenêtre de pondération de 20 à 30ms avec un taux de recouvrement de 50% à 75%, puis représenté dans le domaine spectral (cf. figure 2.4 *b*).

Dans le cas d'un signal échantillonné à 8kHz, une fenêtre d'analyse de 256 points correspond à une longueur de 32ms (cf. figure 2.4 *a*). Une fenêtre classiquement utilisée est celle de Hamming.

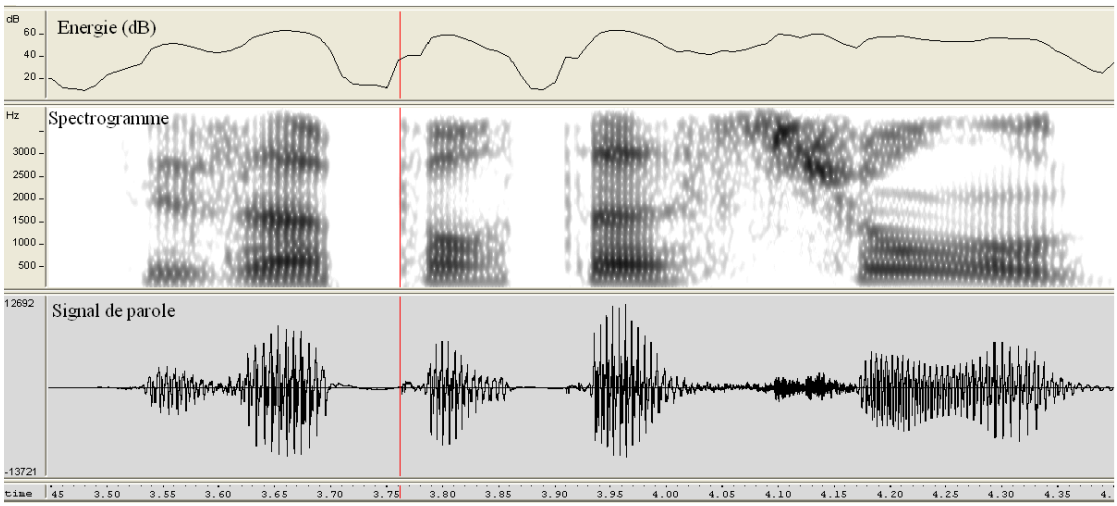


FIG. 2.3 – *Représentation d'un signal de parole, de son spectrogramme et de son énergie*

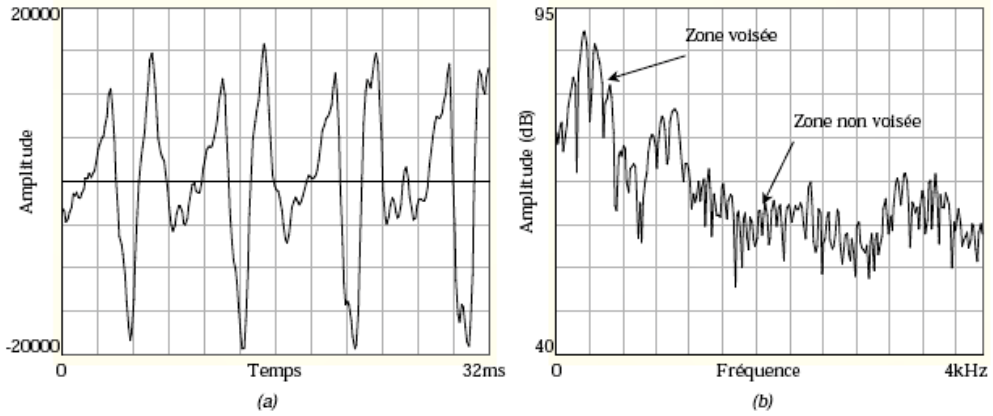


FIG. 2.4 – *Représentation temporelle (a) et spectrale (b) d'un signal de parole voisé et non voisé.*

De par son mécanisme de production, le signal de parole présente une corrélation à court terme, induite principalement par la cavité buccale, et une corrélation à long terme, qui découle directement de la structure périodique du signal. Spectralement ces caractéristiques se traduisent par une structure formantique de l'enveloppe du signal, pour la corrélation à court terme, et par une structure fine en peigne, dite harmonique, pour la corrélation à long terme. La première corrélation conduit à une dépendance des échantillons en fonction des précédents, cette propriété peut être exploitée par l'utilisation d'un filtre de prédiction linéaire. Les sons de parole sont produits par une source d'excitation $G(w)$ qui passe à travers le filtre linéaire qui est la fonction de transfert du conduit vocal $H(w)$ (cf. equation 2.1).

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \qquad (2.1)$$
$$X(\omega) = G(\omega)H(\omega)$$

La corrélation à long terme conduit à une périodicité dans le signal, elle est définie par la détection de la fréquence fondamentale et n'existe que dans le cas d'un son voisé. Les figures 2.4 *a* et *b* présentent le signal temporel et le spectre de deux segments de parole, l'un voisé et l'autre non voisé. Le signal non voisé ne présente pas les mêmes caractéristiques que le signal voisé : la structure harmonique n'existe pas, l'enveloppe spectrale présente une structure formantique moins marquée. De plus, le niveau d'énergie d'un signal non voisé est généralement plus faible que pour un signal voisé.

2.1.3.2 Les paramètres de la parole

Les paramètres du signal de parole décrits dans cette section permettent de discriminer un locuteur des autres individus. Idéalement, ces paramètres doivent avoir une forte variabilité entre les locuteurs et une faible variabilité pour un même locuteur. De plus, il doivent être robustes aux perturbations citées précédemment. Nous citons les paramètres exploitables pour la RAL.

2.1.3.2.1 L'énergie

L'énergie du signal $s(n)$ est calculée à partir du signal temporel suivant l'équation 2.2.

$$E_s = \int_{-\infty}^{\infty} |s(t)|^2 dt \qquad (2.2)$$

L'énergie est généralement exprimée en décibels :

$$E_s(dB) = 10 \log_{10} E_s \qquad (2.3)$$

L'évolution dans le temps du paramètre d'énergie peut déterminer le style d'intonation du locuteur. Utilisé seul, ce paramètre permet essentiellement de classer les trames par ordre énergétique, ce qui constitue une approche majoritaire pour la détection d'activité vocale qui sera abordée par la suite.

2.1.3.2.2 Le cepstre

L'analyse cepstrale utilise le modèle source filtre, avec une source périodique (sons voisés). Le processus consiste en la séparation de la source et du filtre, représentant

le conduit vocal, par déconvolution homomorphique. Le cepstre permet de séparer la structure harmonique de la structure formantique du signal. Le cepstre réel ou coefficient cepstral est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale de puissance du signal, sur la fenêtre d'analyse [Noll, 1964]. Il est souvent nommé *FFT cepstrum* pour le différencier du *LPC cepstrum* décrit plus bas. Dans le domaine de Fourier, le calcul du cepstre est décomposé en une somme de deux composantes, selon la relation décrite par l' équation 2.1.

$$c(\tau) = TF^{-1}log\|X(\omega)\| = TF^{-1}log\|G(w)\| + TF^{-1}log\|H(w)\| \tag{2.4}$$

où $TF^{-1}log\|X(\omega)\|$ est la transformée inverse du logarithme du module de la transformée de Fourier du signal x de parole, et où $TF^{-1}log\|G(w)\|$ est la transformée inverse du logarithme du module de la transformée de Fourier de la source et $TF^{-1}log\|H(w)\|$ est la transformée inverse du logarithme du module de la transformée de Fourier du filtre représentant le conduit vocal.

En pratique les coefficients cepstraux se calculent à partir du module de la transformée de Fourier du signal de parole $\|X(\omega)\|$. Une analyse en bancs de filtres transforme le module $\|X(\omega)\|$ en coefficients d'énergie par bandes de fréquences. L'échelle des bancs de filtres peut être linéaire pour le calcul des *LFCC (Linear Frequency Cepstral Coefficient)* ou suivre une échelle de Mel pour le calcul des *MFCC (Mel Frequency Cepstral Coefficient)*. L'échelle de Mel a été créée pour approximer la réponse du système auditif humain. La répartition des fréquences suit une loi logarithmique.
Les coefficients cepstraux sont finalement obtenus par l'application d'une transformation *DCT (Discrete Cosine Transform)* sur le logarithme des coefficients d'énergie.

Les pics de haute fréquence de l'analyse cepstrale représentent la fréquence fondamentale (la source), tandis que l'enveloppe cepstrale est déterminée par les pics de fréquences basses. La séparation de la source et du filtre dans le domaine cepstral s'appelle liftering ou liftrage (filtrage dans le domaine cepstral).

2.1.3.2.3 Analyse par prédiction linéaire

Nous avons précisé auparavant (cf. section 2.1.3) que la corrélation à court terme du signal de parole permet d'estimer les échantillons du signal de parole à partir des échantillons précédents.
L'analyse par prédiction linéaire ou analyse *LPC (Linear Predictive Coding)*, utilise cette propriété. Le signal est alors remplacé par une source, un train d'impulsions périodiques pour les sons voisés ou bruit blanc pour les sons non voisés [Oppenheim et Schaffer, 1975; Atal, 1974]. Le filtre qui représente la fonction de transfert du conduit vocal est un filtre tout pôle variant dans le temps.

$$\hat{x}(n) = \sum_{i=1}^{p} \alpha_i x_{n-i} \tag{2.5}$$

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^- i} \tag{2.6}$$

L'estimée du signal x est ainsi représentée par une somme pondérée des échantillons précédents x_{n-i} (cf. équation 2.5). Les coefficients de pondération utilisés sont les paramètres du filtre (α_i).

La réponse en fréquence du filtre LPC (cf. équation 2.6) suit les pics du spectre du signal de parole. Cette analyse est donc naturellement utilisée pour déterminer les formants. La figure 2.5 présente le spectre en fréquence issu de l'analyse LPC d'un signal de parole. Les pics du spectre représentent les formants.

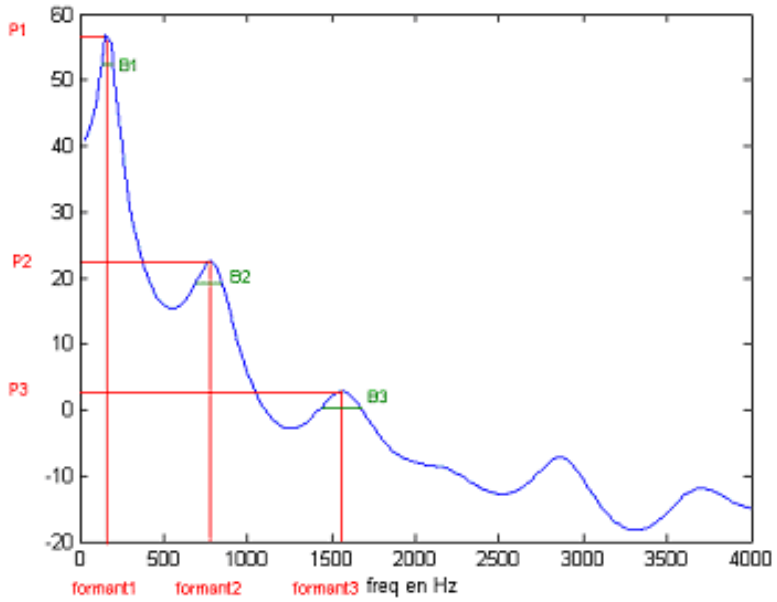


FIG. 2.5 – *Spectre de l'analyse LPC à l'ordre 10.*

L'analyse LPC permet de représenter l'enveloppe spectrale du signal à partir des coefficients de pondérations (α_i). Il existe des formules mathématiques qui permettent de transformer les coefficients de pondération en coefficients cepstraux. Les coefficients cepstraux calculés à partir des coefficients de pondération de l'analyse LPC sont nommés *LPCC, Linear Predictive Cepstral Coefficients*. Le calcul de ces coefficients est décrit dans les formules 2.7,2.9 et2.9.

$$\widehat{c_1} = -\alpha_1 \tag{2.7}$$

$$\widehat{c_n} = -\alpha_n - \sum_{m=1}^{n-1}(1-\frac{m}{n})\alpha_m \widehat{c}_{n-m} \qquad (1 \leq n \leq p) \tag{2.8}$$

$$\widehat{c_n} = -\sum_{m=1}^{p}(1-\frac{m}{n})\alpha_m \widehat{c}_{n-m} \qquad (p \leq n) \tag{2.9}$$

où α_n est le coefficient de pondération d'ordre i, et où $\widehat{c_n}$ est le coefficient LPCC d'ordre n.

Les coefficients cepstraux, qu'ils soient calculés à partir de la transformée de Fourier du signal ou de l'analyse LPC, représentent l'influence du conduit vocal et de la source sur le signal de parole émis. Le nombre de coefficients cepstraux calculés détermine le niveau de lissage de l'estimation de l'enveloppe spectrale. Les coefficients cesptraux d'ordre faible sont très utilisés en RAL. Ils caractérisent un trait anatomique de l'individu, principalement le conduit vocal. Une autre propriété intéressante des coefficients cepstraux est leur faible corrélation, introduite par la transformation DCT. Enfin, le calcul des coefficients cepstraux est indépendant de l'énergie du signal d'entrée, ce qui permet de réduire en conséquence la variabilité du signal.

2.1.3.2.4 La prosodie

La prosodie est définie comme une information qui n'est pas localisée dans un segment de parole spécifique ou une information qui ne change pas l'identité des segments de parole [Childers et al., 1998]. Les informations de pitch (fréquence fondamentale) et ses variations, de durée, d'énergie, de stress ou encore d'accentuation sont des informations classées comme prosodiques. Les informations prosodiques peuvent être un moyen de discriminer les locuteurs, et des systèmes de RAL basés sur ce type de paramètres réapparaissent [Helander et Nurminen, 2007; Nitin et Raina, 2004; Adami et al., 2003; Dehak et al., 2007; Shriberg et al., 2005]. Originellement, les systèmes de RAL étaient basés sur l'utilisation de la fréquence fondamentale. Les faibles performances de ces systèmes ont poussé la recherche à déterminer de nouveaux paramètres aujourd'hui utilisés. Les performances des systèmes actuels, basés sur la prosodie, n'égalent toujours pas celles des systèmes basés sur l'utilisation des coefficients cepstraux.
De tels systèmes sont donc souvent utilisés en complément des approches état de l'art. Leurs résultats sont souvent décorrélés des approches standards et permettent d'améliorer les performances dans certaines conditions d'utilisation, par fusion avec des systèmes basés sur les coefficients cepstraux.

Ces paramètres sont difficiles à estimer. Une difficulté majeure réside dans le choix de la taille de la fenêtre d'analyse du signal pour l'estimation des paramètres prosodique. Pour discriminer les variations de paramètres prosodiques entre chaque phonème, les paramètres sont généralement estimés sur des fenêtres de 50 à 60 ms. Mais les approches les plus robustes utilisent une segmentation en phonèmes pour déterminer les bornes de l'analyse. Une autre approche [Arciénega, 2006] propose de segmenter

l'espace acoustique en catégorie d'évènements à l'aide d'un classifieur et d'utiliser les paramètres prosodiques localement sur ces zones de l'espace pour discriminer les locuteurs.

2.1.3.2.5 La fréquence fondamentale ou pitch et le voisement

La fréquence fondamentale ou pitch est souvent classée dans la catégorie des paramètres prosodiques. C'est un paramètre très difficile à estimer. Le pitch ne peut être estimé que sur des trames voisées (structure harmonique). La technique du *zero crossing* (le nombre de changements de signes par secondes) est la plus ancienne technique de calcul du pitch. Les résultats obtenus par cette technique dans le cas bruité ne permettent cependant pas de conclure avec précision sur l'estimation du pitch. Une seconde technique, plus répandue, consiste à utiliser l'autocorrélation du signal temporel. L'estimation du pitch correspond alors à la recherche du retard donnant la corrélation maximale. A partir du cepstre, il est aussi possible de déterminer la fréquence fondamentale de la source en détectant les pics périodiques au-delà du $c(0)$. Ce paramètre est intéressant car il présente une faible variation pour le locuteur [Ezzaidi et al., 2001]. Néanmoins les variabilités du pitch entre locuteurs ne sont pas suffisantes pour baser la RAL sur ce paramètre. Ce paramètre peut être complémentaire. La fréquence fondamentale des femmes est par exemple généralement plus élevée que celle des hommes. Dès lors on peut classer les locuteurs par genre grâce au seul paramètre de pitch. La figure 2.6 présente les distributions des valeurs de pitch pour une population d'hommes et de femmes. La séparation nette des distributions permet une classification aisée des genres.

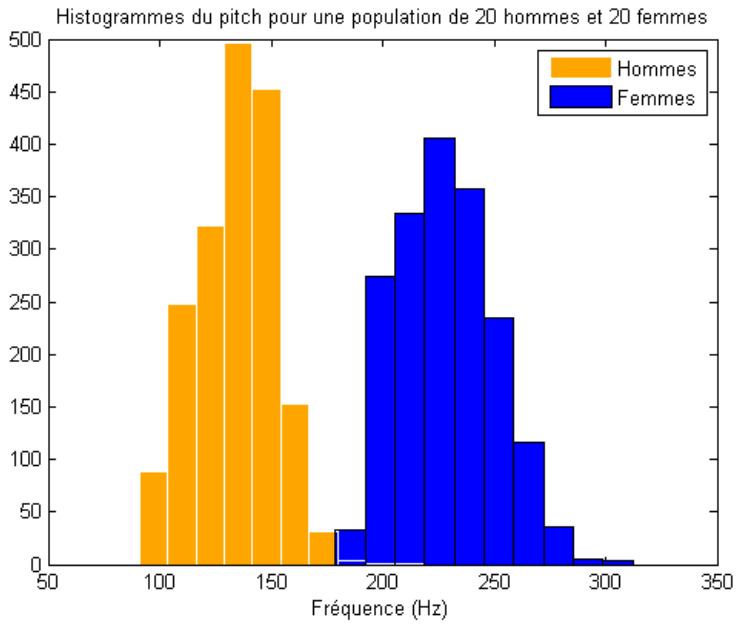


FIG. 2.6 – *Evolution du paramètre de pitch pour des signaux de parole prononcés par une population de femmes et d'hommes (base BREF).*

2.1.3.2.6 Trajectoires des paramètres

L'information dynamique d'évolution des paramètres de parole est aussi considérée comme un paramètre intrinsèque de l'élocution d'une personne. Elle permet de prendre en compte les variations temporelles des coefficients et, ainsi, de mieux prendre en compte la dynamique d'élocution [Fredouille, 2000; Furui, 1978a; Song et Rosenberg, 1986]. Il s'agit d'une dynamique à court terme (moins de 10 trames) représentant les variations d'enchaînement des différents phonèmes. Les dérivées premières et secondes des paramètres sont calculées par approximation polynomiale. Le plus souvent les dérivées d'ordre 1 et 2 des coefficients cepstraux et de l'énergie du signal sont utilisées. Les variations des coefficients cepstraux représentent la dynamique de modification du conduit vocal lors de l'élocution. La variation de l'énergie du signal est quant à elle liée à l'intonation. Il s'agit d'un paramètre pouvant être classé dans la catégorie des paramètres prosodiques.

2.2 La Reconnaissance Automatique du Locuteur

L'objectif de la reconnaissance du locuteur est de reconnaître l'identité d'une personne à l'aide de sa voix. Les applications de la RAL sont principalement liées aux problèmes d'authentification ou de confidentialité.

2.2.1 Les différentes tâches

La reconnaissance du locuteur est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs sont regroupés en trois catégories principales :

- l'identification de locuteurs,
- la vérification du locuteur,
- l'indexation en locuteurs ou le suivi de locuteurs.

Chacune de ces catégories propose son protocole de reconnaissance selon que l'identité du locuteur à reconnaître soit proclamée, ou que les locuteurs à reconnaître soient connus ou non du système de RAL. Le système de RAL peut valider une identité pour la vérification du locuteur, proposer une identité à partir d'un ensemble de locuteurs, déterminer les durées de parole d'un locuteur, compter le nombre de locuteurs présents dans un signal.

Une seconde classification à l'intérieur de ces catégories repose sur le niveau de dépendance au texte. La reconnaissance peut être indépendante du texte ou dépendante du texte. En mode dépendant du texte la reconnaissance bénéficie de la connaissance du contenu linguistique prononcé (fixe ou prompté). L'estimation des paramètres caractéristiques du locuteur est alors plus robuste. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

2.2.1.1 Identification automatique du locuteur

L'identification du locuteur consiste à déterminer l'identité d'un individu parmi une population de personnes connues. A partir d'un échantillon de voix enregistré, il faut déterminer quel locuteur de la base a parlé. Deux modes sont distingués : le fonctionnement en milieu fermé et le fonctionnement en milieu ouvert. En milieu fermé le locuteur est supposé faire partie de la population connue. Le système retourne l'identité du locuteur le plus probable parmi la population. En milieu ouvert le locuteur peut ne pas être connu du système. Dans ce cas le système associe au locuteur le plus probable un indice de fiabilité. Le locuteur le plus probable peut aussi être le locuteur « inconnu ».

2.2.1.2 Vérification automatique du locuteur

La vérification du locuteur consiste à déterminer si l'identité proclamée d'un message vocal correspond à la véritable identité du locuteur. En pratique la réponse est binaire, acceptation ou rejet. Les éléments mis en jeu sont donc une identité proclamée et la référence associée à un échantillon connu de l'identité proclamée. Une mesure de similarité entre le signal à vérifier et cette référence est calculée. Cette mesure est comparée à un seuil de vérification. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.

2.2.1.3 Indexation automatique en locuteurs

L'indexation en locuteurs permet de déterminer les temps de parole des individus dans un signal audio.
La spécificité de cette tâche réside dans le fait que le système ne détienne pas de référence pour les locuteurs présents dans le signal audio. Un mécanisme d'apprentissage aveugle et adaptatif est alors mis en place [Meignier, 2002]. Il est possible de segmenter un signal audio par prise de parole des intervenants, étiqueter des données audio pour permettre des recherches de documents audio par locuteurs ou, enfin, identifier le nombre de locuteurs présents dans le signal.
Le suivi de locuteur est similaire à l'indexation en locuteur, à ceci près que les locuteurs présents dans le signal sont connus par le système de RAL. Il s'agit donc d'une simplification de la tâche d'indexation en locuteur mais qui reste néanmoins, une tâche très complexe.

2.2.2 Scénarios

Ces différentes tâches de reconnaissance du locuteur permettent de mettre en place de nombreux scénarios applicatifs distincts. La détermination du scénario s'effectue en fonction des besoins et contraintes exprimés :

- **la tâche** : vérifier une identité, segmenter les tours de parole des locuteurs, trouver une identité,
- **le mode de dépendance au texte** : prompté, fixé au préalable, libre ;
- **les références connues des locuteurs** : population inconnue, taille de la population connue, le sexe de la population, la ou les langue(s) parlée(s) par la population, la durée des références, la qualité d'enregistrement ;
- **le signal audio disponible pour la tâche** : sa taille, un locuteur présent ou plusieurs, sa qualité d'enregistrement ;
- **le matériel disponible** : la puissance de calcul et de stockage. Le traitement est-il différé ou en en temps réel ?

Une telle définition du scénario est nécessaire pour considérer certains principes généraux de RAL :

- selon le niveau de dépendance au texte les méthodes mises en oeuvre sont différentes,
- la paramétrisation doit être adaptée à la langue et au genre des locuteurs, comme par exemple avec le retrait de l'échelle de Mel pour les signaux féminins [Mason et Thompson, 1993] ou l'utilisation de la prosodie pour les langues tonales [Bin et Meng, 2004; Auckenthaler et al., 2001; Kleynhans et Barnard, 2005],
- la qualité et la quantité des enregistrements d'apprentissage et de test sont déterminantes,
- l'augmentation du nombre de locuteurs à reconnaître diminue les performances en identification du locuteur [Furui, 1978b].

Les méthodes de RAL mises en oeuvre peuvent être très différentes selon le scénario ainsi défini. La serrure vocale est un exemple d'utilisation de l'identification du locuteur en milieu ouvert. Les locuteurs ayant accès au bâtiment protégé sont connus. De manière coopérative ils ont enregistré un message (fixé ou libre) pour servir de référence au système de RAL. La vérification du locuteur peut, quant à elle, remplacer l'utilisation des mots de passe pour sécuriser les transactions bancaires, ou les accès aux systèmes informatiques. Sur les systèmes informatiques, l'utilisateur doit proclamer son identité par un « login » et s'authentifier avec son mot de passe. La vérification du locuteur peut alors authentifier l'utilisateur en comparant un échantillon de sa voix avec une référence, enregistrée au préalable et associée au « login ». Les systèmes d'indexation du locuteur sont particulièrement utiles pour le traitement des bases de données audio. Une recherche des documents audio dans lesquels un locuteur est intervenu devient possible.

2.3 Les approches classiques pour la RAL

L'architecture d'un système de RAL est décomposée en différents modules de traitement. La plupart des approches standards en RAL utilisent une structure similaire :

- un module de paramétrisation : il extrait du signal les éléments permettant une discrimination des locuteurs,
- un module de création des références de locuteurs : à partir des données du locuteur, extraites par le module de paramétrisation, une référence du locuteur est

créée. Elle sert d'élément de référence pour la RAL, où elle est comparée avec le signal de test,
- un module de test : il effectue la comparaison entre la référence (tâche de vérification) ou les références (tâche d'identification) et le signal de test,
- un module de décision : à partir du résultat du module précédent, cet étage rend la décision (le nom d'un locuteur en identification, un rejet ou une acceptation en vérification) et prenant en compte différents éléments comme le niveau de sécurité souhaité.

Différentes méthodologies sont utilisées en RAL pour réaliser les références de locuteurs. Les approches génératives regroupent des méthodes qui utilisent les données d'apprentissage pour modéliser les densités de probabilité de chaque classe, par une famille de fonctions paramétriques. L'approche générative dominante pour représenter la référence du locuteur, en RAL indépendante du texte, est le modèle de mélanges de Gaussiennes (*GMM, Gaussian Mixture Model*). Elle a été introduite par [Reynolds et Rose, 1995; Reynolds et al., 2000] et constitue l'état de l'art des systèmes de VAL. Cette approche est détaillée dans le chapitre 3.3.

Il existe d'autres approches génératives comme les modèles de Markov cachés (HMM, Hidden Markov Model). Les HMM sont très employés en RAL dépendante du texte car ils sont capables de capturer les dépendances temporelles entre différentes variables aléatoires. Dans le cas de la RAL dépendante du texte, la modélisation des variations temporelles, des distributions des paramètres acoustiques, permet de très bonnes performances [Rosenberg et Soong, 1992].

Les approches à base de quantification vectorielle ont été utilisées en RAL. Elles proposent une représentation minimale d'une classe de paramètres observés : un représentant (dans un dictionnaire) pour chaque classe [Soong et al., 1985]. Chaque classe de paramètres est déterminée par un algorithme de classification du type K-moyennes. Cette représentation est choisie en minimisant la distance entre le centroïde et les paramètres de la population observée. Ces approches ne sont plus très employées depuis l'apparition des GMM en RAL.

L'approche discriminante la plus employée en RAL sont les *Support Vector Machine (SVM)* [Wan et Campbell, 2000]. A l'origine, ils ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes. Ils démontrent aujourd'hui des performances similaires à l'approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM/SVM Super-Vecteur [Campbell et al., 2006] qui profite des capacités génératives du GMM et discriminantes du SVM.

2.3.1 La prise de décision

La prise de décision en RAL est basée sur le formalisme probabiliste. Elle est différente pour l'identification et la vérification du locuteur.

2.3.1.1 Décision pour l'identification automatique du locuteur

Considérons une population de locuteur $i = 1, ..., N$ avec M_i la référence associée au locuteur i. L'identité retournée M, présente dans le signal X, est alors celle qui maximalise la probabilité :

$$M = argmax_i P(M_i|X) \tag{2.10}$$

Sans informations *a priori* sur l'apparition des locuteurs, $P(M_i)$, et en appliquant la règle de Bayes la relation devient :

$$M = argmax_i P(M_i|X) = argmax_i \frac{p(X|M_i).P(M_i)}{P(X)} = argmax_i p(X|M_i) \tag{2.11}$$

où $p(S|M_i)$ est la fonction de vraisemblance du locuteur i qui approxime la densité de probabilité des observations du locuteur i. Lorsque le nombre de locuteurs augmente dans la base de référence, des proximités entre locuteurs apparaissent. Il est plus difficile de différencier les locuteurs et les performances se dégradent [Furui, 1978b]. Il faut aussi noter que, dans ce cas, les ressources nécessaires et les temps de traitement augmentent.

2.3.1.2 Décision pour la vérification du locuteur

Considérons une identité proclamée M. Selon l'approche probabiliste, le calcul de la probabilité que le signal $X = \vec{x_1}, .., \vec{x_T}$ ait été prononcé par le locuteur M repose sur le test d'hypothèse suivant :

- H_0 : X est une occurrence prononcée par le locuteur M ;
- H_1 : X n'a pas été prononcé par le locuteur M mais par un autre locuteur que M.

Une des deux hypothèses doit être validée par le système de VAL. L'hypothèse H_0 est représentée par la fonction de vraisemblance $p(X|H_0)$ et l'hypothèse H_1 est représentée par la fonction de vraisemblance $p(X|H_1)$. Le problème de vérification est résolu en comparant le rapport de ces deux hypothèses à un seuil de décision. Dans le cadre de la théorie de la décision bayésienne, le rapport de vraisemblance des deux hypothèses (*likelihood ratio*) est défini par :

$$LR(X, H_0, H_1) = \frac{P(H_0|X)}{P(H_1|X)} \tag{2.12}$$

En appliquant la règles de Bayes :

$$P(H_i|X) = \frac{p(X|H_i)P(H_i)}{P(X)} \tag{2.13}$$

$$LR(X, H_0, H_1) = \frac{p(X|H_0)P(H_0)}{p(X|H_1)P(H_1)} \tag{2.14}$$

$$\begin{aligned} LR(X, H_0, H_1) &< \theta \quad \text{l'hypothèse } H_0 \text{ est rejetée} \\ LR(X, H_0, H_1) &> \theta \quad \text{l'hypothèse } H_0 \text{ est validée} \end{aligned}$$

où θ est le seuil de décision. En pratique les probabilités *a priori* $P(H_0)$ et $P(H_1)$ sont reportées dans le calcul du seuil de décision θ. Le rapport de vraisemblance devient alors :

$$LR(X, H_0, H_1) = \frac{p(X|H_0)}{p(X|H_1)} \lessgtr \theta . \frac{P(H_1)}{P(H_0)} \tag{2.15}$$

La modélisation de l'hypothèse de l'imposture H_1 est réalisée à l'aide d'un modèle du «non-locuteur». Il représente l'ensemble des locuteurs autres que M. Son estimation est une tâche difficile. Différentes approches sont proposées. La première approche consiste à utiliser une cohorte de locuteurs. Les locuteurs peuvent être sélectionnés selon un critère de proximité avec le locuteur M [Rosenberg et al., 1992]. La vraisemblance de l'hypothèse H_1 est alors une fonction (somme, max, ...) des vraisemblances du signal sur les modèles des locuteurs de la cohorte $(\overline{M_1}, ..., \overline{M_n})$: $LR(X|H_1) = f(LR(X|\overline{M_1}), ..., LR(X|\overline{M_n}))$.

Une seconde approche consiste à utiliser un modèle unique pour le modèle du «non-locuteur» [Carey et Parris, 1992; Reynolds et Rose, 1995]. Ce modèle, dénommé modèle du monde ou *UBM (Universal Background Model)*, est estimé sur une grande quantité d'enregistrements de locuteurs. Il représente toute la variabilité de la parole [Reynolds et Rose, 1995]. La modélisation de l'hypothèse H_0 utilise quant à elle les données disponibles du locuteur.

2.4 Evaluation d'un système de VAL

2.4.1 Le score de vérification

Le rapport de vraisemblance (*Likelihood Ratio*), ou son logarithme (utilisé pour des éviter les problèmes de précisions arithmétiques) est utilisé comme score de vérification pour la VAL. L'hypothèse d'indépendance temporelle des observations acoustiques $X = \vec{x_1}, .., \vec{x_T}$ permet d'écrire le *LLR (Log Likelihood Ratio)* comme la somme des LLR pour chaque trame de l'énoncé X. Cette somme est normalisée par le nombre de trames présentes dans l'énoncé. La décision de vérification s'effectue en comparant le LLR à un seuil θ, déterminé empiriquement.

2.4.2 Mesures de performances

Les performances d'un système de VAL s'évaluent en fonction de deux taux d'erreurs. La probabilité de faux rejets (FR) ou de rejet du client à l'identité proclamée et la

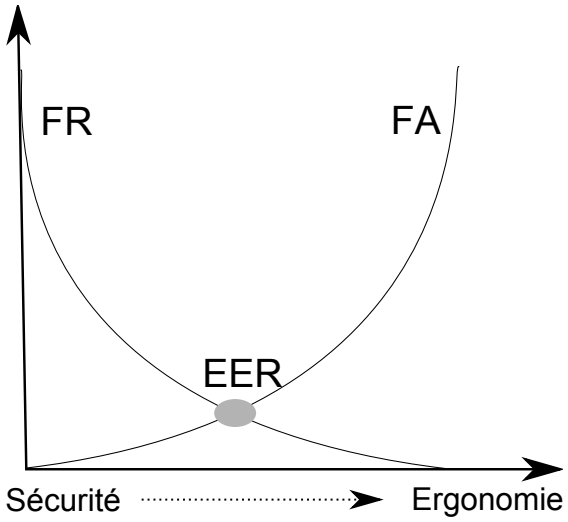


FIG. 2.7 – *Evolution des taux FA et FR.*

probabilité de fausses acceptations (FA) ou d'acceptations d'impostures. Ces taux sont étroitement liés. Au point de fonctionnement, pour un certain seuil de vérification, ces deux taux sont définis. En fonction du type d'application souhaitée, le seuil de vérification peut être choisi pour minimiser le taux de fausses acceptations : application de sécurité, ou minimiser le taux de faux rejets pour augmenter l'ergonomie d'utilisation. Il n'est pas possible de minimiser conjointement ces deux taux (cf. figure 2.7).

2.4.3 Les courbes DET

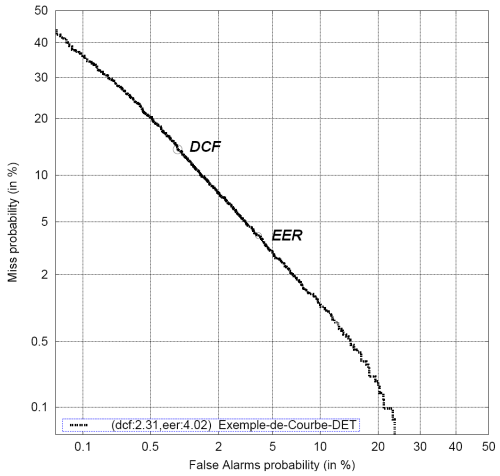


FIG. 2.8 – *Exemple de courbe DET (False alarms : FA, Miss probability : FR).*

La représentation la plus communément utilisée pour évaluer la pertinence du seuil de décision en fonction de ces deux taux d'erreurs est la courbe DET (*Detection Error Tra-*

deoff [Martin et al., 1997]). Les échelles des axes suivent la répartition d'une loi normale contrairement à leurs prédécesseurs, les courbes *ROC (Receiver Operating Characteristic)*, qui utilisent une échelle linéaire.
L'échelle logarithmique est utilisée pour rendre la courbe DET linéaire quand les scores des systèmes suivent une distribution Gaussienne. La courbe DET permet d'évaluer, pour chaque seuil de vérification, les valeurs du couple (FA, FR). La figure 2.8 illustre un exemple de courbe DET.
D'autres solutions ont été proposées pour la représentation des performances d'un système de RAL :

- la courbe EPC *(Expected Performance Curve)* [Bengio et Mariethoz, 2004],
- la courbe APE *(Applied Probability of Error)* [van Leeuwen et Brummer, 2007].

2.4.4 Les points de fonctionnement

Pour comparer les systèmes de RAL deux points de fonctionnement sont extraits pour caractériser plus simplement ces courbes. Le taux d'erreurs égales ou EER (*Equal Error Rate*) défini comme le point de fonctionnement où FA = FR. A ce point de fonctionnement aucune priorité n'est donnée à la minimisation des FA ou de FR. Cette mesure est très utilisée pour comparer les performances des systèmes de RAL.
Pour introduire une pondération pour chacun de ces taux, en fonction du contexte applicatif, une fonction de coût de détection (*DCF, Decision Cost Function*) peut être appliquée. Cette DCF s'exprime sous la forme :

$$DCF = C_{FA}\tau_{FA}P_{false} + C_{FR}\tau_{FR}P_{true} \tag{2.16}$$

où :
τ_{FA} est le taux de fausses acceptations ;
τ_{FR} est le taux de faux rejets ;
C_{FA} est le coût associé à une fausse acceptation ;
C_{FR} est le coût associé à un faux rejet ;
P_{true} est la probabilité *a priori* d'un accès client ;
P_{false} la probabilité d'une imposture.

Une autre mesure, dénommée HTER ou *Half Total Error Rate*, est définie comme la distribution du taux d'erreur moyen pour chaque seuil de décision [Bengio et Mariethoz, 2004].

$$HTER = \frac{1}{2}(FA + FR) \tag{2.17}$$

Les taux d'erreurs sont liés au point de fonctionnement d'utilisation. Le réglage du seuil de décision est effectué sur une population de tests, *a priori*. La calibration de

ce seuil est très importante. Une variation du seuil entre la phase de calibration et de fonctionnement éloigne le système du point de fonctionnement optimal souhaité.

Le point de fonctionnement réel peut être déterminé *a posteriori*. C'est notamment le cas lors de campagnes d'évaluations des système de VAL. Le point de fonctionnement optimal qui minimise le critère DCF est comparé au point de fonctionnement fixé *a priori*. Cette mesure, nommée minDCF, permet d'évaluer l'erreur de calibration du seuil de décision.

En général, pour comparer les performances des systèmes de RAL, le pourcentage relatif de gain/perte, pour les mesures DCF et EER, est utilisé :

$$\% \text{ relatif} = \frac{V_1 - V_2}{V_1} \tag{2.18}$$

où, V peut être la mesure EER ou DCF.

2.4.5 Les corpus utilisés

Un système de VAL s'évalue sur des données de développement. Ces données sont choisies pour leur proximité avec les données réelles que le système de VAL va devoir analyser. Cette phase de développement joue un rôle essentiel. Elle va notamment permettre de calibrer un seuil de vérification et d'évaluer les performances du système. La base de données de développements doit représenter au mieux les variabilités de la parole qui seront présentes dans le système en fonctionnement réel.

Ainsi les variabilités intra-locuteur et inter-session, la variabilité due à l'environnement, la variabilité due au canal d'enregistrement (depuis le combiné/micro jusqu'à la chaîne de transmission du signal), la variabilité inter-locuteur (genre mais aussi les locuteurs montrant des particularités, voir la « ménagerie » de Doddington, [Doddington et al., 1998]) doivent être représentées. La connaissance *a priori* des conditions réelles de fonctionnement est alors nécessaire. Ainsi si le système de VAL est destiné à effectuer de l'authentification sur le réseau téléphonique, la base de données doit être majoritairement composée d'enregistrements téléphoniques.

Il n'existe pas de moyen théorique permettant d'estimer la fiabilité des performances d'un système de VAL sur une base de données de développement [Dass et al., 2006]. Une règle empirique, la « règle des 30 » [Porter, 2000], stipule qu'une erreur est bien modélisée lorsque 30 exemples de cette erreur sont présents dans les tests. Par exemple, pour valider un taux d'erreur de faux rejet de 1%, la règle précise que 30 erreurs de faux rejets doivent être représentées, soit $30 * 100 = 3000$ tests clients. Cette règle permet d'obtenir un pourcentage sur la fiabilité du taux d'erreur du système de VAL, en fonction des types de tests représentés dans la base de développement. Mais les nombreuses variabilités de la parole ne peuvent toutes être reproduites dans une base de données d'enregistrements. Aussi, l'évaluation d'un système de VAL sur des données de développements constitue une simulation.

Chapitre 3

L'approche statistique GMM-UBM pour la vérification du locuteur

Sommaire

L'approche statistique est majoritairement utilisée en VAL. Elle permet de définir une mesure de similarité entre une référence du locuteur et un ensemble de données de test. La référence du locuteur est un modèle statistique qui prend en compte les variabilités du signal de parole. Ce modèle décrit la distribution statistique des observations acoustiques issues des données d'apprentissage. Il doit néanmoins être appris à partir de données représentatives du locuteur et des variabilités de son signal de parole. La mesure de similarité entre un ensemble de paramètres caractérisant le signal audio, S, et la référence du locuteur i, M_i, est représentée par une probabilité, la probabilité de S sachant M_i ($P(S|Mi)$).

Dans ce cadre statistique, les observations acoustiques des signaux d'enregistrements sont utilisées pour :

1. générer un modèle statistique du locuteur,
2. calculer une mesure de similarité entre le signal de test et le modèle statistique du locuteur.

La modélisation des locuteurs repose sur les modèles à base de mélanges de Gaussiennes (GMM). L'hypothèse inverse dans la théorie bayésienne est réalisée à l'aide du GMM du monde (UBM). Cette méthode est plus communément appelée GMM-UBM [Bimbot et al., 2004]. Ce chapitre présente l'application de la méthode GMM-UBM pour la VAL indépendante du texte.

3.1 Schéma général

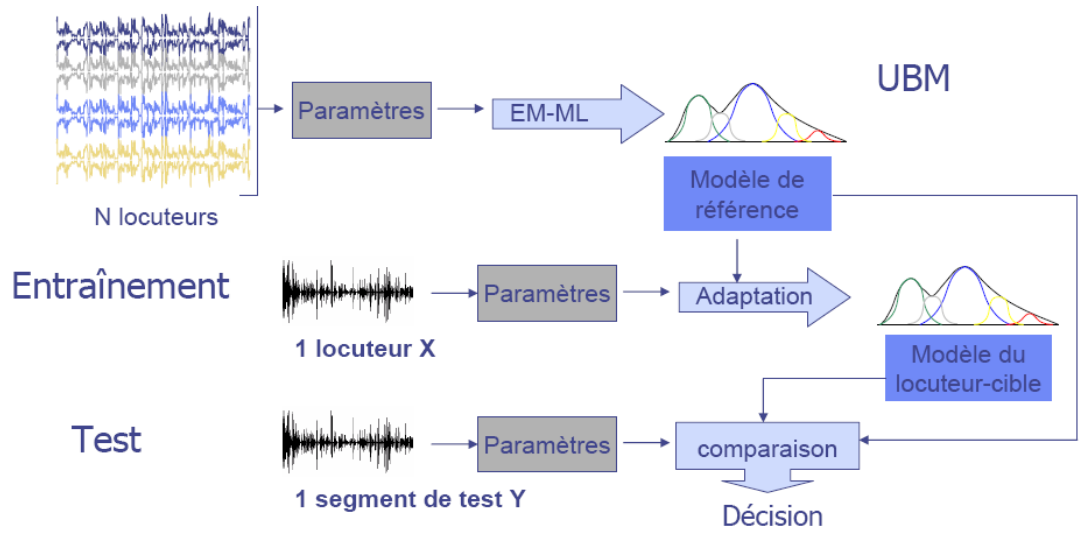


FIG. 3.1 – *Schéma de la méthode GMM-UBM pour la VAL indépendante du texte.*

Le schéma de fonctionnement est représenté sur la figure 3.1. Les différents modules représentés sont :

- le module « Paramètres ». Il permet d'extraire les paramètres du signal de parole pertinents pour la VAL.
- les modules « Modèle de référence et modèle du locuteur-cible », qui estiment, à partir des données d'apprentissage, les modèles statistiques des locuteurs.
- le module « Comparaison », qui calcule la mesure de similarité entre l'échantillon de test et le modèle de locuteur cible. Il fournit la décision de vérification.

La suite de ce chapitre décrit chacun de ces modules.

3.2 La paramétrisation du signal de parole

Le module de paramétrisation de la parole permet d'extraire les paramètres de représentation du signal de parole. Nous avons présenté, au chapitre précédent, les pa-

ramètres permettant de mettre en évidence les informations discriminantes du point de vue du locuteur. Comme vue en section 2.1.3.2, la paramétrisation en VAL est très proche de la paramétrisation utilisée en reconnaissance de la parole. Elle est majoritairement fondée sur l'analyse cepstrale (cf. section 2.1.3.2.2).

3.2.1 L'extraction des coefficients cepstraux

Nous avons précisé que les coefficients cepstraux peuvent être déterminés en utilisant une méthode non paramétrique, l'analyse cepstrale (MFCC ou LFCC), ou une méthode paramétrique, l'analyse LPC du signal (LPCC). Le calcul des coefficients cepstraux est détaillé au paragraphe 2.1.3.2.2.

L'analyse cepstrale est l'approche la plus utilisée en VAL, notamment parce qu'elle présente une plus grande robustesse d'estimation sur des signaux bruités [Tierney, 1980]. En VAL, entre 15 et 20 coefficients cepstraux sont utilisés pour modéliser un locuteur. Ils sont généralement extraits toutes les 10 ms (hypothèse de pseudo-stationnarité), calculés sur une fenêtre d'analyse de type Hamming ou Hanning de 20 à 30 ms.
Une analyse en banc de filtre à échelle linéaire ou échelle de Mel est utilisée dans le calcul des coefficients cepstraux (*LFCC ou MFCC*). Les dérivées premières, ou coefficients Δ (vitesse), et parfois secondes, ou coefficients $\Delta\Delta$(accélération), des coefficients cepstraux sont ajoutés au vecteur de paramètres pour modéliser leurs trajectoires dans le temps. Les coefficients dynamiques (Δ et $\Delta\Delta$) sont généralement estimés sur des fenêtres d'analyse de 5 à 9 trames.
L'énergie du signal joue aussi un rôle important en RAL tant au niveau de la sélection des données utiles que comme paramètre. En effet, ce paramètre est souvent utilisé pour la détection d'activité vocale, et sa trajectoire (Δ-log-énergie) est souvent ajoutée au vecteur de paramètres.

3.2.2 La détection d'activité vocale

Dans la production d'un énoncé de parole, toute la séquence n'est pas utilisable pour identifier un locuteur [Besacier et Bonastre, 1998; Weddin et Winther, 2006]. Il est essentiel de retirer les trames de silence ou de bruits qui ne reflètent pas les caractéristiques du locuteur et qui diminuent les performances de VAL. Le lecteur pourra se référer à la section 4.3.3 pour des résultats d'expériences qui illustrent l'influence du nombre de trames sélectionnées sur les performances de VAL.
Le critère d'énergie peut être utilisé pour sélectionner les trames de parole (haute énergie) et éliminer les trames de silence (énergie très faible) ou de bruits d'ambiance (énergie moyenne) (cf. figure 2.3). En général, les DAV sont basées sur une classification des trames suivant le critère d'énergie, à partir d'une modélisation GMM (à 2 ou 3 composantes) du paramètre d'énergie [Besacier et al., 2000; Bonastre et al., 2004]. Une ou deux composantes regroupent les trames de bruit ou de moyenne énergie et une composante les trames de forte énergie. La figure 3.2 illustre le principe de la modélisation de l'énergie des trames par un mélange de Gaussiennes. La sélection des trames peut s'effectuer

par seuillage ou par sélection des trames appartenant à la Gaussienne de plus hautes énergies.
D'autres procédés ont été envisagés pour la DAV en VAL, comme l'apprentissage d'un

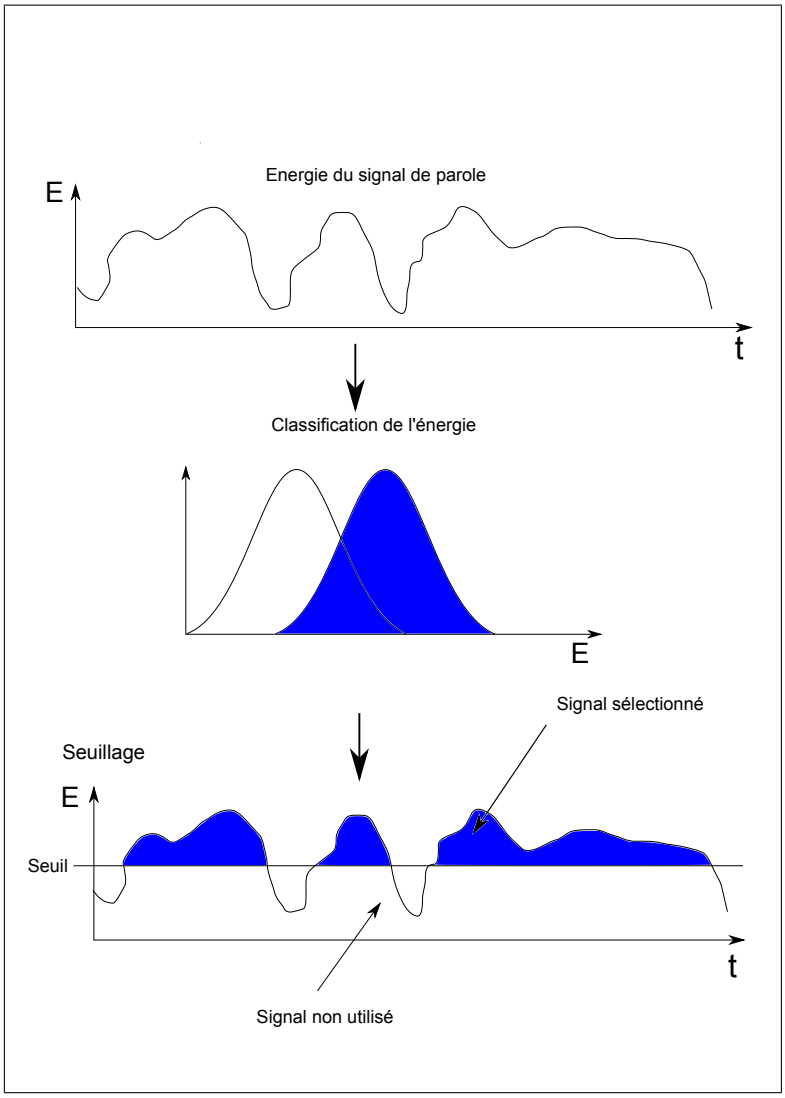


FIG. 3.2 – *Principe de la modélisation de l'énergie à base de Gaussiennes pour la DAV.*

HMM à deux états : parole, non parole, ou simplement l'utilisation d'un seuil sur l'énergie des trames. Les trames d'énergie en dessous de ce seuil sont éliminées. Ce seuil est déterminé empiriquement.

La forte corrélation entre le voisement de la trame et l'énergie permet d'utiliser des DAV basées sur le critère de voisement [Matejka et al., 2007; Adami et al., 2003; Lovekin et al., 2001]. La détermination du taux de voisement d'une trame s'effectue, le plus souvent, par la détection de la fréquence fondamentale (cf. section 2.1.3.2.5). Nous démontrons dans le chapitre 6 que l'utilisation d'une DAV, basée sur la sélection des

trames voisées, est une alternative efficace à la DAV basée sur la modélisation du critère d'énergie.

3.2.3 La normalisation des paramètres pour la compensation canal

Afin de réduire les perturbations du signal de parole, engendrées par les bruits additifs ou convolutifs, induites par la prise de son et la transmission sur un réseau de communication, des techniques de normalisation de paramètres acoustiques ont été développées.

3.2.3.1 Normalisation des variations long-terme du canal

Le canal de transmission, défini comme le trajet du signal de parole entre le locuteur et l'acquisition sur le combiné, ajoute une composante perturbatrice au signal de parole. Les méthodes de normalisation des variations à long-terme du canal font l'hypothèse que cette composante varie lentement sur un enregistrement et qu'elle est convolutive. Dans l'espace cepstral, les bruits de convolutions, variant lentement dans le temps, sont représentés par une composante additive, presque constante sur un enregistrement de parole. Retirer la moyenne à moyen ou long terme permet de supprimer la composante additive dans le domaine cepstral (convolutive dans le domaine temporel). La normalisation, par soustraction de la moyenne cepstrale (*CM, Cepstral Mean Susbtraction*) [Atal, 1974; Furui, 1981], consiste à retirer la moyenne de la distribution de chacun des paramètres cepstraux (composante continue), et permet d'atténuer l'influence du canal de communication. L' extension de la CMS, la normalisation *CMVN, Cepstral Mean and Variance Normalization*, est majoritairement employée en VAL [Viikki et Laurila, 1998]. Elle consiste, après soustraction cepstrale, à réduire la variance des paramètres en divisant par la variance globale des paramètres de l'enregistrement.

3.2.3.2 Normalisation des variations court-terme du canal

Pour prendre en compte les variations à court-terme du canal de transmission, généralement introduites par les déplacements du combiné d'acquisition qui changent la fonction de transfert du microphone, des techniques plus évoluées que celles citées précédemment ont été développées. Pour prendre en compte ces variations à court-terme, la CMVN est utilisée sur des fenêtres glissantes. Les paramètres de moyennes et de variances sont par exemple estimés sur une courte fenêtre temporelle, puis remis à jour.

Une autre technique, la « Gaussianisation » ou *feature warping* [Pelecanos et Sridharan, 2001], transforme chaques paramètres du vecteur de coefficients, pour que leur distribution -sur la fenêtre d'analyse- suivent une distribution Gaussienne, de moyenne nulle et de variance unité. Ce procédé est illustré dans la figure 3.3. La taille de la fenêtre d'estimation des paramètres de moyenne et variance des coefficients cepstraux est un

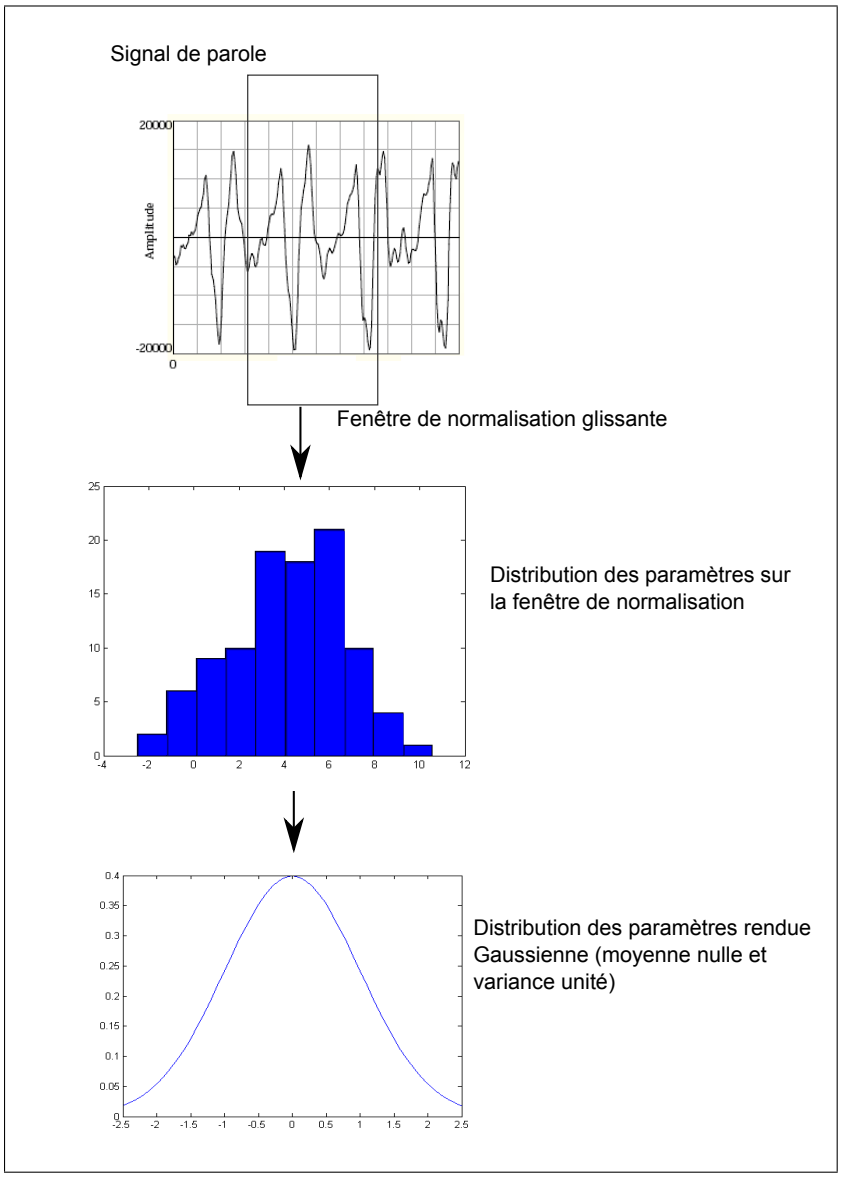


FIG. 3.3 – *Illustration de la méthode de gaussianisation.*

facteur important pour ces techniques [Pelecanos et Sridharan, 2001; Xiang et al., 2002; Viikki et Laurila, 1998]. En général des fenêtres de 300 trames sont utilisées.

Le filtrage RASTA [Hermansky et Morgan, 1994] introduit la compensation des paramètres cepstraux directement au niveau de leur calcul. Celle-ci consiste à ajouter un filtre passe-bande afin de réduire l'influence des bruits convolutifs dus au canal de transmission. Cette technique est très proche de la CMVN appliquée sur une fenêtre glissante.

L'égalisation aveugle s'affranchit du problème de l'horizon d'estimation des paramètres [Benveniste et Goursat, 1984; Mokbel et al., 1996; Shynk, 1992]. La méthode

consiste à utiliser les statistiques du signal à long terme, pour calculer l'erreur entre le vecteur de coefficients cepstraux du signal de parole et un cepstre de référence (issu d'un signal de parole long-terme [Mokbel et al., 1993, 1996]). Les coefficients d'un filtre adaptatif sont mis à jour pour filtrer les vecteurs de coefficients cepstraux, le plus souvent, par l'algorithme du gradient (LMS, Least Mean Square) [Shynk, 1992]. L'égalisation aveugle estime la compensation à appliquer à chaque vecteur de coefficients cepstraux ; la normalisation n'est pas appliquée coefficient par coefficient.

3.3 Modèles statistiques pour la VAL

[Reynolds et al., 2000] a proposé d'approximer la distribution complexe des coefficients acoustiques par des modèles de mélange de Gaussiennes en VAL . La modélisation des locuteurs par *GMM (Gaussian Mixture Model)* est aujourd'hui la méthode la plus performante et la plus répandue des approches statistiques en VAL.
Dans cette approche, la densité de probabilité $p(\vec{x}|\lambda)$, avec λ un GMM à M composantes pour des vecteurs acoustiques $\vec{x}$ de dimension D ($\vec{x} = x_1, .., x_D$) est définie de la façon suivante :

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\vec{x}) \tag{3.1}$$

La densité de probabilité $p(\vec{x})$ est une somme pondérée de M distributions Gaussiennes, chacune définie par un vecteur de paramètres de moyennes $\vec{\mu_i}$ de dimension D, et une matrice de covariance Σ_i de dimension $D * D$ (cf. équation 3.2) et où w_i est le poids associé à la Gaussienne dans le mélange. Les paramètres qui caractérisent le GMM sont : $\lambda = (w_i, \mu_i, \Sigma_i)$.

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu_i})' \Sigma_i^{-1} (\vec{x} - \vec{\mu_i})\right\} \tag{3.2}$$

La somme des poids w_i du mélange est égale à 1.

Les GMM sont utilisés pour leur capacité à modéliser la distribution de probabilités des coefficients cepstraux. Généralement les matrices de covariances utilisées sont diagonales. L'utilisation de matrices de covariance pleines, pour chaque composante Gaussienne dans le mélange, pose le problème de la grande quantité de paramètres à estimer à partir de peu de données. L'utilisation de matrices de covariance diagonales est acceptable car les paramètres cepstraux sont localement non corrélés entre eux et parce que le nombre de composantes utilisées dans le mélange est important (1000 à 2000 Gaussiennes). La figure 3.4 illustre la modélisation de la distribution d'un paramètre par un GMM à deux composantes.

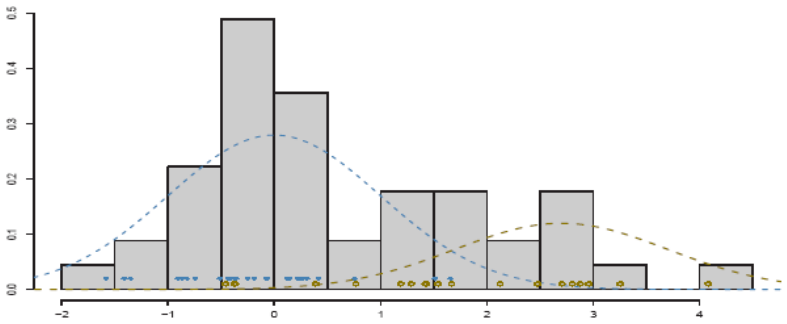


FIG. 3.4 – *Illustration de l'utilisation des GMM pour modéliser des distributions (extrait de [Marin et al., 2005]*

3.3.1 L'apprentissage des modèles GMM

Pour créer un modèle statistique du locuteur, il est nécessaire de déterminer les paramètres de ce modèle (w_i, μ_i, Σ_i). Cette étape est généralement réalisée à partir d'un jeu de données dit d'apprentissage. Un algorithme est utilisé pour estimer ces paramètres en maximisant un critère choisi, vis à vis des données d'apprentissage. Le critère les plus utilisé pour l'apprentissage des modèles GMM, est le critère de maximum de vraisemblance *ML (maximum likelihood)*. L'estimation des paramètres du GMM consiste à trouver ceux qui maximisent la fonction de vraisemblance des données d'apprentissage.

$$\widetilde{\lambda_X} = argmax_\lambda p(X|\lambda) \tag{3.3}$$

où X est l'ensemble des trames d'apprentissage : $X = \vec{x_1}, ..., \vec{x_T}$ et $p(X|\lambda)$ la vraisemblance de X sachant le modèle GMM λ :

$$p(X|\lambda) = \prod_t p(\vec{x_t}|\lambda) \tag{3.4}$$

Il est très complexe de résoudre l'équation 3.4 à cause du manque d'information des données d'apprentissage. En effet, il est difficile de savoir quelle Gaussienne dans le mélange a généré une trame d'apprentissage donnée. Pour résoudre ce problème, appelé problème des données manquantes, l'algorithme *Expectation Maximisation (EM)* [Dempster et al., 1977] est communément utilisé. L'étape *Expectation* détermine les probabilités *a posteriori* que les Gaussiennes aient généré les trames d'apprentissage. Ensuite, l'étape *Maximisation* modifie les paramètres du modèle pour maximiser le critère choisi. Cette algorithme est itératif, il garantit l'augmentation de la vraisemblance des données sachant le modèle, $p(X|\lambda)$.

En pratique, il existe plusieurs critère pour l'apprentissage des modèle GMM de locuteurs : le critère de maximum de vraisemblance *ML (maximum likelihood)* présenté, le critère *MMI (maximum mutual information)* [1] et le critère *MAP (Maximum a posteriori)*.

[1] Le critère MMI est le plus souvent utiisé pour la reconnaissance de la langue ou de la parole.

Le choix du critère d'apprentissage dépend de la quantité de données d'apprentissage disponible. Lorsque cette quantité est limitée, l'estimation au sens du maximum de vraisemblance pose le problème du sur-apprentissage. Le modèle résultant est alors trop spécifique aux données et perd sa capacité de généralisation.
L'estimation au sens du critère MAP permet d'introduire les densités de probabilités des paramètres *a priori* du GMM λ :

$$\widetilde{\lambda_X} = argmax_\lambda p(X|\lambda)p(\lambda) \tag{3.5}$$

Le critère *MMI (Maximum Mutual Information)* [Woodland et Povey, 2002] vise à intégrer un critère discriminant lors de l'apprentissage. Les paramètres du GMM sont alors modifiés selon une fonction objective, dont le but est de diminuer l'influence des densités de probabilité qui modélisent des informations communes, entre le modèle d'apprentissage et des modèles de contre-exemple. Le lecteur pourra se référer, en annexe, à un travail effectué dans le cadre de cette thèse, sur l'intégration du critère *MMI* dans l'apprentissage des GMM pour la VAL.

La phase d'initialisation est très importante lors de l'apprentissage d'un modèle GMM. Les techniques les plus courantes sélectionnent aléatoirement des données dans l'ensemble d'apprentissage pour initialiser les moyennes, la matrice de variance est la matrice unité, et les poids suivent la loi d'équiprobabilité. Les moyennes initialisées du GMM peuvent être réactualisées par l'utilisation de l'algorithme de classification de type *k-means*.

Quand le nombre de composantes dans le mélange est élevée, des Gaussiennes de faible variance modélisent des éléments anormaux ou très peu représentés. Ce phénomène est appelé sur-apprentissage. Le calcul de vraisemblance est une somme normalisée des vraisemblances de chacune des trames. Si une trame est associée à une Gaussienne très spécifique de variance très faible, sa vraisemblance est très élevée et fausse le calcul de vraisemblance moyen. Des techniques de seuillage de la variance des distributions, *variance flooring*, permettent de contraindre l'apprentissage du GMM et limiter le problème. Cette technique consiste à définir une borne inférieure pour les variances du modèle GMM. Comme l'algorithme EM ne peut affecter de faibles variances aux composantes, la capacité de généralisation du modèle est augmentée. Dans ce cas le critère du maximum de vraisemblance n'est plus vérifié, mais le risque de sur-apprentissage est diminué.

3.3.2 Le modèle du non locuteur ou modèle du monde

Les approches statistiques nécessitent des données d'apprentissage suffisamment représentatives de l'espace d'élocution du locuteur, mais aussi des diverses conditions d'enregistrements. Ce postulat implique de connaître l'espace acoustique global de l'individu à reconnaître, pour chaque condition environnementale. A titre indicatif, il faut environ quatre heures d'enregistrement du même individu pour créer une synthèse correcte de sa voix. Ce chiffre donne une idée de la quantité de données nécessaires pour bien modéliser un locuteur. En général, quelques dizaines de secondes, dans le

meilleur des cas quelques minutes, sont disponibles. Avec cette quantité de données, il est difficile d'obtenir un modèle de locuteur robuste au sens du critère de ML. Puisque la quantité de données d'apprentissage des locuteurs est trop faible, d'autres données, en grande quantité, sont utilisées pour estimer un modèle. Ce modèle est ensuite adapté au locuteur par l'adaptation MAP, grâce aux données d'apprentissage du locuteur. L'adaptation MAP permet d'obtenir une meilleure estimation des modèles de locuteur, à condition de disposer d'une bonne estimation des informations *a priori*.

Le modèle du monde est utilisé comme *a priori* pour l'estimation des modèles de locuteur. Il représente la couverture exhaustive des paramètres acoustiques d'un grand nombre de locuteurs enregistrés dans diverses conditions. Ce modèle est appris par maximum de vraisemblance avec une très grande quantité de données. Ainsi les enregistrements collectés pour la création du modèle du monde doivent représenter :

1. les différentes langues de la population de locuteurs à reconnaître,
2. les genres considérés (homme, femme, les deux),
3. les conditions d'enregistrements (type de microphone, type de canal de transmission, conditions de bruits),
4. le type de parole : lue, spontanée, conversationnelle.

Ce modèle constitue une représentation moyenne d'un ensemble de locuteurs et des conditions d'enregistrement. Nous pouvons observer que le modèle du monde intervient à différents niveaux dans un système de VAL, il représente l'hypothèse inverse dans le test de vérification, et sert d'initialisation pour l'estimation des modèles de locuteur. Le lecteur pourra se référer aux travaux de [Scheffer, 2006] pour une description plus complète de son importance dans un système de VAL.

3.3.3 Estimation des modèles de locuteur

Pour palier le manque de données d'apprentissage pour réaliser une estimation robuste des GMM de locuteurs, l'adaptation MAP du modèle du monde est utilisée. Cette approche constitue l'état de l'art en VAL indépendante du texte.

3.3.3.1 L'adaptation MAP

L'adaptation d'un modèle GMM consiste à modifier les paramètres d'un modèle initial vers un modèle plus spécifique, par rapport aux données d'adaptation. L'adaptation MAP [Gauvain et Lee, 1994] est utilisée en VAL, pour adapter le modèle du monde aux données d'apprentissage des locuteurs. Seules les moyennes du GMM sont adaptées en VAL. La maximisation du paramètre de moyenne $\tilde{\mu}$, pour une Gaussienne i du GMM, s'exprime sous la forme :

$$\tilde{\mu_i} = \alpha_i \mu_i^c + (1 - \alpha_i)\mu_i^w \tag{3.6}$$

avec μ_i^c la moyenne i du GMM client et μ_i^w la moyenne i du GMM du monde. α est un coefficient de pondération qui permet d'affecter plus ou moins de poids aux paramètres *a priori* par rapport aux paramètres estimés sur les données d'apprentissage. Il est défini par :

$$\alpha_i = \frac{n_i}{n_i + \tau} \tag{3.7}$$

$$n_i = \sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda) \tag{3.8}$$

$$Pr(i|\vec{x_t}, \lambda) = \frac{w_i p_i \vec{x_t}}{\sum_{j=1}^{M} w_j p_j(\vec{x_t})} \tag{3.9}$$

avec n_i est le nombre de trames associées à la Gaussienne i défini par l'équation 3.8 et où $Pr(i|\vec{x_t}, \lambda)$ est la probabilité que la Gaussienne i, du GMM λ, ait généré le vecteur $\vec{x_t}$.

τ est le *relevance factor*. Il contrôle le degré d'adaptation de chaque Gaussienne en terme de trames attribuées. Un *relevance factor* de 14 est couramment utilisé en VAL. 14 signifie qu'une confiance égale est accordée au modèle de référence et aux données d'apprentissage, lorsque 14 trames d'apprentissage sont associées à une composante du GMM. En dessous de 14 trames, le facteur alpha devient très faible et l'influence des paramètres *a priori* est très forte. L'adaptation MAP permet ainsi de faire varier l'influence des données *a priori*, en fonction de la représentativité des données d'apprentissage pour chaque Gaussienne du modèle. Les paramètres du modèle du monde sont utilisés comme *a priori*. Cela permet d'initialiser les paramètres du modèle de locuteur de façon robuste. Une représentation de l'adaptation MAP des Gaussiennes du modèle du monde est proposée dans la figure 3.5. Cette figure illustre notamment le fait que les Gaussiennes du modèle du monde, pour lesquelles aucune donnée d'apprentissage n'est associée, ne sont pas adaptées.

3.3.4 Estimation robuste des modèles de locuteurs

Deux problèmes majeurs subsiste en VAL pour estimer correctement un modèle de locuteur : la faible quantité de données d'apprentissage et les variations du canal de transmission entre les enregistrements. Des techniques ont été développées pour répondre à ces problèmes.

3.3.4.1 La faible quantité de données pour estimer un modèle de locuteur

Nous avons introduit l'utilisation du modèle du monde, pour compenser le problème de la faible quantité de données d'apprentissage et pour l'estimation des GMM de locuteurs. Malgré l'utilisation de l'adaptation MAP du modèle du monde, la qualité des modèles de locuteurs n'est pas optimale. En effet, l'adaptation MAP modifie les moyennes de Gaussiennes les unes indépendamment des autres. De cette manière,

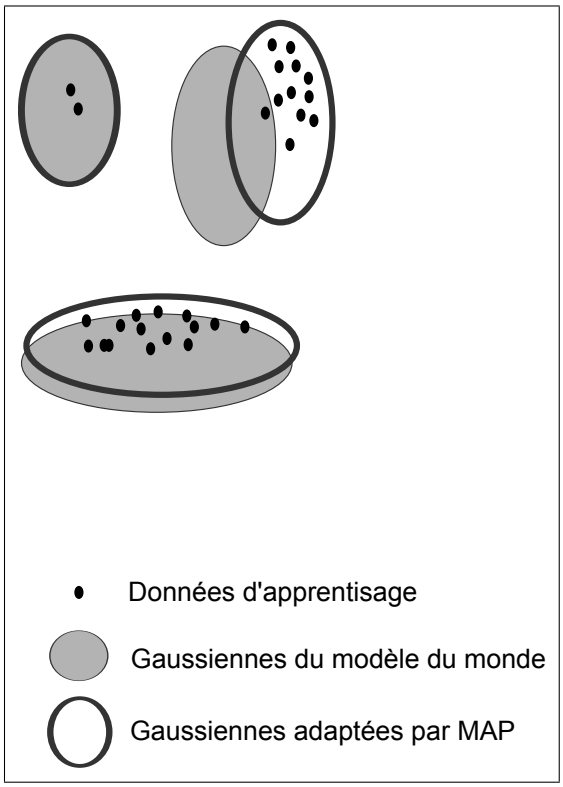


FIG. 3.5 – *Illustration de l'adaptation MAP d'un GMM à 3 composantes.*

certaines Gaussiennes peuvent rester inchangées du fait de l'absence de données d'apprentissage. Une solution consiste à faire en sorte que toutes les Gaussiennes (même celles sans données associées) soient adaptées en utilisant toutes les données d'apprentissage.

Pour cela, [Kuhn et al., 1998] émet l'hypothèse qu'un GMM de locuteur peut être décomposé comme étant une combinaison linéaire d'un certain nombre de GMM de locuteurs de base. Cette méthode est une extension de l'adaptation MAP, appelée *eigenvoice MAP*, en référence à la technique originelle utilisée en reconnaissance de visage *eigenfaces* [Turk et Pentland, 1991], dont elle s'inspire. En général, l'adaptation MAP du modèle du monde est réalisée sur les paramètres de moyenne du GMM. La décomposition en *eigenvoices* est réalisée sur les super-vecteurs[2] de moyennes des GMM locuteurs. La décomposition d'un supervecteur SV_s du locuteur est définie comme :

$$SV_s = SV_w + V.y_s \tag{3.10}$$

où SV_W est le supervecteur du modèle du monde, y_s est un vecteur regroupant les coefficients de la combinaison, et où les colonnes de la matrice V sont les vecteurs correspondant aux locuteurs de base, appelés voix propres. L'intérêt de ce formalisme réside dans deux observations : la première consiste à remarquer que la matrice V, ayant

[2]Le supervecteur de moyenne d'un GMM $SV(\lambda)$ est la concaténation des paramètres de moyenne d'un GMM λ. Sa dimension est : $dim(SV_\lambda) =$ Nb Gaussiennes (M) $*$ Nb Paramètres (D) $= MD$.

beaucoup de paramètres, est estimée en utilisant une large quantité de données provenant d'un grand nombre de locuteurs (indépendamment du locuteur à modéliser). Et la deuxième, où le vecteur y_s, contenant peu de paramètres, est estimé de manière robuste en utilisant les données d'apprentissage du locuteur. Les colonnes de la matrice V sont les vecteurs propres de la matrice de covariance des supervecteurs de moyennes (plusieurs centaines en pratique). Elles peuvent être calculées par maximum de vraisemblance [Kenny et al., 2005a] ou par l'analyse en composantes principales (ACP). Le vecteur y_s est estimé avec les données d'apprentissage du locuteur par maximum de vraisemblance une fois la matrice V calculée.

3.3.4.2 La compensation de canal

Les modèles de locuteur, réalisés par adaptation MAP du modèle du monde, sont adaptés au locuteur mais aussi aux conditions d'enregistrements. La variabilité inter-session est intégrée dans les modèles. Par exemple, si les données d'apprentissage d'un modèle de locuteur proviennent d'une conversation téléphonique, le modèle du locuteur est spécifique au locuteur, mais aussi au canal téléphonique. Ceci pose problème lors du calcul de vraisemblance entre des données enregistrées sur des types différents de canaux de transmission, et ce modèle de locuteur.

L'approche *feature mapping*, introduite par [Reynolds, 2003], exploite la connaissance du type de canal d'enregistrement, pour projeter les trames vers un espace non influencé par le canal de transmission. Une transformation est appliquée pour compenser cette projection et reprojeter les paramètres dans un espace indépendant du canal de transmission. Pour cela deux modèles GMM sont utilisés, le modèle du monde, indépendant du canal et un modèle dépendant du canal (*CD, Channel Dependent*), appris sur une quantité de données suffisantes, ce dernier pouvant être considéré comme un sous espace du modèle du monde. En pratique, il est réalisé par adaptation MAP du modèle du monde avec des données dépendantes du canal considéré. La normalisation du vecteur de paramètres est alors réalisée suivant l'équation :

$$\vec{x}_{norm} = (\vec{x} - \mu_i^{CD})\frac{\sigma_i^W}{\sigma_i^{CD}} + \mu_i^W \tag{3.11}$$

ou i est l'indice de la Gaussienne de plus forte vraisemblance dans le modèle du monde et σ_i et μ_i les moyennes et variances de Gaussiennes d'indice i dans les modèles indépendant (W) et dépendant (CD) du canal. La connaissance du canal d'enregistrement pour le choix du modèle dépendant du canal n'est pas toujours disponible. Un test de maximum de vraisemblance entre le signal et plusieurs modèles dépendant du canal permet de déterminer le canal le plus probable pour l'enregistrement considéré. L'inconvénient de cette méthode est qu'elle ne prend en compte qu'un certain nombre de canaux de transmission connus *a priori*.

[Kenny et Dumouchel., 2004] introduit la méthode *eigenchannel* pour caractériser l'information spécifique du canal de transmission (inter-session) dans la modélisation GMM des locuteurs. Cette technique peut être vue comme une extension du *feature*

mapping au cas continu (dans ce cas les canaux ne sont plus en nombre fini mais un espace continu). Les deux méthodes, *eigenvoices* et *eigenchannel*, sont conjuguées dans le formalisme du *joint factor analysis* [Kenny et al., 2005b; Kenny et Dumouchel., 2004]. Ce formalisme propose une décomposition du modèle du locuteur selon trois composantes :

- une indépendante du locuteur et de la session,
- une dépendante du locuteur,
- une dépendante de la session.

La composante indépendante du locuteur et de la session est introduite par l'utilisation du modèle du monde. Les données d'apprentissage du locuteur, spécifiques au locuteur, mais aussi à la session d'enregistrement, introduisent les deux autres composantes. Pour une session h d'enregistrement du locuteur s, le super-vecteur du locuteur dépendant de la session et du locuteur s'exprime :

$$SV_{h,s} = SV_W + V.y_s + U.x_{h,s} \tag{3.12}$$

où V est la matrice contenant les *eigenvoices*, y_s un vecteur contenant les *speaker-factors*, U est la matrice contenant les *eigenchannel*, et $x_{h,s}$ un vecteur contenant les *channel-factors*.

En pratique, les *eigenchannel* peuvent être estimés par l'analyse en composante principale ou par maximum de vraisemblance [Kenny et al., 2005a], à partir de la matrice de covariance des supervecteurs de plusieurs locuteurs, et avec plusieurs sessions par locuteur. Lorsque seule la décomposition en *eigenchannel* est réalisée (appelée *LFA, Latent Factor Analysis*), seule la variabilité inter-session est estimée. L'ajout de la décomposition selon les *eigenvoices* a prouvé ses bons résultats lors des dernières évaluations NIST SRE (2008)[3]. Le retrait de la composante, induite par le canal de transmission par la technique LFA, a permis de réduire grandement les taux d'erreurs des systèmes de VAL[4]. Les performances de la méthode réside dans une bonne estimation de la matrice U (généralement de rang faible : de l'ordre de 40). De très nombreux enregistrements sont alors nécessaires (en général une vingtaine de session pour une centaine de locuteurs).

Cette méthode a été appliquée de différentes façons dans les système de VAL. Ainsi [Vogt et al., 2005] propose d'estimer l'influence du canal dans le signal de test. Le canal du modèle d'apprentissage est ensuite remplacé par celui estimé sur le test, pour projeter l'apprentissage et le test dans le même espace. Il est à noter que le rapport de vraisemblance est à modifier pour compenser le fait que le modèle générique n'est pas décomposé selon le formalisme du *joint factor analysis*.
[Vair et al., 2006] et [P. Kenny et al., 2006] proposent de normaliser l'espace des paramètres en retirant la contribution du canal directement sur les vecteurs acoustiques.
Une autre méthode, appelée *symmetrical compensation* par ses auteurs [Matrouf et al., 2007], propose de retirer les variations du canal dans l'espace des paramètres, pour les énoncés de test, et dans l'espace des modèles pour l'apprentissage. Cette méthode a

[3] Pour un rang élevé de la matrice V, de l'ordre de 300.
[4] La méthode a régulièrement été évaluée lors des évaluations NIST SRE

démontré de meilleurs résultats que la compensation de canal complètement réalisée dans l'espace des paramètres, lorsqu' elle est utilisée dans le système LIA08.

3.4 Le test de vérification

Le test de vérification permet d'obtenir la mesure de similarité entre un modèle de locuteur et un signal de test. Cette mesure est appelée score de vérification.

3.4.1 Calcul du score vérification

En prenant l'hypothèse d'indépendance des réalisations du vecteur $\vec{x}$, la vraisemblance pour que le test $X = (\vec{x_1}, .., \vec{x_t})$ ait été généré par le GMM λ, est définie comme :

$$p(X|\lambda) = \prod_{t=1}^{T} w_i p_i(\vec{x_t}|\lambda) \tag{3.13}$$

où p_i est une distribution Gaussienne de moyenne $\vec{\mu_i}$ (cf. équation 3.2), avec une matrice de covariance Σ_i et où w_i est le poids associé à la Gaussienne dans le mélange.

Le test de vérification repose sur le rapport d'hypothèse défini en section 2.3.1.2. En utilisant le modèle du monde comme modèle de l'hypothèse inverse, le rapport d'hypothèse s'écrit sous la forme d'un rapport de vraisemblance (*likelihood ratio*) :

$$LR(X|S) = \frac{p(X|S)}{p(X|UBM)} \tag{3.14}$$

En pratique, on utilise le logarithmique des vraisemblances, ce qui donne le logarithme du rapport de vraisemblance *Log Likelihood Ratio (LLR)*, pour éviter les problèmes de précision numérique dus aux multiplications de faibles valeurs. Le score de vérification utilisé en VAL est alors défini comme :

$$Score(X|S) = LLR(X|S) = \frac{1}{T}(log(p(X|S)) - log(p(X|UBM))) \tag{3.15}$$

Une technique nommée *N-best scoring* a été introduite pour réduire le coût de calcul du LLR. Cette technique utilise une propriété issue de l'adaptation MAP du modèle du monde, pour créer les modèles de locuteur ; chaque Gaussienne d'indice i du modèle du monde peut être associée à la Gaussienne i des modèles clients. Le calcul de vraisemblance d'une trame x_t sur le modèle du monde permet de déterminer N composantes présentant les plus hautes valeurs en terme de vraisemblance. Le calcul de vraisemblance de la trame x_t sur le modèle client n'est alors effectué que pour ces composantes, en estimant que la contribution des autres composantes est négligeable dans le calcul du LLR.

3.4.2 La normalisation des scores

La variabilité du canal de transmission est un facteur important de perte de performance en VAL [Bin et al., 2007; Li et Porter, 1988]. Elle est souvent nommée variabilité inter-session, car c'est la différence de contexte entre plusieurs enregistrements qui la caractérise. Les différences de contexte d'enregistrement entre les sessions de test et d'apprentissage introduisent des disparités entre les données. Les paramètres, extraits du signal, sont influencés différemment par le canal de transmission et projetés dans des espaces différents. Ceci implique notamment une grande variabilité des scores de vérification. Le seuil de décision, fixé empiriquement, est commun pour toutes les conditions de test rencontrées. Il ne peut être optimal au regard de la variabilité des données. Pour renforcer la robustesse d'un système de VAL, des techniques de compensation au niveau des scores ont été proposées.

La normalisation de scores a pour but de proposer un score optimal pour chaque locuteur, Z-norm[Li et Porter, 1988; Reynolds et al., 2000], ou pour chaque test, T-norm [Auckenthaler et al., 2000], plus approprié à la comparaison avec un seuil de décision global. Les techniques de normalisation sont essentiellement basées sur l'analyse des distributions de scores clients et imposteurs du système de VAL, supposées distribuées selon une loi normale.

En général il est plus aisé d'estimer la moyenne et la variance des scores imposteurs, car les accès imposteurs sont facilement réalisables, alors qu'il est rare de posséder, *a priori*, plusieurs jeux de données étiquetés d'un même locuteur, pour pouvoir estimer correctement sa distribution de scores client. Les techniques de normalisation consistent à retrancher la moyenne de la distribution des scores imposteurs aux scores de vérification, puis à les diviser par la variance :

$$\widetilde{Score}(\vec{x}|S) = \frac{Score(\vec{x}|S) - \mu_{imp}}{\sigma_{imp}} \tag{3.16}$$

où μ_{imp} et σ_{imp} sont respectivement la moyenne et la variance des scores imposteurs.

Il est à noter que le rapport de vraisemblance peut être considéré comme une première étape de normalisation des scores [Rosenberg, 1992; Reynolds et Rose, 1995]. L'étape de division par la vraisemblance de l'énoncé de test sur le modèle du monde permet de diminuer la variance des scores. Les deux techniques Z-norm et T-norm sont les plus couramment utilisées en VAL, et se différencient par l'estimation des paramètres μ_{imp} et σ_{imp}.

3.4.2.1 Z-norm

Dans le cas de la Z-Norm, μ_{imp} et σ_{imp} sont les scores des séquences imposteurs comparées au modèle de locuteur cible (cf. figure 3.6). Il s'agit d'une normalisation dépendante du locuteur. Des énoncés imposteurs sont utilisés comme accès de test.

Les scores obtenus par ces accès permettent d'estimer la moyenne μ_{imp} et la variance σ_{imp}. Cette normalisation ne nécessite pas de connaissance sur les énoncés de test. Les paramètres de normalisation peuvent être estimés dès l'apprentissage d'un modèle de locuteur.
La Z-norm a pour but de définir un seuil de décision, dépendant du locuteur et de la qualité de son modèle.

3.4.2.2 T-norm

L'approche de normalisation T-norm repose également sur l'utilisation d'énoncés imposteurs. La différence réside dans le fait qu'ils sont utilisés pour apprendre des modèles imposteurs. Les paramètres μ_{imp} et σ_{imp} sont estimés par les scores de l'énoncé de test sur cet ensemble de modèles de locuteurs imposteurs (cf. figure 3.7). Il s'agit d'une normalisation dépendante du test. La T-norm permet de compenser les variations de conditions d'enregistrements observées dans les tests. Généralement l'effet de cette normalisation est d'améliorer les performances pour les points de fonctionnement à faible taux de fausses acceptations. En pratique, les durées des séquences imposteurs de test et d'apprentissage des modèles GMM imposteurs sont choisies de même durée que les séquences envisagées dans le système de VAL, pour coller au mieux aux distributions réelles de scores de fonctionnement. Ces méthodes sont d'autant plus performantes que les imposteurs choisis sont proches des locuteurs cibles [Sturim et Reynolds, 2005]. Ces deux normalisations sont souvent combinées, pour compenser à la fois les variabilités de l'apprentissage et des conditions de test. Elles sont appliquées l'une après l'autre pour former ZT-norm ou TZ-norm.

3.4.2.3 La normalisation WMAP

Une autre normalisation WMAP ou *World + MAP* a été proposée par [Fredouille et al., 1999]. L'idée ici est de projeter les scores dans un espace probabiliste. L'intervalle de mesure de vérification est restreint d'un intervalle non borné (LLR) à un intervalle [0,1]. Le rapport de vraisemblance est remplacé par la probabilité *a posteriori* d'être en présence d'un accès client, sachant le rapport de vraisemblance. Posons Y le locuteur de l'énoncé test y et X le modèle de l'identité proclamée. Alors le rapport de vraisemblance $LLR(y|X)$ est remplacé, selon l'approche WMAP, par $P(X = Y|LLR(y|X))$. En posant $LLR(y|X) = S_y$ et en appliquant la règle de Bayes cette probabilité s'exprime sous la forme :

$$P(X = Y|S_y) = \frac{p(S_y|X = Y)p(X = Y)}{p(S_y|X = Y)p(X = Y) + p(S_y|X \neq Y)p(X \neq Y)} \tag{3.17}$$

où $p(S_y|X = Y)$ est la probabilité du score de l'accès y, sachant que l'accès est client et, respectivement, $p(S_y|X \neq Y)$ la probabilité du score de l'accès y, sachant que l'accès est imposteur. Les probabilités $p(S_y|X = Y)$ et $p(S_y|X \neq Y)$ sont estimées à partir des distributions de LLR obtenues sur une base de développement. $p(X \neq Y)$ et

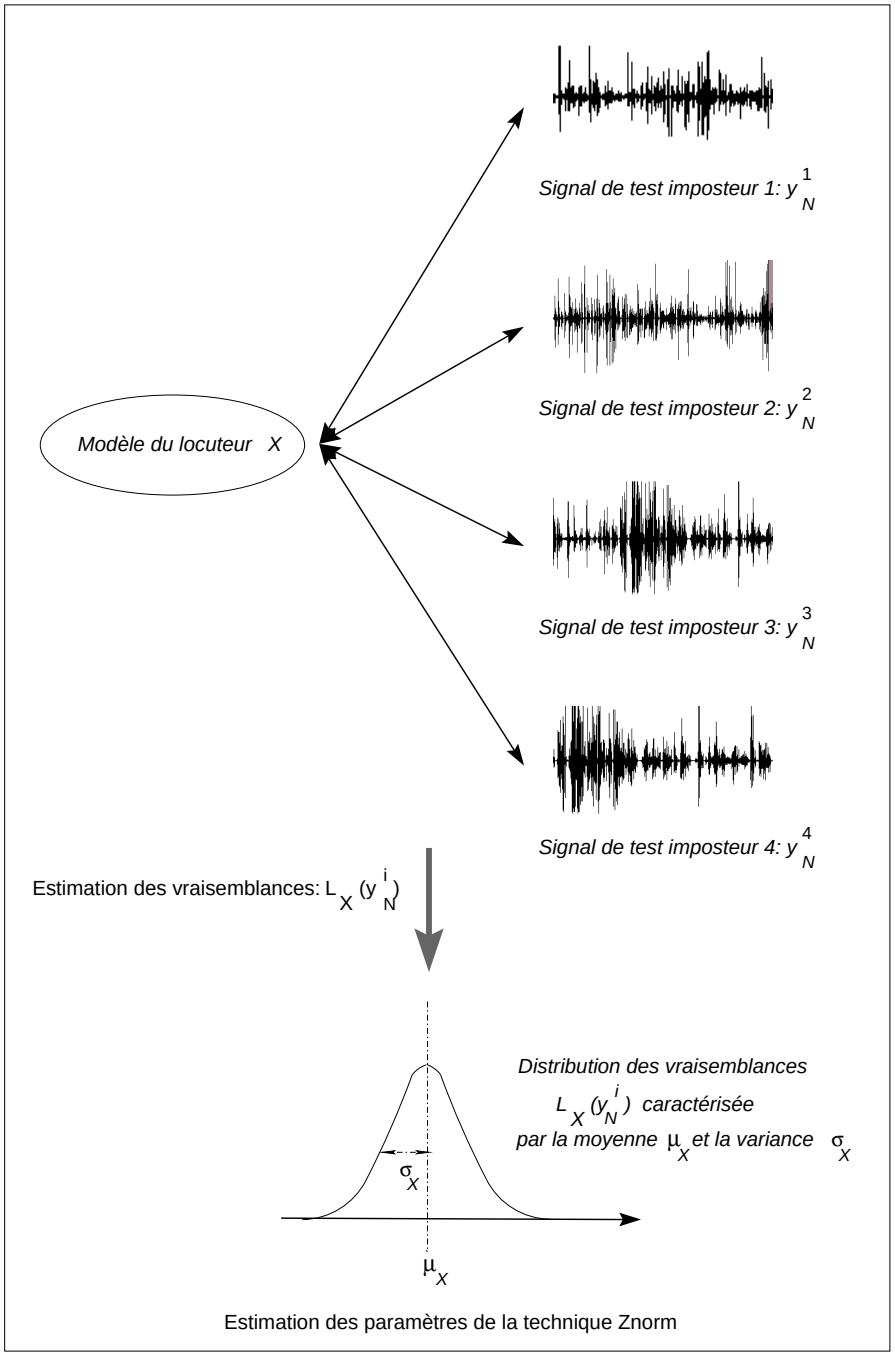


FIG. 3.6 – *Illustration z-norm, extrait de [Fredouille, 2000]*

$p(X = Y)$ sont les probabilités *a priori* que l'accès soit un accès imposteur et, respectivement, client. Elles sont constantes et fixées pour une application donnée. A partir de ces paramètres, une fonction WMAP est déterminée. Elle associe une probabilité à chaque valeur de LLR. Les résultats obtenus par cette technique de normalisation sont compa-

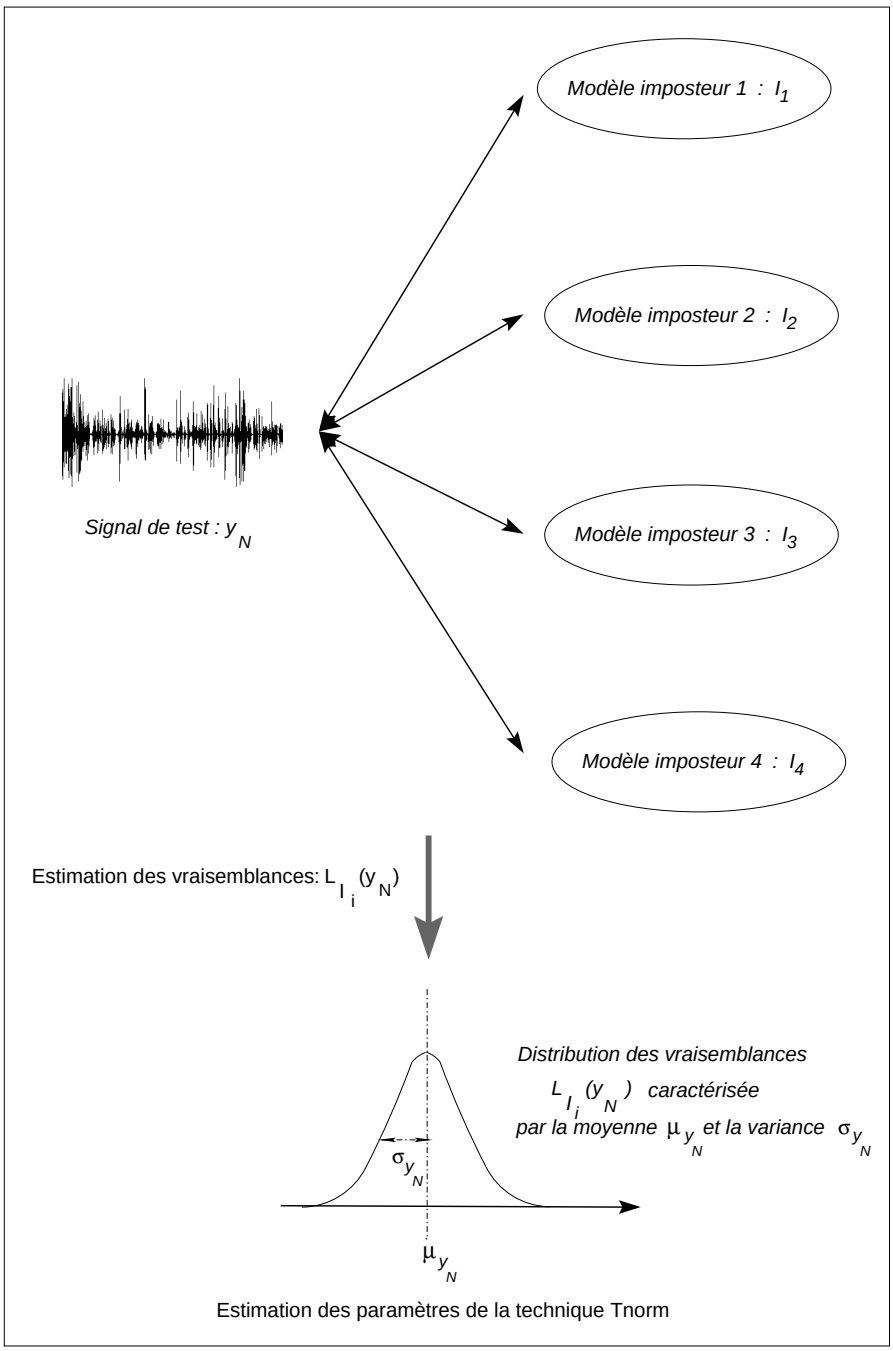


FIG. 3.7 – *Illustration T-norm, extrait de [Fredouille, 2000]*

rables à ceux obtenus par les approches état de l'art décrites auparavant. Néanmoins, les auteurs ont démontré que l'apprentissage des paramètres de la fonction de normalisation WMAP détermine son efficacité. Elle est alors dépendante de la similitude des deux jeux de données considérés (développement et test).

3.4.3 La fusion des scores

La fusion de systèmes est une technique très répandue en VAL. Elle met en jeu différents systèmes, le plus souvent basés sur une extraction de paramètres différents, mais aussi sur l'utilisation de divers classifieurs (*SVM, GMM, Neural Nets*). Chacun des systèmes fournit un score de vérification. Il est alors possible de les combiner (fusion) pour ne former qu'un seul score. La fusion se base sur l'hypothèse de complémentarité des scores [Campbell et al., 2001]. La fusion essaie de tirer parti des informations divergentes entre les systèmes. En effet, s'il existe des discordances entre systèmes, un apprentissage supervisé peut déterminer les zones de confiance dans la distribution des scores des systèmes, et leur attribuer un poids. Une combinaison linéaire (régression logistique [Brummer, 2005], SVM [Garcia-Romero et al., 2003], combinaison linéaire apprise par validation croisée [Scheffer et Bonastre, 2006]) ou non linéaire (*neural nets* [Campbell et al., 2004]) des scores est alors estimée par apprentissage supervisé. Les

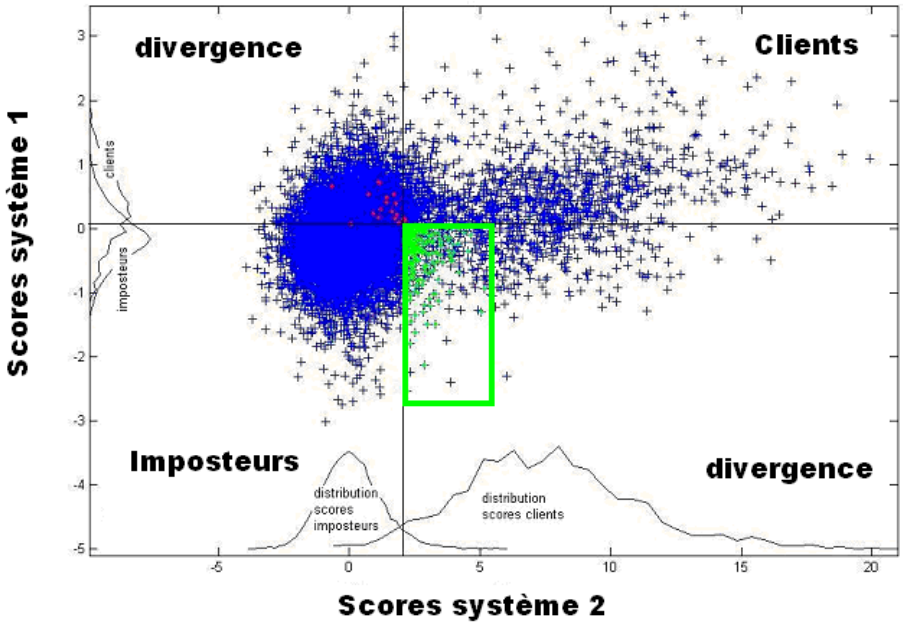


FIG. 3.8 – *Illustration des distributions des scores de deux systèmes. Les seuils de décision des deux systèmes sont représentés par les lignes. Les zones de divergences des deux systèmes sont respectivement en haut à droite et en bas à gauche. Ce type de représentation permet de déterminer l'importance des divergences entre les systèmes, en vue de les fusionner.*

approches de fusion sur les scores sont les plus répandues, car les plus simples à mettre en oeuvre. Il est alors facile de multiplier les systèmes, soit en multipliant les types de paramètres extraits du signal pour un même classifieur, soit en multipliant les classifieurs pour un même jeu de paramètres. Les scores fusionnés sont le plus souvent Z-normés et/ou T-normés au préalable. Des systèmes plus complexes intègrent la fusion au niveau de la modélisation statistique. Les différentes paramétrisations sont réunies en un seul vecteur dont chaque composante va être modélisée de façon classique. Il

faut néanmoins respecter l'hypothèse d'indépendance inter-paramètres, pour utiliser des matrices de covariance diagonales dans la modélisation GMM. Les performances des deux méthodes sont difficiles à comparer. [Conrad et Paliwal, 2001] montre, par exemple, que la fusion au niveau des scores n'est pas optimale.

La figure 3.8 illustre les scores clients et imposteurs de deux systèmes de VAL utilisant deux extractions de paramètres différentes. La figure est divisée en quatre parties. En haut à droite sont projetés les scores des accès clients pour les deux systèmes. Les scores des tests imposteurs sont représentés en bas à gauche. Les parties haute à gauche et basse à droite représentent les scores pour lesquels les systèmes sont en désaccord. Les points dans la zone encadrée sont les scores de tests imposteurs qui sont acceptés par un système (abscisses), mais rejetés par le deuxième (ordonnées). La fusion optimale consiste à utiliser les scores du système qui ne se trompe pas, seulement pour ces tests. La difficulté réside dans l'apprentissage de cette fusion. Ces méthodes sont très populaires, mais l'inconvénient qui en résulte est que les ressources à mettre en oeuvre peuvent être multipliées par le nombre de systèmes à fusionner. Des techniques de mutualisation d'informations et de calculs sont alors nécessaires.

Deuxième partie

Adaptation d'un système de RAL à la surveillance de réseaux professionnels de communication

Introduction

Dans la première partie de cette thèse, nous avons présenté l'approche statistique majoritaire, GMM-UBM, pour la VAL indépendante du texte. Cette deuxième partie expose les contributions majeures de ce travail de thèse, consacrées à la surveillance des réseaux professionnels de communication (PMR) : une application de la VAL dédiée à la vérification d'identité des locuteurs en cours de communication, sur ce type de réseaux.
Une mise en oeuvre de l'approche GMM-UBM est détaillée dans le chapitre 4. Le système ALIZE/SpkDet, servant de système de référence pour toute cette partie, y est décrit. Une évaluation des performances générales de ce système, basée sur les protocoles d'évaluations internationales NIST SRE, est également proposée. Le chapitre 5 s'attache à évaluer les contraintes de mise en oeuvre de la RAL sur un réseau PMR et à démontrer que pour un certain scénario applicatif, il existe des contraintes fortes sur les signaux de parole et sur les traitements algorithmiques associés. Nous proposons ensuite, dans le chapitre 6, une solution de paramétrisation optimisée pour prendre en compte le codage source, le bruit ambiant et le traitement en ligne. Les chapitres 7 et 8 exposent une nouvelle méthode pour l'adaptation non supervisée des modèles de locuteurs. La méthode proposée est détaillée dans le chapitre 7. Puis, une analyse de ses résultats, basée sur les protocoles d'évaluations NIST SRE, ainsi que des propositions d'améliorations de la méthode sont l'objet du chapitre 8.

Chapitre 4

Présentation du système GMM-UBM de référence SPKDET

Sommaire

Les différents résultats des contributions présentées dans ce travail de thèse sont évaluées comparativement à un système de référence, le système GMM-UBM ALIZE/SpkDet. Nous détaillons dans ce chapitre les différents modules qui le constituent. Nous présentons les résultats de ce système sur les bases d'évaluation internationales NIST SRE.

4.1 Historique du projet ALIZE

Le LIA a initié en 2003 un projet « logiciel libre » dédié à la RAL, « ALIZE » [Bonastre et al., 2005]. L'objectif principal du projet ALIZE est de proposer un recueil d'algorithmes à l'état de l'art pour la reconnaissance automatique du locuteur.
Elle a été développée dans le cadre du projet TECHNOLANGUE AGILE/ALIZE. Ce projet a été soutenu par le consortium ELISA, composé de plusieurs laboratoires francophones (ENST, IDIAP, IRISA, LIA et EPFL), qui existe depuis 1999. Un des principaux

objectifs du consortium a été de participer aux évaluations internationales de reconnaissance du locuteur, NIST SRE. Ceci permet notamment de faire progresser les techniques et de comparer les performances des systèmes de RAL. Cette volonté de proposer une plateforme logicielle, accessible à tous[1], permet de promouvoir la reconnaissance du locuteur (et ses applications), et de faciliter le transfert et la valorisation des connaissances entre les laboratoires académiques et le monde industriel. La communauté d'utilisateurs du système ALIZE ne cesse de grandir. De nombreux systèmes de vérification du locuteur sont basés sur cette plateforme.

4.2 Le système de RAL SpkDet

Le système SpkDet se décompose en différents modules qui abordent des thématiques différentes. Nous ne nous intéresserons qu'à la partie permettant de traiter la vérification automatique du locuteur, et plus précisément le système GMM-UBM disponible [Bonastre et al., 2008]. D'autres systèmes état de l'art sont disponibles dans SpkDet, comme un classifieur SVM intégrant les derniers noyaux et les techniques de compensation de canal (*NAP, Nuisance Attribute Projection*).

4.2.1 Le système GMM-UBM

Nous reprenons chacune des sections étudiées dans le chapitre 3.3, pour en décrire l'implémentation associée dans le système SpkDet.

4.2.2 L'extraction des paramètres acoustiques

L'extraction des paramètres cepstraux est réalisée avec SPRO[2]. L'extraction du vecteur de paramètres complet s'effectue toutes les 10 ms sur une fenêtre d'analyse de Hamming de 20 ms. 19 dérivées premières, ainsi que 11 dérivées secondes, sont ajoutées au vecteur de paramètres. Ces valeurs, déterminées empiriquement, ont prouvé qu'elles avaient de bonnes performances pour un système GMM-UBM, lors des évaluations NIST SRE [Fauve et al., 2007].

4.2.3 La détection d'activité vocale

Une détection parole/non parole est appliquée sur les vecteurs de paramètres [Bonastre et al., 2004]. L'élimination des trames, de silence, de bruit ou de peu d'information, a prouvé son efficacité pour les systèmes de RAL (cf. section 4.3.3). Elle est basée sur une modélisation de l'énergie des trames par un GMM à trois composantes. Le GMM est utilisé ici comme outil de classification. Les paramètres d'énergie des trames

[1] http ://alize.univ-avignon.fr
[2] Projet SPRO `http://gforge.inria.fr/projects/spro`

de tout l'enregistrement sont utilisés. Un paramètre permet de régler le pourcentage de trames sélectionnées dans la Gaussienne de moyenne intermédiaire. Ce paramètres est fixé à 0 pour sélectionner seulement les trames de la Gaussienne de plus haute moyenne. Environ 60% du signal sont sélectionnés par la DAV. Nous nous référons plus tard à cette DAV par le terme DAV LIA.

4.2.4 La compensation de canal

Les techniques de compensation de canal CMVN, *Feature mapping* et *Latent Factor Analysis* (cf. section 3.2.3) sont intégrées dans le système SpkDet.

4.2.4.1 La normalisation moyenne variance

La normalisation moyenne variance (CMVN) est appliquée globalement sur les trames de tout l'enregistrement. Les paramètres de moyenne et variance globales des paramètres cepstraux, après retrait de trames, sont estimés sur la totalité des trames, puis chacun des paramètres est normalisé par un centré-réduit.

4.2.4.2 *Feature mapping*

Le *Feature mapping* a été utilisé dans le cadre des évaluations NIST SRE 2005 et 2006. Trois modèles de canal, représentant le canal GSM, le canal RTC et le canal sans-fil ont été appris par adaptation MAP du modèle du monde, avec des données de développement. Ces trois modèles sont utilisés pour déterminer le canal présent dans le fichier de test (calcul de vraisemblance), et pour appliquer la technique de projection vers un canal indépendant décrite dans la section 3.3.4.2.

4.2.4.3 *Latent Factor Analysis*

La méthode de compensation *Latent Factor Analysis* (LFA) appliquée, est la méthode symétrique[3] proposée dans [Matrouf et al., 2007]. L'estimation de la matrice U est réalisée avec un nombre important de locuteurs et de sessions pour chaque locuteur. Ces enregistrements proviennent de bases de données de développement très corrélées avec les données d'évaluations (les bases NIST SRE des années antérieures à l'évaluation). Les techniques de compensation de canal *feature mapping* ainsi que *Latent Factor Analysis* (LFA) sont intégrées dans SpkDet. Depuis 2007 seule la technique de LFA est utilisée.

[3] ALIZE/SpkDet inclut aussi la méthode de compensation dans l'espace des paramètres [Vair et al., 2006].

Système	Compensation canal
LIA06	*feature mapping*
LIA08	*symmetrical LFA*
LIA-THL06, 07	CMVN seule
LIA-THL08	*feature space LFA*

TAB. 4.1 – *Type de compensation de canal pour chaque système GMM-UBM présenté (Une description complète des systèmes est fournie en annexe A).*

4.2.5 Modélisation

Dans ce travail, la taille des modèles GMM est de 512 composantes Gaussiennes. Ce choix est motivé par la réduction de la complexité de calcul. L'utilisation de GMM à 2048 composantes présente de meilleures performances, mais la réduction de la quantité de calcul permet à des techniques comme le LFA d'être plus rapidement mises en oeuvre. L'apprentissage du modèle du monde est effectué grâce à l'algorithme EM par maximum de vraisemblance. Le seuillage des paramètres de variance est appliqué pour éviter le sur apprentissage. Pour chaque composante Gaussienne, les matrices de covariances sont seuillées à 50% de la variance globale. En d'autres termes, pour la composante i du modèle et le paramètres d :

$$\begin{aligned} \text{Si } \Sigma_i(d,d) &< 0.5 * \text{Variance globale} \\ \text{alors } \Sigma_i(d,d) &= 0.5 * \text{Variance globale} \end{aligned}$$

L'initialisation du modèle du monde est réalisée par tirage aléatoire des vecteurs acoustiques de l'ensemble d'apprentissage. Les données d'apprentissage du modèle du monde sont constituées de plusieurs sous-ensembles, différenciés par le canal de transmission utilisé, la langue des locuteurs .... Chacun de ces sous-ensembles est pondéré par un poids qui permet de régler son influence pour le tirage aléatoire des données et l'optimisation du modèle.

Les modèles de locuteurs sont dérivés par adaptation bayésienne MAP des moyennes de l'UBM. Une itération de l'algorithme EM est utilisée pour calculer les statistiques clients, puis les moyennes sont mises à jour par adaptation MAP, avec un *relevance factor* de 14. Le score de vérification est calculé en utilisant les 10 meilleures Gaussiennes. Les normalisations Z-norm et T-norm sont disponibles et très souvent utilisées.

4.3 Evaluation des performances du système

Nous présentons les performances du système ALIZE/SpkDet sur les bases de données d'enregistrements téléphoniques issues des campagnes NIST SRE. Nous avons aussi utilisé une autre base de données, la base BREF 120, qui nous a permis d'isoler certains facteurs de variation du signal de parole, comme le codage à bas-débit ou l'ajout de bruit ambiant.

4.3.1 Les corpus d'évaluation

4.3.1.1 Les campagnes d'évaluation NIST SRE

L'institut américain NIST (National Institute of Standards and Technology) organise annuellement des campagnes d'évaluation des systèmes de VAL (*SRE speaker recognition evaluation*) depuis 1996 [Przybocki et Martin, 1998]. Des laboratoires scientifiques du monde entier y participent. Ces campagnes ont pour but de fournir un cadre d'évaluation strict des systèmes de VAL. Un corpus de données conséquent, qui adresse différentes problématiques de VAL, est fourni à chaque participant avec le protocole d'évaluation à suivre. Celui-ci propose différentes tâches qui diffèrent par les durées d'apprentissage et de test disponibles.
Ces évaluations ont permis d'identifier de nouvelles problématiques en VAL. La collecte de la base de données est réalisée avec l'objectif d'adresser une ou plusieurs de ces problématiques. Un effort est ainsi engagé pour collecter des données couvrant au mieux les variations du signal de parole (décrite en section 2.1.2). Depuis ces dernières années, les enregistrements sont collectés sur de multiples canaux de communications (téléphone, GSM, microphones de marque diverses), ainsi que dans diverses langues. 15 langues sont aujourd'hui représentées, des enregistrements de locuteurs dont la langue n'est pas la langue maternelle sont aussi présents.
En mettant l'accent sur une ou plusieurs problématiques, lors de la collecte de la base de données, le NIST oriente le développement des nouvelles technologies de VAL[4]. Les résultats de ces évaluations sont partagés entre les participants. Ils sont classés par problématiques. Les systèmes sont évalués dans les multiples combinaisons que propose l'évaluation, performances sur les sous-ensembles de langues, de canaux, etc... Les mesures des performances s'effectuent selon la mesure DCF (cf. equation 2.16) avec les paramètres suivants : $C_{FA} = 1; \tau_{FA}; C_{FR} = 10; P_{true} = 0.01$. La mesure DCF s'exprime selon l'équation 4.1 :

$$DCF = 1.\tau_{FA}0.99 + 10\tau_{FR}.0.01 \tag{4.1}$$

Les résultats en termes d'EER sont aussi fournis. Le coût d'une fausse acceptation est très pénalisant dans ce calcul de la DCF. Le point de fonctionnement choisi est en adéquation avec les applications où la sécurité prime. La probabilité *a priori* d'observation d'un test client est fixée à 0.01. Les campagnes d'évaluations NIST sont une compétition. Les participants développent leurs systèmes de VAL dans le but de proposer un système bien classé lors de cette évaluation. Les techniques appliquées sont alors destinées à la seule application potentielle que vise le protocole d'évaluation. [Bonastre, 2008] met en garde la communauté sur les performances des systèmes de VAL sur ces bases de données. Ces performances sont biaisées lorsque l'on parle d'appliquer de tels systèmes dans d'autres cadres applicatifs.

[4]En fait, ces choix sont plus guidés par le sponsor des évaluations, le département de la défense américaine, *DoD*, que par le NIST

4.3.1.2 Critiques des bases de données NIST

L'application visée dans ce travail de thèse ne relève pas des mêmes problématiques que celles abordées dans le cadre de ces évaluations. Les bases de données NIST proposent une première approche pour évaluer les variabilités du signal de parole. Ainsi divers canaux de communications sont utilisés, et les enregistrements sont, pour certains, espacés dans le temps. De plus il s'agit de discours spontanés dans diverses langues. Nous pouvons citer les contraintes suivantes, propres à la mise en place d'un système de RAL sur les réseaux professionnels de communications :

1. les durées de test et d'apprentissage sont variables et plus courtes. La phase d'apprentissage peut être coopérative, mais la phase de test, quant à elle, doit être transparente pour l'utilisateur, et ne peut se baser que sur de courts énoncés dans un langage « opérationnel ».
2. les signaux sont codés à bas-débit ;
3. le bruit d'environnement peut être très élevé (le RSB (Rapport Signal à Bruit) moyen sur la base NIST SRE 2005 est aux alentours de 25dB) ;
4. les accès sont majoritairement des test clients (la probabilité d'un test client sur les bases NIST est de 0.1)
5. la vérification d'identité est effectuée « à la volée » (le traitement des bases NIST s'effectue fichier par fichier) ;
6. le nombre de clients est plus faible.
7. les ressources de calculs sont limitées.

Pour évaluer les techniques que nous proposons, nous avons défini notre propre protocole complémentaire d'évaluation, en utilisant la base de données BREF.

4.3.1.3 La base BREF 120

La base d'enregistrements BREF [Lamel et al., 1991] est composée de phrases en langue française lues dans un environnement non bruité. Les signaux sont échantillonnés à 16KHz. Elle se compose d'enregistrements provenant de 120 locuteurs, 65 femmes et 55 hommes ayant été enregistrés pendant de longues durées, environ 20 minutes pour l'enregistrement le plus court. La couverture en phonème des enregistrements lus est considérée comme riche, les textes lus proviennent du journal Le Monde. Cette base a été choisie pour mener des expériences sur des conditions bien particulières comme le codage des signaux à bas-débit ou l'ajout de bruit. L'avantage majeur de l'utilisation de cette base de données réside dans le fait que la qualité des signaux enregistrés est presque optimale. Les signaux sont échantillonnés à 16kHz, sans bruit ambiant et sans codage de la parole. Il est possible d'effectuer des expériences pour isoler les influences de divers facteurs sur la RAL :

- le bruit ambiant par bruitage de la base,
- le codage à bas débit de la parole par type de codeur.

De plus, la grande quantité d'enregistrements disponible par locuteur permet de faire varier aisément les durées d'apprentissage et de test. La contrainte majeure de l'utilisation d'une telle base de données est la non spontanéité des enregistrements (textes lus). Les résultats proposés sur cette base de données sont donc des simulations. Une phase de test avec des enregistrements réels est nécessaire pour les valider.

4.3.2 Les résultats de référence

Les résultats présentées dans de document de thèse sont majoritairement représentés par des courbes DET avec les pourcentages EER et minDCF associés. Le protocole, ainsi que le système de RAL utilisé, sont indiqués sous la forme **(protocole, système - type de normalisation de scores)**. Les définitions des protocoles ainsi que des systèmes sont fournies en annexe A.

Base de données	Système	DCF	EER
NIST SRE 2005	LIA06-tnorm	3.08	8.11
NIST SRE 2006	LIA06-tnorm	2.71	5.18
NIST SRE 2008	LIA08-ztnorm	1.52	3.87

TAB. 4.2 – *Résultats de référence du système Alize/SpkDet sur les bases de données NIST SRE 2005, 2006 et 2008 (hommes seulement).*

Le tableau 4.2 présente les performances des systèmes LIA06 et LIA08 (décrits en annexe A) sur les bases de données NIST SRE 2005, 2006 et 2008. La différence majeure entre le système LIA06-tnorm et LIA08-ztnorm réside dans l'utilisation du LFA. L'évolution du système LIA est détaillée dans l'annexe A.

4.3.2.1 Influence du genre

Les performances de RAL sont dépendantes du genre des locuteurs considéré. La figure 4.1 présente les courbes DET du système LIA08 sur la base de données NIST SRE 2008 pour les signaux masculins et féminins. Il apparaît que les performances sur les signaux féminins sont plus faibles de 35%, pour la mesure DCF, et 39%, pour la mesure EER, relativement aux performances sur les signaux masculins.

4.3.3 Influence de la détection d'activité vocale

Des travaux [Bonastre et al., 2004] ont démontré qu'effectuer une sélection plus restrictive, basée sur un critère d'énergie, apporte un gain de performances en RAL. La figure 4.2 illustre les courbes DET de performances d'un système de VAL pour différentes configurations de sélection de trames. Le critère de sélection est le critère d'énergie. Le pourcentage de sélection des trames est d'environ 60% pour la configuration ED0.0. Le facteur, permettant de sélectionner des trame classées dans la Gaussienne de moyenne

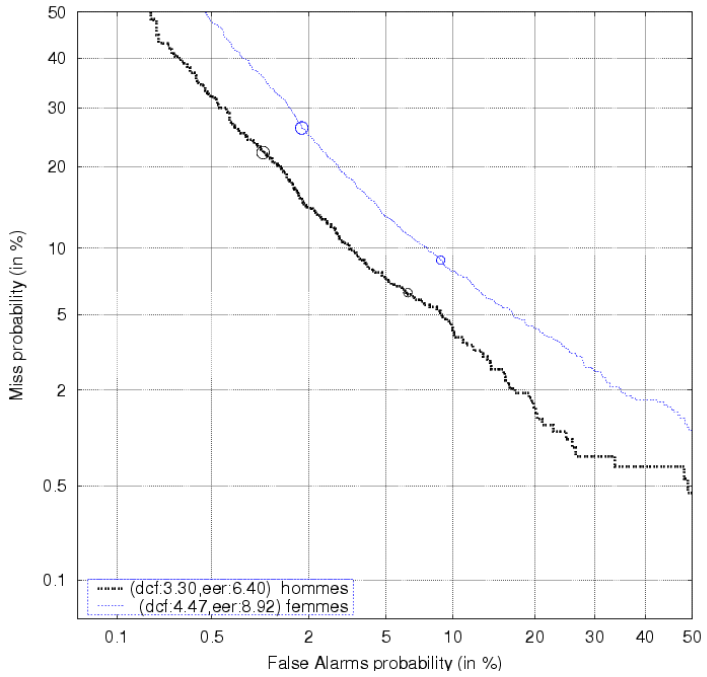


FIG. 4.1 – *Influence du genre (NIST SRE 2008, LIA08-ztnorm). Les signaux masculins bénéficient d'un EER plus favorable (6.40%) que les signaux fémninins (8.92%).*

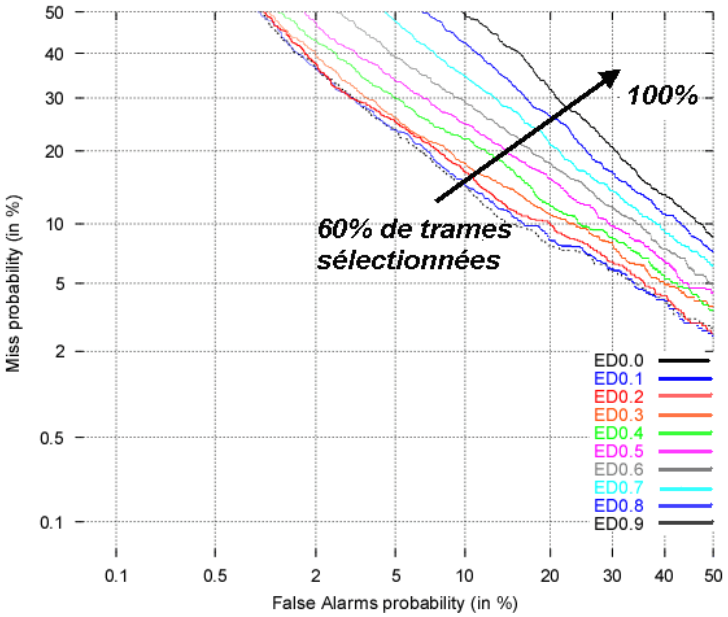


FIG. 4.2 – *Influence de la sélection de trames (NIST SRE 2004).*

intermédiaire, est progressivement augmenté. Le taux de trames sélectionnées est maximal pour la configuration ED0.9. Les résultats démontrent la forte corrélation entre la sélection de trames et les performances du système de VAL. Le meilleur résultat est

obtenu pour le système ED0.0 qui n'utilise que les trames de forte énergie (Gaussienne de moyenne la plus élevée).

4.3.3.1 Influence du canal

Pour illustrer les différences de performances, quand le canal d'enregistrement utilisé entre les sessions d'apprentissage et de test n'est pas le même, nous présentons les résultats du système de RAL LIA08 sur les données de l'évaluation NIST SRE 2008. La figure 4.3 *a* présente la courbe DET de performance du système de RAL, lorsque les données d'apprentissage et de test sont enregistrées sur le réseau téléphonique. La figure 4.3 présente la courbe DET de performance du système de RAL, lorsque différents microphones ont été utilisés pour la session de test, alors que les données d'apprentissage sont enregistrées sur le réseau téléphonique. Le système de RAL, utilisé pour ces expériences, implémente des mécanismes de compensation de la variabilité du canal de transmission (LFA). Pourtant, une perte relative de 32% est observée à l'EER, lorsqu'il existe une différence de conditions entre l'apprentissage et le test.

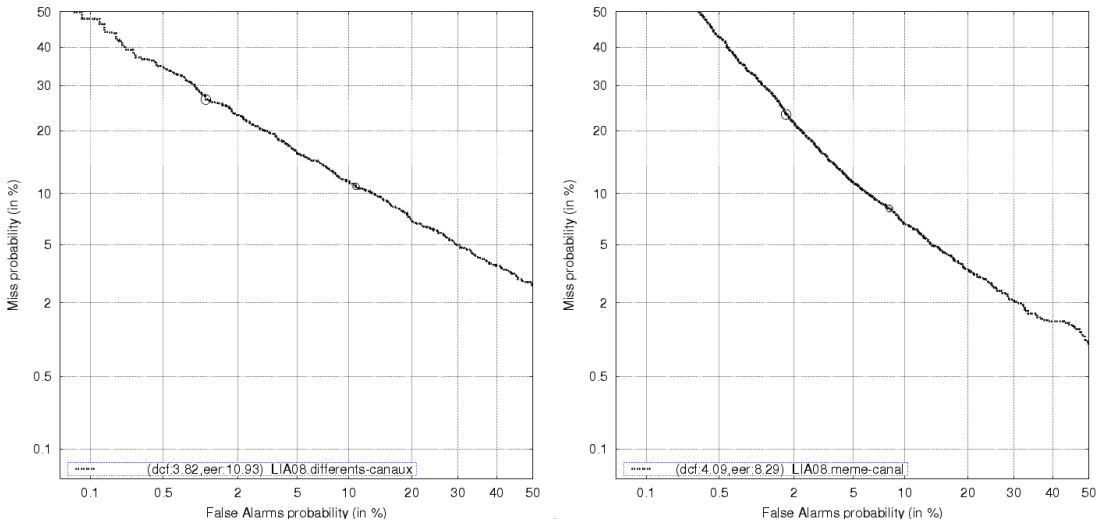


FIG. 4.3 – *Expériences sur des signaux d'apprentissage et de test enregistrés : a) sur le même canal (téléphone) , b) sur des canaux différents (microphones) (NIST SRE 2008, LIA08-ztnorm)*

4.3.3.2 Influence du *mismatch* entre les langues

Depuis 2004, dans les bases de données d'évaluation NIST SRE, les langues représentées sont multiples. Le modèle du monde doit être appris en fonction de la ou des langues de test considérées. Pour évaluer l'influence de l'utilisation de la même langue dans le modèle du monde et dans les tests nous proposons deux expériences. Dans un premier temps le modèle du monde est appris avec des données de la base TIMIT

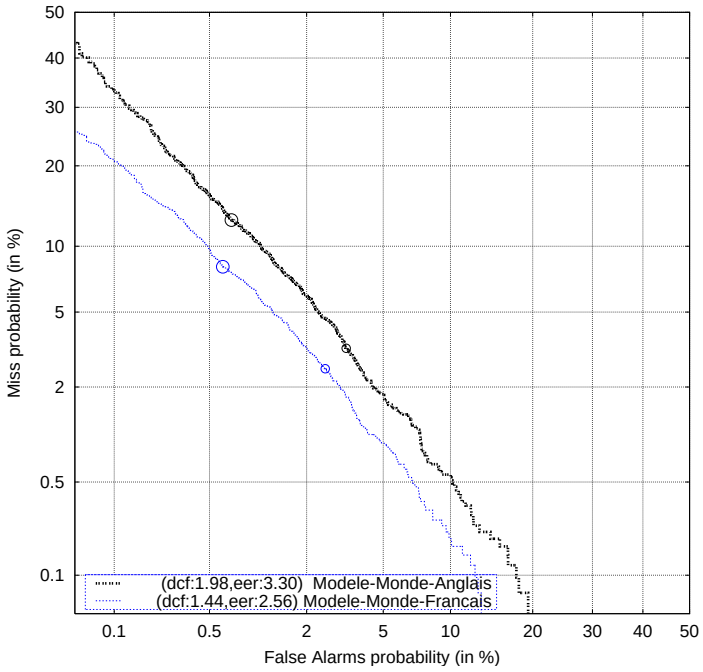


FIG. 4.4 – *Exemple de mismatch apprentissage-test, évaluation sur des données de parole de langage français (base BREF). Courbes DET avec l'utilisation d'un modèle du monde utilisant des données anglaises (base TIMIT) et françaises (base BREF).*

(anglais). Ensuite nous utilisons des données de la base BREF (français). La lange des modèles de locuteur et des signaux de test est le fançais. La figure 4.4 présente les résultats de ces deux expériences. Les résultats démontrent que l'utilisation d'un modèle du monde, créé à partir d'enregistrements de langue anglaise, différente de la langue de test, engendre une perte de 29% pour la mesure EER et 37 % pour la mesure DCF, relativement à l'utilisation d'enregistrements de langue française.

4.3.3.3 Influence de la compensation de canal

Nous présentons ici les résultats de l'implémentation de deux approches très répandues, pour la compensation de la variabilité canal :

- le *Feature mapping*,
- le *Latent Factor Analysis*.

4.3.3.3.1 Utilisation du *Feature mapping*

La figure 4.5 présente les courbes DET d'un système du RAL LIA06-tnorm avec et sans l'utilisation du *feature mapping*. Un gain de 10% relatif pour la mesure DCF et de 20% relatif pour la mesure EER est observé lorsque la technique est utilisée.

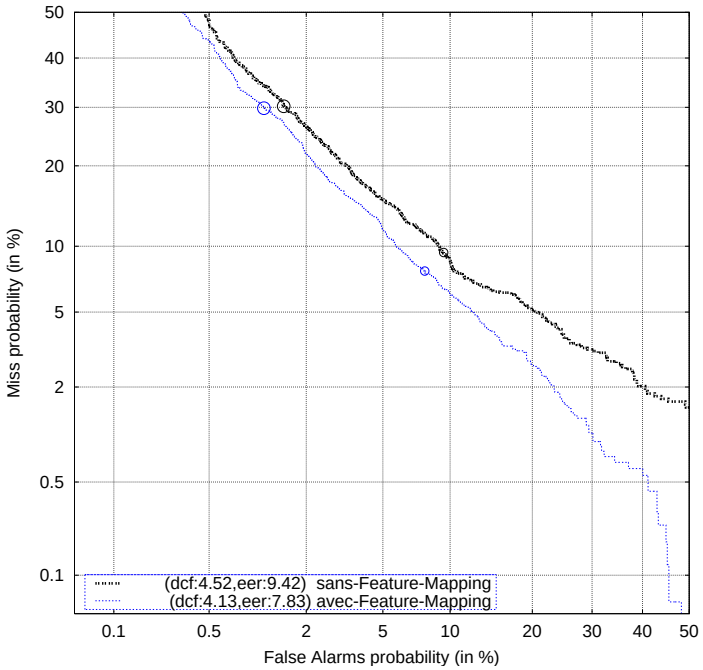


FIG. 4.5 – *Expériences avec et sans utilisation de la technique de feature mapping (NIST SRE 2006, LIA06-tnorm). La méthode apporte un gain significatif.*

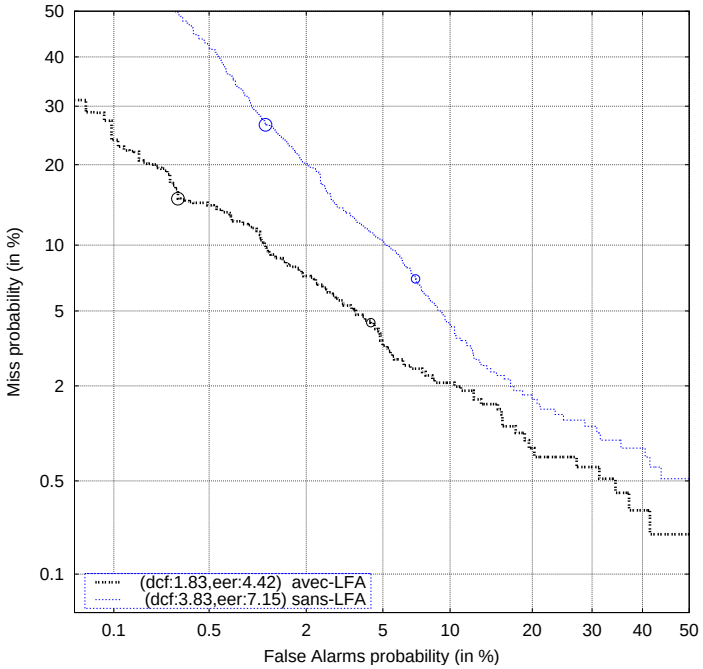


FIG. 4.6 – *Influence du Latent Factor analysis (NIST SRE 2005, LIA08-ztnorm). La méthode apporte un gain conséquent au niveau de la DCF et de l'EER.*

4.3.3.3.2 Utilisation du *Latent Factor Analysis*

La figure 4.6 présente les courbes DET du système LIA08-ztnorm qui implémente la compensation de canal LFA. Le résultat du système de référence sans méthode de compensation est donné pour comparaison. La méthode du LFA apporte dans ce cas une amélioration de 36% relatif pour la mesure EER. La figure 4.7 présente les résultats de l'utilisation de la méthode LFA appliquée dans l'espace des paramètres et selon la méthode de *symmetrical compensation* sur le système LIA08-ztnorm. La méthode symétrique surpasse la méthode de LFA appliquée dans l'espace des paramètres de 15% pour la mesure DCF.

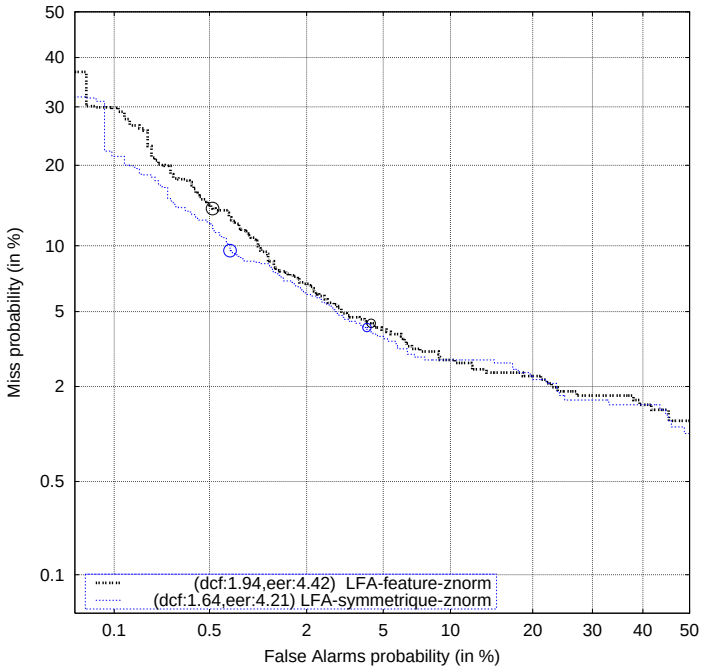


FIG. 4.7 – *Comparaison de l'application de la méthode LFA sur les paramètres (feature) et par la méthode symétrique. Z-norm est appliquée (NIST SRE 2005, LIA08-ztnorm). Il apparaît que la méthode LFA sur les paramètres est légèrement moins performante que la méthode LFA symétrique.*

Chapitre 5

La surveillance de réseaux professionnels de communication

Sommaire

Nous présentons dans ce chapitre une description des réseaux professionnels de communication et du scénario applicatif envisagé.
Nous exposons les diverses contraintes auxquelles nous devons répondre, notamment les dégradations du signal de parole, introduites par l'environnement d'utilisation particulier de ces réseaux (facteur peu évalué en RAL). Ces contraintes sont induites par le matériel et les conditions d'utilisation spécifiques aux réseaux PMR, mais elles sont aussi induites par le scénario applicatif. Nous illustrons l'impact de ces contraintes sur notre système de RAL de référence. Enfin, quelques solutions à ces différentes contraintes sont détaillées.

5.1 Description des réseaux professionnels de communications

La structure des réseaux professionnels de communication (*PMR, Private Mobile Radio networks*) est normalisée. Ces derniers reposent sur une architecture maillée commune à tout type de réseaux de communication. Les terminaux mobiles se connectent au réseau via des points d'accès sans-fil, appelés stations de base. Les liens entre les stations de base sont majoritairement filaires. Ceci permet notamment l'interception du

flux audio en de multiples points du réseau. Le schéma 5.1 représente ce type d'architecture. La chaîne complète de transmission, entre l'acquisition du signal sur le

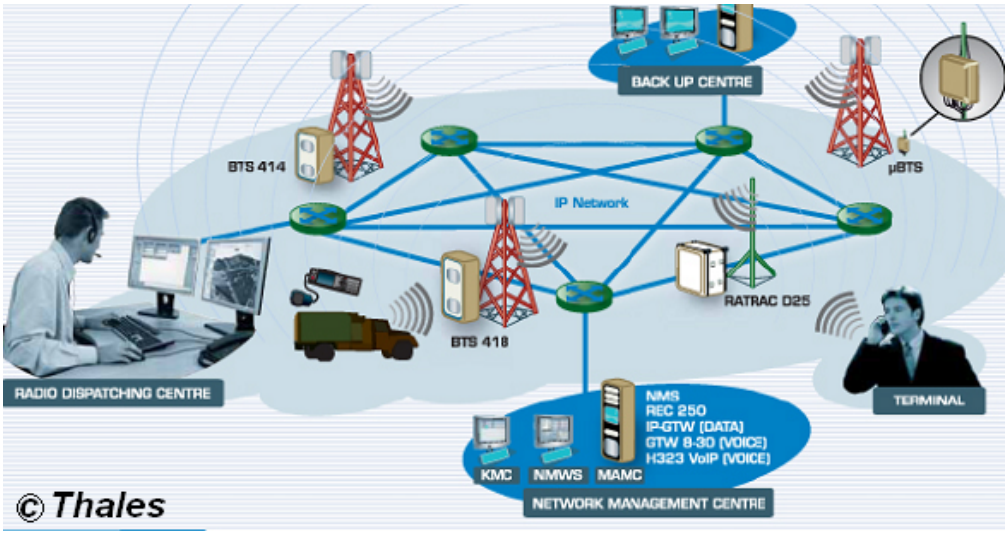


FIG. 5.1 – *Illustration de l'infrastructure d'un réseau de communication professionnel type TETRA Digicom 25.*

terminal mobile et la réception du flux sur le réseau, se décompose selon plusieurs traitements :

- l'acquisition du signal sur le terminal mobile,
- le codage à bas, ou très bas débit de la parole (entre 4,6kbit/s et 600 bits/s) ,
- la quantification des paramètres des codeurs,
- le codage canal pour la protection contre les erreurs.

Une spécificité singulière des réseaux PMR est le codage à bas débit de la parole. Le réseau GSM utilise un codeur à 13 kbit/s. Aujourd'hui, un codeur très employé sur les réseaux professionnels est le codeur MELP à 2.4 kbit/s [Supplee et al., 1997](décrit en section 6.1.2.1.2). Des versions de ce codeur sont même disponibles en 1200 bit/s et 600 bit/s ; ce codage s'explique par la volonté de ne pas surcharger le réseau. D'autres traitements, comme le cryptage, sont souvent ajoutés et consomment alors plus de bande passante. Les faibles débit permettent aussi une plus grande portée du signal.

5.2 Description du scénario opérationnel

L'objectif principal d'un système de surveillance des réseaux professionnels de communication par la RAL, est de réaliser une vérification d'identité au vol, sans contrainte ergonomique pour l'utilisateur, afin de lever une alarme en cas d'intrusion sur le réseau ou de détournement d'un terminal. Une telle application peut être décomposée en plusieurs phases :

- l'authentification, l'utilisateur est reconnu par l'identifiant de son terminal et son mot de passe pour accéder au réseau,

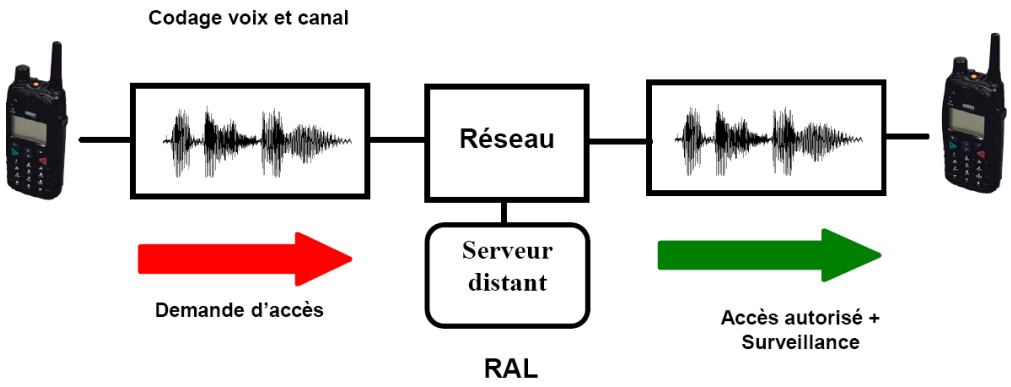


FIG. 5.2 – *Principe du monitoring de communications par la RAL*

– la surveillance en continue de la communication pour détecter un changement d'utilisateur.

La figure 5.2 représente schématiquement l'application de surveillance. Pour référencer les clients du système (phase d'apprentissage) deux possibilités sont envisageables :

– effectuer une collecte d'enregistrements pour chaque utilisateur,
– utiliser les premières secondes d'une communication de l'utilisateur, juste après son authentification ; la probabilité d'un vol ou d'un échange de terminal est alors considérée très faible.

Pendant la surveillance, un indicateur doit permettre à un opérateur, basé dans un centre de commandement distant, d'identifier la communication qui pose problème et, le cas échéant, de prendre la décision de couper la communication, ou de redemander l'authentification de l'utilisateur. La vérification d'identité par la RAL doit être continue pour un maximum de réactivité du système.

Sur les réseaux professionnels de communications le terminal peut être personnalisé selon l'utilisateur. Il existe notamment la possibilité de télécharger, sur le terminal, un carnet d'adresses spécifique. Cette personnalisation est, aujourd'hui, difficile à mettre en place, car l'authentification sur le réseau s'effectue grâce à l'identifiant du terminal. L'utilisateur n'est pas connu. Ainsi, la RAL permet, en plus de la surveillance, de proposer des services de personnalisation des terminaux.

5.3 La surveillance de réseaux professionnels de communications par la RAL

Le scénario applicatif envisagé impose diverses contraintes sur le système de RAL que nous devons mettre en place. Nous les classons selon deux catégories :

1. les contraintes introduites par le matériel et les conditions spécifiques d'utilisation des réseaux PMR,
2. les contraintes introduites par le scénario applicatif.

5.3.1 Spécificités des réseaux professionnels de communication

L'architecture des réseaux PMR est normalisée. Bien que ceci permette leur interopérabilité, la possibilité d'ajouter de nouveaux traitements pour la RAL est presque impossible. Nous présentons l'influence des spécificités matérielles, pour une application de RAL.

5.3.1.1 Chaîne de transmission : altération du signal de parole

La chaîne de transmission introduit des altérations du signal de parole. Ainsi, on considère les dégradations induites par les étages suivants :

- la prise de son,
- le codage à bas débit de la parole (sur le terminal) [Lilly et Paliwal, 1996],
- la quantification des paramètres (sur le terminal) [Tsuge et al., 2002],
- la transmission sur le canal.

Le codage à bas-débit de la parole est une source majeure d'altérations du signal de parole. La figure 5.3 présente les courbes DET de performance du système LIA-THL07-nonorm lorsque les signaux sont clairs et codés à bas débit (codeur TETRA à 4.6kbit/s). Le codage à bas-débit de la parole, considéré pour l'expérience, engendre une perte de performance de 60% relativement à l'utilisation de signaux non codés.

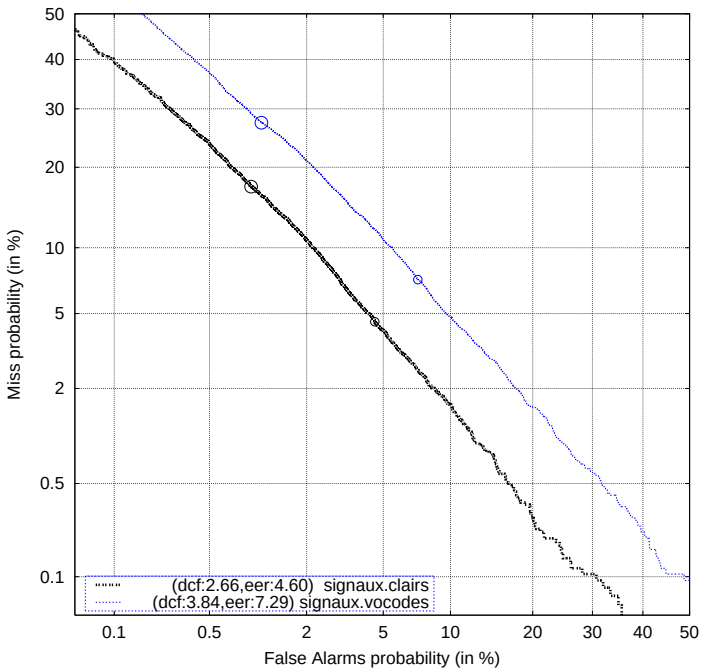


FIG. 5.3 – *Courbes DET d'un système de RAL (LIA-THL07-nonorm) avec l'utilisation de signaux de parole codés à bas débit (codeur TETRA[1]) et des signaux clairs. Le codage à bas débit de la parole utilisé pour cette expérience introduit une perte relative de 60% à l'EER.*

Les solutions les plus répandues, visant à éviter les différents étages de dégradation du signal de parole, sont basées sur le principe de l'architecture terminal-serveur distribuée. Ces solutions permettent d'éviter d'utiliser le signal de parole codée à bas débit, en proposant une extraction des paramètres dans le terminal. Dans ce cas une procédure de quantification et d'émission des paramètres est mise en place. Cette solution présente les inconvénients suivants :

- un traitement supplémentaire doit être implanté dans le terminal,
- l'inter opérabilité de tous les terminaux mobiles n'est assurée que si un standard est défini,
- une charge réseau supplémentaire est induite pour la transmission des paramètres.

Une architecture distribuée de paramétrisation, Aurora, a été standardisée par l'ETSI dans ce sens. Une description détaillée de ce standard est disponible dans le chapitre 6.1.1.

5.3.1.2 Capacité de calcul : pas de traitement sur les terminaux

La complexité des traitements algorithmiques, présentés dans le chapitre 3.3, peut en limiter l'utilisation. Nous proposons de classer les modules de la RAL en fonction de leurs contraintes vis à vis des ressources consommées :

- les traitement « hors-ligne » : ce sont les traitements qui peuvent être réalisés au préalable de la mise en fonctionnement du système, comme la création du modèle du monde et des modèles de locuteurs,
- les traitement « en-ligne » : ce sont les traitements qui doivent être réalisés en cours de fonctionnement, la paramétrisation et le test de vérification.

Les traitements « hors-ligne » peuvent être réalisés sur des serveurs de calcul indépendants de l'architecture de fonctionnement. La consommation en ressources, ainsi que les temps de calcul, ne sont pas des limitations. Ils permettent notamment d'estimer le modèle du monde (très consommateur de ressources), et d'en dériver les modèles de locuteurs, si une campagne d'enregistrement à été réalisée au préalable. Certains paramètres, liés aux méthodes de compensation de canal, peuvent être aussi pré-calculés comme les modèles dépendant des canaux (*feature mapping*) ou la matrice de covariance (*Latent Factor Analysis*).

Les traitements « en-ligne » sont réalisés lors du fonctionnement du système. La paramétrisation et le test de vérification doivent être repensés pour satisfaire les contraintes du scénario applicatif envisagé. En reprenant le scénario décrit en section 5.2, nous introduisons la contrainte d'horizon de calcul des paramètres. La plupart des méthodes algorithmiques, citées précédemment dans ce document, ont pour base un traitement sur un signal audio d'une durée déterminée (un fichier). Pour introduire le concept de vérification en continue, il faut assurer un traitement complet de RAL sur des horizons temporels restreints. Les méthodes de paramétrisation (extraction des paramètres acoustiques, détection d'activité vocale et normalisation pour la compensation canal) doivent être réalisées au « fil de l'eau ». Le test de vérification qui, lui, est basé sur une somme moyennée des LLR de chaque trame, est compatible avec le traitement recherché. Néanmoins, la décision ne peut s'effectuer à la trame, un horizon de décision est à

définir. Le tableau 5.1 illustre les temps de traitements des différents modules de RAL.

Type	Traitement	Temps
En ligne	Extraction des MFCC	10 ms
	DAV	2.5 min
	Normalisation CMNV	2.5min
	Normalisation des scores	0
Hors ligne	Estimation des paramètres du LFA	24h
	Modèle du monde	6h [2]
	Modèle client	2s

TAB. 5.1 – *Temps de traitement pour chaque module de RAL (calculs basés sur des enregistrements de longueur 2 minutes 30 secondes et sur l'utilisation d'un PC bi-processeurs Intel Xeon 3Ghz de 2Go de RAM.*

Certains traitements nécessitent la connaissance de tout le fichier d'enregistrement :

- la détection d'activité vocale,
- la normalisation moyenne-variance des paramètres cepstraux,
- l'apprentissage des modèles de locuteurs.

Dans ce cas, la durée de traitement peut être définie égale à la longueur de l'enregistrement (2.5 minutes dans cet exemple). Les traitements hors-ligne sont très gourmands en calculs. Souvent plusieurs heures de calculs, sur des machines puissantes, sont nécessaires pour l'estimation de certains paramètres. Les problèmes d'intégration d'un système de VAL sur un terminal mobile ne sont pas abordés. Notons simplement que la taille des modèles GMM, la population de locuteurs et la consommation de ressources des algorithmes sont des problématiques majeures.

5.3.1.3 Bruits ambiants : altération du signal de parole

Les conditions d'acquisition peuvent être très difficiles (bruits...) sur les réseaux PMR, et un facteur important de dégradation du signal de parole réside dans le bruit additif de l'environnement extérieur [Openshaw, 1994]. Des expériences menées sur la base BREF (protocole BREF2), démontrent que dans des conditions d'enregistrements très favorables, sans aucun bruit d'ambiance, l'EER de notre système de VAL de référence est de 0.32%. Le même système, utilisé sur des signaux de test bruités (RSB de 0 dB) à l'aide d'un bruit de communication hautes-fréquences, présente un EER de 17%. Cette expérience démontre la très forte corrélation entre les performances d'un système de VAL et le RSB des signaux de test. Les signaux acquis sur les réseaux de communication sont souvent très bruités. Sur ce type de réseaux à vocation militaire, les contraintes de bruits peuvent êtres sévères (champ de bataille, feu, ...). L'intégration de techniques de compensation des bruits est nécessaire pour diminuer la perte de performance due aux bruits ambiants.

[2] Pour 4 millions de trames utilisées.

5.3.2 Spécificités du scénario

Le scénario envisagé introduit des contraintes :
- de réactivité : un évènement anormal (vol ou échange d'un terminal) doit être rapidement détecté,
- d'ergonomie : la surveillance doit s'effectuer sans interférer avec les activités des utilisateurs. Il ne faut pas, par exemple, leur demander de longues sessions d'apprentissage.

5.3.2.1 Réactivité : peu de données de test

Nous avons précisé, dans le scénario, que la RAL doit permettre d'authentifier les utilisateurs en cours de communication. Ceci implique de définir une période de calcul pour le score de vérification. Les systèmes état de l'art sont basés sur des traitements

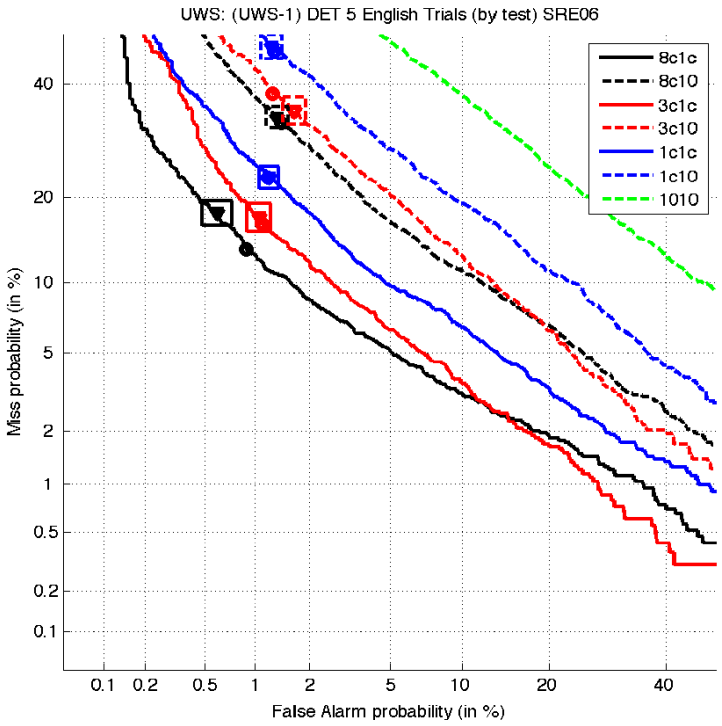


FIG. 5.4 – *Expériences en utilisant différentes durées d'apprentissage (NIST SRE 06, Système Swansea)*

par fichier : dans le cadre des évaluation NIST SRE, un fichier correspond à un enregistrement de 2 minutes 30 secondes. Les méthodes de RAL sont très performantes aujourd'hui avec cette quantité de données pour l'apprentissage et le test. Mais 2 minutes 30 secondes représentent une période trop longue pour surveiller une communication. Sur les réseaux PMR les « conversations » se limitent à quelques phrases. Nous pouvons imaginer cumuler les segments de parole, pour obtenir un enregistrement équivalent

à 2 minutes 30 secondes. Mais il faut imaginer que la période de temps pour cumuler cette durée peut s'échelonner de 2 minutes 30 secondes, dans le meilleur des cas, à plusieurs heures. Il est donc nécessaire d'utiliser des segments de test beaucoup plus courts. La figure 5.4 présente les courbes DET du système de VAL de Swansea[3], utilisé pour l'évaluation NIST SRE 2006, selon différentes conditions d'apprentissage et de test (nc : n conversations de 2 minutes 30 secondes). Les courbes en pointillés représentent les expériences pour des durées de test de 10 secondes (nc10). Comparativement à des durées de test de 2 minutes 30 secondes, lorsque seulement 10 secondes sont disponibles, le taux d'erreur est doublé (de 9% d'EER à 20%).

5.3.2.2 Ergonomie : peu de données d'apprentissage

L'apprentissage des modèles de locuteurs repose sur un échantillon de référence, les données d'apprentissage. Cette quantité de données est souvent très faible, c'est pour cela que l'adaptation MAP du modèle du monde est utilisée. Cette quantité est faible pour respecter une contrainte d'ergonomie du système de RAL. Il ne faut pas soumettre l'utilisateur à de longues périodes fastidieuses d'apprentissage. Sur les réseaux PMR,

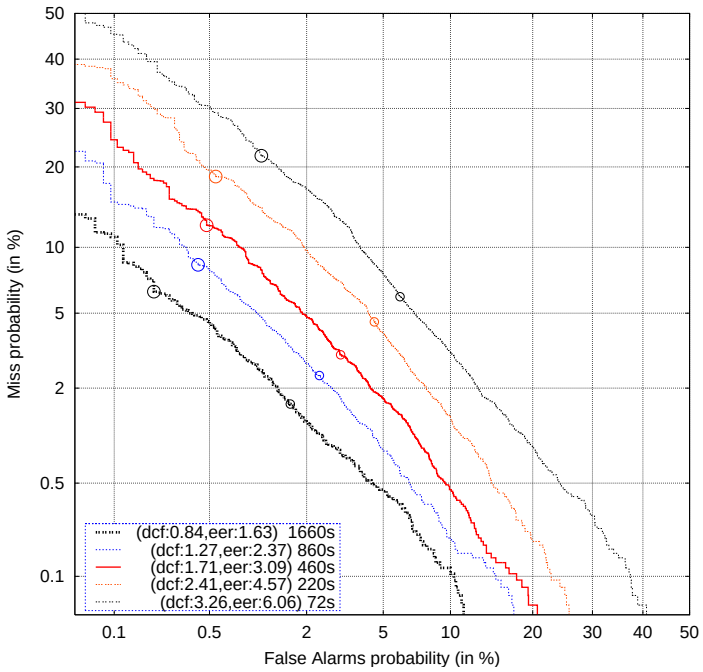


FIG. 5.5 – *Expériences en utilisant différentes durées d'apprentissage (BREF1, LIA-THL07-nonorm)*

les utilisateurs sont coopératifs. Il est possible de demander des séquences d'apprentissage à la mise en route du système de RAL. Néanmoins, ces durées ne peuvent approcher les durées d'apprentissage généralement fournies par le NIST. Les utilisateurs

[3]Ce système de RAL utilise ALIZE et était très proche du système LIA06

sont dans un cadre de travail professionnel et ne peuvent pas perdre de temps avec le processus d'apprentissage. Dès lors, il faut se pencher sur des méthodes permettant de collecter de nouvelles données du locuteur, pour améliorer son modèle.

Pour illustrer l'importance des durées d'apprentissage, une expérience est réalisée sur la base BREF. Les modèles de locuteur sont appris avec différentes durées d'apprentissage. Les courbes DET de performances du système de VAL par durée d'apprentissage sont proposées sur la figure 5.5). Les performances du système de VAL augmentent avec la quantité de données d'apprentissage. L'utilisation de la base BREF permet de mettre en évidence l'intérêt de la quantité de données d'apprentissage. L'influence du canal de transmission est presque nulle et la variabilité intra-locuteur très faible (textes lus dans un environnement calme). Mais lorsque différents canaux de transmission sont utilisés, et que les conditions d'enregistrements sont libres (base NIST SRE), l'augmentation de la durée d'apprentissage permet aussi d'améliorer les performances. Les courbes DET d'expériences démontrant cette observation sont illustrées dans la figure 5.4.

5.4 Quelques approches envisagées

Nous avons exposé les problèmes majeurs liés à l'intégration d'un système de RAL sur un réseau PMR. Les spécificités matérielles de ce type de réseau peuvent être prises en compte dans un schéma de RAL. En effet, des traitements particuliers peuvent être apportés à la paramétrisation, pour tenter de diminuer les effets du codage à bas débit de la parole. Par exemple, les techniques d'extraction des paramètres cepstraux à partir des paramètres internes des codeurs de parole ont démontré de bons résultats [Raj et al., 2001; Huerta et Stern, 1998; Petracca et Servetti, 2006; Quatieri et al., 2000]. Il est aussi possible d'utiliser une architecture distribuée, pour transmettre les paramètres cepstraux calculés directement sur le terminal. Au niveau du problème de la puissance de calcul, là encore, l'architecture distribuée peut être une réponse, les traitements de RAL peuvent être effectués sur un serveur distant.

L'altération du signal dû aux bruits ambiants a fait l'objet de nombreuses recherches. Il existe des solutions au niveau du traitement du signal de parole, majoritairement basées sur le filtrage de Wiener [Wiener, 1949]. cette méthode consiste à calculer une estimation du signal de parole non bruité, par filtrage du signal bruité. A partir des statistiques du signal de parole, il est possible de déterminer un filtre optimal afin d'obtenir une estimée du signal de parole utile. Cette technique a prouvé son efficacité pour la RAL [Xie et al., 2006]. Nous privilégions ce type d'approche car elle peut être directement intégrée dans le calcul des paramètres cepstraux. Certains terminaux ou codeurs de parole intègrent même leur propre réduction de bruit. Il existe aussi la technique de masquage de bruit [Klatt, 1976] qui utilise le fait que certaines bandes de fréquences du signal de parole sont plus ou moins affectées par le bruit. Par seuillage sur l'énergie en sortie du banc de filtre, il est possible de sélectionner les paramètres non corrompus. [Drygajlo et El-Maliki, 1998] introduit la notion de *missing feature theory* pour la RAL qui consiste à ne pas utiliser les trames considérées comme trop bruitées.

La combinaison de modèle, *PMC (Parallel Model Combination)* [Gales et Young, 1993; Matrouf, 1997], propose d'utiliser la décomposition du signal de parole en une composante parole propre et une composante bruit. Un modèle de bruit et un modèle de parole sont estimés puis combinés. La difficulté majeure de cette méthode réside dans l'estimation du modèle de bruit. [Ming et al., 2007] propose d'utiliser l'apprentissage multi-conditions des modèles de locuteurs, *multi style training* [Lippmann et al., 1987]. Plusieurs modèles de locuteurs bruités à différents niveaux de RSB avec différents bruits sont créés. La sélection du modèle de locuteur bruité, utilisé pour le calcul du score de vérification, se fait par maximum de vraisemblance avec les données de test.
Notons enfin que [Ming et al., 2007] propose d'étendre les techniques de *multi style training* et *missing feature theory* en les combinant. Le *multi style training* est alors utilisé pour générer de multiples modèles clients bruité à différents niveaux de RSB. Le critère de maximum de vraisemblance est utilisé pour déterminer un sous-ensemble des paramètres acoustiques sélectionné, pour le calcul du LLR. L'évaluation de cette technique est réalisée sans connaissance *a priori* sur les bruits additifs. Les auteurs démontrent alors qu'avec un nombre limité de données d'apprentissage des conditions de bruits, la technique permet de compenser une grande variété de conditions de bruits.

Enfin, les techniques de compensation des variabilités inter-session, les plus performantes, sont aujourd'hui les méthodes de Latent Factor Analysis (LFA) [Kenny et al., 2005b]. Le LFA estime la variabilité induite par les canaux de transmission pour les éliminer de la modélisation statistique des locuteurs. Cette technique a été évaluée dans le cadre des évaluations NIST SRE et est considérée comme la méthode la plus performante pour compenser les effets du canal de transmission. Nous avons illustrés les résultats de cette méthode sur notre système de référence dans le paragraphe 4.3.3.3.2.

Au niveau du traitement en ligne des paramètres, nous pouvons citer les travaux de normalisation à court-terme du canal acoustique sur fenêtre glissantes. Les méthodes de *Feature warping* [Pelecanos et Sridharan, 2001] ou de normalisation CMNV [Xiang et al., 2002; Viikki et Laurila, 1998] appliquées sur des fenêtres glissantes présentent des performances comparables aux normalisations appliquées sur des enregistrements entiers.

Pour répondre aux problèmes des durées courtes d'apprentissage, seules les méthodes d'adaptation non supervisée permettent de collecter automatiquement de nouvelles données de locuteurs. Ces méthodes utilisent les informations de nouvelles sessions d'enregistrements pour mieux couvrir la variabilité du locuteur et des environnements d'acquisition [Preti et al., 2007] [Hansen et al., 2006; Van Leeuwen, 2004; Heck et Mirghafori, 2000; Yin et al., 2006; McLaren et al., 2008].
Nous proposons, dans les chapitres suivants, des solutions inspirées de ces techniques.

Chapitre 6

Contraintes applicatives : paramétrisation et test de vérification en ligne

Sommaire

Un système de RAL évalué dans un cadre donné ne peut pas être optimisé pour un tout autre type d'application. Pour répondre aux contraintes de la chaîne de transmission des réseaux PMR, nous présentons dans ce chapitre, une solution optimisée de paramétrisation. L'étape de traitement des paramètres est habituellement composé du calcul des coefficients acoustiques, d'une détection d'activité vocale et d'une normalisation des paramètres. Chacun de ces éléments fait l'objet d'une optimisation pour répondre aux contraintes, de codage à bas-débit de la parole, des bruits ambiants et de traitement en ligne. Une méthode pour évaluer les scores de vérifications sur de courtes périodes est aussi évaluée. Tout ceci permet de proposer un système de RAL, à la fois

robuste et en ligne.

6.1 Choix de la paramétrisation

La chaîne de transmission est une des caractéristiques majeures des réseaux professionnels de communication. Les terminaux mobiles accèdent au réseau via des stations de bases. Les stations de bases sont reliées entre elles grâce à un réseau *IP, Internet Protocol*. Sur ce type d'architecture, le signal de parole est disponible en différents endroits :

1. sur le terminal, lors de l'acquisition,
2. dans le réseau, sous la forme d'un train binaire compressé,
3. au niveau du récepteur.

Il est possible de réaliser l'extraction des paramètres en chacun de ces points. Au niveau de l'acquisition, le calcul des paramètres doit être intégré dans le terminal, et un protocole d'émission vers un serveur distant de reconnaissance doit être défini. Il y a alors transmission du flux audio pour la communication et du flux données pour les paramètres. Cette configuration est considérée comme la référence en termes de performance puisque le signal originel, non codé, est utilisé. Le standard Aurora a été créé pour répondre à ce besoin. Ce standard est décrit au paragraphe suivant.

L'extraction des paramètres peut également être réalisée après le décodage bas débit de la parole, au niveau du récepteur. L'altération de la parole due au codage n'est pas évitée dans ce cas de figure. Cette configuration est généralement celle utilisée pour les campagnes d'évaluation NIST SRE. La collecte de données s'effectue en réception du flux audio après décodage des paramètres transmis et la resynthèse du signal de parole.

Enfin, une configuration optimisée consiste à extraire les paramètres dans le domaine compressé, à partir du train binaire émis. Cette configuration ne nécessite pas le signal décodé, elle se montre économe en termes de bande passante comme de ressources de calcul.
Nous présentons dans cette section les résultats de méthodes de paramétrisation pour chacune de ces trois configurations.

6.1.1 La solution Aurora pour une architecture distribuée

Le standard ETSI Aurora [ETSI, 2005a] a été créé à l'origine pour effectuer la reconnaissance de la parole sur des architectures distribuées. Nous le présentons en détails dans cette section.

6.1.1.1 Définition du standard

Les performances de reconnaissance sont altérées par le codage bas débit de la parole mis en place sur les réseaux de communication, mais aussi par les erreurs de trans-

mission. Sur ce type d'architecture, le terminal a alors pour charge d'extraire les paramètres cepstraux et de les transmettre sur un canal de données protégé, après compression. Le flux compressé est ensuite reçu par le serveur distant pour effectuer la reconnaissance. Les dégradations dues au codage bas débit de la parole et aux pertes de transmission sont ainsi évitées. L'architecture distribuée, entre le terminal mobile et un serveur d'applications distant, permet notamment de s'affranchir des problèmes dus au codage bas-débit de la parole [Pearce, 2000; Grassi et al., 2002].
Le standard Aurora est issu d'un groupe de travail de l'ETSI et a été publié pour la première fois en février 2000, dénommé Aurora-1. Le standard a régulièrement bénéficié d'améliorations. Le standard Aurora-2 intègre ainsi un étage de réduction de bruit, basée sur le filtrage de Wiener, avant le calcul des paramètres cepstraux et une méthode d'égalisation aveugle des paramètres cepstraux [Kuroiwa et Tsuge, 2003; Mauuary, 1998]. Elle permet de compenser les variations, à court terme, du canal d'acquisition acoustique. L'évolution Aurora-3 permet la reconstruction du signal au niveau du serveur d'application distant (cf. figure 6.1). Pour reconstruire le signal, à partir des paramètres transmis, et améliorer la reconnaissance sur les langues tonales (Mandarin, Cantonais, et Thai), le standard utilise les paramètres de fréquence fondamentale (pitch) et de classes de voisement.
Destiné aux réseaux de communication téléphonique, le standard est compatible avec des signaux d'une fréquence d'échantillonnage de 8kHz, mais aussi de 11kHz et de 16kHz. L'étage d'extraction des paramètres cepstraux s'effectue comme suit :

- le signal d'entrée est débruité par un filtrage de Wiener,
- les paramètres cepstraux sont calculés avec 23 filtres, de fréquence centrale fixée selon l'échelle de MEL,
- les paramètres cepstraux sont égalisés par un algorithme du gradient (*Least Mean Square*, LMS), à partir d'un cepstre de référence estimé sur un bruit blanc.

Le standard Aurora utilise 12 paramètres cepstraux statiques et un paramètre d'énergie, calculés toutes les 10 ms sur une fenêtre d'analyse de 25 ms (Hamming). Il est important de noter que ce schéma d'extraction des paramètres permet un traitement du flux audio en temps réel ; *i.e.* le vecteur de paramètres acoustiques est disponible toutes les 10 ms avec un retard de 6 trames seulement (60 ms). L'étage de quantification est ensuite appliqué aux paramètres cepstraux avant transmission. Le débit total utilisé pour la transmission est alors de 5.6 kbit/s.
A la réception, le serveur déquantifie les paramètres et extrait les dérivées, premières et secondes, des cepstres et de l'énergie. Un horizon de 9 trames (90 ms) est utilisé pour le calcul des dérivées. Au total un vecteur de 39 paramètres est disponible au niveau du serveur :

$$log(E), C_1....C_{12}, \Delta C_1, ..., \Delta C_{12}, \Delta log(E), \Delta\Delta C_,..., \Delta\Delta C_{12}, \Delta\Delta log(E)$$

Le standard Aurora ne se contente pas seulement d'effectuer une analyse cepstrale, il propose sa propre détection d'activité vocale. Elle utilise la mesure de l'accélération de l'énergie, calculée sur la trame, sur une sous-région du spectre qui contient la fréquence fondamentale et sur la bande basse du spectre [ETSI, 2005a]. L'accélération de l'énergie est basée sur un calcul de variance spectrale.
L'état de voisement de la trame est aussi disponible. Au total quatre classes, dont trois sur l'état de voisement et une sur la présence de parole, de sont définies :

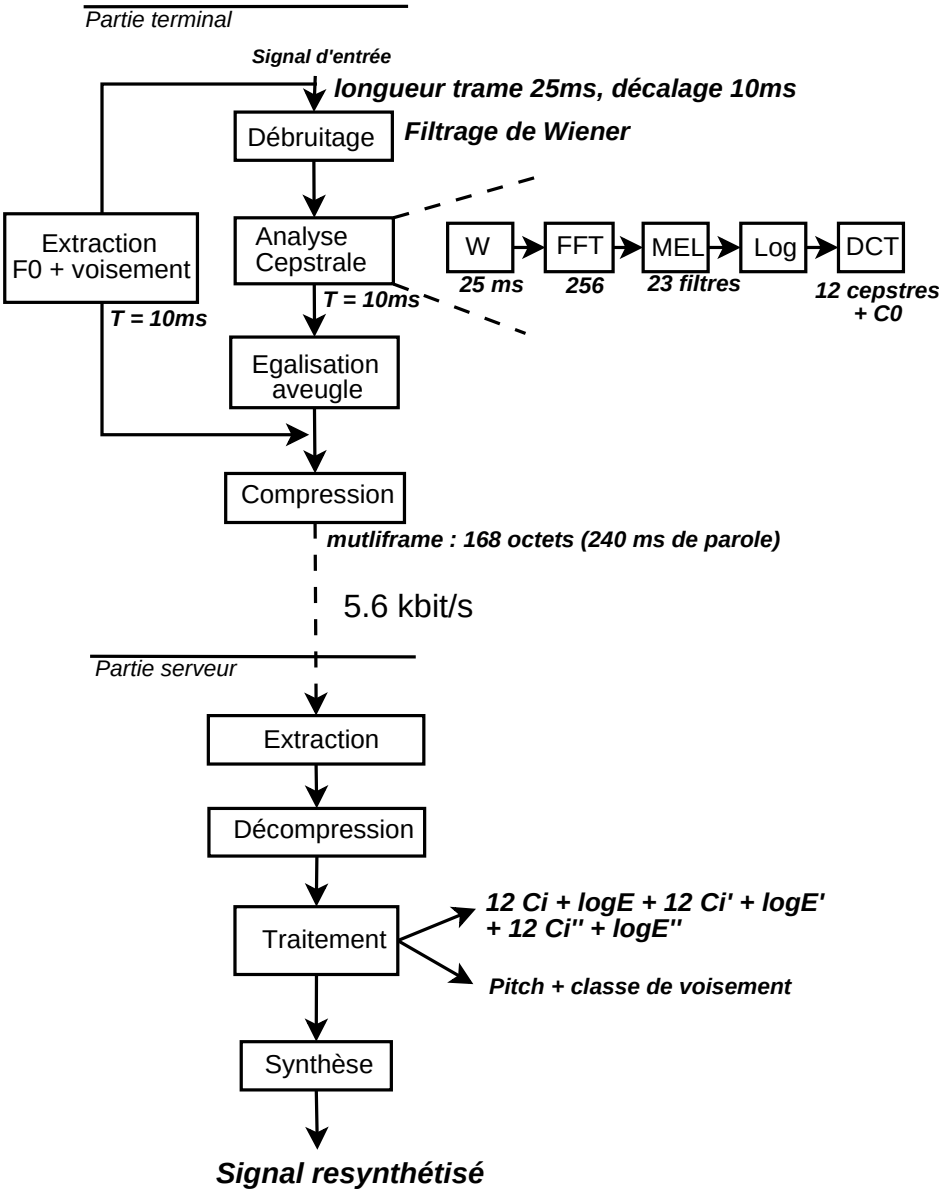


FIG. *6.1* – *Architecture terminal-serveur du standard Aurora.*

– bruit/silence,
– non voisée,
– moyennement voisée,

– voisée.

L'information de voisement est obtenue par différents indicateurs :

– la mesure du taux de passage par zéro du signal (*ZCM, zero Crossing Measure*),
– un seuil sur l'énergie, dans la sous-bande haute du spectre,
– l'indicateur d'activité vocale.

Si le traitement au niveau du serveur peut être facilement implémenté, l'utilisation de ce standard implique une compatibilité avec les terminaux mobiles, *i.e.*, les routines algorithmiques du standard doivent être embarquées dans les terminaux. C'est aujourd'hui de plus en plus le cas, mais il est encore difficile de proposer une solution basée sur le standard Aurora, sur les réseaux de communication existants.

L'utilisation du standard Aurora présente de multiples avantages. Nous proposons une synthèse des avantages et des inconvénients du standard.

Avantages :

– utilisation du signal originel acquis sur le terminal pour l'extraction des paramètres,
– DAV incluse,
– débruitage et égalisation,
– architecture standardisée,
– traitement en ligne (60ms de retard maximum au niveau du terminal),
– paramètres de voisement et pitch disponibles pour chaque trame.

Inconvénients :

– débit supplémentaire nécessaire pour la transmission,
– 12 premiers cepstres disponibles seulement (19 sont utilisés dans les systèmes LIA06, 07, 08).

6.1.1.2 Robustesse de la paramétrisation Aurora : le débruitage

Nous avons, précisé auparavant, que le standard Aurora utilise un étage de débruitage du signal, avant le calcul des paramètres cepstraux. Nous évaluons, dans cette section, les performances de l'étage de débruitage. Les paramètres Aurora ont déjà été utilisés en RAL [Broun et al., 2001]. Le standard Aurora-1, alors utilisé, ne disposait pas de l'étage de débruitage.

Nous évaluons la robustesse aux bruits des paramètres Aurora, calculés après débruitage et égalisés. Pour cela, nous avons mené des expériences sur des signaux bruités avec et sans l'étage de débruitage. Au total un vecteur de 39 paramètres est disponible mais nous avons choisi de n'utiliser que 37 paramètres :
$C_1 ....C_{12}, \Delta C_1, ..., \Delta C_{12}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{12}$
Nous avons utilisé la base BREF et le protocole BREF2.
Trois bruits différents ont été choisis pour bruiter les signaux de parole :

– bruit de communication sur un canal radio haute fréquence (HF),
– bruit de char Leopard 2 roulant à 70 km/h,
– bruit de babillage ou *babble noise*.

Ces bruits sont extraits de la base de données *NOISE-ROME-0* [Institute for Perception-TNO, 1990]. Elle a été réalisée par *Institute for Perception-TNO* dans le cadre du groupe

de recherche de l'OTAN, sur la reconnaissance de la parole et la qualité des communications dans un environnement militaire. Les spectres des bruits sélectionnés sont illustrés dans les figures 6.3 *a*, *b* et *c*. Les signaux de la base BREF sont enregistrés à

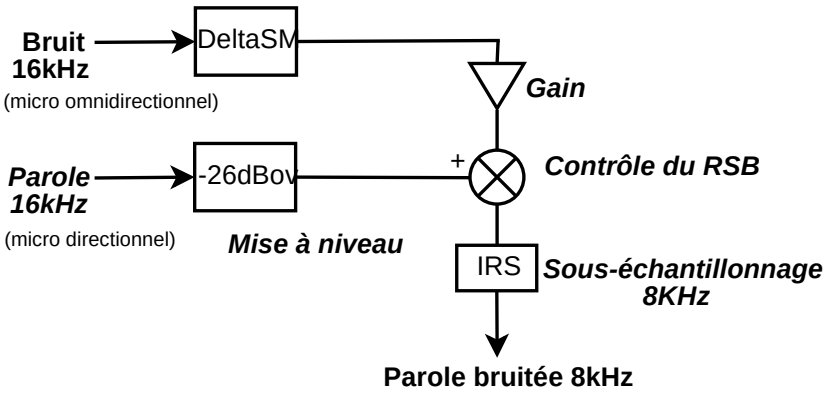


FIG. 6.2 – *Procédure de bruitage de la base de données BREF.*

16kHz, et ceux de la base ROME à 19.98kHz. La procédure de bruitage décrite dans la figure 6.2 a été suivie. Cette procédure de bruitage de signaux est standardisée. Elle est utilisée pour la création de base de données pour l'évaluation des systèmes de transmission. Les outils utilisés pour suivre ce protocole sont fournis par l'ITU sous la forme d'une librairie, la STL2000 [De Campos Neto, 1999]. Différentes étapes constituent ce protocole :

- sous-échantillonnage à 16kHz des signaux de bruits,
- filtrage des bruits enregistrés sur un micro omnidirectionnel, par un filtre (ΔSM), pour introduire les distorsions d'un microphone directionnel,
- mise à niveau des signaux de parole, à -26dBov[1] en dessous de la limite de saturation de gain nominal par l'algorithme P.56 [ITU, 2006],
- addition des signaux de bruit et de parole au niveau temporel, avec un contrôle du gain pour obtenir le RSB considéré pour l'expérience (saturation évitée),
- sous-échantillonnage à 8kHz des signaux.

Des expériences suivant le protocole BREF2 (40 locuteurs, 1 minute d'apprentissage et 2.5 minutes de test, cf. annexe A) ont été menées pour déterminer l'influence du débruitage. Les résultats d'expériences menées sur différents signaux, à différents niveaux de

[1] Il s'agit du niveau nominal mesuré par rapport à la saturation (*overload*). Généralement la plan de test prévoit l'utilisation d'un niveau faible, -36dBov et d'un niveau fort -16dB.

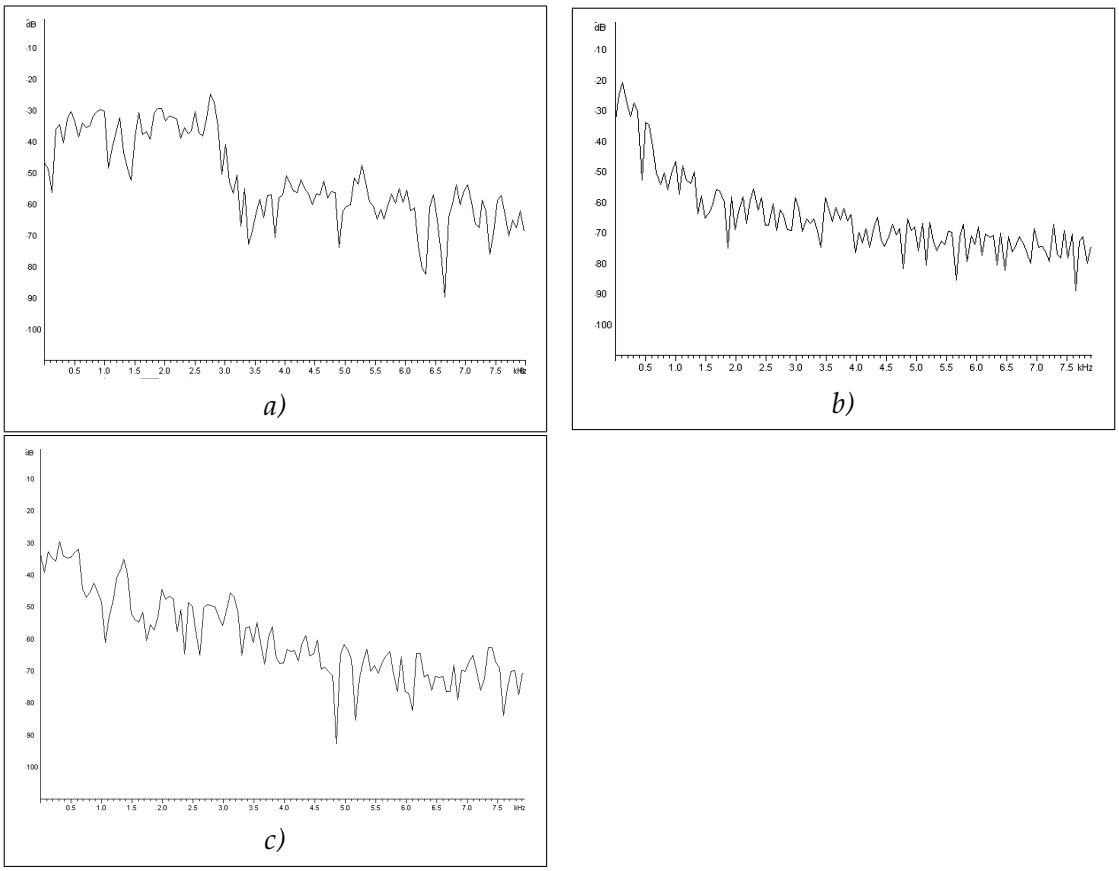


FIG. 6.3 – *Spectres des différents signaux de bruits utilisés. (a) bruit de communication HF, (b) babble noise, (c) bruit du char Leopard 2 roulant à 70 km/h.*

RSB, avec et sans débruitage, confirment que les paramètres Aurora permettent de réduire les taux d'erreurs de la VAL en environnement bruité (cf. table 6.1). La table 6.1 et les graphiques 6.4, 6.5 et 6.6 présentent les résultats d'expériences menées avec et sans débruitage selon le protocole BREF2, avec le système LIA-THL07-nonorm. Nous pouvons tout d'abord noter, que pour toutes les expériences menées sur des signaux à un RSB de 15dB, les résultats avec et sans débruitage sont très proches. A ce niveau de RSB, le débruitage est peu activé. Ensuite nous pouvons remarquer que, pour toutes les conditions de bruits, le débruitage Aurora améliore significativement les performances de reconnaissance.
Ces résultats démontrent aussi que le bruit de communication HF est le plus perturbateur. Le spectre en fréquence de ce bruit, représenté sur la figure 6.3 *a*, démontre qu'il s'agit du bruit de plus haute puissance sur les fréquences de la bande [0 :3000] Hz. C'est donc le bruit qui affecte le plus les fréquences de parole. Dans cette condition de bruit, pour un RSB de 0dB, le débruitage est très actif. Le gain relatif apporté est alors de 47 %, de 17% à 9% d'EER.

Bruits	Débruitage	RSB (dB)	DCF	EER
Bruit de communication HF	ON	0	4.40	9.03
	OFF		6.25	17.10
	ON	5	1.43	2.08
	OFF		2.86	6.19
	ON	15	0.78	0.98
	OFF		1.64	3.54
Bruit de foule	ON	0	2.36	5.73
	OFF		2.57	6.10
	ON	5	1.46	3.67
	OFF		1.58	3.67
	ON	15	0.80	1.96
	OFF		0.81	2.58
Bruit char Leopard 2	ON	0	2.50	6.46
	OFF		2.60	7.06
	ON	5	1.72	4.63
	OFF		1.82	5.00
	ON	15	1.1	3.2
	OFF		1.37	3.56

TAB. *6.1* – *Évaluations du standard Aurora sur des signaux bruités à différents RSB. Pour les signaux d'apprentissage le débruitage est toujours activé. Nous formulons ici l'hypothèse que sur les données d'apprentissage non bruitées le module de débruitage ne modifie pas le calcul des cepstres (les résultats à haut RSB le prouve). Le SNR est calculé sur les trames sélectionnées par la DAV. Expériences menées selon le protocole BREF2, avec le système LIA-THL07-nonorm.*

6.1.2 Extraction des paramètres dans le domaine compressé

Dans le but d'évaluer un autre schéma d'extraction des paramètres, compatible avec les caractéristiques de ces réseaux, nous proposons une solution utilisant l'extraction des paramètres cepstraux dans le domaine compressé, i.e, à partir des paramètres internes des codeurs de parole. Cette solution ne tient pas compte de la contrainte d'utilisation en ligne jusqu'alors abordée.

6.1.2.1 Le codage de la parole

Le codage de la parole est très utilisé sur les réseaux de communication. Il permet de diminuer la bande passante nécessaire au transfert de la parole. Sur les réseaux PMR, le débit alloué à la parole est encore plus faible que sur les réseaux de communications standards. Alors qu'il est de 13 kbit/s pour le GSM, les débits alloués aux codeurs de parole sur les réseaux PMR sont toujours inférieurs à 5kbit/s. Les distorsions apportées par les codeurs de parole sont d'autant plus importantes que le débit alloué est faible.
Pour exploiter le flux audio qui transite sur les réseaux, dans le but d'effectuer la RAL, la méthode la plus simple est de calculer les paramètres cepstraux à partir du signal

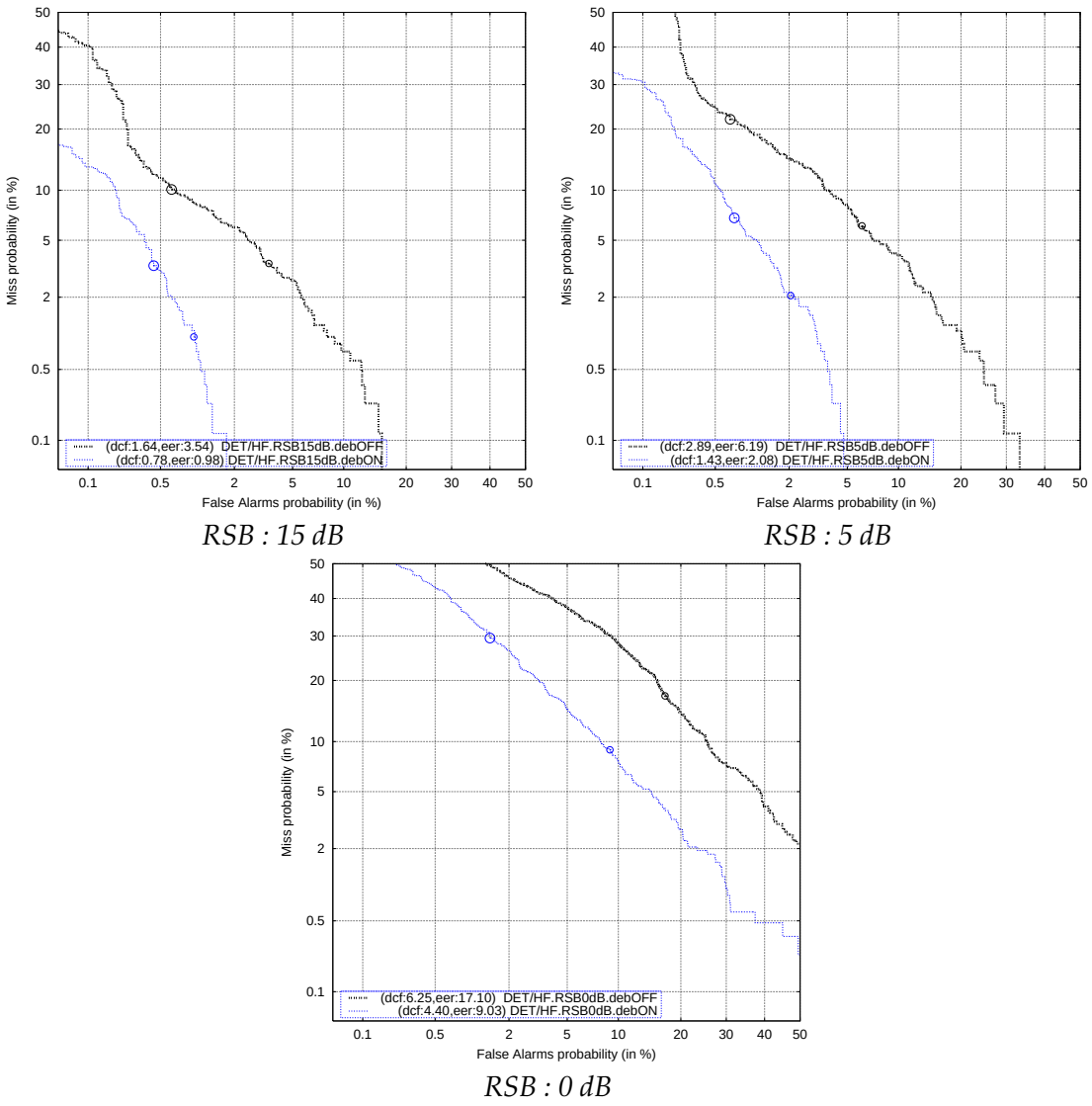


FIG. 6.4 – *Résultats en terme de DCF et EER pour les expériences menées avec et sans débruitage sur des signaux bruités par le bruit de communication HF (BREF2, LIA-THL07-nonorm).*

resynthétisé à partir des paramètres transmis. Ceci introduit alors des distorsions dues à la resynthèse (décodage) du signal.

Des travaux ont été menés pour utiliser des paramètres utiles à la RAL, à partir des paramètres internes des codeurs [Raj et al., 2001; Huerta et Stern, 1998; Petracca et Servetti, 2006; Quatieri et al., 2000]. Ceci permet notamment d'éviter les distorsions dues à la resynthèse du signal et la consommation de ressources supplémentaires.
Les codeurs utilisés comme références dans ces travaux (GSM) sont des codeurs CELP (*Code Excited Linear Prediction*), basés sur le modèle LPC (défini en section 2.1.3.2.3). Les auteurs proposent de dériver les coefficients cepstraux, à partir des coefficients de prédiction, selon les équations 2.7, 2.8 et 2.9. L'analyse LPC fait intervenir un résidu d'ana-

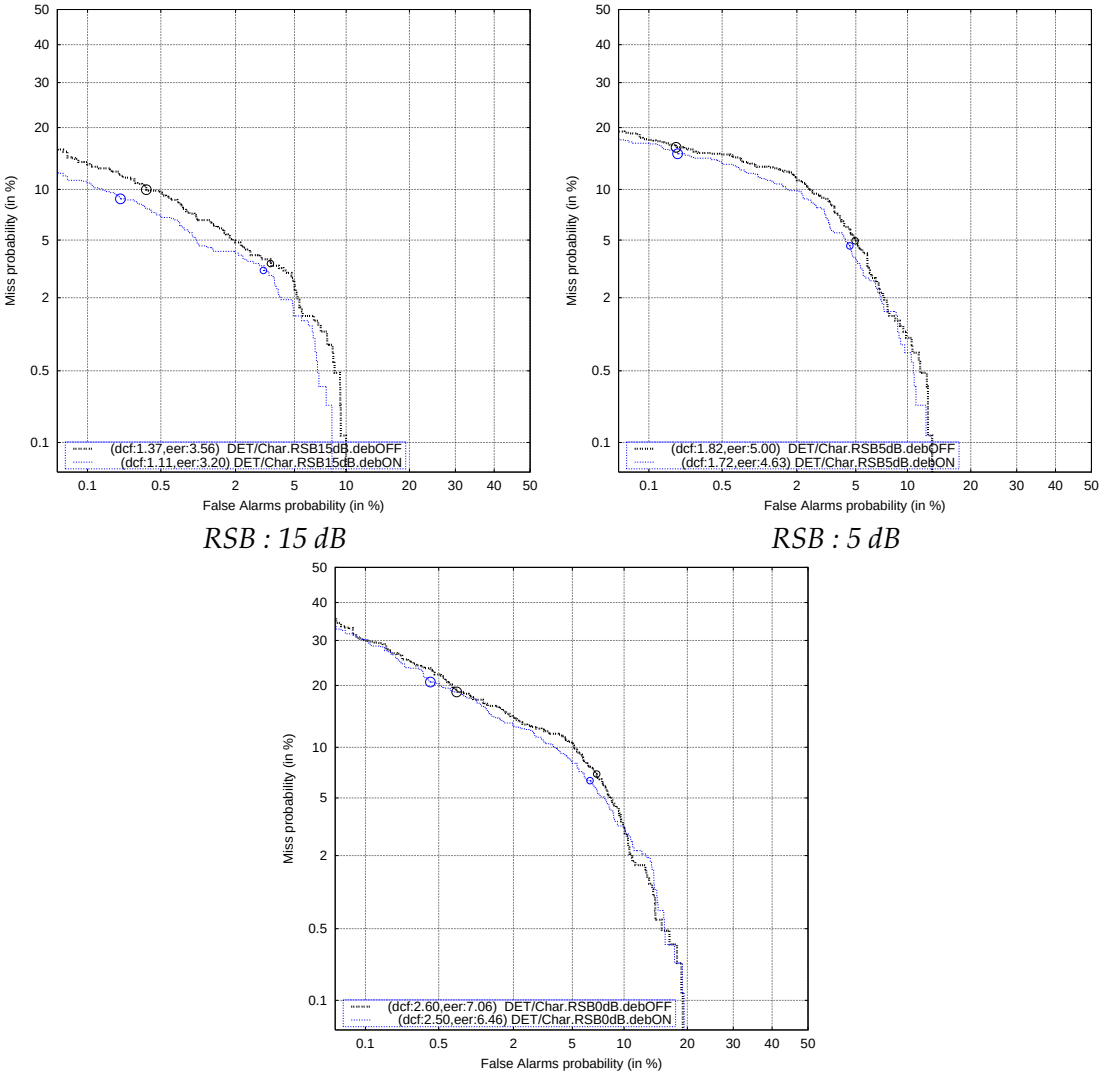


FIG. 6.5 – *Résultats en terme de DCF et EER pour les expériences menées avec et sans débruitage sur des signaux bruités par le bruit de char Leopard 2 (BREF2, LIA-THL07-nonorm).*

lyse : l'erreur de prédiction du modèle LPC. [Huerta et Stern, 1998] propose d'utiliser la combinaison des paramètres cepstraux, extraits des paramètres LPC, et de ce résidu. Les résultats prouvent que le fait de ne pas décoder le signal améliore sensiblement les performances.

Les codeurs de parole, qui seront détaillés dans le cadre de cette thèse, sont les codeurs TETRA et MELP. Ce sont aujourd'hui les codeurs les plus répandus sur les réseaux de communication de type PMR.

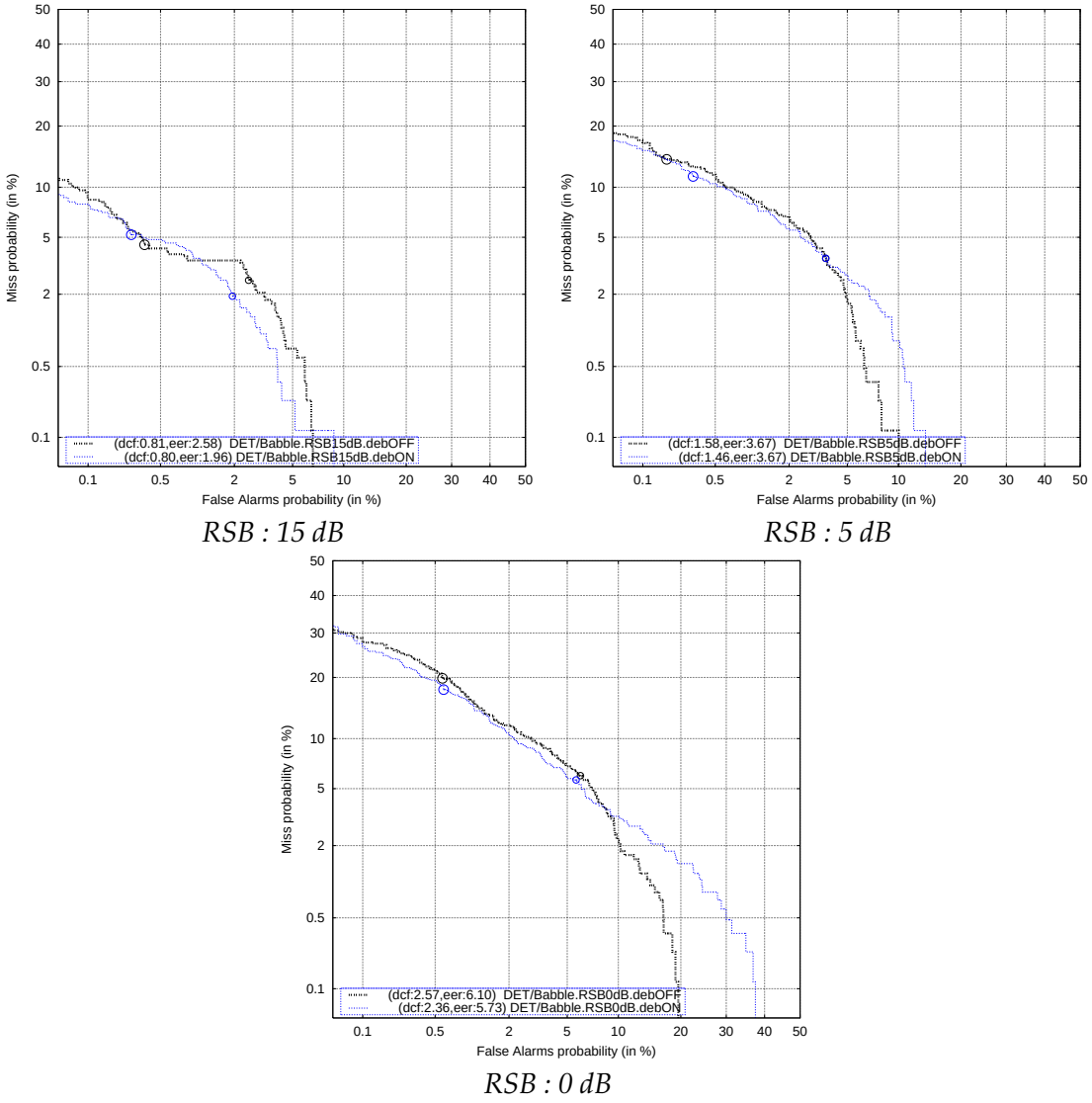


FIG. 6.6 – *Résultats en terme de DCF et EER pour les expériences menées avec et sans débruitage sur des signaux bruités par le bruit de babillage (BREF2, LIA-THL07-nonorm).*

6.1.2.1.1 Le codeur de parole TETRA

Le codeur *TETRA, TErrestrial Trunked RAdio* [ETSI, 2005b], est un codeur de parole bas-débit très répandu dans les réseaux PMR. Il a notamment été développé par l'entreprise Thales.
Son débit est de 4.6kbit/s. Il est basé sur le modèle de l'analyse LPC du signal. L'analyse LPC est réalisée sur des fenêtres de 30 ms de signal de parole. Les coefficients de pondération sont transformés en coefficients *LSP (Line Spectral Pair)*. Ce choix s'explique par la robustesse des LSP à la quantification. Au niveau du décodeur, les LSP sont interpolés. Un vecteur de LSP est disponible toutes les 7.5ms. Le schéma bloc de codage-décodage de ce codeur de parole est décrit en annexe.

6.1.2.1.2 Le codeur de parole MELP

Le codeur de parole OTAN STANAG-4591 *MELP, Mixed Excitation Linear Prediction* [NSA, 2006; McCree et al., 1996] est considéré comme le codeur à bas-débit état de l'art, pour les applications bas-débit [Supplee et al., 1997]. Il est majoritairement utilisé dans les équipements militaires et professionnels.
Son débit est de 2.4 kbit/s. Il est aussi basé sur le modèle classique de l'analyse LPC. Le codeur intègre un étage de réduction de bruit. L'analyse LPC est réalisée sur des fenêtres de 22.5 ms de signal de parole. Les coefficients de pondération sont transformés en coefficients *LSF (Line Spectral Frequency)* avant quantification. L'interpolation est aussi utilisée au niveau du décodage. Les LSF sont interpolés sur toutes les périodes égales à la fréquence fondamentale. Le schéma bloc de codage-décodage de ce codeur de parole est décrit en annexe. Une étude sur l'influence des codeurs bas-débit pour la RAL, est introduite par [Petracca et Servetti, 2006], avec notamment l'utilisation du codeur MELP à 2.4kbit/s.

6.1.3 Comparaison des méthodes de paramétrisation

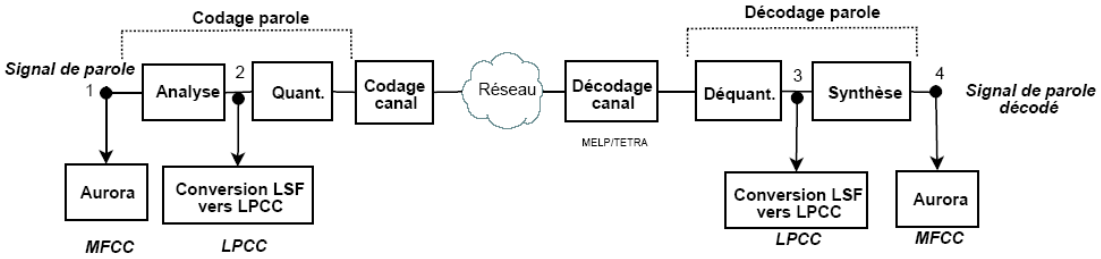


FIG. 6.7 – *Schéma de l'architecture distribuée d'un réseau PMR. Les méthodes d'extraction des paramètres possibles sont précisées. Méthode 1 : au niveau du terminal, méthode 2 : au niveau du codeur de parole, méthode 3 : au niveau du décodeur de parole et méthode 4 : après la synthèse du signal méthode 4.*

Nous évaluons, dans cette section, les différentes méthodes de paramétrisation disponibles aux différents niveaux de l'architecture d'un réseau PMR, décrites sur la figure 6.7. Les codeurs de parole TETRA et MELP sont les codeurs considérés dans les expériences présentées. Nous détaillons les différentes méthodes de paramétrisation :

- méthode 1 : le standard Aurora est utilisé. L'extraction des MFCC est réalisé sur le signal non codé.
- méthode 2 : niveau codage de la parole. L'analyse LPC permet d'extraire les LPCC directement dans le codeur avant quantification.
- méthode 3 : réception du flux binaire quantifié au niveau du récepteur. L'extraction des LPCC est réalisée après déquantification.
- méthode 4 : décodage de la parole. Le signal est resynthétisé au niveau du récepteur, Aurora est utilisé pour extraire les MFCC.

La méthode 1 est considérée comme la référence car le signal de parole n'est pas altéré par le codage. Cette configuration présente néanmoins des inconvénients (cf. 6.1.1.1). La méthode 2 n'est pas réalisable : elle imposerait de modifier les codeurs de voix implémentés dans les terminaux pour utiliser les paramètres. Néanmoins elle permet d'évaluer l'effet de la quantification des paramètres sur les performances de la RAL. La méthode 3 est la solution optimisée que nous proposons, car elle présente de multiples avantages :

1. elle évite l'altération du signal de parole et la consommation supplémentaire de ressources liées à la resynthèse du signal,
2. le flux binaire peut être intercepté n'importe où dans le réseau.

Comme exposé dans le chapitre 5, un traitement en ligne de toute la chaîne de paramétrisation est nécessaire pour proposer un application de surveillance réactive. Cette contrainte est vérifiée dans le cas de l'utilisation de la méthode 3. En effet, les codeurs TETRA et MELP sont des codeurs à la trame ; i.e, que les paramètres des codeurs sont calculés par périodes d'analyse, 22.5 ms pour le codeur MELP et 30 ms pour le codeur TETRA. Il est alors possible de calculer les LPCC toutes les 22.5 ou toutes les 30 ms.

6.1.3.1 Résultats des différentes paramétrisations

Nous évaluons les performances des différentes méthodes de paramétrisation exposées pour la RAL. Les protocoles suivants ont été suivis pour réaliser les expériences selon les méthodes 1 à 4 :

- pour la méthode 1 : le standard Aurora-3 est utilisé sur les signaux de parole de la base BREF ; le module de quantification des paramètres est utilisé. Les paramètres cepstraux sont déquantifiés.
- pour la méthode 2 : les signaux sont codés par le vocodeur considéré pour l'expérience. Les LPCC sont calculés avant la quantification des paramètres, au niveau de l'analyse LPC dans le vocodeur,
- pour la méthode 3 : les signaux sont codés par le vocodeur considéré pour l'expérience ; le train binaire obtenu est ensuite décodé. Les LPCC sont calculés après déquantification des paramètres du codeur.
- pour la méthode 4 : les signaux sont codés par le vocodeur considéré pour l'expérience ; le train binaire obtenu est ensuite décodé. Le standard Aurora-3 est utilisé sur les signaux de parole décodés.

Contrairement à l'extraction des paramètres Aurora (cf. paragraphe 6.1.1.1), l'échelle de Mel n'est pas appliquée aux LPCC. La DAV LIA, sur fichier, a été utilisée. Les expériences sont menées suivant le protocole BREF1 (40 locuteurs, 1 minute d'apprentissage et 8 scondes de test, cf. annexe A). Les résultats sont présentés dans la figure 6.8.

L'analyse de ces résultats démontre que, lorsque les paramètres cepstraux (LPCC) sont extraits du train binaire des codeurs, le codeur TETRA obtient de meilleures performances que la solution utilisant le signal décodé (issu de ce codeur). Le gain s'élève à 25% en terme d'EER par rapport à la méthode 4, soit un EER de 5.57 % (IdC à 95% [5.29 ;5.88]) contre un EER de 7.29% (IdC à 95% [6.97 ;7.7]).

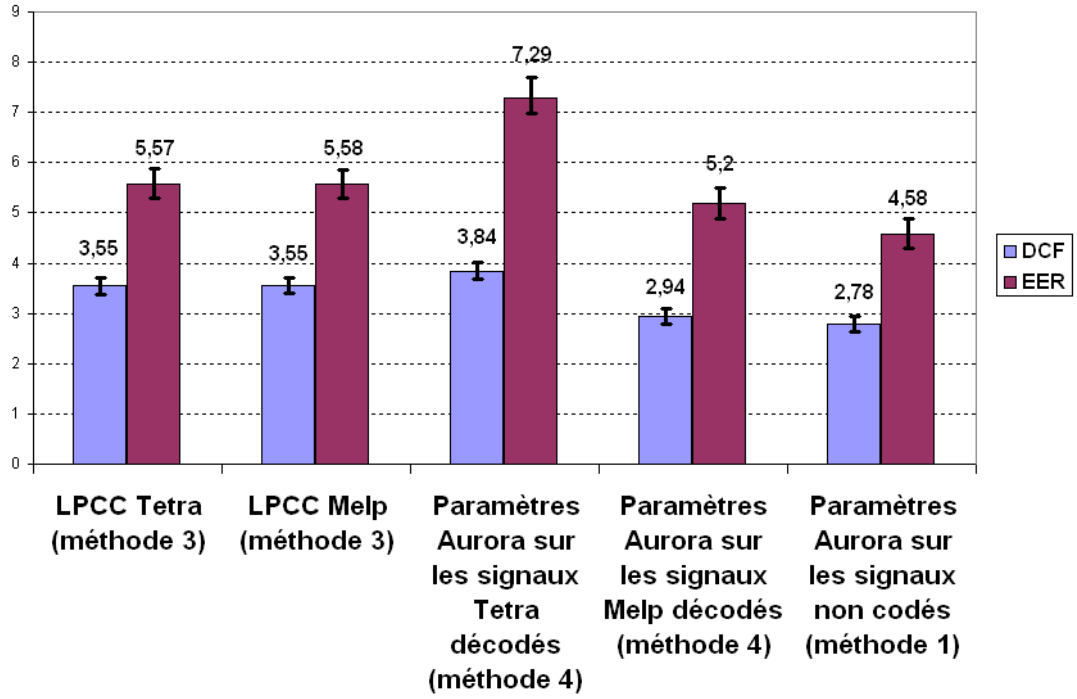


FIG. 6.8 – *Résultats des différentes méthodes de paramétrisations en terme d'EER et de DCF (protocole BREF1, DAV LIA, LIA-THL07-nonorm). Les intervalles de confiance à 95% des mesures sont représentés par les barres d'erreur.*

Le codeur MELP montre un profil différent ; il obtient en effet légèrement de meilleurs résultats en travaillant sur le signal décodé (5.2% d'EER, IdC à 95% [4.88 ;5.5]) qu'en utilisant le train binaire (5.58% d'EER, IdC à 95% [5.29 ;5.86]). Notons que le MELP sur le signal décodé permet d'approcher les performances du système de référence (méthode 1) avec 13% relatif de perte en terme d'EER seulement (contre 21% relatif de perte pour la meilleure configuration avec le codeur TETRA). Nous pouvons émettre l'hypothèse que la reconstruction du signal, en sortie du décodeur MELP, ajoute de l'information utile à l'analyse cepstrale, notamment avec la transmission des dix premiers modules de la transformée de Fourier du signal.

Enfin, on peut noter que la perte relative due au codage bas débit de la parole est évaluée à 20% pour les mesures EER et DCF (différences mesurées entre les expériences sur le signal non codé, point 1 figure 6.7 et les expériences dans le domaine compressé, point 3 figure 6.7).

Nous avons aussi évalué la perte due à la quantification, en extrayant les paramètres LPCC à la sortie du codeur de parole (point 2 figure 6.7). Les résultats sont présentés dans la table 6.2. Comparée aux expériences menées au point 3 de la figure 6.7, la perte due à la compression des paramètres est estimée à 20 % en terme d'EER. Le tableau 6.2 résume ces résultats d'expérience. Il apparaît que les paramètres LPCC extraits au niveau des codeurs MELP et TETRA ont les mêmes performances pour la RAL. Ces deux

Type de codeur	DCF	EER
MELP	3	4.56
TETRA	2.9	4.55

TAB. 6.2 – *Evaluation de la perte due à la quantification des paramètres des codeurs de parole. Expériences menées sur les LPCC extraits au niveau du codeur de parole (méthode 2).*

codeurs utilisent l'analyse LPC du signal de parole ; nous pouvons émettre l'hypothèse que les LPCC extraits sont très proches.

6.1.3.2 Résultats additionnels

Cette section présente des résultats complémentaires aux résultats précédents. Tout d'abord nous avons précisé que les codeurs TETRA et MELP utilisent l'interpolation des trajectoires de paramètres. Nous avons mené des expériences, sur le codeur de parole MELP, pour évaluer l'influence de l'interpolation pour la tâche de VAL. Le codeur de parole MELP interpole des paramètres *LSF, Line Spectral Frequencies* issus de l'analyse LPC. Une interpolation linéaire est appliquée entre les paramètres de deux trames consécutives : il s'agit d'une interpolation pitch synchrone. Les paramètres sont interpolés toutes les périodes de pitch.

La table 6.3 présente les résultats d'expériences menées sur le sous-ensemble masculin du protocole BREF1. Lorsque l'interpolation n'est pas mise en oeuvre, la fréquence

Interpolation	DCF	EER
Oui	2.52	4.72
Non	3.97	7.35

TAB. 6.3 – *Influence de l'interpolation des paramètres dans le codeur de parole MELP pour la tâche de VAL (BREF1, LIA-THL07-nonorm).*

de calcul des paramètres LPCC est de 22.5 ms. L'interpolation introduit une augmentation de la fréquence de calcul des paramètres d'un facteur 5 (environ une trame toute les 5 ms). L'interpolation étant générée par périodes de fréquence fondamentale (pitch), des expériences sur des signaux féminins montrent que deux fois plus de trames sont alors générées (pitch plus élevé). Pour les signaux masculins, le gain introduit par l'utilisation de l'interpolation est de 35% relatif pour les mesures DCF et EER, soit un EER de 4.72% (IdC à 95% [4.29 ;5.21]) avec interpolation, contre une EER de 7.35% (IdC à 95% [6.82 ;7.88]) sans interpolation. Les paramètres représentent plus finement la trajectoire de la réponse du conduit vocal.

Ensuite nous avons évalué l'utilisation des paramètres extraits dans le domaine compressé, sur des signaux réels. Une base de données, d'enregistrement réels, a été collectée (cf. protocole BREFVOC, annexe A) sur un réseau PMR TETRA. Ces enregistrements ont été collectés avec de vrais terminaux. Les enregistrements de simulation, utilisés pour les expériences précédentes, ne prenaient pas en compte la prise de son par

un terminal. Pour évaluer l'influence de la prise de son sur le terminal, nous avons réa-

Expérience sur le codeur TETRA	DCF	EER
pseudo-réel (BREFVOC)	2.86	5.43
simulé (BREF1)	3.55	5.57

TAB. 6.4 – *Expérience sur une base de données pseudo-réelles codées TETRA.*

lisé une expérience à partir de ces signaux « pseudo-réels ». Les résultats sont résumés dans le tableau (cf. table 6.4). Les performances de VAL en utilisant les données simulées et les données collectées sont proches, 5.43% d'EER (IdC à 95%[4.25 ;8.68]) pour l'expérience « pseudo-réelle » et 5.57% (IdC à 95% [5.29 ;5.88]) d'EER pour l'expérience de simulation. Bien que les protocoles d'évaluations soient différents, nous pouvons confirmer l'hypothèse que le terminal d'acquisition n'introduit pas de distorsion, dans les conditions d'enregistrements considérées.

6.1.4 Conclusion sur la paramétrisation

Nous avons évalué les performances de différentes méthodes de paramétrisation, et démontré que l'utilisation du standard Aurora est une solution robuste de paramétrisation. L'étage de débruitage implémenté dans Aurora réduit significativement l'influence des bruits ambiants sur les performances de VAL[2]. De plus, la perte due au codage bas débit de la parole, estimée à 20% en terme d'EER, est évitée et le standard est parfaitement compatible avec un traitement en ligne des paramètres.

Néanmoins cette solution n'est pas toujours disponible, comme dans le cadre applicatif qui nous intéresse. Nous avons détaillé les autres types de paramétrisation pouvant être mis en place. Pour le codeur MELP, l'utilisation du standard Aurora sur les signaux décodés, au niveau du récepteur, permet d'approcher les résultats de référence, sur le signal non codé. Pour le codeur TETRA, l'extraction des paramètres dans le domaine compressé amène un gain relatif de 25% pour rapport à l'utilisation du signal décodé. Similairement au standard Aurora, les paramètres cepstraux sont disponibles périodiquement au niveau du récepteur. Un traitement en ligne du flux audio est possible.

Ces résultats permettent de déterminer la meilleure solution de paramétrisation à choisir, en fonction du type de réseau *PMR* considéré. Les expériences menées sur les différents types de paramétrisation n'ont pas été effectués dans le cadre de signaux bruités. A des fins de comparaison avec le standard Aurora, il serait intéressant de mener des expériences pour évaluer les performances du débruitage du MELP, et la dégradation induite pour le codeur TETRA, qui n'intègre pas de débruitage.

[2]Ces résultats sont satisfaisants et il ne nous semble pas utile d'approfondir d'autres méthodes de débruitage comme la *PMC* ou le *multistyle training* (cf. paragraphe 5.4), dans le cadre de cette thèse.

6.2 Optimisation de la détection d'activité vocale

Nous avons précisé l'importance de la DAV sur les performances de VAL (cf. 4.3.3). Le critère d'énergie est majoritairement utilisé pour sélectionner les trames utiles. Dans la DAV de référence (DAV LIA), décrite en section 4.2.3, une modélisation à base de trois Gaussiennes de l'énergie du signal est appliquée sur un enregistrement entier.
Ce processus utilise la totalité du fichier d'enregistrement pour estimer le niveau d'énergie minimal des trames à sélectionner. Ceci ne permet pas un traitement en ligne. De plus ce type de DAV n'a pas été évalué dans le cadre de signaux fortement bruités. Le critère d'énergie, utilisé seul, peut alors sélectionner des trames de bruits. Nous proposons une solution de DAV, compatible avec un traitement en ligne, basée sur le standard Aurora-3.

6.2.1 Une solution de Détection d'Activité Vocale

Les techniques de détection d'activité vocale (DAV) basées sur un critère d'énergie ont prouvé leur efficacité pour les systèmes de VAL (cf. section 4.3.3). Les DAV utilisées dans les réseaux de communication, pour le codage de la parole [Vlaj et al., 2005], éliminent les silences non considérés comme des pauses linguistiques. Ces DAV n'éliminent pas de trames de parole pour en conserver son intelligibilité. [Vlaj et al., 2005] compare les différentes DAV des codeurs de parole pour une application distribuée de reconnaissance de la parole. Les auteurs classent ces DAV (G.729, G.723 et Aurora) par ordre de performances. La DAV implémentée dans le standard Aurora est, en moyenne, la plus performante dans les conditions de RSB considérées pour les expériences[3].
Le fonctionnement de la DAV Aurora utilise une logique à plusieurs étages. Tout d'abord, un indicateur sur l'énergie de la trame, permet de sélectionner cette trame. Puis, des mesures de l'accélération de l'énergie de la trame, sur une sous-région du spectre qui contient la fréquence fondamentale, et sur la bande basse du spectre [ETSI, 2005a], permettent de déterminer si la trame est une trame de parole.
La DAV Aurora sélectionne en moyenne 80% des trames des signaux considérés dans ce travail (base BREF). Par référence à la DAV LIA (60% de trames sélectionnées), cette DAV n'est pas assez sélective.

Pour rendre plus restrictif le critère de sélection des trames, nous proposons de combiner les informations de DAV et de voisement de la trame, disponible dans le standard Aurora-3. Le critère de voisement est associé au module de débruitage. L'information de voisement, basée majoritairement sur la détection de la fréquence fondamentale, doit alors assurer que la trame est une trame de parole. Pour obtenir un pourcentage de trames sélectionnées de l'ordre de 40%, nous choisissons de n'utiliser que la classe de plus fort voisement. L'information de voisement présente les mêmes avantages que la DAV classique d'Aurora :

– cette information est transmise au serveur,

[3] Les auteurs proposent leur technique de DAV qui surpasse toutes les DAV testées.

– l'information est disponible en même temps que le vecteur de MFCC (traitement en ligne).

6.2.2 Résultats

Pour évaluer l'intérêt de l'approche de DAV basée sur le standard Aurora, présentant un taux de trames sélectionnées comparable à la DAV LIA de référence, nous comparons deux expériences menées sur la base BREF. La figure 6.9 présente les courbes

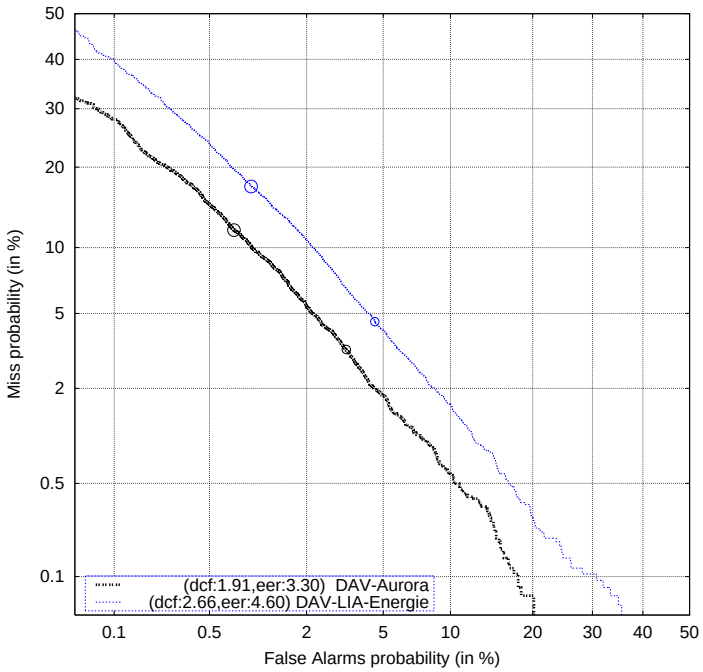


FIG. 6.9 – *Courbes DET d'expériences menées en utilisant la DAV Aurora et la DAV LIA (protocole BREF1, LIA-THL07-nonorm). La DAV Aurora surpasse les performances de la DAV LIA, avec un EER de 3.30 % (IdC à 95% [3.08 ;3.55]) contre un EER de 4.60% pour la DAV LIA (IdC à 95% [4.35 ;4.88])*

DET de performances de la DAV basée sur le voisement et celles de la DAV de référence (LIA), selon le protocole BREF1 (cf. annexe A), avec le système LIA-THL07-nonorm. Les meilleures performances sont obtenues avec la DAV Aurora+voisement, compatible avec un traitement en ligne. De plus, l'utilisation du critère de trame voisée du standard Aurora apporte un gain en performance de 28% relatif, pour les mesures DCF et EER, comparativement à la DAV LIA (traitement sur fichiers) avec un EER de 3.30 % (IdC à 95% [3.08 ;3.55]) contre un EER de 4.60% pour la DAV LIA (IdC à 95% [4.35 ;4.88]).

6.2.3 Conclusion sur la DAV

Les trames voisées sont généralement de plus haute énergie. Néanmoins, une modélisation globale de l'énergie, sur l'ensemble d'un enregistrement, ne peut tenir compte des fines variations sur un horizon temporel à court terme. Des trames de bruits, de forte énergie, peuvent être sélectionnées par une DAV classique, basée sur une classification de l'énergie. La DAV Aurora implémente des mécanismes robustes pour la détection des sons voisés. L'étage de débruitage, dont nous avons prouvé l'efficacité sur les signaux bruités, associé à l'information de voisement, permet de mettre en place une DAV robuste pour la VAL. De plus, cette DAV est compatible avec un traitement en ligne de VAL. L'information de voisement est disponible toutes les 10 ms après un retard à l'exécution de seulement 6 trames (soit 60ms).

6.3 La normalisation « en ligne » des paramètres

Dans les systèmes de VAL, l'étage de traitement des paramètres acoustiques est complété par la normalisation de la moyenne et de la variance des paramètres, pour réduire l'effet des bruits de convolution. Les variations du canal acoustique à long terme sont estimées sur des durées longues d'enregistrement. Généralement, la normalisation moyenne et variance des paramètres cepstraux (*CMVN*) est appliquée par fichier (cf. 3.2.3.1). Les estimateurs de moyenne et variance sont alors estimés de façon robuste. Les variations à court-terme du canal acoustique sont majoritairement compensées par des méthodes de normalisation moyenne-variance sur des fenêtres glissantes [Pujol et al., 2006; Viikki et Laurila, 1998; Pelecanos et Sridharan, 2001; Xiang et al., 2002]. Ainsi, les résultats des méthodes de compensation du canal acoustique sur fichier et sur fenêtres glissantes présentent généralement les mêmes performances sur des signaux de type conversation téléphonique.

Pour proposer un étage de paramétrisation complet, compatible avec un traitement en ligne, nous nous basons sur les méthodes de compensation de canal à court-terme pour définir notre solution. Nous introduisons une méthode de normalisation des paramètres, basée sur une réestimation des paramètres de moyenne et variance, par un facteur d'oubli [Mauler et Martin, 2006; Pujol et al., 2006].

6.3.1 Solution proposée

La composante introduite par le canal de transmission est difficile à caractériser sur des tailles de fenêtres courtes. Pour caractériser la moyenne et la variance d'une population de façon robuste, il faut un nombre conséquent d'observations. L'utilisation d'un facteur d'oubli permet de capitaliser sur les estimation précédentes (estimation à long terme), et de suivre les variations des paramètres (variation à court terme).
Cette procédure consiste en l'utilisation d'une fenêtre d'initialisation de N trames sélectionnées par la DAV. Les estimateurs μ et σ sont initialisés à partir des N premières

trames (sélectionnées par la DAV). Ensuite les paramètres sont normalisés, trame à trame, sans aucun délai (une puissance de calcul suffisante doit pouvoir rattraper le retard du au calcul des estimateurs sur les N premières trames).
Les paramètres de normalisation sont calculés sur la fenêtre d'initialisation, puis mis à jour continuellement pour chaque nouvelle trame i, selon les équations 6.1 et 6.2, avec N le nombre de trames nécessaires à l'initialisation, et β le facteur d'oubli.

$$\hat{\sigma}_i^2 = \beta\hat{\sigma}_{i-1}^2 + (1-\beta)\sigma_i^2 \quad (6.1)$$

$$\hat{\mu}_i = \beta\hat{\mu}_{i-1} + (1-\beta)\sigma_i \quad (6.2)$$

$$\beta = \begin{cases} \frac{N-1}{N} & \text{si } i > N \\ 1 & \text{sinon} \end{cases}$$

6.3.2 Résultats

Les résultats d'expériences menées en utilisant la normalisation en ligne sont comparées avec les résultats de la normalisation de référence sur fichier. Les résultats obtenus, selon le protocole BREF1 (cf. annexe A), avec la normalisation en ligne et avec la normalisation de référence, fonctionnant en mode « fichier », sont présentés sur la figure 6.10. Deux expériences avec deux quantités de trames d'initialisation (150 et 300 trames) ont été réalisées. Nous avons choisi ces quantités de trames pour l'initialisation car 150 trames représentent la limite acceptable en terme de perte de performance de VAL (20% de perte relativement à la référence). Avec un retard de 300 trames, les résultats approchent la normalisation sur fichier. Nous retrouvons ici les tailles majoritairement utilisées dans les schémas de normalisation sur fenêtre glissante [Pujol et al., 2006; Xiang et al., 2002]. Pour les expériences, les fichiers d'apprentissage, du modèle du monde et des locuteurs, ont été normalisés de façon classique, car ce traitement peut être considéré comme « hors-ligne » (cf. 5.3.1.2). Les résultats obtenus avec un horizon d'initialisation de 300 trames (3.49% EER, IdC à 95% [3.3 ;3.74]) sont très proches de ceux obtenus avec une normalisation sur fichiers (3.30% EER, IdC à 95% [3.08 ;3.55]). Cependant il est à noter que comme la durée des enregistrements est de seulement 8 secondes, soit environ 500 trames après DAV, 60% des trames sélectionnées se trouvent dans la fenêtre d'initialisation. L'horizon d'estimation est très proche de celle du fichier.

Pour évaluer la méthode de normalisation en ligne, en éliminant ce biais, nous avons utilisé la base NIST SRE 2005 où les durées de test sont de 2 minutes 30 secondes. Les résultats sont donnés dans la table 6.5. Les résultats démontrent que les performances de la normalisation « en ligne » sont similaires à celles de la normalisation sur fichier, lorsque 300 trames sont utilisées pour la fenêtre d'initialisation, pour des fichiers de test de 2 minutes 30 secondes. Nous pouvons donc conclure que cette méthode est robuste à la longueur des enregistrements (8 secondes et 2 minutes 30 secondes).

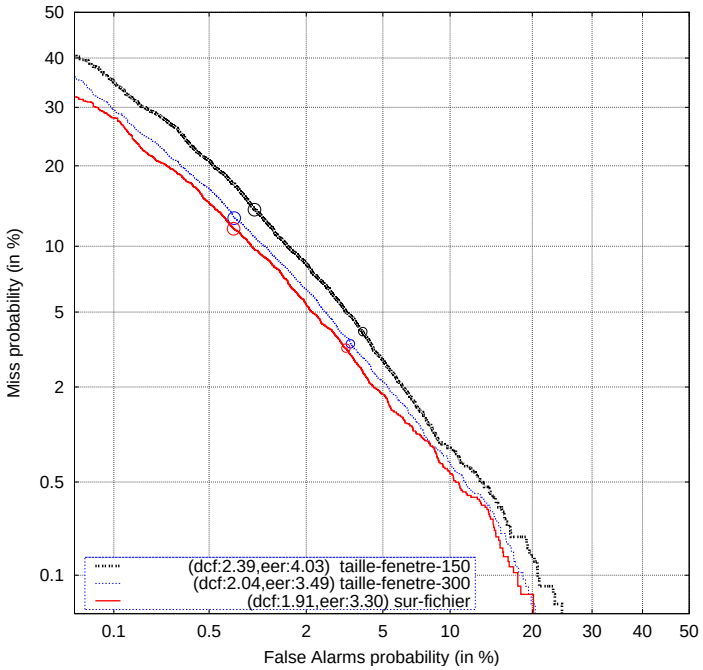


FIG. *6.10* – *Résultats de la normalisation en ligne et sur fichier (BREF1, LIA-THL07). La normalisation en ligne avec une initialisation de 300 trames (EER de 3.49%, IdC à 95% [3.3 ;3.74]) approche les performances de la normalisation de référence sur fichier (EER de 3.30, IdC à 95% [3.08 ;3.55]).*

6.3.3 Conclusion sur la normalisation

L'estimation des paramètres du canal de transmission à long terme n'est pas réalisable avec très peu de données. Les mécanismes d'estimation à court-terme, le *Feature Warping* par exemple, ont démontré de bonnes performances en VAL. Nous avons implémenté une normalisation compatible avec des durées courtes de traitement. L'utilisation d'un facteur d'oubli permet une estimation robuste des paramètres de normalisation. Elle permet aussi de prendre en compte la dynamique des paramètres de moyenne et variance sur un horizon à court terme. Les résultats en VAL de cette normalisation sont similaires à la normalisation sur fichier, lorsque suffisamment de trames sont utilisées pour calculer les estimées initiales des paramètres. L'utilisation de 300 trames

Méthode de normalisation	**DCF**	**EER**
Sur fichier	4.14	8.13
Initialisation 300 trames	4.24	8.13
Sur fichier + TNORM	3.37	8.93
Initialisation 300 trames +TNORM	3.38	8.79

TAB. *6.5* – *Effets des différentes méthodes de normalisation sur la base NIST SRE 2005 (LIA-THL07-nonorm et LIA-THL07-tnorm). Les paramètres et la DAV Aurora sont utilisés. Les modèles imposteurs utilisés pour TNORM sont normalisés par la méthode sur fichier.*

pour la fenêtre d'initialisation semble être un bon compromis entre le retard d'initialisation et les performances obtenues. Le délai initial, pour le traitement en ligne, est de 3 secondes. Une fois ce délai passé, les trames sont normalisées une à une. Ce fonctionnement permet un traitement en ligne car un retard de 300 trames peut aisément être envisagé dans notre scénario de surveillance.

6.4 Adapter la décision de vérification à un fonctionnement « en ligne »

Nous avons présenté des solutions de paramétrisation, ainsi qu'une DAV et une normalisation des paramètres, qui répondent aux contraintes d'utilisation « en ligne ». Nous savons que les durées d'apprentissage et de test sont déterminantes en RAL. Un test de vérification, basé sur de très petites périodes de signal, ne présente pas de bonnes performances de RAL [Magrin-Chagnolleau et Bonastre, 1995].
Une méthode pouvant être envisagée est de cumuler et de moyenner les scores à la trame. Le score de vérification obtenu à un instant t bénéficierait ainsi des scores des $t-1$ trames précédentes. Néanmoins, cette méthode ne permet pas de détecter un changement de locuteur. En effet, un changement de valeur du LLR sur quelques trames n'influence pas suffisamment le calcul moyen des LLR, lorsqu'il est calculé sur beaucoup de trames.

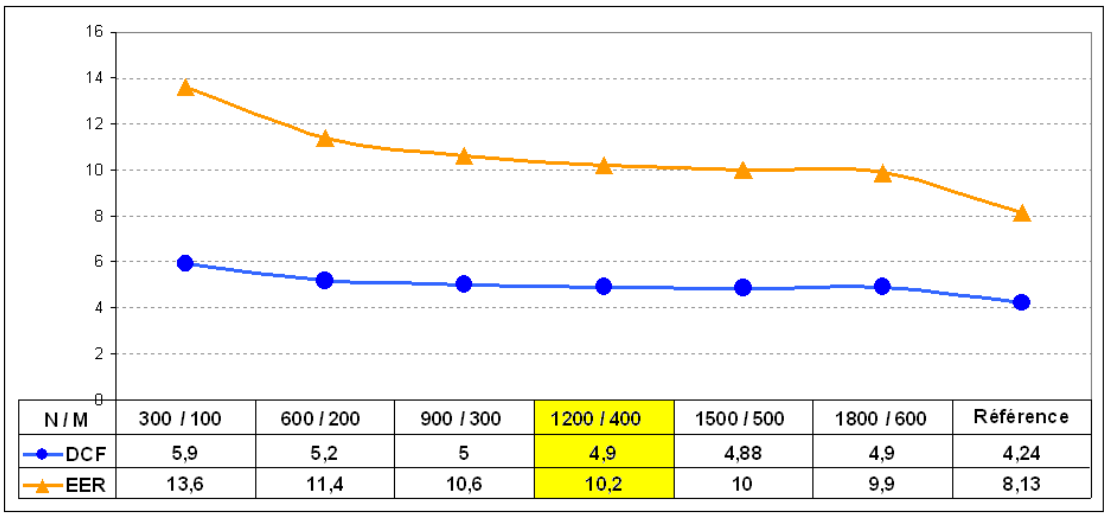
Nous avons choisi de proposer une mise à jour par facteur d'oubli des LLR calculés sur de courtes périodes. Ce processus permet de capitaliser sur les LLR des trames précédentes, mais sur un horizon à plus court terme, permettra probablement de détecter un changement de locuteur.
La méthode proposée utilise une phase d'initialisation de N trames pour estimer un LLR. Ensuite, les LLR des M trames suivantes sont utilisées pour mettre à jour l'estimation initiale du LLR :

$$\begin{aligned} \widehat{LLR}(F_i) &= \lambda * LLR(F_{i-1}) + (1-\lambda) * LLR(F_i) \\ \lambda &= \frac{N-1}{N} \end{aligned} \tag{6.3}$$

où $LLR(F_i)$ représente le LLR cumulé sur M trames, et $\lambda = \frac{N-1}{N}$ le facteur d'oubli. A l'initialisation $i = 0$, $\lambda = 0$ et $LLR(F_0)$ représente le LLR cumulé sur les N premières trames. Deux durées en nombre de trames définissent les paramètres de notre méthode :

1. N pour l'estimation du LLR de la phase d'initialisation,
2. M qui représente la période où un score de vérification est disponible.

Le score de vérification, disponible par période de M trames après la phase d'initialisation, permet un fonctionnement en ligne. Nous évaluons les performances de la méthode selon différentes durées pour les paramètres N et M. Ces expériences ont été menées sur la base NIST SRE 2005. La figure 6.11 présente les taux d'erreurs associés à des durées d'estimation des LLR différentes, en terme de trames sélectionnées par la DAV.



N / M	300 / 100	600 / 200	900 / 300	1200 / 400	1500 / 500	1800 / 600	Référence
DCF	5,9	5,2	5	4,9	4,88	4,9	4,24
EER	13,6	11,4	10,6	10,2	10	9,9	8,13

FIG. *6.11* – *Résultats en terme de mesure DCF et EER pour des expériences menées avec différentes quantités de trames N pour l'initialisation du LLR et M trames pour la mise à jour du LLR. Le score de vérification est obtenu par périodes de M trames. Il s'agit ici de durées de test en nombre de trames sélectionnées par la DAV. La normalisation en ligne (initialisation de 300 trames) a été utilisée (NIST SRE 2005, LIA-THL07-nonorm).*

La référence est le traitement sur fichier : le calcul du LLR est la somme pondérée des LLR à la trame, sur tout le fichier de test (2 minutes 30 secondes). Nous pouvons ainsi observer que le taux d'erreurs est élevé pour de faibles durées de test (13.6% d'EER). Il diminue lorsque l'horizon de calcul du LLR augmente. Le taux d'erreur atteint une asymptote, pour notre configuration, aux alentours de 400 trames sélectionnées avec un EER de 10.2%, soit 12 secondes de parole voisée pour initialiser le LLR, puis une décision de vérification toutes les 4 secondes. Un traitement en ligne de la décision de vérification peut être envisagé avec une perte relative de 25% pour l'EER. Ces résultats démontrent que l'étage de décision induit des contraintes plus fortes que la normalisation des paramètres, en terme de trames à traiter, pour déterminer une décision de vérification fiable. Le retard du à l'étage de paramétrisation est le moins limitant.

6.5 Conclusion

Dans ce chapitre, nous avons analysé les contraintes induites par l'architecture des réseaux de communication, pour la mise en place d'un système de surveillance basé sur la VAL.
Le choix d'une paramétrisation appropriée a été abordée. La solution basée sur le standard Aurora permet d'éviter les pertes dues au codage bas-débit de la parole. L'étage de débruitage permet de réduire la perte de performances lorsque les signaux de parole sont très fortement bruités. Pour des conditions de RSB de 0dB, le gain estimé est de 47% relatif pour l'EER.

Nous avons aussi évalué des méthodes alternatives de paramétrisation. Aujourd'hui, le standard Aurora n'est pas disponible sur les réseaux *PMR*. Dans ce cas, l'extraction des paramètres de VAL, à partir des paramètres internes des codeurs de parole, peut permettre de limiter la perte due à la resynthèse du signal. Lorsque cette technique de paramétrisation est appliquée au codeur de parole TETRA, le gain se porte à 25% relativement à l'utilisation d'une paramétrisation standard sur le signal décodé. Le codeur MELP ne profite pas de l'utilisation des paramètres internes des codeurs. Il présente de meilleurs résultats de VAL lorsque le signal resynthétisé, au niveau du décodeur, est utilisé.

Le choix de la DAV est un élément déterminant des performances de VAL. Dans l'optique qui consiste à proposer un traitement de VAL « en ligne », nous avons évalué l'utilisation d'une DAV basée sur le voisement. La DAV proposée apporte un gain relatif de 28% pour les mesures EER et DCF, comparée à la DAV de référence (DAV LIA).

L'étage de paramétrisation est composé de l'extraction des paramètres acoustiques, de la sélection des trames utiles (DAV) et d'une étape de normalisation des paramètres, pour réduire l'influence des bruits de convolutions apportés par la chaîne de transmission. Une normalisation des paramètres, opérant sur de petites périodes temporelles, est nécessaire pour envisager un traitement de VAL « en ligne ». La normalisation présentée égale les performances de la normalisation « sur fichier ».

Pour surveiller une communication en continu, il faut proposer un score de vérification sur des échantillons temporels les plus petits possibles. Les traitements considérés « hors-ligne » (cf. section 5.3.1.2) n'imposent pas de modifications particulières de l'architecture de VAL. Le test de vérification doit cependant fournir une décision sur une durée temporelle minimale, qui permettra au système d'être réactif à une intrusion (vol de terminal). En se basant sur une méthode de mise à jour des scores de vérification, nous avons évalué que 12 secondes de parole voisée sont nécessaires pour l'initialisation de la méthode, puis que 4 secondes sont suffisantes pour proposer une décision avec une fiabilité de 10% en terme d'EER.

En résumé, les méthodes proposées dans ce chapitre encouragent la mise en place d'un système de VAL, en ligne, sur les réseaux professionnels de communications, sans perte significative de performance. La paramétrisation, la DAV et la normalisation moyenne variance, sont compatibles avec un traitement en ligne et présentent des résultats proches de notre système de référence sur fichier. La décision de vérification a été évaluée comme l'étage le plus contraignant. Néanmoins, la méthode que nous proposons réduit le nombre de trames nécessaires pour une décision fiable. Nous espérons que cette méthode permettra aussi de détecter les changements de locuteur.

Chapitre 7

L'adaptation non supervisée continue des modèles de locuteur

Sommaire

Pour répondre aux contraintes ergonomiques du scénario envisagé et, notamment, aux durées courtes d'apprentissage, nous introduisons dans ce chapitre la technique connue sous le nom d'adaptation non supervisée des modèles de locuteurs.
Cette méthode de collecte de données, en aveugle, appartenant aux locuteurs, à partir des signaux de test de vérification permet d'améliorer la modélisation du locuteur.
Nous exposons ici les techniques d'adaptation non supervisée les plus communément utilisées. Les difficultés majeures existantes dans l'utilisation de telles techniques sont détaillées.
Enfin, nous présentons une méthode qui s'affranchit des contraintes des systèmes classiques d'adaptation.

7.1 Principe de l'adaptation non supervisée

Nous avons évoqué qu'une grande partie des efforts de recherche a été axée sur la compensation des variabilités intra-locuteur et intersession, afin de traiter le problème de variabilité entre enregistrements. Ce phénomène est la cause de la plus grande partie des pertes de performances en RAL.
La méthode de *Joint Factor Analysis*, développée récemment dans [Kenny et al., 2005a] permet de réduire les problèmes de variabilités locuteur et du canal de transmission. Cette méthode permet de retirer les composantes de variabilités d'un enregistrement, pour ne conserver que la partie importante du locuteur. Cette méthode apporte un gain très significatif (cf. section 4.3.3.3.2).
L'adaptation non supervisée des modèles de locuteur est une autre solution à ce problème (cf. section 5.3.2.2). Elle consiste à utiliser les tentatives d'accès au système comme nouvelles données pour, par exemple, augmenter la quantité de données utiles pour la modélisation des locuteurs. En effet, durant le fonctionnement du système de reconnaissance de locuteur, certains accès au système sont reconnus comme étant prononcés par un client du système. Ils peuvent être alors utilisés comme données d'adaptation (cf. figure 7.1). Ces données permettent d'améliorer la modélisation des variabilités de la parole.
Ces techniques peuvent s'avérer être très intéressantes pour les applications de sur-

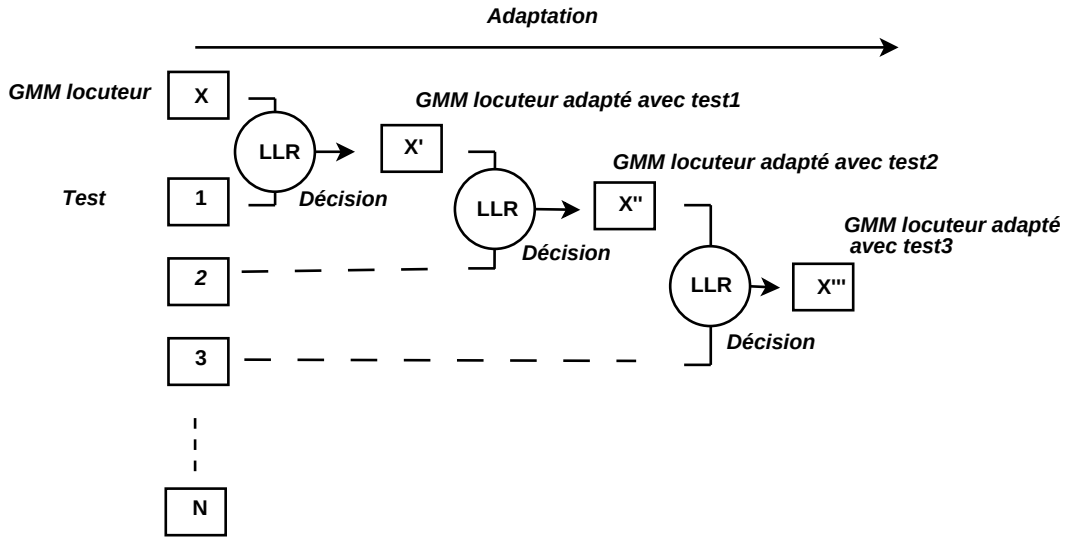


FIG. 7.1 – *Principe de l'adaptation non supervisée des modèles de locuteurs.*

veillance. La surveillance étant permanente, il est utile de pouvoir enrichir le modèle client, en continu, avec des données récentes.
[Wu et al., 2003] utilise les méthodologies de l'adaptation non supervisée, pour la segmentation en locuteur. Le problème récurrent en segmentation du locuteur, le manque de données étiquetées peut être compensé.

De plus, cette technique d'adaptation en ligne peut répondre au problème de collecte de données d'apprentissage. Elle permet de démarrer le système de RAL avec peu de données d'apprentissage, pour ne pas contraindre les utilisateurs à de longues périodes d'enregistrement. L'enrichissement des données est automatique.

La sélection des données utiles pour l'adaptation est basée sur le test de vérification. En pratique, les seuils de décision de vérification et de sélection pour l'adaptation sont différents. Le seuil de sélection est souvent plus restrictif car, choisi trop faible, des accès imposteurs peuvent être sélectionnés et dégrader le système. [Fredouille et al., 2000] évalue le comportement de l'adaptation selon différents scénarios d'attaques du système par des imposteurs. Le taux de FA est critique pour l'adaptation ; il peut engendrer des erreurs dans la sélection des données d'adaptation, et permettre à des imposteurs d'accéder au système. Si le seuil de sélection est choisi trop élevé, peu de données vont être sélectionnées. Dans ce cas, l'ajout de données très bien reconnues par le système de vérification (scores élevés) ne permet pas d'améliorer sensiblement les performances. En effet, l'information additionnelle contenue dans ces données est très faible et même redondante. Ces observations exposent la problématique majeure pour l'adaptation non supervisée des modèles de locuteurs : la difficulté de fixer le seuil de sélection.

7.2 Les solutions basées sur un seuil de sélection

7.2.1 Principes

La décision de sélection des accès pour l'adaptation est basée sur un seuil fixé *a priori* sur les scores [Vair et al., 2007; Hansen et al., 2006; Van Leeuwen, 2004]. Déterminé empiriquement, il est conditionné par la minimisation du taux de fausses acceptations (FA). Les données sélectionnées sont ensuite utilisées pour adapter le modèle d'apprentissage du locuteur. La technique communément adoptée est l'adaptation MAP du modèle du monde avec les données d'apprentissage, et les données de test sélectionnées.
Les gains en performance de ces systèmes sont limités. Les résultats présentés sur les bases de données d'évaluations NIST SRE, où le nombre d'accès client est faible (1 accès client pour 10 tests imposteurs en moyenne, annoncé par le NIST), sont très loin de l'optimal pouvant être atteint par les méthodes d'adaptation supervisée (cf. section 5.3.2.2).
Les récents résultats de Loquendo aux évaluations NIST SRE 2008, montrent un gain significatif, avec l'utilisation d'une telle technique, sur une tâche difficile : seulement 10

secondes pour l'apprentissage et le test. Il faut noter que le protocole d'évaluation introduit une grande quantité d'accès clients. [Vair et al., 2007] a évalué la même technique sur la base de données NIST SRE 2006, avec cette fois-ci un gain non significatif. Les auteurs précisent qu'avec un seuil de sélection conservatif, environ un tiers du nombre total des données de test clientes sont sélectionnées pour l'adaptation. 8% d'entre elles sont des accès imposteurs.

Pour tenir compte de la confiance accordée au score de vérification, des méthodes basées sur l'utilisation de mesure de confiance dans le processus d'adaptation ont été développées. [Heck et Mirghafori, 2000; Mirghafori et Heck, 2002; Heck et Mirghafori, 2001] utilisent une mesure de confiance basée sur la probabilité *a posteriori* du test d'appartenir au modèle de locuteur. La mise à jour des moyennes du GMM par adaptation MAP est alors pondérée par cette probabilité. Elle est transformée en un poids par projection grâce à une fonction déterminée. [Heck et Mirghafori, 2000] utilise une fonction non linéaire basée sur une distribution de Rayleigh cumulée (cf. equation 7.1). Elle s'apparente a une séparatrice (sigmoïde).

$$w(Score(X)) = 1 - exp(\frac{-(Score(X) - \tau)^2}{2b^2}) \tag{7.1}$$

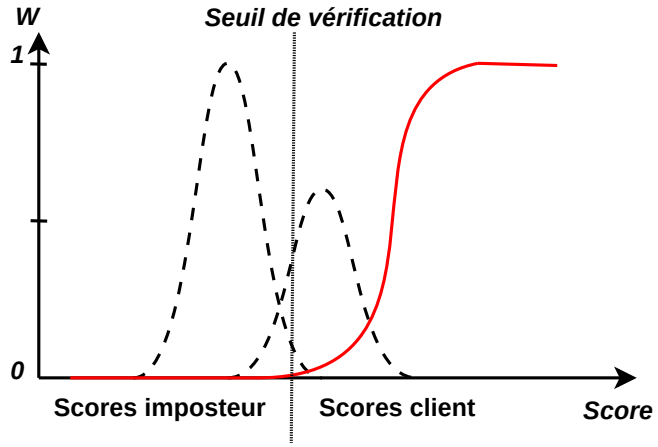


FIG. 7.2 – *Illustration de l'utilisation d'une distribution de Rayleigh cumulée pour déterminer un poids d'adaptation.*

où τ est le seuil de décision du système de VAL et b un coefficient qui contrôle la pente de la sigmoïde.

On peut voir sur la figure 7.2 que le poids augmente à mesure que le score augmente, pour tendre vers 1. Les statistiques μ, σ, w des GMM clients sont alors mises à jour selon un facteur d'oubli F :

$$\widehat{y} = y(n-1)(1-F) + Score(X)\beta_y y(n) \tag{7.2}$$

où y représente alternativement les paramètres du GMM μ, σ, w. $Score(X)$ est le score de vérification du test X, et où β_y est un facteur de lissage de différentes valeurs pour l'estimation de chacun des paramètres du GMM. n représente l'indice de l'itération d'adaptation. Les auteurs ont choisi de fixer le poids d'adaptation à zéro, au niveau du seuil de décision.

[Mengusoglu, 2003] introduit une mesure de confiance binaire pour aider à la décision de sélection d'un test. Elle est basée sur une mesure de corrélation par la transformation inverse de Fischer [Fischer, 1990], entre le score de vérification, et les distributions de scores client et imposteur déterminées sur un ensemble de développement.

[Barras et al., 2004] propose d'introduire le score de vérification dans l'estimation de la moyenne du GMM locuteur par adaptation MAP. La moyenne est alors mise à jour selon l'équation :

$$\widehat{\mu}_i = \frac{\tau \mu_i + p(\lambda|S) * O_i(X)}{\tau + p(\lambda|S) * n_i} \tag{7.3}$$

ou i est l'indice de la Gaussienne dans le GMM client, $p(\lambda|S)$ est la probabilité *a posteriori* du modèle λ sachant le score de vérification S, n_i est définie selon l'équation 3.8 et $O_i(X)$ est l'occupation cumulée sur le test X de la Gaussienne i.

Les auteurs font alors varier le poids d'adaptation MAP pour déterminer empiriquement le meilleur taux. Ils estiment aussi la dégradation des performances due à l'ajout de données imposteurs proches du locuteur.
Ceci souligne, que l'utilisation d'une mesure de confiance basée sur le score de vérification, ne peut être le meilleur moyen de sélectionner les données d'adaptation. Les imposteurs acceptés par le système de RAL seront donc toujours sélectionnés pour l'adaptation. La solution optimale serait d'utiliser un système présentant un taux de FA nul. Un tel système n'existe pas et, dans le cas où il existerait, un système d'adaptation serait inutile.

7.2.2 Evolution du seuil optimal de vérification

La quantité de données d'apprentissage et de test conditionnent le choix du seuil de décision et, de ce fait, du seuil de sélection. Les distributions des scores clients évoluent en fonction de la quantité de données d'apprentissage (cf. figure 7.3). La figure 7.3 présente la variation des moyennes des distributions de scores clients en fonction de la quantité d'accès sélectionnés pour l'adaptation. La moyenne de la distribution augmente. Les seuils optimaux, de décision de vérification et de sélection pour l'adaptation, évoluent. Les techniques de normalisation de scores, qui permettent de déterminer un seuil optimal pour chaque couple modèle de locuteur/accès, sont nécessaires (cf. section 3.4.2).

Pour éviter les problèmes de variations du seuil de sélection, le modèle d'apprentissage de référence est utilisé pour la décision de sélection. Le modèle adapté est utilisé pour la décision de vérification. Le seuil de vérification doit alors être redéterminé

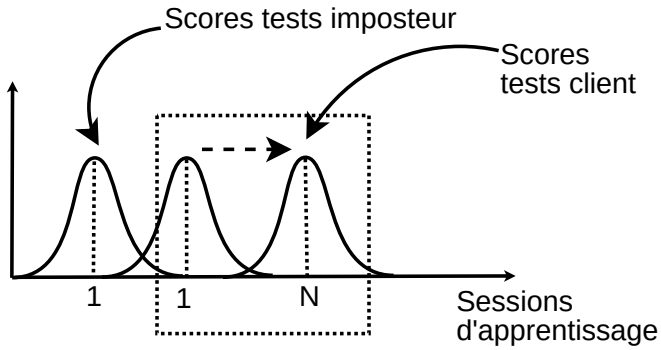


FIG. 7.3 – *Illustration du décalage dans les scores clients lorsque la quantité de données d'apprentissage augmente.*

empiriquement ; son estimation est alors liée à la corrélation entre la base de développement et la base de test. Ainsi [Van Leeuwen, 2004] démontre le problème de la calibration des seuils entre les bases de données NIST SRE 2002 et 2004. Il met aussi en évidence la relation entre les performances de l'adaptation et la quantité de test clients disponibles pour un locuteur. Par analogie à la T-norm Adaptative, [McLaren et al.,

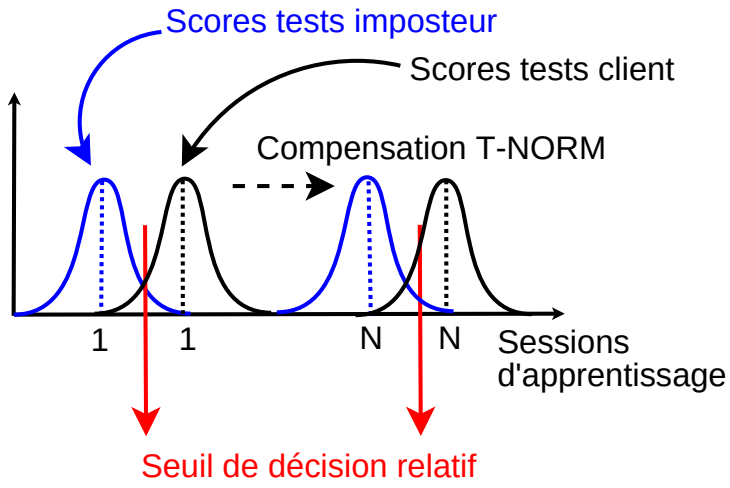


FIG. 7.4 – *Illustration de la compensation par T-norm.*

2008] propose d'utiliser la Z-normalisation adaptative.

Pour répondre au problème de variabilité de la distribution des scores client [Yin et al., 2006] propose d'utiliser la T-normalisation, afin d'obtenir une stabilité dans le

seuil de sélection et de vérification. La méthode, alors employée, est la compensation de l'ajout de données d'adaptation par l'ajout d'une quantité de données similaire dans les modèles imposteurs de la cohorte. L'auteur montre alors une stabilité dans la moyenne des scores client normalisés (cf. figure 7.4).

Le procédé de Z-normalisation des scores est défini comme la normalisation d'un score par la distribution des scores d'un cohorte de tests imposteur, calculée sur le modèle client concerné. La Z-normalisation Adaptative consiste en la ré-estimation de cette distribution sur le modèle client, une fois adapté, et pour chaque nouvelle adaptation (cf. figure 7.5). Cette méthode présente un avantage par rapport à la T-normalisation adaptative : il est plus aisé de réestimer la distribution des scores imposteurs sur le modèle client adapté, plutôt que d'adapter les modèles de la cohorte d'imposteurs. [Hansen et al., 2006] introduit une technique de compensation du seuil de sélection,

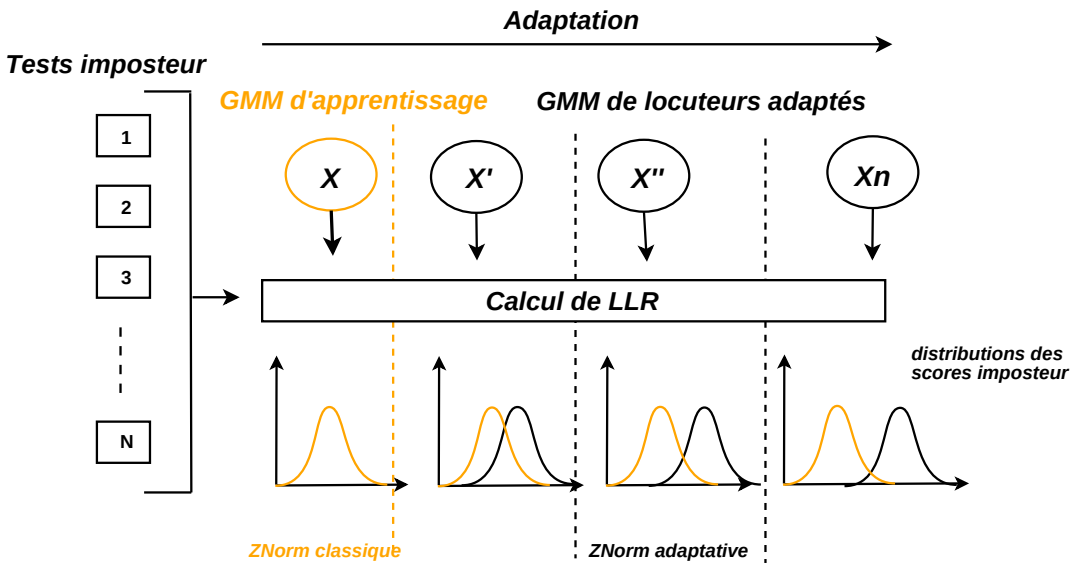


FIG. 7.5 – *Illustration du principe de la Z-normalisation adaptative.*

par la modification du *relevance factor* utilisé dans l'adaptation MAP. Le *relevance factor*, lié par un lien de proportionnalité avec le nombre de trames utilisées pour créer le modèle client, permet de réguler l'écart entre les moyennes des Gaussiennes du GMM du monde et du client. L'auteur propose alors de le modifier, en fonction du nombre de trames sélectionnées pour l'adaptation, pour engendrer le même écart qu'une seule session d'apprentissage. Des résultats très intéressants sont proposés avec des améliorations significatives en termes de performances.

7.2.3 Conclusion

Les méthodes état de l'art pour l'adaptation non supervisée, démontrent les difficultés de réalisation de système robuste. Ainsi, l'utilisation d'un seuil de sélection -approche majoritairement utilisée- fait apparaître la difficulté de calibration pour ne pas dégrader le système par l'ajout de données imposteur.
La mise à jour des modèles de locuteur entraîne aussi des variations dans la distribution des scores de vérification. La calibration des seuils de vérification et de sélection devient alors problématique.
Les méthodes utilisant des mesures de confiance ou poids d'adaptation se servent de paramètres à déterminer empiriquement. La variation entre les bases de données de développement et de test, explique les variations dans les performances de ces techniques.
Enfin, la sélectivité de l'adaptation est difficile à calibrer. Si elle est trop sélective, peu de données sont sélectionnées, et si elle est trop faible, des données imposteurs vont faire diverger les modèles. Ces approches utilisent donc un seuil de sélection à un point de fonctionnement qui minimise les FA. Nous pouvons remettre en question ce choix, étant donné qu'il est admis que très peu de données sont sélectionnées.

7.3 Utilisation de mesures de confiance pour une adaptation continue non supervisée des modèles de locuteur

Nous présentons, dans cette partie, une nouvelle méthode concernant l'adaptation non supervisée des modèles de locuteur.

7.3.1 Motivations

L'utilisation de la méthode d'adaptation non supervisée est motivée par les importants gains en performances observés lors de l'utilisation de la méthode d'adaptation supervisée. Néanmoins, il reste des problèmes à adresser pour proposer une technique d'adaptation robuste. Nous attacherons une importance particulière à déterminer une méthode permettant de s'affranchir d'un seuil de décision, à rendre la calibration robuste aux variations entre les données de développement et les données d'évaluation et, enfin, à sélectionner des données pertinentes d'adaptation.
L'idée directrice est de définir un système d'adaptation automatique, sans seuil, qui utilise les tests clients pour lesquels la décision de vérification n'est pas évidente. Nous émettons l'hypothèse que les tests clients pour lesquels le taux d'erreurs est maximum, sont les tests qui portent de l'information nouvelle sur le locuteur (cf. figure 7.6). Pour s'affranchir du problème de seuillage, nous proposons une solution automatique qui ne différencie pas l'appartenance des tests. Basée sur le principe des mesures de confiance, introduit en section 7.2, cette technique associe un poids d'adaptation à chaque test, en fonction de sa probabilité *a posteriori* d'appartenir au locuteur.

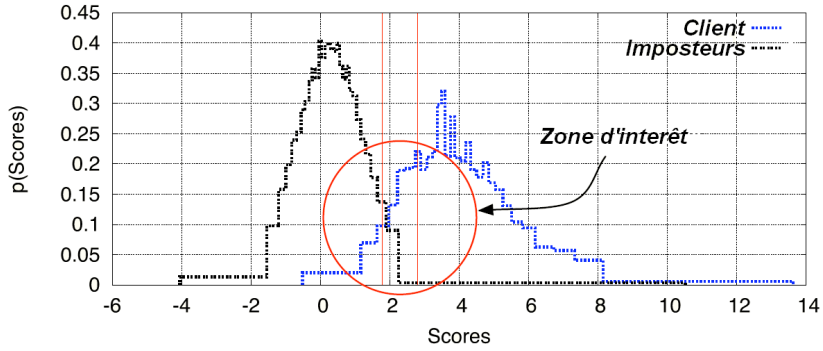


FIG. 7.6 – *Distributions des scores T-normés client et imposteur et zone d'intérêt (NIST SRE 2005).*

Un premier travail a été l'utilisation d'une mesure de remplacement du LLR pour sélectionner les tests [Preti et Bonastre, 2006]. Nous avons choisi le *Generalized Likelihood Ratio* [Solomonoff et al., 1998]. La mesure GLR est définie comme :

$$GLR(X_0, X_1) = \frac{p(X_1|M(X_1))p(X_2|M(X_2))}{p((X_1X_2)|M(X_1X_2))} \tag{7.4}$$

où X_1 etX_2 sont les segments de parole à comparer, $M(X_i)$ le GMM généré par ML à partir des données X_i et $M(X_1X_2)$ le GMM généré par ML avec les deux segments de parole X_1 et X_2.

Cette mesure a un pouvoir de classification inférieur au LLR (cf. table 7.1). Nous espérions alors que la mesure GLR apporterait une information nouvelle, décorrélée du score de vérification classique (LLR).

Mesure	**DCF**	**EER**
système de référence (LLR)	4.92	9.67
système de référence (GLR)	7.20	19.82
système de adapté (LLR)	4.80	9.83
système de adapté (GLR)	4.79	9.34

TAB. 7.1 – *Mesure LLR comparée à la mesure LLR pour le système de référence et l'adaptation non supervisée (NIST SRE 2005, LIA06-tnorm (128 GD)). Le seuil utilisé pour le choix des tests est le seuil optimal qui minimise la DCF, calculé a posteriori.*

Le tableau 7.1 présente les résultats d'expériences de RAL basées sur les mesures LLR et GLR pour le système de référence (non adapté) et le système d'adaptation non supervisée, sur la base NIST SRE 2005, avec le système LIA06-tnorm modifié avec l'utilisation de 128 composantes[1]. Le seuil pour le choix des tests utilisés dans l'adaptation est le seuil optimal qui minimise le critère *DCF*, calculé *a posteriori*. Ces résultats démontrent le pouvoir de classification bien inférieur de la mesure GLR pour la RAL, avec

[1]Pour limiter les temps de calcul

un EER de 19.82% contre 9.67% pour la mesure LLR. Au niveau de l'utilisation des mesures LLR et GLR pour l'adaptation non supervisée, il apparaît clairement que aucun gain significatif n'est apporté. Ces expériences prouvent que les méthodes d'adaptation non supervisée, basées sur un seuil de sélection, ne collectent que peu de données pertinentes pour améliorer la modélisation des locuteurs.

7.3.2 Une nouvelle mesure de confiance pour une adaptation continue sans seuil

Nous introduisons une solution pour l'adaptation non supervisée, indépendante d'un seuil de décision [Preti, 2007; Preti et al., 2006, 2007]. Nous proposons de pondérer les données d'adaptation par un poids, fonction de leur probabilité *a posteriori* d'appartenir à la classe client. Ici aucune différenciation entre classe d'appartenance des données n'est établie, *i.e.*, toutes les données, y compris les données imposteur sont utilisées pour l'adaptation. L'adaptation est continue.
Le principe de cette méthode d'adaptation se base sur le fait que les données imposteur seront associées à une faible mesure de confiance, et donc, n'auront aucune influence lors de l'adaptation.
De plus, il est supposé que des informations relatives au canal de transmission seront « captées » par l'adaptation, pour des valeurs de mesure de confiance faibles ou moyennes.

7.3.2.1 La mesure de confiance

La probabilité *a posteriori* ou score du test client est transformée en un poids d'adaptation par la fonction *World Maximum A Posteriori* (WMAP). A l'origine WMAP a été défini pour la normalisation dans l'espace des scores (cf. section 3.4.2.3). Il s'agit d'un détecteur Bayésien bi-classe (client : x ; imposteur : y) qui introduit les probabilités *a priori* des classes ($P(x = y)$ et $P(x \neq y)$) .
La fonction WMAP calcule la vraisemblance du score s d'appartenir à la classe client puis sa vraisemblance d'appartenir à la classe imposteur. Ces vraisemblances sont transformées en une probabilité par l'ajout des probabilité $apriori$ des classes dans le calcul. Ainsi, pour un score s donné en entrée de la fonction WMAP, la probabilité de ce score d'appartenir à la classe client, $P(x = y|s)$, est obtenue en sortie (cf. equation 3.17). Cette probabilité est ensuite utilisée comme poids lors de l'adaptation du modèle client (cf. figure 7.7). Les variations de la courbe WMAP en fonction des scores sont définies comme suit :

- le poids est nul lorsque le score de vérification est proche de la moyenne de la distribution des scores imposteurs, dans l'intervalle [-0.5 ;0] sur la figure 7.7 ;
- le poids est proche de 1 sur l'intervalle [0.2 ;1] ; le score est alors proche de la moyenne de la distribution de scores client ;
- entre ces deux intervalles, la mesure de confiance augmente à mesure que le score se rapproche de la moyenne cliente ;

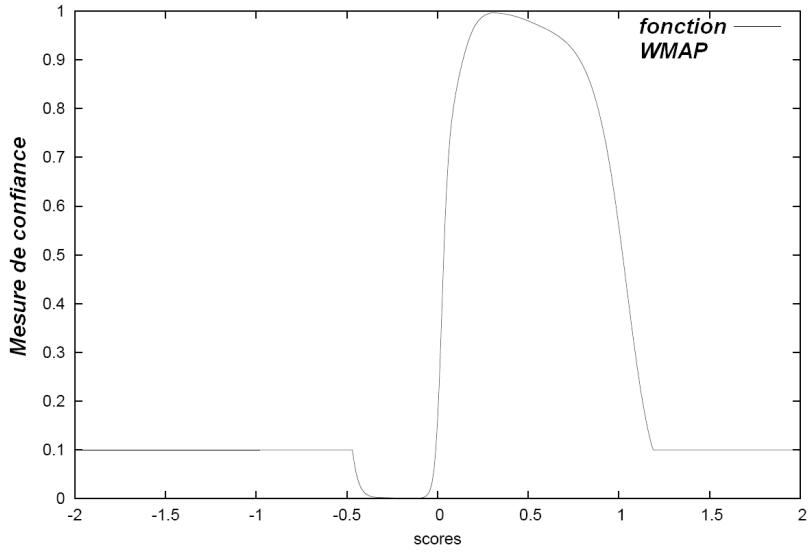


FIG. 7.7 – *Courbe WMAP calculée à partir de scores de l'évaluation NIST SRE 2005.*

– le poids égale la probabilité *a priori* de la classe «score client», ici 0.1, pour les scores non observés dans les données de développement (intervalles [-2 ;-0.5] et [1.2 ;2]).

Cette mesure de confiance affecte un poids nul pour les tests imposteur et un poids élevé pour les tests client. Lorsque le poids est de 0.5, le test associé peut être considéré comme client et imposteur. Deux cas de figures sont envisagés :

– ce test est client : il est très utile pour améliorer la couverture du modèle client,
– ce test est imposteur : les informations de sessions peuvent améliorer le modèle client.

De plus pour des taux d'erreurs du système de RAL très faibles, la fonction WMAP est assimilée à un seuil. La pente de la courbe est très forte, le poids WMAP devient presque une décision binaire 0 ou 1. Les taux d'erreurs de notre système de référence ne nous ont pas encore permis de valider cette remarque.

7.3.2.2 Mise en oeuvre de la fonction WMAP

[Fredouille, 2000] a souligné dans ses travaux, l'importance de la modélisation des classes «score client» et «score imposteur», dans la réalisation de la fonction WMAP. Cette phase consiste en l'utilisation de données de développements. Les scores client et imposteur, récoltés sur la base de développement, sont utilisés comme évènements des deux classes. La base de développement doit être très corrélée avec la base d'évaluation.

L'estimation des distributions de LLR est confrontée au problème de la longueur des signaux de test, en nombre de trames, qui peut varier d'un test à l'autre. La normalisation des LLR, par le nombre de trames du test, ne permet pas de diminuer suf-

fisamment la variance des LLR moyens [Fredouille, 2000]. En conséquence, la variance des distributions des scores est introduite dans la modélisation de ces distributions. Dans [Fredouille, 2000], les auteurs utilisent des LLR calculés sur de court segments de parole, de durée constante, pour éviter les variations des LLR. La variance des LLR, calculée sur de courts segments de parole, est grande et, de plus, le taux d'erreur d'un système de RAL dépend de la durée du test. Pour obtenir une bonne estimation des classes client et imposteur, il faut utiliser un nombre de trames suffisant pour le calcul du LLR.
La plus petite unité que nous utilisons dans ce travail, pour modéliser les distributions de scores, est la durée d'un test. Le plus souvent notre méthode est évaluée sur les bases NIST SRE. La durée de test est de 2 minutes 30 secondes. Chaque score utilisé pour modéliser ces distributions a été calculé sur cette durée[2]. Nous représentons les classes des scores client et des scores imposteur, pour les calculs de $p(s|x = y)$ et $p(s|x \neq y)$, par des GMM. 6 composantes sont utilisées.

7.3.2.3 La fonction d'adaptation

A chaque nouvel accès, le modèle client est recalculé avec les données d'apprentissage ainsi que les données des accès, successivement sélectionnés, et leurs poids respectifs σ. La fonction d'adaptation proposée est basée sur l'adaptation MAP, où seulement les moyennes du GMM sont adaptées. Les statistiques du modèle client, initialisées avec le modèle du monde, sont alors calculées par l'algorithme EM par maximum de vraisemblance (cf. section 3.3.1). Les statistiques d'occupations (cf. equation 3.8) sont alors modifiées selon l' équation suivante :

$$n_i = \sum_{t=1}^{T} Pr(i|\vec{x_t}, \lambda)\sigma_{\vec{x_T}} \tag{7.5}$$

$\sigma_{\vec{x_T}}$ est le poids d'adaptation associé au test $\vec{x_T}$ et où $Pr(i|\vec{x_t}, \lambda)$ est la probabilité que la Gaussienne i, du GMM λ, ait généré le vecteur $\vec{x_t}$. Les probabilités d'occupation pour chaque Gaussienne sont pondérées par le poids σ associé à chaque trame. Pour les données d'apprentissage, il est fixé à 1. Les tests, qui ont une faible mesure de confiance σ, sont aussi utilisés pour adapter le modèle. Ensuite, les moyennes adaptées pour chaque Gaussienne sont calculées par adaptation MAP des moyennes empiriques et des moyennes du GMM du modèle du monde, selon la formule MAP 3.6. Le *relevance factor* est fixé à 14.

7.3.3 Évaluation expérimentale de l'approche

Nous évaluons notre approche d'adaptation sur les différentes bases de données des évaluations internationales NIST SRE 2005, 2006 et 2008 (cf. annexe A). Le tableau 7.2 illustre les performances de l'adaptation, sur les bases de données NIST SRE 2005, 2006

[2]Les variations du LLR dues aux nombres de trames sélectionnées par la DAV ne sont pas compensées.

Evaluation	DCF	EER
NIST SRE 2005	-27%	-37%
NIST SRE 2006	+123%	+69%
NIST SRE 2008	+18%	+12%

TAB. 7.2 – *Tableau de performances moyennes relatives de l'adaptation non supervisée par rapport au système de référence. Evaluations sur les bases NIST SRE 2005, 2006 et 2008. Le pourcentage de gain/perte relatif est proposé (négatif pour un gain, positif pour une perte).*

et 2008 relativement au système de référence. Les performances sont exposées en terme de gain ou de perte relative (négatif pour un gain, positif pour une perte). Il apparaît clairement que cette méthode présente de très nettes différences de performances, en fonction de l'évaluation NIST SRE considérée.

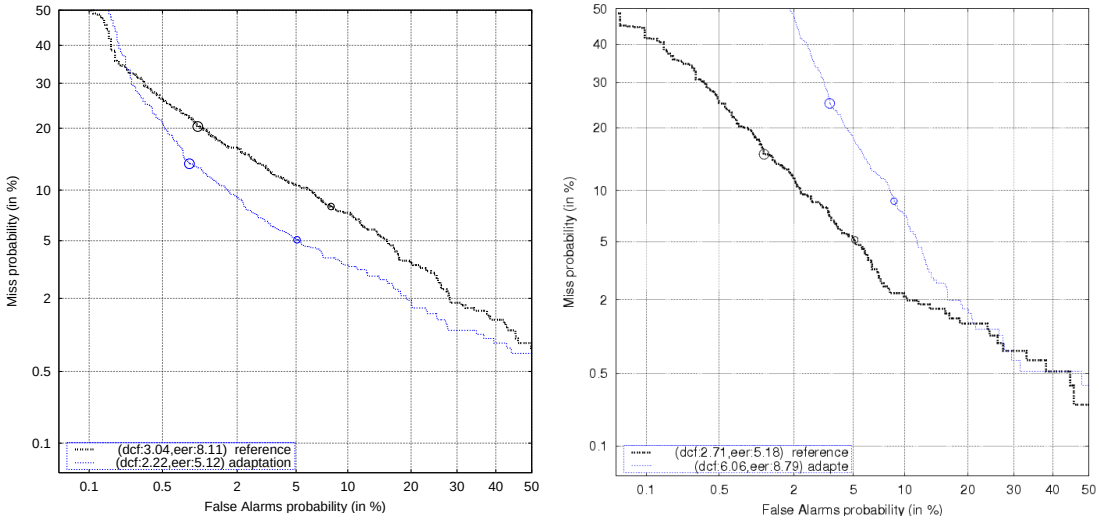


FIG. 7.8 – *Courbe DET pour le système de référence et le système adapté : a) NIST SRE 2005, LIA06-tnorm ; b) NIST SRE 2006, LIA06-tnorm*

Le tableau 7.3 présente les bases de développements utilisées en fonction des bases de test considérées. Sur la base de données NIST SRE 2005, le gain s'élève à 27% pour la mesure DCF et 37% en terme d'EER (cf. figure 7.8 *a*). Une importante perte est observée sur la base NIST SRE 2006, jusqu'à 123 % relatifs pour la mesure DCF (cf. figure 7.8 *b*). Sur la base NIST SRE 2008, la perte est moins prononcée, avec 18% pour la mesure DCF.

7.3.4 Conclusion

La méthode d'adaptation non supervisée que nous proposons a démontré des résultats très positifs sur la base de données NIST SRE 2005, avec une amélioration de 27% en termes de DCF et 37% en termes d'EER. Sur les bases de données NIST SRE 2006 et

Estimation WMAP	Base de test
NIST SRE 2005	NIST SRE 2006
NIST SRE 2006	NIST SRE 2005
NIST SRE 2008	NIST SRE 2005

TAB. 7.3 – *Base de développement pour l'estimation de la fonction WMAP en fonction de la base de test considérée.*

2008, la méthode altère les performances du système de référence. Nous émettons différentes hypothèses pour expliquer les variations de performances de notre méthode :

1. une mauvaise estimation de la mesure WMAP ; la corrélation entre les bases de développement et de test est importante pour le calcul de la fonction WMAP [Fredouille, 2000],
2. la méthode n'est pas robuste aux quantités de tests client et imposteur, notamment aux taux de fausses acceptations,
3. les variations du seuil optimal de décision, pendant l'adaptation, sont trop importantes.

Une analyse détaillée de la méthode d'adaptation fait l'objet du chapitre suivant. Ce chapitre est notamment consacré à déterminer la validité de ces hypothèses.

Chapitre 8

Analyse détaillée et améliorations de la méthode d'adaptation non supervisée proposée

Sommaire

Ce chapitre est dédié à l'analyse de la méthode d'adaptation originale que nous proposons. Nous décrivons les résultats de la méthode sur les bases NIST SRE 2005 et 2006. Nous illustrons les performances de cette méthode selon différents protocoles. De fortes disparités, en terme de performances, sont remarquées suivant la base de données et le protocole utilisés. Ces résultats nous ont amené à préciser certains facteurs pouvant expliquer de tels écarts de performances :

1. l'estimation de la mesure de confiance,
2. le rapport du nombre de tests client sur imposteur,
3. le taux de fausses acceptations.

Nous proposons également des solutions pour améliorer notre méthode d'adaptation avec, notamment, l'utilisation de normalisations de scores adaptatives pour contrôler les variations du seuil de décision.

8.1 Résultats sur la base NIST SRE 2005

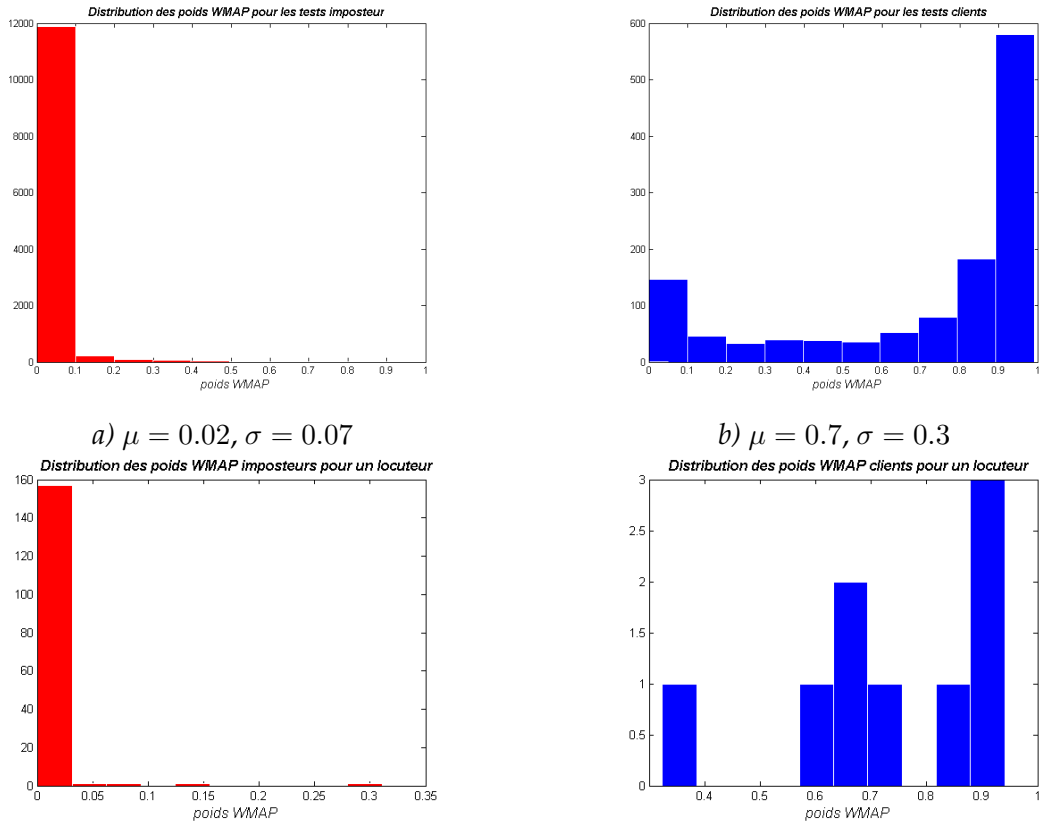
Dans un premier temps, nous nous focalisons sur les performances de l'adaptation sur la base de données NIST SRE 2005. Le but de ce paragraphe est de déterminer les causes qui affectent le comportement du système d'adaptation proposé, en utilisant un contexte favorable (NIST SRE 2005) pour élucider les problèmes intervenant sur la base NIST SRE 2006. Les premiers résultats, présentés dans la figure 7.8 *a*, illustrent les performances de la méthode d'adaptation sur la base de données NIST SRE 2005. Le gain relatif apporté par l'adaptation est de 27% pour la mesure DCF et 37% pour la mesure EER, comparé à la référence. La fonction WMAP est calculée à partir :

- des scores T-normés de la base NIST SRE 2006,
- des probabilités *a priori* fixées par le NIST : $P(client) = 0.1$ et $P(imposteur) = 0.9$,
- du système GMM-UBM de référence LIA06-tnorm.

8.1.1 Utilisation de la zone d'intérêt pour l'adaptation

Nous illustrons l'influence du poids d'adaptation, pour valider l'hypothèse de l'utilisation de la zone d'intérêt (cf. figure 7.6). Les histogrammes, globaux et pour un locuteur choisi, du poids WMAP pour les accès imposteur et client sont illustrés dans les figures 8.1 a, b, c et d.

Nous pouvons observer que les poids associés à des tests imposteurs (cf. 8.1 *a* et *c*) ont majoritairement des valeurs très faibles, entre 0 et 0.1. En revanche, les poids associés à des tests clients (cf. 8.1 *b* et *d*) sont répartis sur l'intervalle [0 ;1]. Ceci illustre le bien fondé de notre hypothèse. De nombreux tests client, associés à des poids WMAP moyens, vont permettre à l'adaptation de collecter de nouvelles données déterminantes. Rappelons que cette caractéristique est liée à notre hypothèse : un système classique d'adaptation non supervisée, basé sur un seuil, n'utilise pas ces données.



a) $\mu = 0.02, \sigma = 0.07$ *b)* $\mu = 0.7, \sigma = 0.3$

c) $\mu = 0.007, \sigma = 0.027$ *d)* $\mu = 0.73, \sigma = 0.2$

FIG. 8.1 – *Histogrammes des poids WMAP pour a) les accès imposteur b) les accès client c) les accès imposteur pour 1 locuteur d) les accès client pour 1 locuteur (NIST SRE 2005). Les paramètres de moyenne et variance des poids WMAP sont présentés à titre indicatif.*

Le choix d'adapter avec tous les tests, sans discernement, peut paraitre osé. Nous décrivons quelques expériences pour démontrer qu'utiliser des tests, avec des mesures de confiance moyennes ou petites, est bénéfique pour la modélisation d'un locuteur. Nous mettons en oeuvre des expériences sur la base NIST SRE 2005, en utilisant notre méthode d'adaptation mais en la contraignant au niveau de l'utilisation des mesures de confiance :

- en utilisant que les poids supérieurs à 0.8,
- en utilisant les poids compris entre 0.1 et 0.5

Nous proposons, à des fins de comparaison, deux expériences utilisant un mode « Oracle » (adaptation supervisée). Elles sont différenciées par le poids appliqué aux tests client. Dans un premier cas, le poids d'adaptation est fixé à 1 (Oracle standard). Dans un deuxième cas il est fixé par la fonction WMAP (oracle WMAP). Ceci permet d'évaluer si les poids WMAP des accès clients sont proches de l'optimal. Le tableau 8.1 présente les résultats des expériences sur la base de données NIST SRE 2005, selon les différentes contraintes appliquées au poids d'adaptation.
L'expérience qui n'utilise que les poids d'adaptation compris entre 0.1 et 0.5 apporte un

gain relatif de 20% pour la mesure EER par rapport à la référence. L'expérience n'utilisant que les poids supérieurs à 0.8 (ce qui est très proche de l'utilisation d'un seuil de sélection) apporte, quant à elle, un gain de seulement 14% relatif pour la mesure EER. Il apparaît que le seuillage des poids d'adaptation apporte un gain pour toutes les configurations. Ces résultats valident l'hypothèse que de l'information utile est contenue dans les tests associés à une mesure de confiance de poids faibles ou moyens.
En ce qui concerne les performances des modes « Oracle » (Oracle classique et Oracle WMAP), les résultats relevés sont très proches. Les différences en termes de DCF et d'EER pour l'Oracle WMAP sont respectivement de 1% et 11% relativement à l'Oracle classique. Nous pouvons en conclure que les poids d'adaptation sont proches de l'optimal, lorsqu'ils sont calculés par la fonction WMAP.

Expérience	**DCF**	**EER**
Référence	3.04	8.11
$0.1 < w < 0.5$	3.08	6.50
$w > 0.8$	2.58	6.98
Pas de contrainte sur w (*)	2.22	5.12
Oracle *WMAP*	1.72	4.40
Oracle classique	1.70	3.97

TAB. *8.1* – *Tableau de performances de l'adaptation non supervisée selon différentes contraintes sur le poids d'adaptation (w). (*) dénote la méthode d'adaptation proposée.*

8.1.2 Chronologie des tests et robustesse face aux accès imposteurs

La chronologie des tests peut également jouer un rôle important dans la robustesse de notre approche. Si une grande quantité de tests client est présentée au début de l'adaptation, les modèles de locuteurs sont très vite robuste aux accès imposteurs. Dans le cas contraire, si la majeure partie des tests présentés en premier sont des tests imposteurs, les modèles de locuteurs vont être dégradés. Nous proposons ensuite d'évaluer cet aspect à l'aide de deux nouvelles expériences :

- la première pour évaluer si l'ordre chronologique de présentation des tests n'influence pas le résultat final,
- la seconde pour évaluer le comportement du système, lorsque aucun test client n'est présenté.

Ces résultats sont décrits dans la table 8.2. Le changement d'ordre de présentation des

Expérience	**DCF**	**EER**
Tirage aléatoire des tests	2.35	6.01
Pas de test client	4.37	10.81
Pas de contrainte sur w (*)	2.22	5.12

TAB. *8.2* – *Tableau de performances de l'adaptation non supervisée selon différentes contraintes sur la nature des tests. (*) dénote la méthode d'adaptation proposée.*

tests influence peu le gain de la méthode. Les mesures EER et DCF présentent un gain

de respectivement 22% et 26% relatifs, par rapport à la référence. A l'inverse, lorsque uniquement des tests imposteur sont présentés, les performances se dégradent fortement et ceci bien que, en moyenne, les poids soient faibles pour les tests imposteurs. Cette dégradation s'explique aussi par les fausses acceptations, qui constituent des données d'adaptation imposteur associées à des poids élevés. Il apparaît que l'influence des tests imposteurs n'est pas nulle dans ce schéma d'adaptation, néanmoins, le taux de fausses acceptations est un problème pour tous les systèmes d'adaptation. Nous pouvons désormais éliminer certains facteurs qui pourraient avoir eu une influence sur les différences de résultats entre les bases NIST SRE 2005 et 2006 :

- l'ordre des tests ;
- une mauvaise estimation des poids WMAP, pour les accès clients ;
- l'utilisation de l'ensemble des mesures de confiance, sans seuillage.

8.2 Etude des résultats sur la base NIST SRE 2006

Toutes les expériences présentées dans le paragraphe précédent, menées sur la base de données NIST SRE 2005, montrent que l'adaptation apporte un gain significatif. En revanche, les courbes DET illustrées dans la figure 7.8 exposent des performances nettement moins convaincantes sur la base NIST SRE 2006. La perte relative est alors de 123% pour la DCF et 69% pour l'EER, comparativement à la référence (sans adaptation).

8.2.1 Estimation des distributions de scores pour le calcul de la fonction WMAP

L'estimation des distributions de scores client et imposteur détermine la qualité de la fonction WMAP [Fredouille, 2000]. Ces distributions sont calculées globalement sur tous les scores d'une évaluation car le système GMM-UBM est relativement stable ; cependant il existe une variabilité des scores entre locuteurs. Pour évaluer si une différence existe entre les estimations de la fonction WMAP, nous calculons la fonction WMAP à partir des scores obtenus sur la base NIST SRE 2006 et la comparons avec la fonction WMAP, calculée sur les scores obtenus sur la base NIST SRE 2005. Ces courbes

Évaluation	moyenne	variance
NIST SRE 2005 scores client	7.61	3.35
NIST SRE 2008 scores client	7.48	3.17
NIST SRE 2005 scores imposteur	0.10	1.13
NIST SRE 2008 scores imposteur	0.34	1.26

TAB. 8.3 – *Moyenne et variance des distributions de scores obtenues sur les bases NIST SRE 2005 et 2008 (LIA-THL08, ztnorm).*

WMAP sont représentées sur la figure 8.2. Il apparaît que les scores du système de RAL sont très proches sur ces deux bases de données, menant à une estimation très peu différente des deux courbes. Nous précisons dans le tableau 8.3, les moyennes et variances

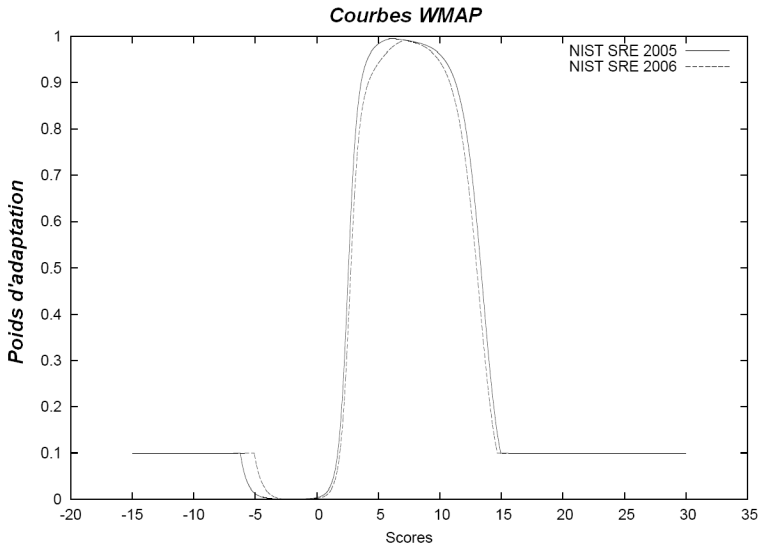


FIG. 8.2 – *Différences entre les courbes WMAP calculées à partir de scores des évaluations NIST SRE 2005 et 2006.*

des distributions de scores client et imposteur obtenues sur les bases NIST SRE 2005 et 2008. Il apparaît que les différences sont minimes. Les distributions de scores client et imposteur sont stables entre les différentes bases de données.

Ces observations démontrent que le changement de comportement ne s'explique pas par un décalage dans les distributions de scores, entre les bases de données.

8.2.2 Influence de la zone d'intérêt pour l'adaptation

Expérience	DCF	EER
Référence	2.71	5.18
Oracle *WMAP*	2.27	4.01
Oracle classique	1.85	3.58
Adaptation WMAP	6.06	8.79

TAB. 8.4 – *Tableau de performances de deux types d'adaptation supervisée sur la base NIST SRE 2006.*

Pour déterminer les causes de la perte de performance sur la base de données NIST SRE 2006, nous avons procédé aux mêmes types d'expériences d'adaptation supervisées que celles présentées dans le tableau 8.1. Les résultats sont présentés dans la table 8.4. La fonction WMAP, pour ces expériences, est calculée à partir :

- des scores T-normés de la base NIST SRE 2005,
- des probabilités *a priori* fixées par le NIST : $P(client) = 0.1$ et $P(imposteur) = 0.9$,
- du système GMM-UBM de référence LIA06-tnorm.

Les deux types « d'Oracle » apportent un gain conséquent, par rapport au système de référence sans adaptation. Toutefois, l'écart entre les deux types d'adaptation est plus élevé que sur la base de données NIST SRE 2005 (cf. table 8.1) : l'écart pour l'Oracle WMAP est de 1% pour la mesure DCF, et 11% pour la mesure EER sur la base NIST SRE 2005, contre 23% pour la mesure DCF et 12% pour la mesure EER, sur la base NIST SRE 2006. Nous pouvons en déduire que certaines mesures de confiance sont faibles pour des tests client.

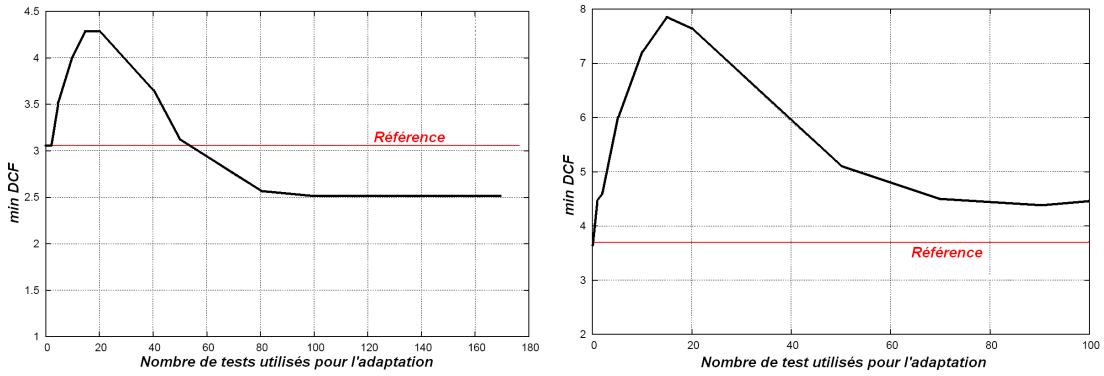
8.2.3 Evolution pas à pas des taux d'erreurs

Les dégradations engendrées par l'adaptation peuvent se caractériser par une augmentation du taux de fausses acceptations ou du taux de faux rejets. Pour caractériser l'évolution de ces taux pendant la phase d'adaptation, nous réévaluons les taux d'erreurs du système adapté après chaque adaptation : les scores de **toutes** les données de l'évaluation sont ainsi recalculés sur le modèle adapté. L'algorithme suivant décrit la procédure appliquée :

début
  pour tous les *les modèles de locuteur* j **faire**
    pour tous les *les tests* i *associés au locuteur* j **faire**
      Adapt(modèle(j), test(i), poids(w_i))
      pour tous les *les tests* i *associés au locuteur* j **faire**
        Score(test i | modèle j)
      fin
      Calcul EER et DCF
    fin
  fin
fin

Algorithme 1 : Evolution des taux d'erreurs pas à pas.

Dans les figures 8.3 *a* et *b* est représentée l'évolution des mesures DCF de ces configurations sur les bases NIST SRE 2005 et 2006. Pour les deux bases de données, nous constatons une phase de perte de performance, suivie d'une phase de gain. Il semble qu'une quantité minimale de tests client soit nécessaire pour compenser les erreurs d'adaptation, dues aux fausses acceptations. En termes de FA et FR, il semble que le taux de FA soit soumis à la plus forte variation. La figure 8.4 présente les courbes d'évolution du nombre de bonnes acceptations et de fausses acceptations, après chaque étape d'adaptation, pour le système adapté sur la base NIST SRE 2005. La plus forte variation est observée pour le taux de FA (NON). Ce taux ne cesse d'augmenter, jusqu'à la trentième itération. Le nombre d'accès client accepté reste stable. Après 30 itérations, le nombre d'accès client acceptés augmente et le nombre de FA diminue.
Nous pouvons en déduire que les modèles clients sont d'abord dégradés. Cela se traduit par un taux de reconnaissance des accès client non dégradé mais des accès imposteurs plus facilement acceptés. Après de nombreuses itérations, un nombre suffisant



a) *b)*

FIG. 8.3 – *Courbes de performances en terme de min DCF pour le système d'adaptation après N itérations pour (a) NIST SRE 2005 (b) NIST SRE 2006*

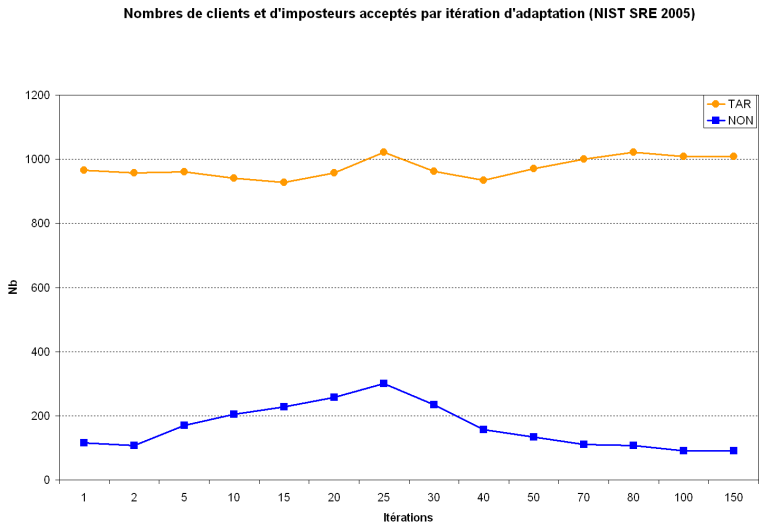


FIG. 8.4 – *Nombre de tests client et imposteur acceptés par le système de RAL pour par itération d'adaptation (NIST SRE 2005).*

d'accès clients a été sélectionné pour l'adaptation. Les modèles de locuteur sont plus robustes, le taux de bonnes acceptations augmente et le taux de FA diminue. Nous expliquons cette phase d'augmentation des FA à partir de deux hypothèses :

1. le taux de FA du système de référence amène des erreurs d'adaptation ; ceci est difficilement évitable et indépendant de la mesure de confiance ;
2. le rapport entre la quantité de tests client et la quantité de tests imposteur peut dégrader les performances : l'expérience d'adaptation avec seulement des accès imposteur le prouve (cf. table 8.2).

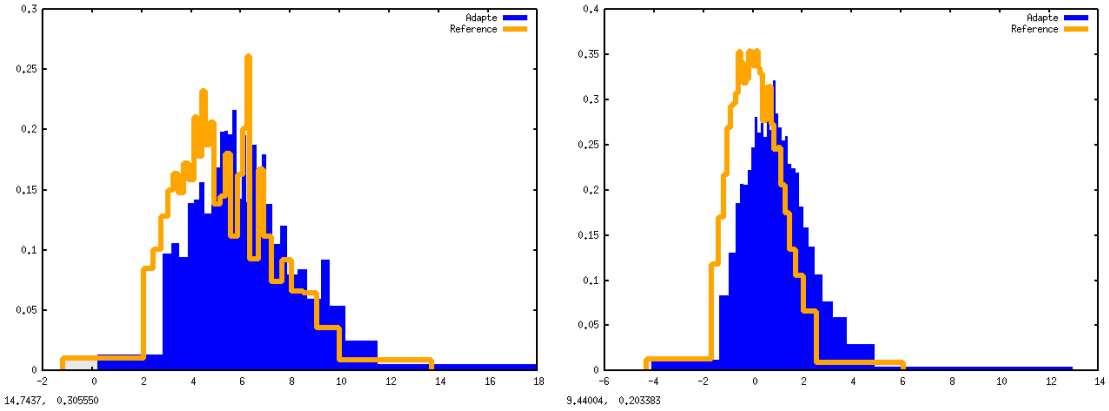
La table 8.5 présente les nombres d'accès client et imposteur potentiels, pouvant être utilisés pour l'adaptation sur les bases NIST SRE 2005 et 2006. Pour la comparai-

son, deux seuils ont été utilisés, le seuil qui minimise la DCF (OS) et le seuil à l'EER (EER score). Il ressort que sur la base NIST SRE 2005 beaucoup plus d'accès clients sont

Base	Seuil	Nb TA	Nb FA
NIST SRE 2005	OS	1172	699
	EER score	1209	2517
NIST SRE 2006	OS	621	97
	EER score	706	479

TAB. 8.5 – *Nombre d'accès potentiels pour l'adaptation en fonction de deux seuils : EER Score, le seuil de décision à l'EER et OS : le seuil de décision qui minimise la DCF.*

présents et acceptés. Ceci permet à la phase de « gain » de dépasser la référence. En revanche, sur la base NIST SRE 2006, la faible quantité de données client ne permet pas de compenser les pertes durant la phase de « gain ».
Les distributions de scores du système de référence, et du système adapté, sur la base NIST SRE 2006, sont représentées dans les figures 8.5 *a* et *b*. Pour le système adapté, les scores imposteur sont globalement supérieurs aux scores imposteurs du système de référence et les modèles clients sont dégradés et acceptent plus facilement des accès imposteurs.



a) *b)*

FIG. 8.5 – *a) Distributions des scores client pour le système de référence et le système adapté, b) Distributions des scores imposteur pour le système de référence et le système adapté. (NIST SRE 2006, LIA06-tnorm)*

Ces résultats nous amènent à approfondir deux facteurs quant à l'explication des variations de performances de la méthode d'adaptation :

1. le rapport entre la quantité de tests client et la quantité de tests imposteur,
2. l'influence du taux de FA du système de référence.

Nous détaillons ces points dans les paragraphes suivants.

8.3 Hypothèse n°1 : Influence du rapport du nombre de tests client sur le nombre de tests imposteur

Ce chapitre détaille notre première hypothèse : une grande quantité de tests imposteur engendre des erreurs d'adaptation ; une grande quantité de test client est alors nécessaire pour amener un gain.

8.3.1 Détails des bases de données NIST SRE

Dans ce paragraphe, nous mettons en évidence les différences, en termes de quantités de tests client et imposteur, entre les bases de données NIST.
Les tableaux 8.6 et 8.7 récapitulent les quantités de tests client, en fonction du nombre total de tests, et du nombre de locuteurs pour les bases de données NIST SRE 2005, 2006 et 2008. *Prior* client correspond à la probabilité *a priori* réelle d'observer un test client :

$$\begin{aligned} Prior(client) &= \frac{Nb(\text{tests client})}{Nb(\text{tests total})} \\ Nb_{NON} &= \text{le nombre d'accès imposteur} \\ Nb_{TAR} &= \text{le nombre d'accès client} \\ Nb_{loc} &= \text{le nombre de locuteurs} \end{aligned} \tag{8.1}$$

Evaluation	Nb_{TAR}	Nb_{NON}	Nb_{loc}	$\frac{Nb_{TAR}}{Nb_{loc}}$	***Prior client réelle***
NIST SRE 2008	439	6176	470	0.93	0.066
NIST SRE 2006	752	89479	219	3.43	0.077
NIST SRE 2005	1231	12393	264	4.67	0.090

TAB. 8.6 – *Tests clients et nombre de locuteurs (1/2).*

Information	**NIST SRE 2005**	**NIST SRE 2006**	**NIST SRE 2008**
Nb clients avec TAR	174	198	188
Nb TAR/loc	7.07	3.74	2.33
Nb clients sans TAR	90	21	282
NB IMP/loc	46	41	13

TAB. 8.7 – *Tests clients et nombre de locuteurs (2/2).*

La première différence flagrante, entre les bases de données, est la faible quantité de données clients sur la totalité de la base, pour les bases NIST SRE 2006 et 2008, comparativement à la base NIST SRE 2005. La probabilité *a priori* de la classe client varie entre les bases de données. Ceci est à prendre en compte dans le schéma de l'adaptation qui intègre cette probabilité au niveau du calcul de la mesure de confiance. Le tableau

8.7 précise quelques détails supplémentaires, concernant les bases de données le constat est encore plus explicite :

- la base de données 2005 propose en moyenne 7 tests client pour 174 locuteurs alors que ce taux est de 3.74 et 2.33 pour les bases NIST SRE 2006 et 2008 ;
- un nombre conséquent de locuteurs ne dispose pas de tests client pour l'adaptation dans les bases NIST 2006 et 2008 (282 pour NIST 2008 !) ;
- le nombre de tests imposteurs par locuteur est élevé sur les bases NIST SRE 2005 et 2006.

Nous pouvons déduire plusieurs conclusions de ces éléments :

- le nombre de tests clients par locuteur doit être élevé (2005) ;
- le nombre de tests imposteur par locuteur peut être élevé si suffisamment de tests clients sont disponibles (2005) ;
- peu de tests client et peu de tests imposteur (2008) = perte limitée ;
- trop d'accès imposteur (2006) amène des FA et des erreurs d'adaptation (cf. tableau 8.2) ;
- les probabilités *a priori* ne sont pas respectées lors des évaluations NIST SRE[1].

8.3.2 Modification de la base NIST SRE 2005 pour valider l'hypothèse 1

Pour évaluer le comportement de notre méthode d'adaptation, nous proposons des expériences sur la base NIST SRE 2005, en modifiant les taux de tests client et de tests imposteur. La proportion entre le nombre d'accès client et le nombre d'accès imposteur est presque la même dans les deux expériences. Cependant, une expérience présente des quantités équivalentes à celles observées sur la base NIST SRE 2006, et une autre expérience présente des quantités deux fois supérieures, similaires à celles observées sur NIST SRE 2005. Les résultats sont listés dans le tableau 8.8. Nous présentons deux expériences avec les mêmes *prior* client. La différence réside dans les quantités de tests client et imposteur (les accès imposteurs supplémentaires sont choisis dans la base NIST SRE 2006). Il apparaît alors que l'adaptation améliore les résultats de référence lorsque une grande quantité de tests client sont présents, même avec de nombreux tests imposteur (expérience 1, tableau 8.8). En effet, le gain relatif s'élève à 22% pour la DCF et l'EER. En revanche, avec un taux plus faible d'accès client et pour la même *prior*, l'adaptation dégrade les performances (expérience 2, tableau 8.8), avec 6% de perte relative pour la DCF et 12 % de perte relative pour l'EER.

Expérience	Nb_{TAR}	Nb_{NON}	**Prior client**	**Système**	**DCF**	**EER**
1	1231	17777	0.065	Référence	2.97	7.98
				Adapté	2.30	6.26
2	593	9186	0.061	Référence	2.55	7.08
				Adapté	2.7	7.98

TAB. 8.8 – *Expériences sur la base NIST SE 2005 modifiée.*

[1] Des expériences tentant de mettre à jour ces probabilités pendant l'adaptation n'ont pas été probantes.

8.4 Hypothèse n°2 : Influence du taux de fausses acceptations

Ce chapitre détaille notre seconde hypothèse : le taux de fausses acceptations du système engendre inévitablement des erreurs d'adaptation ; dans notre cas, des mesures de confiance élevées pour des accès imposteur.

8.4.1 Eviter les fausses acceptations

Le taux de fausses acceptations du système de sélection est le point déterminant des performances d'un système d'adaptation non supervisée. Si ce taux est élevé, des données imposteur sont sélectionnées et les modèles clients vont être dégradés.
[Mezaache et al., 2008] a démontré que les imposteurs qui ont un score de vérification élevé (FA) représentent la moitié des erreurs du système. Ils proposent une méthode, pour diminuer ce taux d'erreurs, basée sur le système *reverse* (cf. figure 8.6).

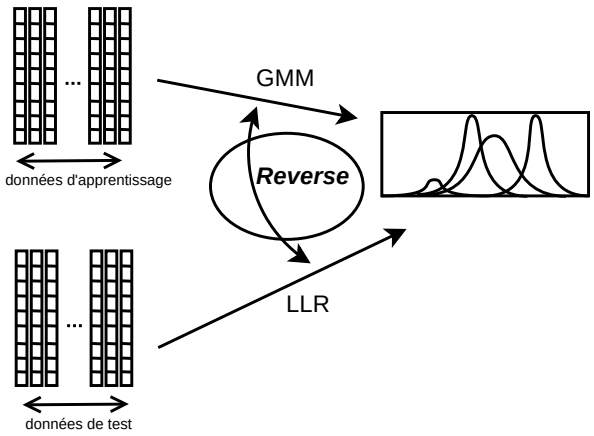
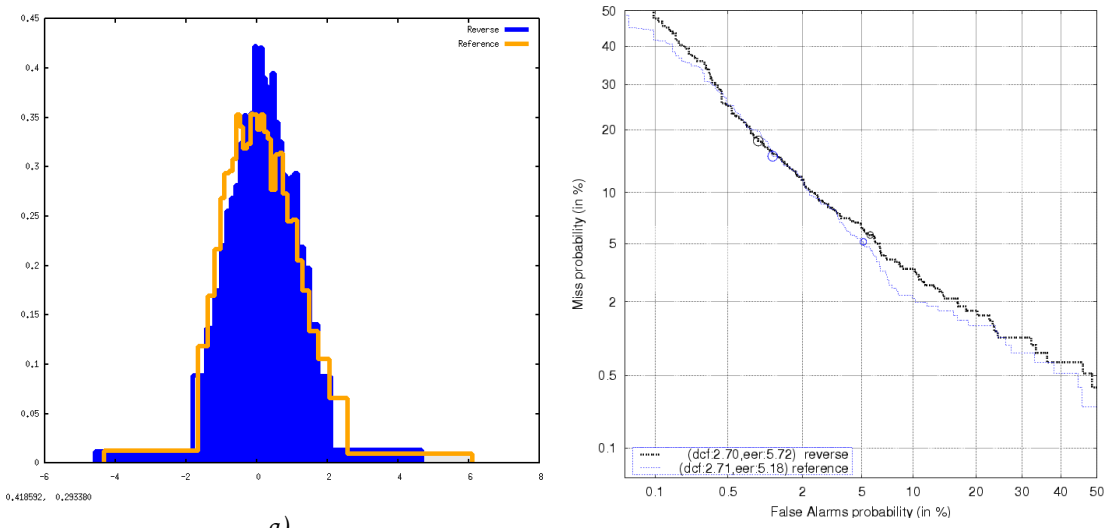


FIG. 8.6 – *Illustration du principe de fonctionnement du système reverse.*

Le système GMM-UBM utilisé en RAL n'est pas symétrique. Cela signifie que les scores de vérification du système ne sont pas les mêmes selon que le LLR soit calculé de manière classique, avec un modèle issu de l'enregistrement de référence et le test considéré, ou en mode *reverse* dans lequel le modèle est appris sur l'enregistrement de test et le LLR calculé en considérant l'enregistrement de référence comme test. Les auteurs ont démontré, qu'utiliser le système « à l'envers » sur les tests qui posent problème, permet de diminuer le taux de fausses acceptations. Cette méthode est efficace lorsque les données de test proposent une meilleure représentation du locuteur et des conditions d'enregistrement que les données d'apprentissage.

La figure 8.7 *a* présente les distributions des scores imposteurs du système de référence et du système *reverse* sur la base de données NIST SRE 2006. Les courbes DET correspondant à ces expériences sont illustrées dans la figure 8.7 *b*). Nous pouvons constater que les performances des deux systèmes sont très proches, cependant le nombre de scores client supérieurs à 2 est plus élevé pour le système de référence. Dans certains cas, si ces données ne sont pas assez spécifiques du locuteur, le GMM du locuteur est proche du modèle du monde. Les tests imposteur sont alors plus facilement acceptés.



a) *b)*

FIG. 8.7 – *a) Distributions des scores imposteur pour le système de référence et le système reverse, b) Courbes DET du systèmes de référence et du système reverse. (NIST SRE 2006, LIA06-tnorm)*

8.4.2 Diminution des mesures de confiance pour diminuer l'influence du test dans l'adaptation

Toujours dans l'optique de réduire le poids d'adaptation des tests imposteur, nous avons ajouté une contrainte supplémentaire dans le schéma d'adaptation.
Les probabilités *a priori* des classes client et imposteur interviennent dans le calcul de la fonction WMAP (cf. section 3.4.2.3). Ces probabilités *a priori* sont très différentes entre les bases NIST. Nous envisageons d'utiliser une probabilité *a priori* de la classe client très faible. Les nouvelles probabilités *a priori* sont désormais : $P_{tar} = 0.01$ et $P_{imp} = 0.99$.
Ce changement implique un décalage de la courbe WMAP (cf. figure 8.8). Il apparaît alors, que le poids d'adaptation atteint son maximum pour des scores plus élevés (cf. figure 8.8). Ceci résulte en une diminution des poids associés aux tests client lors de l'adaptation, mais aussi en une diminution des poids associés aux tests imposteur. Les figures 8.9 *a* à *d* illustrent les histrogrammes des poids WMAP avec et sans l'utilisation de la combinaison. Cette diminution globale peut alors réduire l'influence des tests imposteur dans la modélisation des modèles clients.

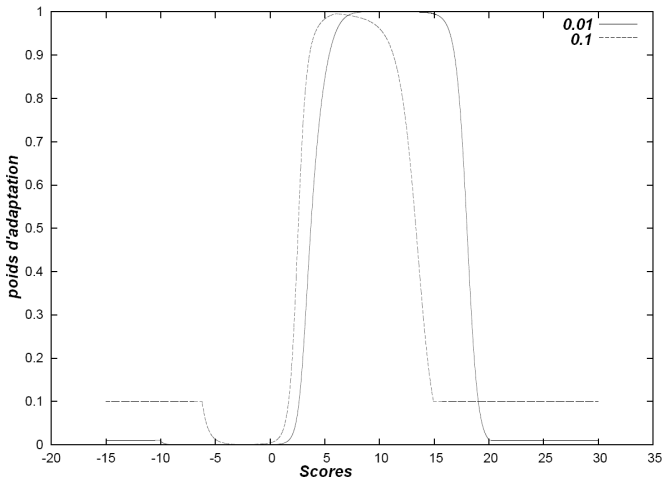


FIG. 8.8 – *Courbes WMAP pour deux probabilités a priori différentes.*

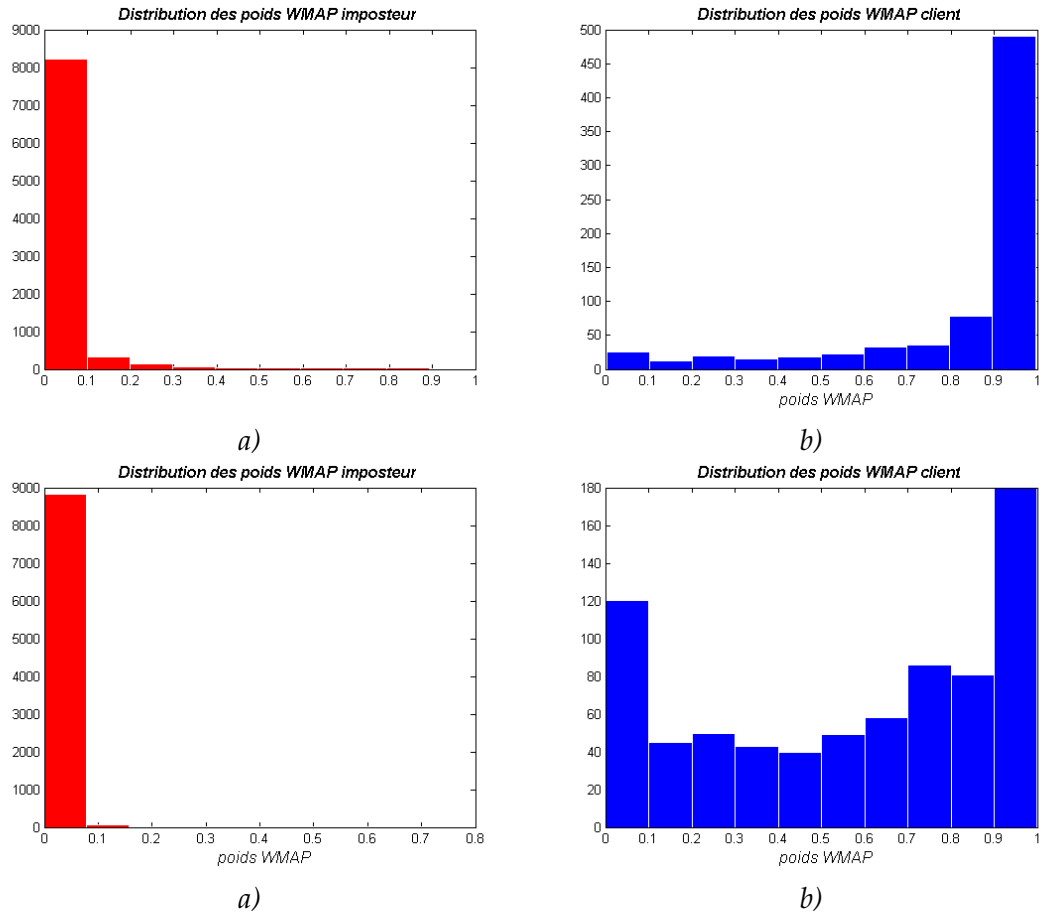


FIG. 8.9 – *a) Histogrammes des poids WMAP pour le système adapté (NON : a, TAR : b) et le système adapté reverse (NON : c, TAR : d) (NIST SRE 2006, LIA06-tnorm).*

8.4.3 Combinaison du *reverse* et du changement de *prior*

Nous évaluons dans ce paragraphe, une combinaison des deux méthodes citées auparavant, pour réduire l'influence des tests imposteur lors de l'adaptation des modèles de locuteurs. Le système reverse est utilisé pour réduire le taux de fausses acceptations. La mesure de confiance est alors calculée à partir du LLR issu du système *reverse*. La réduction de la *prior* client est utilisée pour diminuer les mesures de confiance des accès imposteur. Les résultats de la combinaison de ces deux méthodes sont présentés dans la table 8.9. La combinaison permet de réduire la perte sur NIST SRE 2006 et d'apporter

Evaluation	**DCF**	**EER**
NIST SRE 2006	+123%	+69%
NIST SRE 2008	+18%	+12%
NIST SRE 2006 + Combinaison	+7%	-21%
NIST SRE 2008 + Combinaison	+5%	$<$ +1%

TAB. *8.9* – *Influence de la décision croisée et de la modification des probabilités a priori. Pourcentage de différence relative avec le système de référence sans adaptation.*

un gain relatif de 21% au niveau de la mesure EER.
Sur la base NIST SRE 2008, cette modification permet de conserver les résultats du système de référence sans dégradation.
Les résultats à la DCF, sur ces bases, montrent que l'adaptation non supervisée mène à un taux de fausses acceptations supérieur à la référence, même lorsque les stratégies de décision *reverse* et de pondération spécifique sont utilisées.

Système	**FA**
Référence	106
Reverse	81
Adapté	319
adapté + Combinaison	136

TAB. *8.10* – *Nombres d'imposteurs acceptés (NIST SRE 2006, LIA06-tnorm), calcul basé sur le seuil qui minimise le critère DCF.*

Le tableau 8.10 présente les taux de FA pour les systèmes de référence, *reverse*, adapté et adapté avec combinaison (le seuil choisi minimise le critère DCF). Le plus faible taux de FA, ainsi que la baisse des poids due au changement des *priors*, pour le système *reverse* adapté, permet de réduire les erreurs d'adaptation (poids WMAP élevé pour des accès imposteurs).

8.5 Stabilité du seuil de décision

Le seuil de décision est très lié à la quantité de données utilisée pour l'apprentissage. Dans un schéma d'adaptation non supervisée, la chronologie des tests peut amener des

variations rapides de ce seuil. Par exemple, si les premières données d'adaptation sont des accès clients, la quantité de données utilisée pour adapter le modèle client va être rapidement importante. Le seuil va alors varier fortement.
Au démarrage, seule la session d'apprentissage est disponible. Le seuil optimal de dé-

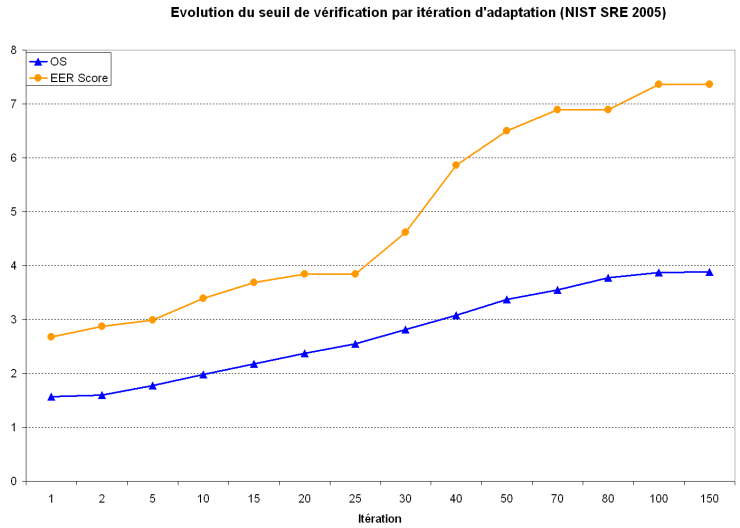


FIG. *8.10* – *Evolution des seuils de vérification à la DCF et l'EER par itération d'adaptation (NIST SRE 2005, LIA06-tnorm). OS est le seuil qui minimise le critère DCF, EER score est le seuil à l'EER.*

cision est celui de la référence (2.55 pour notre référence LIA06-tnorm, NIST SRE 2005). Pour ce seuil, 977 tests client sont acceptés. A la fin du processus d'adaptation, le seuil optimal est devenu 3.9. Pour ce seuil 1053 tests client sont acceptés. La quantité de données disponibles pour l'adaptation n'étant pas connue *a priori*, il est difficile d'estimer la divergence dans le seuil optimal de décision, avant et après les nombreuses opérations d'adaptation.
Si le seuil de décision n'est pas adaptatif, deux solutions existent pour fixer ce seuil :

1. utiliser le seuil *a posteriori* de la référence. Dans ce cas le seuil est optimal jusqu'à ce que les modèles de locuteur soient adaptés. Après plusieurs adaptations, l'utilisation de ce seuil engendre l'acceptation d'un maximum d'accès client (1172) mais aussi d'une grande quantité d'accès imposteur (699 contre 100 pour la référence) ;
2. utiliser le seuil *a posteriori* du système adapté. Dans ce cas, très peu de tests client sont acceptés, le seuil étant trop restrictif. Le seuil devient optimal en fin de processus, lorsque les modèles de locuteur ont été adaptés avec toutes les données d'adaptation.

Pour répondre à la problématique de la variation du seuil de décision au cours de l'adaptation, nous proposons d'utiliser les méthodes de normalisation de scores adaptative basées sur T-norm et Z-norm(cf. section 7.2.2).

8.5.1 Utilisation de la T-normalisation adaptative

Nos premiers travaux se sont portés sur l'utilisation de la T-norm Adaptative, alors introduite par [Yin et al., 2006]. A la différence des auteurs, nous proposons d'estimer le décalage dans la distribution des scores imposteur en utilisant des données de développement. La méthodologie proposée consiste à générer des modèles de locuteur appris avec de multiples sessions d'apprentissage (des multiples de 2 minutes 30 secondes pour les bases NIST SRE). Ensuite, des distributions de scores imposteur sont estimées, globalement, pour chaque durées d'apprentissage. L'écart (δ) entre les moyennes des scores imposteur entre deux durées d'apprentissage est calculé, puis utilisé pour normaliser les scores de l'évaluation de la façon suivante :

$$\widetilde{Score(x|S)} = \frac{Score(x|S) - (\mu_{imp} + \delta)}{\sigma_{imp}} \tag{8.2}$$

Lors de l'évaluation, le nombre de trames de test sélectionnées pour adapter le modèle sert pour définir le δ à utiliser[2].

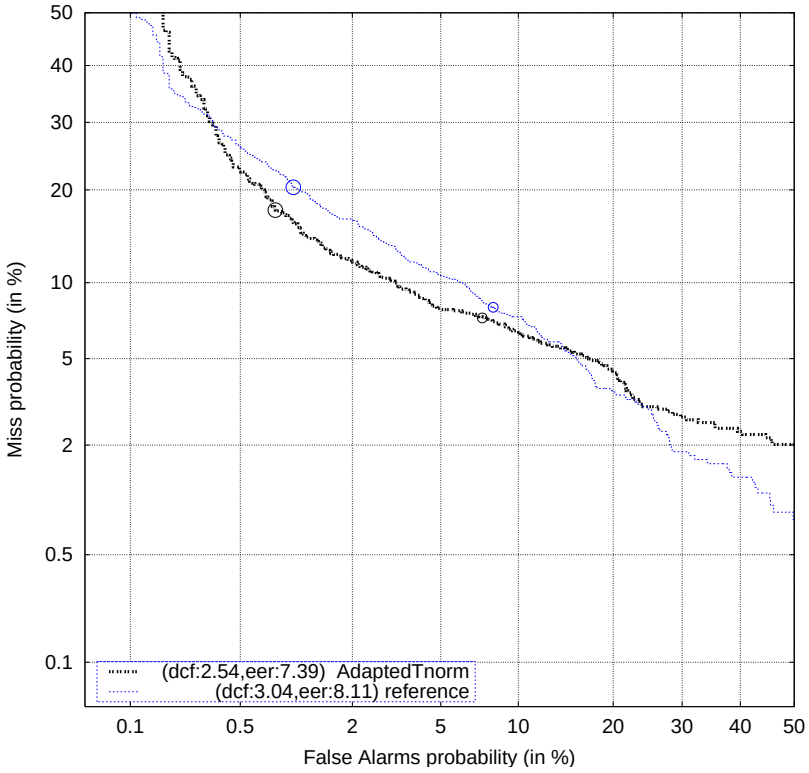


FIG. *8.11* – *Courbes DET du système de référence et du système adapté avec normalisation T-norm Adaptative (NIST SRE 2005, LIA06-tnorm)*

La figure 8.11 présente les résultats de l'utilisation de cette technique sur la base

[2]Ce nombre de trames est égal à : nA + nT * w_test, où nA est le nombre de trames d'apprentissage, nT le nombre de trames du test et w_t le poids WMAP associé au test.

NIST SRE 2005. Le gain observé, par rapport à la référence, est de 16 % relatif pour la DCF et 9% relatif pour la mesure EER. Sans l'utilisation de cette méthode, le gain relatif s'élève à 27% pour la mesure DCF et 37% en terme d'EER (cf. figure 7.8 *a*). En revanche la variation du seuil optimal de décision est très faible. Pour la référence, le seuil qui minimise le critère DCF est de 2.55. Avec adaptation, il est de 3.9. La normalisation adaptative T-norm permet de réduire ce seuil à 2.34, soit proche du seuil optimal sans adaptation.

8.5.2 Utilisation de la Z-normalisation adaptative

La Z-normalisation adaptative a prouvé son efficacité dans un scénario d'adaptation similaire [McLaren et al., 2008]. Cette normalisation, dépendante du client, est plus conforme au principe d'adaptation des modèles de locuteurs. En effet, elle présente l'avantage de mettre en jeu des signaux imposteurs et le modèle client. La distribution des scores imposteur peut être facilement réestimée après chaque adaptation du modèle de locuteur.
Nous utilisons la Z-normalisation adaptative en la combinant avec la T-normalisation classique. La table 8.11 présente les variations du seuil de décision sur les bases de

Système considéré	Seuil optimal	DCF	EER
NIST SRE 2008 référence	3.98	1.90	3.66
NIST SRE 2008 adapté	5.1[1]	2.00	4.10
NIST SRE 2008 adapté ZT-norm Adaptative	4.78	1.75	3.64
NIST SRE 2005 référence	3.944	1.40	3.75
NIST SRE 2005 adapté[2]	6.62	1.07	2.84
NIST SRE 2005 adapté ZT-norm Adaptative[2]	5.10	0.85	2.11

TAB. *8.11* – *Divergence dans les seuils de décision optimaux (NIST SRE 2008, det7 et NIST SRE 2005, det 1).*

données NIST SRE 2005 et 2008. Le système GMM-UBM utilisé pour ces expériences est le système LIA-THL08-ztnorm.
Pour les comparaisons nous utilisons les seuils optimaux qui minimisent le critère *DCF* calculé *a posteriori*.
La variation relative du seuil optimal de décision est réduite sur la base NIST SRE 2008, avec une variation de 20% à une variation de 29% et de moitiée sur la base NIST SRE 2005, avec une variation de 29% au lieu de 68%. En outre, son emploi amène un gain de performance de 17 % et 11% relatifs, respectivement pour les mesures DCF et EER sur la base NIST 2008, par rapport au système d'adaptation sans normalisation adaptative.
Il en résulte un gain de 8% et 0.5% relatifs, pour les mesures DCF et EER, par rapport

[1] La faible variation du seuil optimal de décision sur cette base s'explique par la faible quantité de données client (moins de un test par locuteur).

[2] Notons que les modèles de scores pour les expériences sur la base NIST 2005 avec le système LIA-THL08-ztnorm sont estimés sur la base NIST SRE 2008.

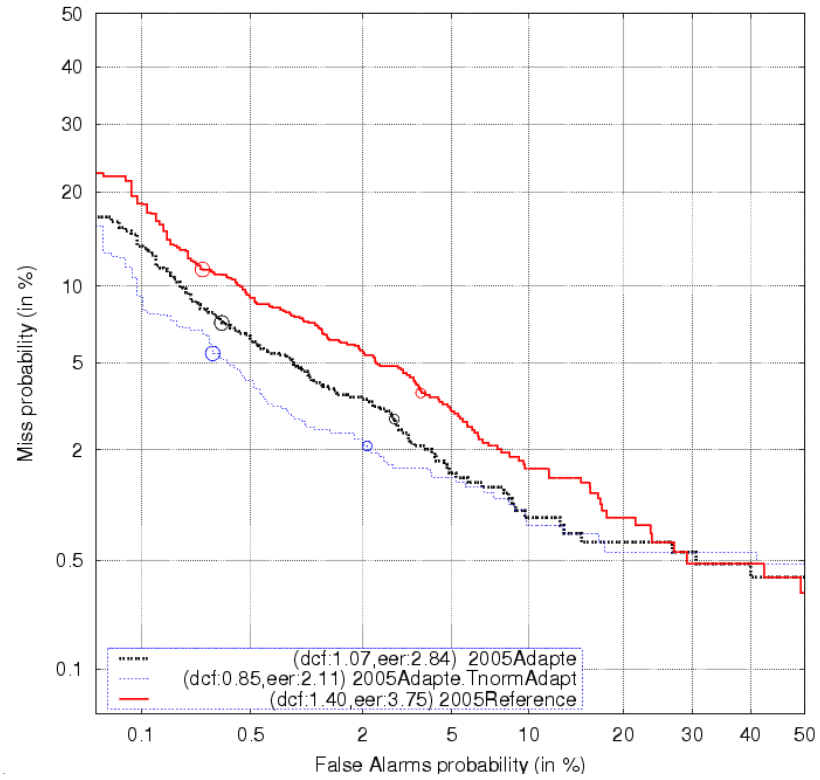


FIG. 8.12 – *Courbes DET des système de référence, adapté et adapté avec normalisation Z-norm Adaptative (NIST SRE 2005, LIA-THL08-ztnorm)*

au système de référence.
L'utilisation de la Z-normalisation adaptative permet au schéma d'adaptation d'être fonctionnel sur l'évaluation NIST SRE 2008.
Sur la base NIST SRE 2005 le gain apporté par la normalisation adaptative est de 20% et 25% relatifs pour les mesures DCF et EER, par rapport à l'adaptation sans normalisation adaptative ; ce qui correspond à un gain de 39% et 44 % relatifs pour les mesures DCF et EER, par rapport à la référence sans adaptation. Les courbes DET de ces expériences sont illustrées dans la figure 8.12.

Les figures 8.13 *a*, *b* et *c* montrent les distributions de scores des trois systèmes, obtenues sur la base NIST SRE 2005. L'utilisation de la Z-normalisation adaptative (figure 8.13 c) a fortement réduit le décalage observé sur les distributions de scores du système adapté (8.13 b).

Il faut noter que, l'écart entre les seuils de décision optimaux reste important, cet écart ne permet pas d'utiliser le score issu modèle adapté pour le calcul de la mesure de confiance (il en résulterait une mauvaise estimation de la mesure de confiance). Nous évaluons les performances obtenues avec un tel schéma d'adaptation. Les résultats d'expériences, utilisant le modèle du locuteur adapté, pour calculer la mesure de confiance, sont donnés dans le tableau 8.12. Il apparaît clairement que cette approche

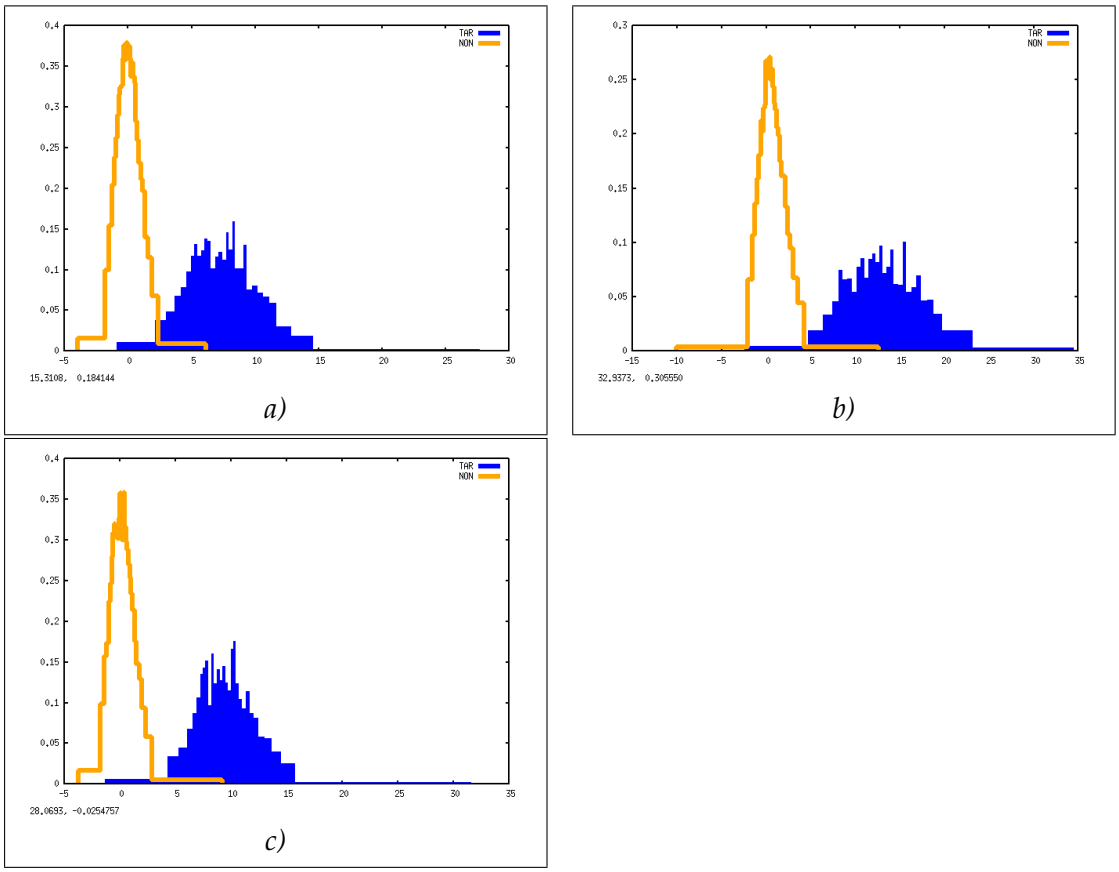


FIG. 8.13 – *Distributions des scores client et imposteur pour a) le système de référence, b) le système adapté et c) le système adapté avec Z-norm Adapatative (NIST SRE 2005).*

introduit des pertes de performances comparativement à l'utilisation du modèle d'apprentissage pour le calcul de la mesure de confiance. Ces pertes sont estimées à environ 30% relatifs sur NIST SRE 2005, 26% et 6% relatifs sur la base NIST SRE 2008, respectivement pour les mesures DCF et EER.

8.6 Complémentarité avec le *Latent Factor Analysis*

Les techniques d'adaptation non supervisée permettent de mieux modéliser les variabilités intra-locuteur et inter-session. La technique du Latent Factor Analysis permet de compenser les effets de la variabilité inter-session. Nous mettons en évidence, dans ce paragraphe, la complémentarité de ces deux approches. La figure 8.14 présente les courbes DET du système de référence LIA-THL07-tnorm sans LFA, du système LIA-THL08-ztnorm avec LFA et du système adapté LIA-THL08-ztnorm avec LFA . L'utilisation de la méthode d'adaptation en addition de la technique du LFA[3], apporte une

[3] Appliquée dans le domaine des paramètres.

Système considéré	Score	DCF	EER
NIST SRE 2008 adapté ZT-norm Adaptative	LLR classique	1.75	3.64
NIST SRE 2008 adapté ZT-norm Adaptative	LLR sur modèle adapté	2.21	3.87
NIST SRE 2005 adapté ZT-norm Adaptative	LLR classique	0.85	2.11
NIST SRE 2005 adapté ZT-norm Adaptative	LLR sur modèle adapté	1.17	2.9

TAB. *8.12* – *Utilisation du modèle adapté pour le calcul des mesures de confiance.*

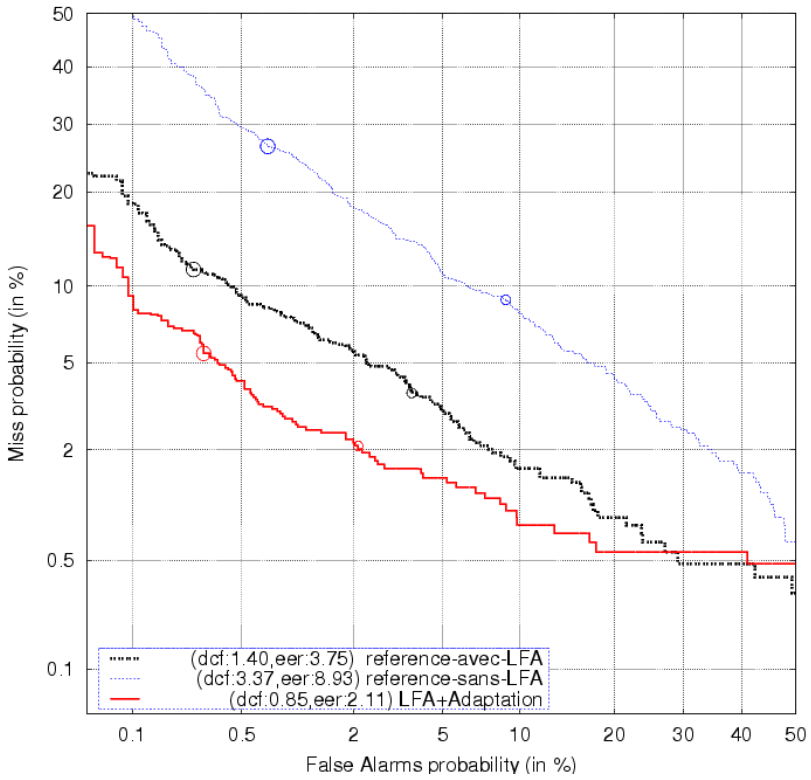


FIG. *8.14* – *Courbe DET pour le système LIA-THL07-tnorm de référence sans LFA, LIA-THL08-ztnorm avec LFA, et le système LIA-THL08-ztnorm adapté avec LFA.*

amélioration par rapport à la référence avec LFA. Le gain relatif est de 39% pour la DCF et 44% pour la mesure EER. L'adaptation ne se porte, dans ce cas, que sur la variabilité intra-locuteur, le LFA ayant retiré en grande partie l'effet du canal de transmission. Ces méthodes apparaissent nettement complémentaires. Le retrait de l'effet canal et l'ajout de données modélisant le locuteur permettent de réduire fortement les taux d'erreurs, avec un gain relatif de 75% pour la DCF et 78% pour l'EER, par rapport au système LIA-THL07-tnorm qui n'implémente aucune de ces deux techniques.
[McLaren et al., 2008] a introduit des méthodes de ré-estimation des paramètres du LFA, prenant en compte les nouvelles données collectées par l'adaptation. Le système SVM-GMM supervecteurs développé est basé sur notre méthode d'adaptation. Le gain alors observé est alors encore plus important. [Yin et al., 2006] a proposé une méthode similaire pour réestimer les paramètres du *factor analysis*. Le système d'adaptation non

supervisée est basé sur un seuil de sélection.

8.7 Conclusion

La couverture de l'espace acoustique d'élocution d'un locuteur ainsi que les conditions d'enregistrements déterminent largement les performances d'un système de RAL. Les conditions d'utilisation, que ce soit dans le cadre des évaluations NIST ou de l'application de surveillance adressée ici, imposent des contraintes fortes sur les durées d'acquisition des signaux de parole. Nous avons pu observer que l'estimation des modèles statistiques de locuteur nécessite des durées conséquentes d'enregistrements.
L'adaptation non supervisée des modèles de locuteur peut être un moyen efficace pour répondre à la contrainte des durées courtes d'apprentissage. Le matériel audio sélectionné pour l'adaptation est alors utilisé pour parfaire l'estimation du modèle client, mais peut servir aussi dans le cadre d'autres techniques comme le LFA.

La difficulté majeure de sélection des données d'adaptation reste néanmoins très liée aux taux d'erreurs du système de RAL. Nous avons décrit les techniques « état de l'art » qui utilisent un seuil de sélection plus sélectif que le seuil de vérification pour éviter les fausses acceptations. Ces techniques ont un intérêt limité car les données pertinentes pour l'adaptation ne sont pas sélectionnées.
Nous avons proposé une solution de sélection des données d'adaptation, sans utilisation de seuil, qui utilise toutes les données de test. Ceci permet d'utiliser les tests de la « zone d'intérêt » lors de l'adaptation. Nous avons démontré le potentiel de cette technique dans des conditions favorables d'exploitation. Cette méthode apporte un gain de 39% et 44 % relatifs, respectivement pour les mesures DCF et EER, par rapport à la référence sur la base NIST SRE 2005. Cependant, une grande variation des performances de l'adaptation a été observée sur les bases NIST SRE 2006 et 2008, avec une perte relative pouvant atteindre 123% pour la mesure DCF (NIST SRE 2006).
Nous avons également analysé le comportement de la méthode proposée au travers d'expériences multiples sur des bases de données différentes et proposé des techniques pour rendre le schéma d'adaptation plus robuste. De fait, les pertes de performances sur les bases les plus difficiles, en terme de quantité de données clientes, ont été fortement réduites en amenant même un gain de de 8% et 0.5% pour les mesures DCF et EER, relativement à la référence sur la base, NIST SRE 2008.

Conclusion et Perspectives

Sur les réseaux professionnels de communication (PMR), aucun moyen de surveillance de l'utilisation des terminaux mobiles en cours de communication n'est aujourd'hui disponible. L'accès au réseau s'effectue sous la forme d'un identifiant (celui du terminal) et d'un mot de passe. C'est donc le terminal qui est authentifié sur le réseau. Aucun contrôle sur l'identité de l'utilisateur n'est réalisé. Le vol ou le prêt d'un terminal ne sont pas détectés. La sécurité du réseau repose sur un opérateur, basé dans un centre de contrôle, qui écoute les communications.
Pour mettre en place une authentification sûre, pendant la communication, la voix seule est disponible. Pour répondre à ce besoin, nous avons traité dans ce document de l'utilisation de la reconnaissance automatique du locuteur (RAL) pour la surveillance de réseaux PMR.
Les problématiques générales en RAL sont abordées dans de nombreux travaux de recherche, mais peu de travaux abordent les contraintes introduites par les conditions spécifiques de fonctionnement de ces réseaux. Ainsi, la chaîne de transmission particulière et les conditions difficiles d'acquisition (bruits ambiants) altèrent la qualité du signal de parole. De plus, les spécifications de l'application de surveillance (réactivité, ergonomie) rendent la tâche plus ardue.
Notre système de référence est régulièrement évalué lors des évaluations internationales NIST SRE. Ces participations permettent de situer les méthodes proposées au niveau international par une comparaison directe. En effet, le calendrier, le protocole ainsi que les données d'évaluations sont les mêmes pour tous les participants. Ce cadre d'évaluation n'est cependant pas assez représentatif de l'environnement auquel sera soumis notre application. Nous évaluons ce système de référence en simulant certaines de ces contraintes, grâce à des protocoles d'évaluations spécifiques (base BREF). Ces évaluations nous ont conduit à définir de nouvelles méthodes pour proposer un système robuste de RAL, adapté à notre cadre applicatif. Nous avons pour cela proposé des solutions répondant aux contraintes de la chaîne de transmission et du bruit ambiant, grâce à des méthodes appropriées de paramétrisation du signal de parole. L'ergonomie et la réactivité du système ont fait l'objet d'une étude spécifique. La réactivité du système est assurée par un test de vérification en ligne. L'ergonomie du système, définie en particulier par les durées courtes d'apprentissage, est obtenue grâce à l'utilisation d'une nouvelle méthode, originale, d'adaptation non supervisée des modèles de locuteur.

La paramétrisation du signal de parole

Le signal de parole est soumis à de nombreuses perturbations lors de son émission et de sa transmission, sur un réseau de communications. Ces sources de variations du signal se succèdent dans la chaîne de transmission du signal, entre son origine (le cerveau du locuteur) et sa réception par le système de RAL. Pour éviter au maximum les perturbations, une solution simple est d'extraire les paramètres au plus près de la source.

Nous proposons une solution basée sur le standard Aurora. Basé sur une architecture distribuée, ce standard permet d'éviter les pertes dues au codage bas-débit de la parole. De plus, il présente d'autres avantages pour le cadre applicatif qui nous intéresse. Ainsi, l'étage de débruitage qu'il intègre réduit très significativement les erreurs de RAL, en conditions bruitées. Il dispose aussi d'une classification des trames par type de voisement. Nous avons utilisé cette information pour définir une nouvelle détection d'activité vocale (DAV) performante pour la RAL, présentant une amélioration de l'EER de 28% relatifs par rapport à notre DAV de référence.

Nous avons également évalué des méthodes alternatives de paramétrisation. Les architectures des réseaux PMR sont normalisées ; aussi, le seul moyen de proposer des améliorations pour la paramétrisation du signal de parole est d'effectuer des traitements au niveau du récepteur. Pour ne pas être limité par la qualité de restitution d'un codeur de parole, nous avons utilisé une méthode de paramétrisation appliquée aux paramètres internes des codeurs. Cette méthode permet d'éviter la resynthèse du signal. Sur un réseau de type TETRA, le gain de cette méthode s'élève à 25% pour l'EER, relativement à l'utilisation du signal de parole décodé.

Réactivité du système de surveillance

Le scénario de surveillance envisagé impose de proposer un système réactif, *i.e.* d'effecteur la RAL sur de courts segments de parole et de retourner des décisions de vérification ou d'identification en continu. Nous répondons à cette contrainte par le traitement « en ligne » de la reconnaissance. Les éléments nécessitant un traitement particulier sont la paramétrisation et le test de vérification (ou d'identification).

Au niveau de la paramétrisation du signal, l'étape de calcul des paramètres cesptraux ne nécessite pas de modification. Le calcul est effectué sur de très courts segments de parole (10 ms). La détection d'activité vocale proposée, bâtie sur le standard Aurora, est aussi compatible avec le traitement en ligne. Le calcul de l'information de voisement est réalisé en même temps que les calculs de paramètres cepstraux.

Nous avons modifié l'étape de normalisation des paramètres de référence, classiquement basée sur l'estimation à long-terme des moyennes et variances des paramètres. Nous proposons l'utilisation d'une méthode qui permet de normaliser chaque vecteur de coefficients cepstraux, dès leur calcul (après une phase d'initialisation). Les

performances obtenues approchent les performances de la méthode de référence.

Au niveau du test de vérification, nous avons appliqué une méthode de lissage temporel des scores. Ceci permet de réduire fortement les durées de tests nécessaires pour obtenir de faibles taux d'erreurs. De plus, cette méthode permet de prendre en compte les variations à court termes des scores des trames de test. Nous faisons l'hypothèse que cela permet de détecter rapidement un changement de locuteur en cours de communication.

Ergonomie du système de surveillance : adaptation des modèles de locuteur

Nous avons abordé le problème d'ergonomie de l'application de surveillance. Pour une plus grande facilité de mise en oeuvre, les périodes d'apprentissage sont réduites. La modélisation des variabilités intra-locuteur ainsi que l'adaptation des modèles de locuteur à l'environnement sont alors limitées. Pour répondre au problème des durées courtes d'apprentissage et des variations intra-locuteur et inter-session, nous nous sommes penchés sur les méthodes d'adaptation des modèles de locuteur. L'adaptation non supervisée convient au type de fonctionnement désiré, cette méthode est automatique et permet de tirer parti du fonctionnement continu du système de surveillance.

Les approches classiques d'adaptation non supervisée démontrent des gains de performance très limités. Nous avons focalisé notre recherche sur une méthode utilisant des tests locuteurs mal reconnus, voire même des tests imposteurs. Nous pensons que le système d'adaptation est d'autant plus performant que les nouvelles données des locuteurs, sélectionnées par l'adaptation, sont complémentaires des données d'apprentissage.
Nous proposons une solution de sélection des données d'adaptation, sans utilisation de seuil, qui utilise toutes les données de test.
Une mesure de confiance est associée à chaque test et détermine son influence dans l'adaptation du modèle de locuteur. Nous avons évalué le potentiel de cette méthode et montré qu'elle apporte un gain de 39% et 44 % relatifs, respectivement pour les mesures DCF et EER par rapport à la référence, sur la base NIST SRE 2005. Néanmoins, la grande variabilité de la méthode est illustrée par ses performances sur la base NIST SRE 2006, où elle conduit à une perte de 123% pour la mesure DCF et 69% pour la mesure EER, relativement à la référence. Nous avons isolé certains facteurs qui limitent les performances de notre adaptation non supervisée. La quantité de tests imposteur par rapport à la quantité de tests client, ainsi que le taux de fausses acceptations, sont des facteurs limitant. Notons tout de même que ces points faibles se retrouvent dans toutes les méthodes d'adaptation publiées.
La compensation des variations du seuil optimal de décision, ainsi que la modification du calcul des mesures de confiance, ont permis de réduire fortement les pertes de performances observées sur les bases NIST SRE 2006. Ceci aussi apporte un gain sur la base NIST SRE 2008, de 8% et 0.5% relatifs pour les mesures DCF et EER.

PERSPECTIVES

Sur le plan général, la première observation que nous pouvons faire est que toutes les contributions, présentées dans ce document, ont été évaluées dans un cadre simulé. Une phase de validation en conditions réelles de fonctionnement est encore nécessaire.

Au niveau des contributions sur la paramétrisation du signal de parole, les tests présentés ne sont pas exhaustifs. Par exemple, le bruitage des signaux n'a pas été appliqué sur les signaux codés à bas débit. Ainsi, l'influence des bruits ambiants, conjuguée au codage à bas débit de la parole, n'a pas été caractérisée. Il serait ainsi intéressant de pouvoir évaluer l'étage de débruitage du codeur de parole MELP. Pour le codeur de parole TETRA, il serait intéressant de caractériser les performances de RAL en milieu bruité, en utilisant les LPCC et le standard Aurora sur les signaux décodés. Le débruitage Aurora pourrait s'avérer plus efficace sur les signaux bruités que l'extraction des LPCC, issus du codeur.

Nous avons démontré l'importance de l'utilisation de la technique LFA dans le domaine des paramètres. Cette technique de compensation de canal apporte un gain très significatif. La compensation peut être appliquée trame à trame. Il est donc possible de l'utiliser dans notre application de surveillance. Cependant, il est difficile de collecter la grande quantité de données de développements nécessaires, pour estimer correctement les paramètres du LFA. Nous pouvons envisager de simuler ces données.

De nombreuses perspectives concernent l'adaptation non supervisée des modèles de locuteurs. Nous avons précisé la difficulté de mettre en place ce type de méthode. Les résultats de la méthode que nous proposons sont très encourageants. Néanmoins la méthode est perfectible ; de nombreux points peuvent encore être adressés, majoritairement au niveau de l'estimation des mesures de confiance.

Le calcul des mesures de confiance est actuellement basé sur le résultat du système de vérification, induisant une forte corrélation avec les erreurs du système de vérification. Nous pensons qu'il est essentiel de décorréler les calculs du score de vérification et de la mesure de confiance. Une illustration simple de la validité de cette hypothèse est que le taux de fausses acceptations (FA) du système engendre inévitablement des erreurs d'adaptation. En considérant que ces erreurs sont dues à des proximités entre les locuteurs, une solution pourrait consister à ne pas adapter les modèles de locuteurs proches. Nous illustrons cette perspective en effectuant une classification des 30 locuteurs extraits de la base Fischer (15 hommes et 15 femmes). La distance entre les locuteurs, utilisée pour la classification, est le score de vérification. Le regroupement en classe s'effectue par la moyenne pondérée des scores des locuteurs de chaque classe. La figure 8.15 représente le dendrogramme de classification. Le haut de l'arbre présente bien la séparation en genres. Certains locuteurs sont globalement distants des autres, le 29 par exemple. Les accès client des autres locuteurs ont une faible probabilité d'être considérés comme des accès clients pour ce locuteur. A l'inverse, les locuteurs 7 et 9 sont très proches. Pour ces locuteurs, le taux de fausses acceptations peut être plus élevé. Nous pourrions envisager de ne pas adapter ces modèles de locuteurs, pour réduire le risque d'introduire des données imposteur dans l'adaptation.

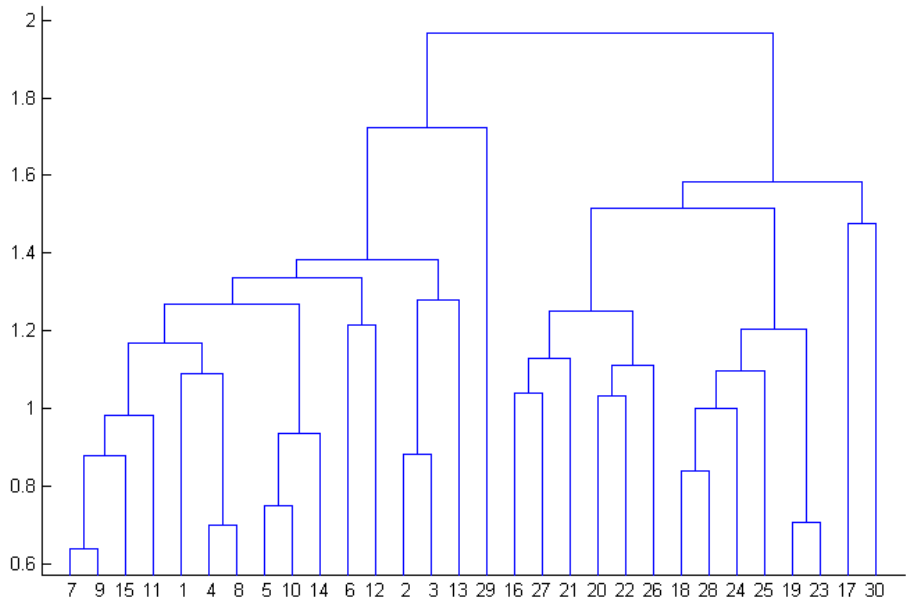


FIG. 8.15 – *Dendrogramme de classification d'une population de 30 locuteurs (15 hommes, 15 femmes). Les distances entre locuteurs sont calculées par logarithme du rapport de vraisemblance (LLR). Le critère de classification est la moyenne pondérée des distances des éléments de chaque classes. Indices des locuteurs masculins : [1-15], indices des locuteurs féminins [16-30] (30 locuteurs de la base Fischer).*

La complémentarité entre les systèmes, utilisée avec réussite dans les fusions de systèmes, peut aussi conduire à une estimation plus robuste des mesures de confiance. Nous avons d'ailleurs réalisé des expériences dans ce sens, où la mesure de confiance résultait de la fusion des scores d'un système SVM et du GMM. La forte corrélation entre les deux systèmes n'a pas amélioré les résultats, démontrant la nécessité de combiner des systèmes très différents.

Cette estimation doit aussi être réalisée dépendamment du modèle de locuteur. L'estimation des distributions de scores client et imposteur, utilisées dans le calcul de la fonction WMAP, ne doit pas être calculée globalement sur une base de données. Les variations de scores propres à un locuteur, que la Z-normalisation tente de réduire, introduisent nécessairement des mesures de confiance mal estimées, pour des locuteurs à la marge. Pour cela nous pouvons envisager de calculer des modèles de scores dépendant du locuteur. Une fonction WMAP par client sera alors calculée.

La difficulté d'une telle approche réside dans l'estimation de la distribution de scores client pour un locuteur. En effet, peu de données clientes sont disponibles *a priori* dans les évaluations NIST, et plus généralement dans un système de RAL. Les normalisations de scores sont d'ailleurs toutes dépendantes des distributions de scores imposteur. Il est plus facile de calculer des distributions de scores imposteur, les données imposteur sont

facilement collectées.
[Poh et Bengio, 2005] introduit la notion de F-normalisation, basée sur la mesure du F-ratio. Les auteurs montrent qu'avec seulement deux accès clients, il est possible de déterminer une normalisation des scores dépendante du client. Cette normalisation surpasse alors les performances de la Z-norm.
Nous avons mené des expériences où la modélisation des scores imposteurs est dépendante du client. Ces modèles ont été appris à partir des données de la normalisation Z-norm. Pour les modèles GMM de scores client, les distributions globales ont été utilisés. Ces expériences n'ont apporté aucun gain.
Nous pensons que le potentiel de cette méthode d'adaptation réside dans l'estimation correcte de la distribution de scores client. Utiliser une adaptation MAP du GMM de scores client global, avec des données de scores client collectées au cours de l'adaptation, semble être une solution envisageable.

Nous avons pu observer que la méthode d'adaptation dégrade dans un premier temps les modèles de locuteur, jusqu'à ce qu'une quantité suffisante de données du locuteur soit recueillie pour permettre d'inverser la tendance. Une idée pour compenser cette phase de perte est de « durcir » l'adaptation jusqu'à obtenir une quantité de données du locuteur suffisante. Nous avons exposé que les *priors* utilisées dans le calcul de la fonction WMAP était déterminantes pour l'estimation des mesures de confiance. Diminuer la *prior* locuteur permet de diminuer globalement les mesures de confiance. L'augmentation progressive de ces *priors*, en fonction des données d'adaptation collectées, peut être une solution pour initialiser la méthode avec des poids d'adaptation faibles, en moyenne, et de progressivement les augmenter.

L'adaptation des modèles de locuteur permet de mieux modéliser les variations inter-session et intra-locuteur. Au niveau de la variabilité inter-session, il serait intéressant, à la manière du *multi-style training*, de créer de multiples modèles de locuteurs pour l'adaptation. En fonction de chaque condition de sessions (bruit, canal) le modèle de locuteur, le plus proche de ces conditions, pourrait être sélectionné et servir de base à l'adaptation.
Cette idée peut amener à proposer l'utilisation de modèles du locuteur « mémorisés » : chacune des adaptations successives engendre un nouveau modèle. Un critère, à déterminer, évaluerait s'il y a divergence entre les modèles mémorisés. Il serait alors possible de revenir en arrière, soit à réinitialiser le modèle courant du locuteur avec un modèle précédent.

Enfin, un problème récurrent en RAL est la difficulté d'estimer la qualité du modèle de locuteur. La qualité du modèle initial influe fortement sur les performances générales du système et de l'adaptation non supervisée. Une première réponse à ce problème a été soulignée au paragraphe 8.4.1, avec l'utilisation du système *reverse*. Une estimation de la qualité du modèle, par une validation croisée de l'apprentissage, pourrait constituer une approche intéressante. Dans ce cas, un sous-ensemble des données d'apprentissage du locuteur est sélectionné pour maximiser le critère de vraisemblance de l'apprentissage. Notons toutefois que cela ne résout en aucun cas le problème de la qualité des enregistrements d'apprentissage.

Les méthodes de RAL sont aujourd'hui peu orientées vers des applications de surveillance de réseaux de communication. Nos travaux permettent pourtant d'envisager de multiples scénarios applicatifs dans ce domaine, notamment sur les réseaux professionnels de communication. Il est d'ailleurs fortement possible que ces applications se diversifient et se développent dans les années à venir.

Annexes

Annexe A

Protocoles d'évaluations

Le protocole, ainsi que le système de RAL utilisé, sont indiqués sous la forme **(protocole, système - type de normalisation de scores)**.

A.1 Expériences menées sur la base BREF 120

La base BREF est une base de données d'enregistrements de textes lus en français (extraits du journal le Monde). Les signaux sont enregistrés sur un microphone de type Shure à une fréquence d'échantillonnage de 16kHz.
Pour les protocoles BREF 1, 2 et BREFVOC, un modèle du monde, indépendant du genre, est créé avec un sous ensemble de 40 locuteurs (20 hommes et 20 femmes).
40 locuteurs sont utilisés comme clients du système de reconnaissance (20 hommes et 20 femmes).
Enfin 35 autres locuteurs sont utilisés comme imposteurs.

A.1.1 Protocole BREF1

Au total, environ huit mille tests clients sont effectués sur un total d'environ quatre-vingt dix mille tests.
Pour respecter au mieux les contraintes des réseaux PMR, nous utilisons des enregistrements de 1 minute pour l'apprentissage des modèles client, et de 8 secondes pour la phase de test.

A.1.2 Protocole BREF2

Au total 820 tests clients sont effectués sur un total d'environ quinze mille tests.
Nous utilisons des enregistrements de 1 minute pour l'apprentissage des modèles client, et de 2.5 minutes pour la phase de test. Ce protocole est notamment utilisé pour les expériences sur les bruits additifs.

A.1.3 Protocole BREFVOC

Nous avons enregistrés 20 locuteurs (17 hommes et 3 femmes) en conditions pseudo-réelles dans une salle machine sur un terminal Tetra Motorolla. 3 minutes d'enregistrements par locuteur ont été collectées, constituées de phrases extraites de la base BREF. Le modèle du monde utilisé est celui dont on se sert dans les protocoles BREF1 et 2, mais les signaux ont été codés à bas débit par le codeur de parole TETRA. Les durées d'apprentissage et de test utilisées ont été de respectivement de 1 minute et de 8 secondes. L'évaluation consiste en 280 tests client et 5320 tests imposteur.

A.1.4 Le système GMM-UBM

Les modèles GMM possèdent 512 composantes.
Les modèles client sont dérivés par adaptation MAP du modèle du monde avec un *relevance factor* de 14.
Paramètres : $C_1....C_{12}, \Delta C_1, ..., \Delta C_{12}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{12}$
Les informations sur la DAV et la normalisation appliquées sont décrites avec les résultats de l'expérience concernée.
Le *Latent Factor Analysis* n'a pas été appliqué sur ces protocoles d'évaluations.
Aucune normalisation de scores n'a été utilisée.

A.2 Expériences menées sur les bases NIST SRE

Les protocoles utilisés pour les expériences sur les bases NIST SRE sont des sous-ensembles ne contenant que les tests hommes.
Les durées d'apprentissage et de test sont de 2 minutes 30 secondes. Le lecteur pourra se référer à [NIST, 2005, 2006, 2008] pour de plus amples détails sur les protocoles d'évaluations NIST SRE.

A.2.1 Protocole NIST SRE 2005

Nous utilisons le sous ensemble det1, constitué des signaux masculins seulement, de la base de données. Les enregistrements sont des conversations téléphoniques et microphoniques, en différentes langues. L'évaluation consiste en 1231 tests client et 12393 tests imposteur (264 locuteurs).

A.2.2 Protocole NIST SRE 2006

Nous utilisons le sous ensemble det3, constitué des signaux masculins seulement, de la base de données. Les enregistrements sont des conversations téléphoniques en

langue anglaise. L'évaluation consiste en 752 tests client et 8960 tests imposteur (219 locuteurs).

A.2.3 Protocole NIST SRE 2008

Nous utilisons le sous ensemble det7, constitué des signaux masculins seulement, de la base de données. Les enregistrements sont des conversations téléphoniques en langue anglaise. L'évaluation consiste en 439 tests client et 6176 tests imposteur (470 locuteurs).

A.3 Systèmes de référence

Cette annexe a pour but de préciser les techniques employées par les systèmes dits de référence dans les expériences et de préciser leur évolution.

Système	**Paramétrisation**	**DAV**	**Compensation canal**	**Normalisation des scores**
LIA06-tnorm	SPRO	LIA	CMVN+ Feature Mapping	T-NORM
LIA-THL06-tnorm	Aurora	LIA	CMVN	T-NORM
LIA-THL07-tnorm	Aurora	Aurora	CMVN	T-NORM
LIA08-ztnorm	SPRO	LIA	CMVN + *symmetrical LFA*	ZT-NORM
LIA-THL08-ztnorm	Aurora	Aurora	CMVN + *feature LFA*	ZT-NORM

TAB. *A.1* – *Description des systèmes utilisés.*

Le tableau A.1 présente une synthèse des éléments clés constituant chacun de ces systèmes. Une présentation plus détaillée de chacun des systèmes est proposée dans les sections suivantes.

Nous présentons dans le tableau A.2 les résultats des différents systèmes chronologiquement. Les taux d'erreurs de ces systèmes sont listés pour le protocole NIST SRE 2005. Tous les systèmes utilisent les mêmes bases de données pour la création du modèle du monde (Fischer[1]) et des modèles imposteurs (NIST SRE 2004).

[1]Fisher English Training Speech Part 1, LDC n° :LDC2004S13

Système	Date	DCF	EER
Système LIA06-tnorm	2006	3.05	7.80
Système LIA-THL06-tnorm	2006	3.63	9.78
Système LIA-THL07-tnorm	2007	3.37	8.93
Système LIA08-ztnorm	2008	1.64	4.21
Système LIA-THL08-ztnorm	2008	1.4	3.5

TAB. A.2 – *Taux d'erreurs des systèmes de référence évalués sur la base NIST SRE 2005.*

A.3.1 Le système GMM-UBM commun

Les modèles du monde dépendant du genre ont été appris avec environ 800 locuteurs provenant de la base Fischer, totalisant environ 10 heures d'enregistrement. 512 composantes Gaussiennes ont été utilisées pour les modèles.
Les modèles client sont dérivés par adaptation MAP du modèle du monde, avec un *relevance factor* de 14.
Les modèles de la cohorte d'imposteur utilisée pour les normalisations Z et T-norm ont été créés à partir de données de la base Fischer d'une durée de 2 minutes 30 secondes.
La sélection des 10 meilleures Gaussiennes est utilisée pour la phase de test.

A.3.2 Système LIA06

Le calcul des paramètres cepstraux est réalisé avec SPRO.
Une fenêtre de Hamming de 20 ms avec un décallage de 10 ms est utilisée.
50 coefficients sont utilisés :
$C_1....C_{19}, \Delta C_1, ..., \Delta C_{19}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{11}$.
La DAV est basée sur une classification des trames par leur énergie (Détecteur à 3 Gaussiennes, l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
La technique du Feature Mapping est ensuite appliquée aux vecteurs de paramètres avant une normalisation moyenne variance (l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
T-norm est utilisée. La cohorte est composée de 160 locuteurs, divisée en parts égales entre des enregistrements du RTC, GSM et DECT provenant de la base Fischer.

A.3.3 Système LIA-THL06

Le calcul des paramètres cepstraux est réalisé avec Aurora.
62 coefficients sont utilisés :
$C_1....C_{20}, \Delta C_1, ..., \Delta C_{20}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{20}, \Delta\Delta log(E)$
Une fenêtre de Hamming de 25 ms avec un décalage de 10 ms est utilisée.
La DAV est basée sur une classification des trames par leur énergie (Détecteur à 3 Gaussiennes, l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
La technique du *Feature Mapping* est ensuite appliquée aux vecteurs de paramètres

avant une normalisation moyenne variance (l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
T-norm est utilisée : la cohorte est composée de 160 locuteurs, divisée en parts égales entre des enregistrements du RTC, GSM et DECT provenant de a base NIST SRE 2004.

A.3.4 Système LIA-THL07

Le calcul des paramètres cepstraux est réalisé avec Aurora.
50 coefficients sont utilisés :
$C_1....C_{19}, \Delta C_1, ..., \Delta C_{19}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{11}$.
Une fenêtre de Hamming de 25 ms avec un décalage de 10 ms est utilisée.
La DAV est basée sur la combinaison de la DAV Aurora et de l'information de voisement Aurora.
La technique du *Feature Mapping* n'est pas appliquée.
Une normalisation moyenne variance est utilisée (l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
T-norm est utilisée : la cohorte est composée de 160 locuteurs, divisée en parts égales entre des enregistrements du RTC, GSM et DECT provenant de la base Fischer.

A.3.5 Système LIA08

Le calcul des paramètres cepstraux est réalisé avec SPRO.
50 coefficients sont utilisés :
$C_1....C_{19}, \Delta C_1, ..., \Delta C_{19}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{11}$.
Une fenêtre de Hamming de 20 ms avec un décalage de 10 ms est utilisée.
La DAV est basée sur une classification des trames par leur énergie (Détecteur à 3 Gaussiennes, l'horizon d'estimation est ici de la durée du fichier d'enregistrement).
La technique du *symmetrical Latent Factor analysis* est utilisée. La matrice de covariance est générée à partir de 124 locuteurs masculins de la base NIST SRE 2004 totalisant 2938 sessions d'enregistrements (rang de la matrice : 40). TZ-norm est utilisée : la cohorte est composée de 160 locuteurs, divisée en parts égales entre des enregistrements du RTC, GSM et DECT provenant de la base NIST SRE 2004.

A.3.6 Système LIA-THL08

Le calcul des paramètres cepstraux est réalisé avec Aurora.
50 coefficients sont utilisés :
$C_1....C_{19}, \Delta C_1, ..., \Delta C_{19}, \Delta log(E), \Delta\Delta C_1, ..., \Delta\Delta C_{11}$.
Une fenêtre de Hamming de 25 ms avec un décalage de 10 ms est utilisée.
L'échelle de MEL n'est pas appliquée sur les signaux féminins (LFCC). La DAV est basée sur la combinaison de la DAV Aurora et de l'information de voisement Aurora.
Une normalisation moyenne variance est utilisée (l'horizon d'estimation est ici de la durée du fichier d'enregistrement).

Le LFA est appliqué dans l'espace des paramètres. Deux matrices de covariance, dépendantes du genre, sont générées. La matrice de covariance femmes est estimée à partir de 133 locutrices des bases NIST SRE 2004 et 2006, totalisant 3856 sessions d'enregistrements (rang de la matrice : 40).
La matrice de covariance hommes est estimée à partir de 136 locuteurs des bases NIST SRE 2004 et 2006, totalisant 3080 sessions d'enregistrements (rang de la matrice : 40).
TZ-norm est utilisée : La cohorte est composée de 160 locuteurs, divisée en parts égales entre des enregistrements du RTC, GSM et DECT provenant de la base NIST SRE 2004.

Annexe B

Apprentissage discriminant des modèles de locuteur

Cette annexe présente les expériences menées sur l'introduction d'un critère discriminant dans la modélisation générative des locuteurs. Le critère de discrimination choisi est le *Maximum Mutual Information* (MMI). Ce critère a démontré de bonnes performances en reconnaissance de la parole et de la langue. Nous proposons dans ce travail deux méthodes d'adaptation des poids du GMM applicables à la RAL. Ce travail a été publié dans [Preti et al., 2006].

B.1 Introduction

L'idée d'intégrer un critère discriminant dans les méthodes génératives a été appliquée avec succès en reconnaissance de la parole. Les classifieurs discriminants, tels les *SVM*, sont désormais très utilisés en RAL. La combinaison d'un critère discriminant et d'une approche générative prend alors tout son sens. Dans le système GMM-UBM, une part importante de la modélisation est acquise grâce au modèle de non locuteur ou modèle du monde. On peut alors le considérer comme un imposteur, et présenter le système GMM-UBM comme un système discriminant, qui s'appuie sur la divergence entre les paramètres d'un locuteur et du modèle du monde. La tentative d'intégrer le critère MMI au sein du GMM se base majoritairement sur ce constat. Nous proposons d'estimer le poids des Gaussiennes du GMM par l'approche MMIE (MMI Estimation). L'idée sousjacente est de diminuer l'influence des Gaussiennes dans le modèle GMM qui porte une forte information commune avec une cohorte de modèles imposteurs. On peut ainsi espérer valoriser l'information spécifique d'un locuteur.

B.2 Adaptation des poids du GMM par MMIE

Différentes techniques d'estimation des poids du GMM existent pour les systèmes de RAL. Nous retiendrons l'adaptation par maximum de vraisemblance et l'adaptation MAP, décrites au chapitre 3.3.1. Nous proposons une fonction d'estimation des poids des Gaussiennes du GMM qui introduit le critère MMI. Le critère MMI permet de réduire l'influence des paramètres d'information mutuelle maximum, entre les paramètres estimés du modèle GMM client (algorithme EM), et les paramètres d'une cohorte de modèles imposteur [Woodland et Povey, 2002; Normandin et Morgera, 1991]. En reconnaissance de la parole, le critère consiste en la maximisation de la fonction objective suivante :

$$F_\lambda = \sum_{r=1}^{R} log\big(\frac{P_\lambda(O_r|M_{w_r}) * P(W_r)}{\sum_{\hat{W}} P_\lambda(O_r M_{\hat{w}}) * P(\hat{W})}\big) \tag{B.1}$$

où O_r représente la séquence de paramètres observée, W_r la transcription correcte, $\hat{W}$ la transcription incorrecte, M_{wr} and $M_{\hat{w}}$ les modèles considérés des transcriptions des classes correcte et incorrecte.

Woodland and Poovey [Woodland et Povey, 2002] introduisent une formule d'estimation des poids pour un état particulier j basé sur la maximisation de la fonction :

$$F_W(\lambda) = \sum_{m=1}^{M} \left[\gamma_{jm}^{num} * log(\hat{c_{jm}}) - \frac{\gamma_{jm}^{den}}{c_{jm}} * \hat{c_{jm}}\right] \tag{B.2}$$

[Levy et al., 2006] a démontré que maximiser cette fonction peut être réalisé en optimisant chaque terme de la somme, qui est une fonction convexe, résultant en :

$$\hat{c_{jm}} = ArgMax_c \left(\gamma_{jm}^{num} * log(c) - \frac{\gamma_{jm}^{den}}{c_{jm}} * c\right) \tag{B.3}$$

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X\in\omega^k} \frac{L(X|G_{jm})}{L(X|S_j)+\sum_{i\neq k} L(X|S_i)}}{\sum_l \sum_{X\in\omega^l} \frac{L(X|G_{lm})}{L(X|S_l)+\sum_{i\neq l} L(X|S_i)}} \tag{B.4}$$

Ils proposent une approximation en :

$$\hat{c_{jm}} = c_{jm} * \frac{c_{jm}}{\sum_k c_{km}} \tag{B.5}$$

Pour adapter ces résultats à la RAL, nous avons considéré que la séquence d'observation est la séquence d'apprentissage, que le modèle de transcription correcte est le modèle du locuteur, et que les observations et modèles de transcription incorrecte seraient issues du modèle du monde, bien que les modèles de contre exemple soient souvent représentés par une cohorte d'imposteurs. Cette approximation est supposée

vraie car le modèle du monde est un imposteur « moyen » pour le modèle client. Un certain pourcentage des observations utilisées pour créer le modèle du monde est donc utilisé ici.

$$DL = \frac{L(X|G_{jc})}{L(X|M_c) + L(X|M_w)} \tag{B.6}$$

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X \in \Omega^c} DL}{\sum_{X \in \Omega^w} DL + \sum_{X \in \Omega^c} DL} \tag{B.7}$$

où G_{jc} est l'indice de la Gaussienne j du client c, M_c est le modèle client, M_w le modèle du monde, Ω^w le *corpus* d'observation imposteur, Ω^c les données d'apprentissage.

Nous adaptons aussi la fonction d'approximation B.5 à la RAL :

$$\hat{c_{jc}} = c_{jc} * \frac{c_{jc}}{c_{jw} + c_{jc}} \tag{B.8}$$

Cette approximation peut être réalisée en RAL car les moyennes et variances des GMM du monde et du client sont identiques.

B.3 Evaluation expérimentale

Type d' adaptation	DCF	EER
MAP	7.21	16.51
MMIE	5.96	15.45
MLE	5.91	16.51
MMIE Approx	5.73	14.83
MMIE Approx (T-norm)	5.08	14.61
MAP (T-norm)	4.76	13.43
MMIE (T-norm)	4.63	13.77
MLE (T-norm)	4.39	13.67

TAB. *B.1* – *Comparaison des adaptations MAP, MLE, MMIE, MMIE Approx pour l'estimation des poids des Gaussiennes du GMM.*

Nous avons confronté cette technique d'adaptation des poids aux techniques d'adaptation des poids MAP et MLE. L'adaptation MAP des poids est réalisée avec une *relevance factor* de 14 et l'adaptation MLE est réalisée par 2 itérations de l'algorithme EM initialisé avec le modèle du monde. L'adaptation des poids MMIE consiste en une première estimation des poids du GMM client par l'algorithme EM, et ensuite par une modification des poids selon l'équation B.8 ou B.6 en fonction de l'expérience. Lors de

l'utilisation de l'équation d'approximation B.8 trois itérations sont réalisées pour diminuer l'effet dû à l'approximation. 5% des séquences utilisées pour créer le modèle du monde ont été sélectionnées comme données de contre exemple (environ 200000 trames). La table B.1 présente les résultats. Les résultats de comparaison entre les méthodes d'adaptation standard MAP et MLE démontrent que la fonction d'approximation améliore les performances de 3% et de 11% relatifs, pour les mesures DCF et EER quand aucune normalisation de scores n'est appliquée. La réestimation des poids par MMIE n'améliore pas les performances. Avec l'utilisation de la T-normalisation les résultats s'inversent. Nous avons noté que la distribution des scores imposteur ne suit plus une loi normale avec l'utilisation du critère MMIE. Ceci peut expliquer les faibles résultats obtenus par la T-normalisation. Nous pouvons noter que ces résultats sont en accord avec [Longworth et Gales, 2006] qui propose d'introduire le critère MMI dans la réestimation des moyennes du GMM.

Annexe C

Schéma bloc des codeurs de parole TETRA et MELP 2400

Cette annexe présente les schémas bloc des codeurs TETRA et MELP. Nous illustrons en pointillés les traitements ajoutés pour extraire les LPCC au niveau du codeur et du décodeur. L'interpolation effectuée au niveau des décodeurs, ainsi que les conversions des coefficients de prédiction vers les LPCC, ont notamment été ajoutées au niveau des codeurs, pour les expériences comparatives d'extraction de paramètres dans le domaine compressé.

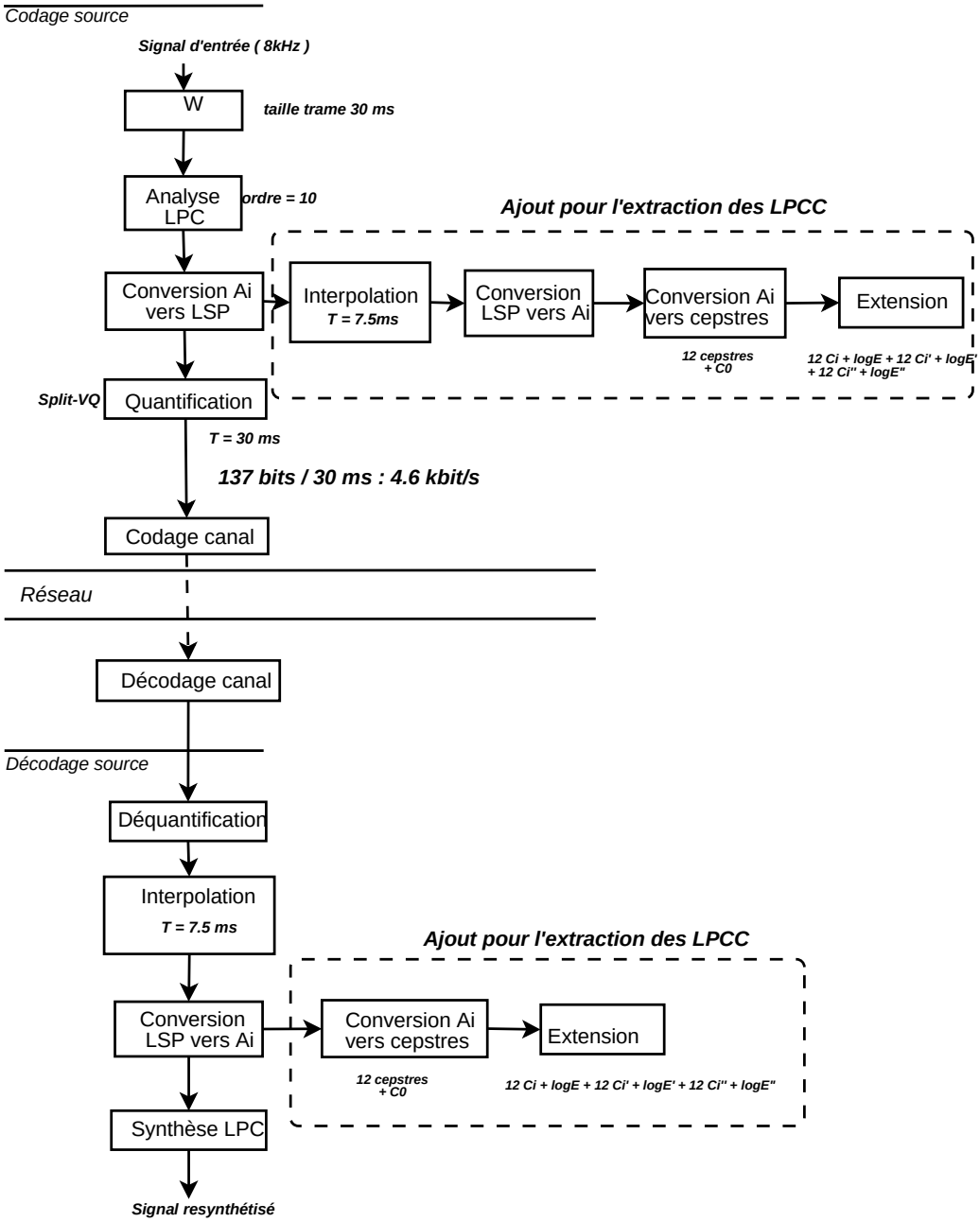


FIG. *C.1* – *Schéma Bloc de l'architecture du codeur de parole TETRA.*

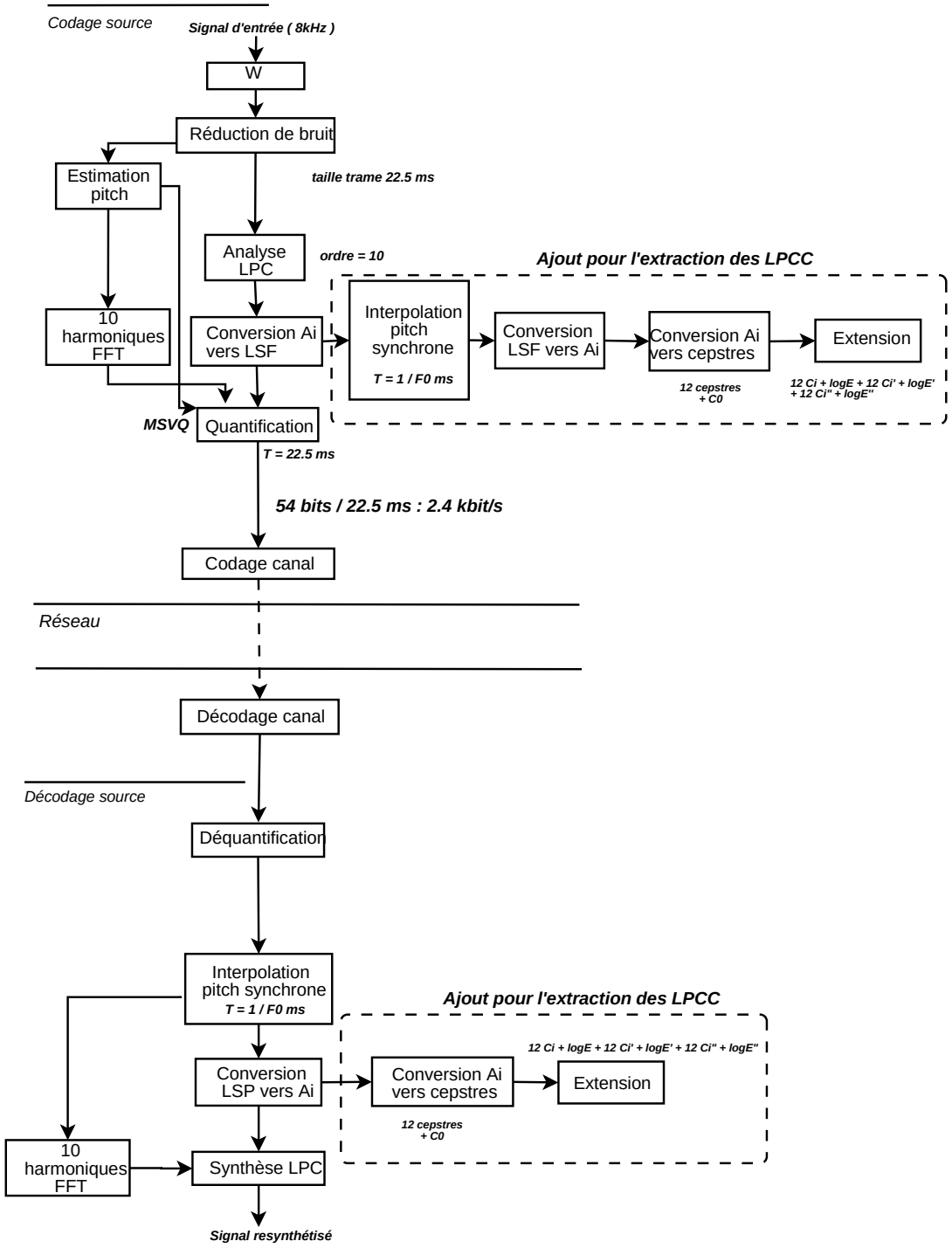


FIG. C.2 – *Schéma Bloc de l'architecture du codeur de parole MELP 2400.*

Annexe D

Démonstrateur de RAL sur le réseau PMR Thales Digicom25

Dans le cadre de cette thèse, un démonstrateur de RAL sur le réseau PMR Tetra digicom 25, commercialisé par Thales, a été réalisé en collaboration avec l'entité Thales ISR. Il a été présenté au TETRA WORLD CONGRESS 2008 par Thales ISR, à Hong Kong, dans le but de promouvoir les services potentiellement disponibles sur le réseau Digicom25.

D.1 Présentation du matériel

a) *b)*

FIG. *D.1* – *Démonstrateur de RAL sur le réseau PMR Digicom 25 : a) Présentation des terminaux et de la station de base ; b) Présentation de l'architecture complète du démonstrateur.*

Les photos D.1 *a* et *b* présentent le matériel utilisé pour mettre en place le démonstrateur. Le réseau PMR est ici constitué de deux terminaux TETRA, d'une station de

base et d'un contrôleur de réseau (le PC fixe). La RAL est effectuée sur un PC distant (PC portable).

D.2 Démonstrateur

Le démonstrateur est un applicatif Windows. L'interface graphique a été développée avec Qt[1]. Deux modes sont disponibles dans le démonstrateur :

- un mode apprentissage
- un mode de test.

D.2.1 L'apprentissage

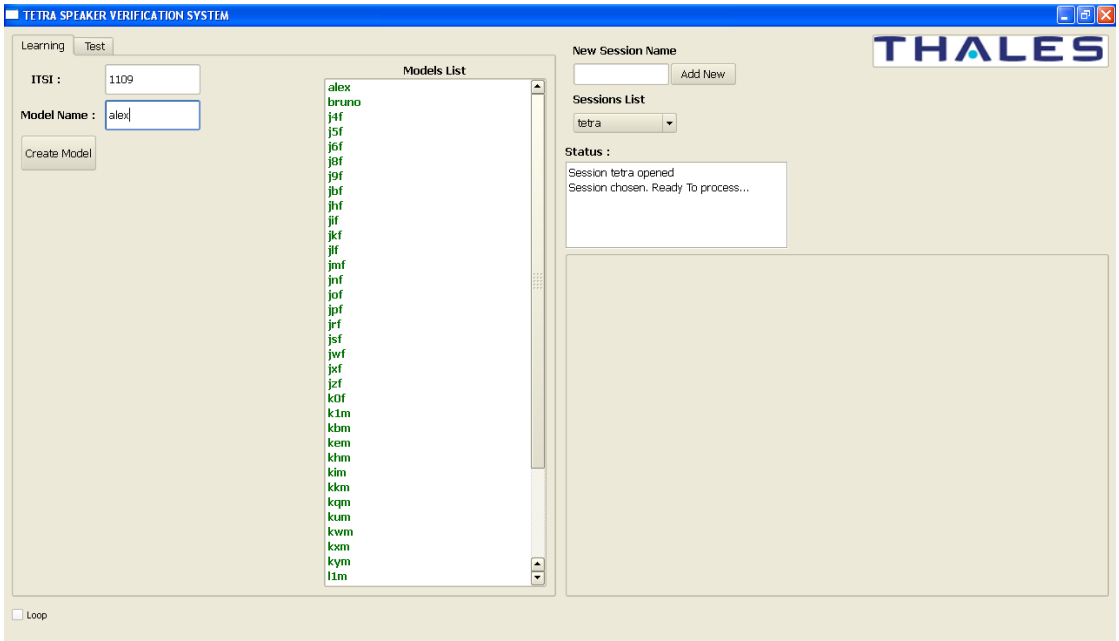


FIG. *D.2* – *Mode apprentissage du démonstrateur.*

Le mode apprentissage est représenté dans la figure D.2. Le choix de l'identifiant du terminal (ITSI) est utilisé pour acquérir le flux audio provenant de ce terminal. Une fois l'acquisition réalisée sur le terminal, le modèle correspondant à l'utilisateur peut être créé (*create model*). Une liste des modèles créés est disponible.

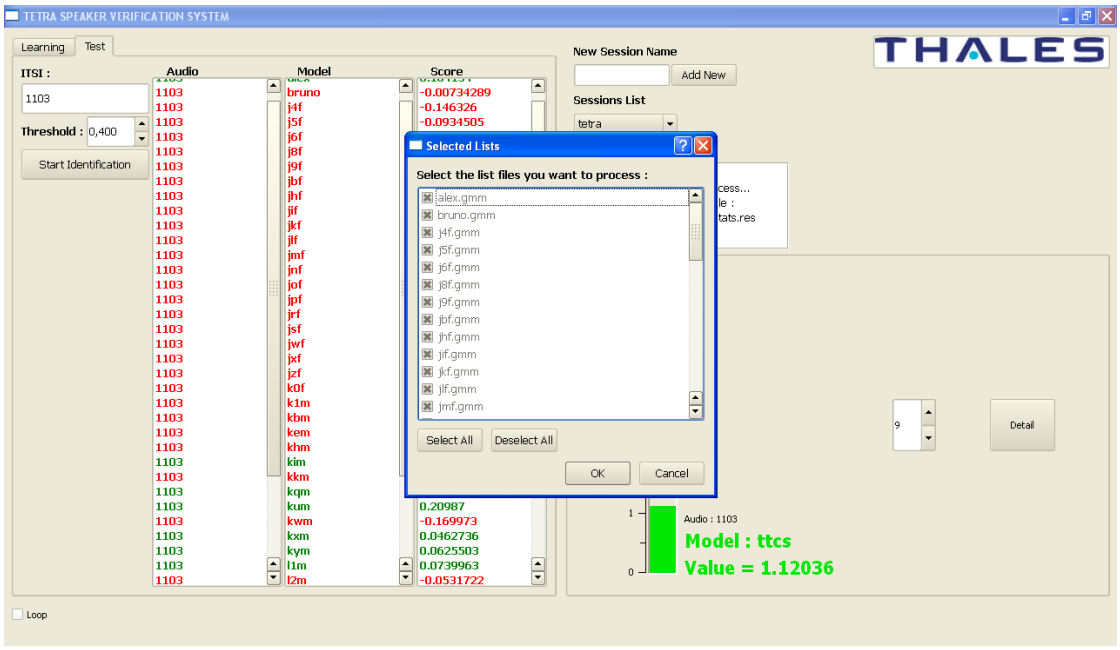


FIG. *D.3* – *Mode de test du démonstrateur.*

D.2.2 Le test

L'onglet « Test » permet d'effectuer une identification, en confrontant un enregistrement à plusieurs modèles, ou une vérification, en confrontant un enregistrement à un modèle.Le mode de test est représenté sur la figure D.3.

Le signal de test est sélectionné par l'identifiant du terminal (ITSI). Une fois l'acquisition réalisée sur le terminal correspondant, il est possible de lancer la vérification ou l'identification. L'appui sur le bouton « Start Identification » fait apparaître une fenêtre permettant la sélection du/des modèle(s) à tester.

Le résultat est affiché sur l'interface. Un tableau présente le score obtenu pour la confrontation entre l'acquisition audio et chaque modèle sélectionné. Le meilleur résultat obtenu est affiché en bas à droite de l'interface sous la forme d'un « bargraph ». Si le score est supérieur au seuil de vérification, il est affiché en vert.
Le bouton « Detail », situé en bas à droite, permet de visualiser les scores du test pour chaque modèle, classés par ordre décroissant. Les modèles acceptés (scores supérieurs au seuil de vérification) sont affichés en vert, les autres sont affichés en rouge.

[1]Librairie développée par la société TROLLTECH

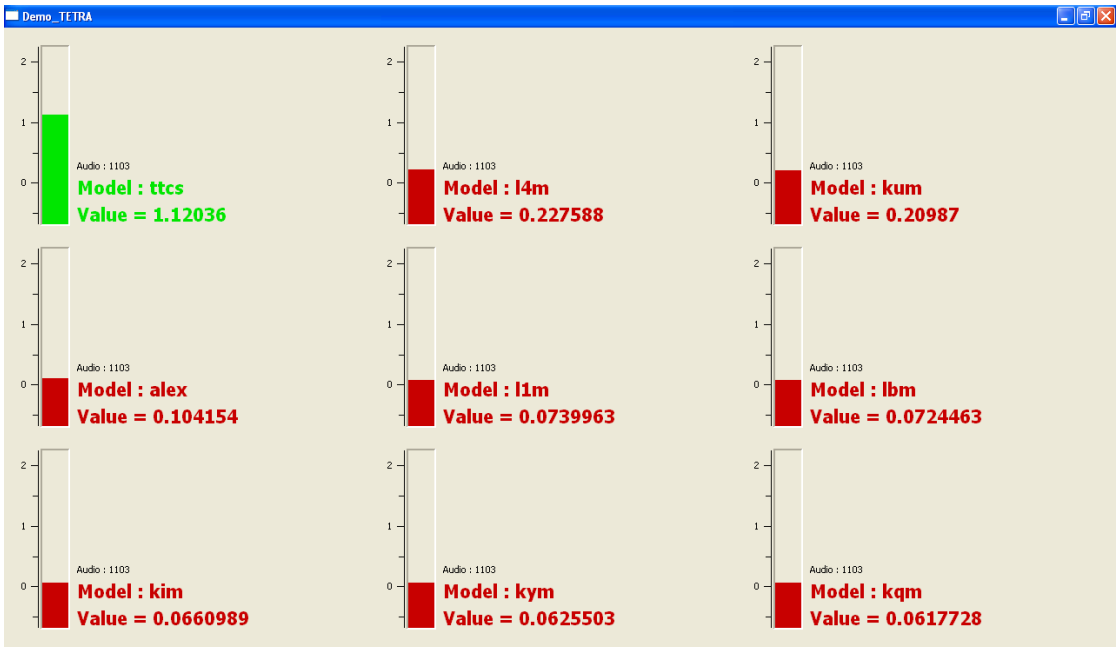


FIG. *D.4 – Affichage des résultats.*

D.3 Le système de RAL utilisé

Le système utilisé dans ce démonstrateur est le système LIA-THL07 (cf. annexe A). Aujourd'hui, le démonstrateur ne bénéficie pas des éléments suivants :

- une normalisation de scores,
- des paramètres internes du codeur de parole TETRA (LPCC),
- des méthodes de traitement en ligne.

Ces éléments sont actuellement en cours d'intégration.

Liste des abréviations

APE	*Applied Probability of Error*
CELP	*Code Excited Linear Prediction*
CMS	*Cesptral Mean Substraction*
CMVN	*Cesptral Mean and Variance Normalization*
DAV	*Détection d'Activité Vocale*
DCF	*Decision Cost Function*
DCT	*Discrete Cosine Transform*
DET	*Detection Error Tradeoff*
EM	*Expectation Maximization*
EPC	*Expected Performance Curve*
ETSI	*European Telecommunications and Standards Institute*
FA	*Fausses Acceptations*
FFT	*Fast Fourier Transform*
FR	*Faux Rejets*
GLR	*Generalized Likelihood Ratio*
GMM	*Gaussian Mixture Model*
GSM	*Global System for Mobile Communications*
HMM	*Hidden Markov Model*
HTER	*Half Total Error Rate*
IdC	*Intervalle de Confiance*
ITSI	*Individual TETRA Subscriber Identity*
JFA	*Joint Factor Analysis*
LFA	*Latent Factor Analysis*
LFCC	*Linear Frequency Cepstral Coefficient*
LIA	*Laboratoire d'informatique d'Avignon*
LLR	*Log Likelihood Ratio*
LMS	*Least Mean Square*
LPC	*Linear Predictive Coding*
LPCC	*Linear Predictive Cepstral Coefficient*
LR	*Likelihood Ratio*
LSF	*Line Spectral Frequency*
LSP	*Line Spectral Pair*
MAP	*Maximum A Posteriori*
MELP	*Mixed Excitation Linear Prediction*

MFCC *Mel Frequency Cepstral Coefficient*
ML *Maximum Likelihood*
MMI *Maximum Mutual Information*
NAP *Nuisance Attribute Projection*
NIST *National Institute of Standards and Technology*
PMC *Parallel Model Combination*
PMR *Private Mobile Radio network*
RAL *Reconnaissance Automatique du Locuteur*
RASTA *Relative Spectral*
RSB *Rapport Signal sur Bruit*
ROC *Receiver Operating Characteristic*
SRE *Speaker Recognition Evaluation*
SVM *Support Vector Machine*
TETRA *Terrestrial Trunked Radio*
TCF *Thales Communication France*
UBM *Universal Background Model*
VAL *Vérification Automatique du Locuteur*
WMAP *World Maximum A Posteriori*

Liste des illustrations

Liste des tableaux

Bibliographie

[Adami et al., 2003] A. Adami, R. Mihaescu, D. Reynolds, et J. Godfrey, 2003. Modeling prosodic dynamics for speaker recognition. Dans les actes de *ICASSP*, 788–91.

[Arciénega, 2006] M. Arciénega, 2006. *Speaker recognition in noisy environments using auxiliary information and bayesian networks*. Thèse de Doctorat, Ecole Polytechnique fédérale de Lausanne.

[Atal, 1974] B. Atal, 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America 55*(6), 1304–1312.

[Auckenthaler et al., 2000] R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. Score normalization for text-independent speaker verification systems. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10*, 42–54.

[Auckenthaler et al., 2001] R. Auckenthaler, M. Carey, et J. Mason, 2001. Language dependency in text-independent speaker verification. Dans les actes de *ICASSP*.

[Barras et al., 2004] C. Barras, S. Meignier, et J.-L. Gauvain, 2004. Unsupervised online adaptation for speaker verification over the telephone. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

[Ben, 2004] M. Ben, 2004. *Approches robustes pour la vérification du locuteur par normalisation et adaptation hiérarchique*. Thèse de Doctorat, Université de Rennes 1.

[Bengio et Mariethoz, 2004] S. Bengio et J. Mariethoz, 2004. The expected performance curve : a new assessment measure for person authentication. Dans les actes de *2001 : A Speaker Odyssey - The Speaker Recognition Workshop*.

[Benveniste et Goursat, 1984] A. Benveniste et M. Goursat, 1984. Blind equalizers. *IEEE transactions on communications 32*(8), 871–833.

[Besacier et Bonastre, 1998] L. Besacier et J.-F. Bonastre, 1998. Frame pruning for speaker recognition. Dans les actes de *ICASSP*, Volume 2, 765 – 768.

[Besacier et al., 2000] L. Besacier, J.-F. Bonastre, et C. Fredouille, 2000. Localization and selection of speaker-specific information with statistical modelling. *Speech Communication 31*, 89–106.

[Bimbot et al., 2004] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, et D. A. Reynolds, 2004. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing 4*, 430–451.

[Bin et Meng, 2004] M. Bin et H. Meng, 2004. English-chinese bilingual text-independent speaker verification. Dans les actes de *ICASSP*.

[Bin et al., 2007] M. Bin, H. Meng, et M. Man-Wai, 2007. Effects of device mismatch, language mismatch and environmental mismatch on speaker verification. Dans les actes de *ICASSP*.

[Bolle et Pankanti, 1998] R. Bolle et S. Pankanti, 1998. *Biometrics, Personal Identification in Networked Society : Personal Identification in Networked Society*. Norwell, MA, USA : Kluwer Academic Publishers.

[Bonastre, 2008] J.-F. Bonastre, 2008. La reconnaissance du locuteur : un problème résolu ? Dans les actes de *Journées d'études sur la Parole (JEP)*.

[Bonastre et al., 2003] J.-F. Bonastre, F. Bimbot, L. Boe, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau, 2003. Person authentication by voice : a need for caution. Dans les actes de *EUROSPEECH*, 33–36.

[Bonastre et al., 2004] J.-F. Bonastre, N. Scheffer, C. Fredouille, et D. Matrouf, 2004. NIST'04 speaker recognition evaluation campaign : new lia speaker detection plateform based on ALIZE toolkit. Dans les actes de *NIST SRE'04 Workshop : speaker detection evaluation campaign*.

[Bonastre et al., 2005] J.-F. Bonastre, F. Wils, et S. Meignier, 2005. ALIZE, a free toolkit for speaker recognition. Dans les actes de *ICASSP*.

[Bonastre et al., 2008] N. Bonastre, J-F. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, et J. Mason, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

[Broun et al., 2001] C. C. Broun, W. M. Campbell, D. Pearce, et H. Kelleher, 2001. Distributed speaker recognition using the ETSI distributed speech recognition standard. Dans les actes de *2001 : A Speaker Odyssey - The Speaker Recognition Workshop*.

[Brummer, 2005] N. Brummer, 2005. Focal, tools for fusion and calibration of automatic speaker detection systems. http ://www.dsp.sun.ac.za/nbrummer/focal/index.htm.

[Campbell et al., 2001] W. Campbell, D. Reynolds, et R. Dunn, 2001. Fusing high- and low-level features for speaker recognition. Dans les actes de *Eurospeech*.

[Campbell et al., 2004] W. Campbell, D. Reynolds, et R. Dunn, 2004. Fusing discriminative and generative methods for speaker recognition : Experiments on switchboard and NFI/TNO field data. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

[Campbell et al., 2006] W. M. Campbell, D. E. Sturim, D. E. Sturim, D. A. Reynolds, et D. A. Reynolds, 2006. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters, IEEE 13*(5), 308–311.

[Carey et Parris, 1992] M. J. Carey et E. S. Parris, 1992. Speaker verification using connected words. Dans les actes de *Proceedings of the Institute of Acoustics*, Volume 146, 95–100.

[Childers et al., 1998] D. Childers, R. V. Cox, R. DeMori, S. Furui, B.-H. Juang, J. J. Mariani, P. Price, S. Sagayama, M. M. Sondhi, et R. Weischedel, 1998. The past, present and future of speech processing. *IEEE Signal Processing Magazine 15*(3), 24–28.

[Conrad et Paliwal, 2001] S. Conrad et K. Paliwal, 2001. Information fusion for robust speaker verification. Dans les actes de *Eurospeech*.

[Dass et al., 2006] S. Dass, Y. Zhu, et A. Jain, 2006. Validating a biometric authentication system : Sample size requirements. *Pattern Analysis and Machine Intelligence, IEEE Transactions 28*, 1902–1319.

[De Campos Neto, 1999] S. De Campos Neto, 1999. The ITU-T software tool library. *International journal of speech technology 2*, 259–272.

[Dehak et al., 2007] N. Dehak, P. Dumouchel, et P. Kenny, 2007. Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 2095–2103.

[Dempster et al., 1977] A. Dempster, N. Laird, et D. et Rubin, 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1–38.

[Doddington, 1985] G. Doddington, 1985. Speaker recognition - Identifying people by their voices. *Proceedings of the IEEE 73*(11), 1651–1664.

[Doddington et al., 1998] G. Doddington, W. Liggett, et M. Przybocki, 1998. Sheep, goats, lambs and wolves : A statistical analysis of speaker performance. Dans les actes de *NIST 1998 Speaker Recognition Evaluation*.

[Drygajlo et El-Maliki, 1998] A. Drygajlo et M. El-Maliki, 1998. Speaker verification in noisy environment with combined spectral substraction and missing data theory. Dans les actes de *ICASSP*.

[ETSI, 2005a] ETSI, 2005a. ETSI ES 202 212 v1.1.2 (2005-11).

[ETSI, 2005b] ETSI, 2005b. TETRA ETSI EN 300 395-1 v1.3.1 (2005-06).

[Ezzaidi et al., 2001] H. Ezzaidi, J. Rouat, et D. OShaughnessy, 2001. Towards combining pitch and MFCC for speaker identification systems. Dans les actes de *Eurospeech*.

[Fant, 1960] G. Fant, 1960. *Acoustic Theory of Speech Production*. Hague : Mouton's Co.

[Fauve et al., 2007] B. Fauve, N. W. D. Evans, N. R. Pearson, J.-F. Bonastre, et J. S. D. Mason, 2007. Influence of task duration in text-independent speaker verification. Dans les actes de *Interspeech*, 794–797.

[Fischer, 1990] R. Fischer, 1990. *Statistical Methods, Experimental Design and Scientific Inference*. Oxford Science Publications.

[Fredouille, 2000] C. Fredouille, 2000. *Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances*. Thèse de Doctorat, Université d'Avignon et des pays de Vaucluse.

[Fredouille et al., 1999] C. Fredouille, J. Bonastre, et T. Merlin, 1999. Similarity normalization method based on world model and a posteriori probability for speaker verification. Dans les actes de *Eurospeech*.

[Fredouille et al., 2000] C. Fredouille, J. Mariethoz, C. Jaboulet, J. Hennebert, J.-F. Bonastre, C. Mokbel, et F. Bimbot, 2000. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. Dans les actes de *ICASSP*.

[Furui, 1978a] S. Furui, 1978a. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE transactions on Acoustics Speech and Signal Processing 29*(3), 342–350.

[Furui, 1978b] S. Furui, 1978b. *Research on Individuality Information in Speech Waves*. Thèse de Doctorat, Tokyo University.

[Furui, 1981] S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. *IEEE transactions on Acoustics Speech and Signal Processing 29*(2), 254–272.

[Furui, 1986] S. Furui, 1986. Research on individuality feature in speech waves and automatic speaker recognition techniques. *Speech Communications 5*, 183–197.

[Gales et Young, 1993] M. Gales et S. Young, 1993. HMM recognition in noise using parallel model combination. Dans les actes de *Eurospeech*.

[Garcia-Romero et al., 2003] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodryguez, et J. Ortega-Garcia, 2003. Support vector machine fusion for idiolectal and acoustic speaker information in spanish conversational speech. Dans les actes de *ICASSP*.

[Gauvain et Lee, 1994] J.-L. Gauvain et C. Lee, 1994. Maximum a posteriori estimation for multivariate Gaussian Mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing 2*, 291–298.

[Grassi et al., 2002] S. Grassi, M. Ansorge, F. Pellandini, et P.-A. Farine, 2002. Distributed speaker recognition using the ETSI Aurora standard. Dans les actes de *Proc. of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*.

[Hansen et al., 2006] E. Hansen, R. Slyh, et T. Anderson, 2006. Supervised and unsupervised speaker adaptation in the NIST 2005 speaker recognition evaluation. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

[Heck et Mirghafori, 2000] L. Heck et N. Mirghafori, 2000. Online unsupervised adaptation in speaker verification. Dans les actes de *ICSLP*.

[Heck et Mirghafori, 2001] L. Heck et N. Mirghafori, 2001. Unsupervised on-line adaptation in speaker verification : Confidence-based updates and improved parameter estimation. Dans les actes de *Adaptation in Speech Recognition*.

[Helander et Nurminen, 2007] X. Helander et X. Nurminen, 2007. On the importance of prosody on speaker identity. Dans les actes de *Eurospeech*.

[Hermansky et Morgan, 1994] H. Hermansky et N. Morgan, 1994. RASTA processing of speech. Dans les actes de *IEEE Transactions on Speech and Acoustics*, Volume 2, 587–589.

[Huerta et Stern, 1998] J. Huerta et R. Stern, 1998. Speech recognition from gsm codec parameters. Dans les actes de *ICSLP*.

[Institute for Perception-TNO, 1990] T. N. Institute for Perception-TNO, 1990. NOISE-ROM-0, NATO : Ac243/(panel 3)/RSG-10, ESPRIT : Project no. 2589 - SAM.

[ITU, 2006] ITU, 2006. 3GPP TS 06.90 : Adaptive multi-rate speech codec ; transcoding functions.

[Jain et al., 2001] A. Jain, R. Bolle, S. Pankanti, S. Liu, et M. Silverman, 2001. A practical guide to biometric security technology. Rapport technique, IEEE Computer Society.

[Kenny et al., 2005a] P. Kenny, G. Boulianne, et P. Dumouchel., 2005a. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing 13*(3), 345–359.

[Kenny et al., 2005b] P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel., 2005b. Factor analysis simplified. Dans les actes de *ICASSP*, 637–640.

[Kenny et Dumouchel., 2004] P. Kenny et P. Dumouchel., 2004. Disentangling speaker and channel effects in speaker verification. Dans les actes de *ICASSP*, 37–40.

[Klatt, 1976] D. Klatt, 1976. Digital filter bank for spectral matching. Dans les actes de *ICASSP*.

[Kleynhans et Barnard, 2005] N. Kleynhans et E. Barnard, 2005. Language dependence in multilingual speaker verification. Dans les actes de *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, 117–121.

[Kuhn et al., 1998] R. Kuhn, P. Nguyen, J. Junqua, et L. Goldwasser, 1998. Eigenfaces and eigenvoices : Dimensionality reduction for specialized pattern recognition. Dans les actes de *Workshop on Multimedia Signal Processing (MMSP 98)*, 71–76.

[Kuroiwa et Tsuge, 2003] S. Kuroiwa et S. Tsuge, 2003. Blind equalization techniques for ETSI standard DSR front-end. Dans les actes de *ICASSP*.

[Lamel et al., 1991] L. Lamel, J.-L. Gauvain, et M. Eskenazi, 1991. BREF, a large vocabulary spoken corpus for french. Dans les actes de *Eurospeech*.

[Levy et al., 2006] C. Levy, G. Linares, P. Nocera, et J. Bonastre, 2006. Embedded mobile phone digit recognition. *Digital Signal Processing for In-Vehicle and Mobile Systems 2*, 345–359.

[Li et Porter, 1988] K.-P. Li et J. Porter, 1988. Normalization and selection of speech segments for speaker recognition scoring. Dans les actes de *ICASSP*.

[Lilly et Paliwal, 1996] B. Lilly et K. Paliwal, 1996. Effect of speech coders on speech recognition performance. Dans les actes de *ICSLP*, Volume 4, 2344–2347.

[Lippmann et al., 1987] R. Lippmann, E. Martin, et D. Paul, 1987. Multi-style training for robust isolated-word speech recognition. Dans les actes de *ICASSP*.

[Longworth et Gales, 2006] C. Longworth et M. J. F. Gales, 2006. Discriminative adaptation for speaker verification. Dans les actes de *ICSLP*.

[Lovekin et al., 2001] J. Lovekin, R. Yantorno, K. Krishnamachari, D. Benincasa, et S. Wenndt, 2001. Developing usable speech criteria for speaker identification technology. Dans les actes de *ICASSP*, 421–424.

[Magrin-Chagnolleau et Bonastre, 1995] I. Magrin-Chagnolleau et J.-F. Bonastre, 1995. Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods. Dans les actes de *EUROSPEECH*.

[Marin et al., 2005] J. Marin, K. Mengersen, et C. Robert, 2005. *Bayesian Thinking, Modeling and Computation*, Chapter 16 : Bayesian modelling and inference on mixtures of distributions. Elsevier.

[Martin et al., 1997] A. Martin, G. Doddington, T. Kamm, M. Ordowski, et M. Przybocki, 1997. The DET curve in assessment of detection task performance. Dans les actes de *Eurospeech*, 1895–1898.

[Mason et Thompson, 1993] J. S. Mason et J. Thompson, 1993. Gender effects in speaker recognition. Dans les actes de *ICSLP*, 733–736.

[Matejka et al., 2007] P. Matejka, L. Burget, P. Schwarz, O. Glembek, M. Karafiat, F. Grézl, J. Cernocky, D. van Leeuwen, D. Brümmer, et A. Strasheim, 2007. STBU system for the NIST 2006 speaker recognition evaluation. Dans les actes de *ICASSP*, 221–224.

[Matrouf, 1997] D. Matrouf, 1997. *Adaptation des modèles acoustiques pour la reconnaissance de la parole bruitée*. Thèse de Doctorat, Université de Paris 11, Orsay.

[Matrouf et al., 2007] D. Matrouf, N. Scheffer, B. Fauve, et J.-F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *ICSLP*.

[Mauler et Martin, 2006] D. Mauler et R. Martin, 2006. Noise power spectral density estimation on highly correlated data. Dans les actes de *IWAENC*.

[Mauuary, 1998] L. Mauuary, 1998. Blind equalization in the cepstral domain for robust telephone based speech recognition. Dans les actes de *EUSIPCO*, 359–362.

[McCree et al., 1996] A. McCree, K. K. Truong, E. George, T. Barnwell, et V. Viswanathan, 1996. A 2.4 kbits/s MELP coder candidate for the new u.s. federal standard. Dans les actes de *ICASSP*.

[McLaren et al., 2008] M. McLaren, D. Matrouf, R. Vogt, et J.-F. Bonastre, 2008. Combining continuous progressive model adaptation and factor analysis for speaker verification. Dans les actes de *ICSLP*.

[Meignier, 2002] S. Meignier, 2002. *Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection.* Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

[Mengusoglu, 2003] E. Mengusoglu, 2003. Confidence measure based model adaptation for speaker verification. Dans les actes de *The 2nd IASTED International Conference on Communications, Internet and Information Technology*.

[Mezaache et al., 2008] S. Mezaache, J. Bonastre, et D. Matrouf, 2008. Analysis of impostor tests with high scores in NIST-SRE context. Dans les actes de *ICSLP*.

[Ming et al., 2007] J. Ming, T. Hazen, J. Glass, et D. A. Reynolds, 2007. Robust speaker recognition in noisy conditions. *IEEE transactions on Audio, Speech, and Language Processing 15*, 1711–1723.

[Mirghafori et Heck, 2002] N. Mirghafori et L. Heck, 2002. An adaptative speaker verification system with speaker dependent a priori decision thresholds. Dans les actes de *ICSLP*.

[Mokbel et al., 1996] C. Mokbel, D. Jouvet, et J. Monne, 1996. Deconvolution of telephone line effects for speech recognition. *Speech Communication 19*, 185–196.

[Mokbel et al., 1993] C. Mokbel, J. Monn, et D. Jouvet, 1993. On-line adaptation of speech recognizer to variations in telephone line conditions. Dans les actes de *Eurospeech*, 1247–1250.

[NIST, 2005] NIST, 2005. The NIST year 2005 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v5.pdf.

[NIST, 2006] NIST, 2006. The NIST year 2006 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2005/sre-06_evalplan-v9.pdf.

[NIST, 2008] NIST, 2008. The NIST year 2008 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2005/sre08_evalplan-release4.pdf.

[Nitin et Raina, 2004] J. Nitin et A. Raina, 2004. Pitch correlogram clustering for fast speaker identification. *EURASIP Journal on Applied Signal Processing*. *2004*(1), 2640–2649.

[Noll, 1964] A. Noll, 1964. Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection. *Journal of the Acoustical Society of America* *36*(2), 430–451.

[Normandin et Morgera, 1991] Y. Normandin et D. Morgera, 1991. An improved MMIE training algorithm for speaker-independent small vocabulary continuous speech recognition. Dans les actes de *ICASSP*.

[NSA, 2006] NSA, 2006. STANAG 4591 ratification draft. Rapport technique, NATO standardization Agency (NSA).

[Openshaw, 1994] J. Openshaw, J. et Mason, 1994. On the limitations of cepstral features in noise. Dans les actes de *ICASSP*.

[Oppenheim et Schaffer, 1975] A. Oppenheim et R. Schaffer, 1975. *Digital signal Processing*. New Jersey : Prentice-Hall.

[P. Kenny et al., 2006] P. P. Kenny, V. Gupta, et G. Boulianne, 2006. Feature normalization using smoothed mixture transformations. Dans les actes de *ICSLP*.

[Pearce, 2000] D. Pearce, 2000. Enabling new speech driven services for mobile devices : An overview of the ETSI standards activities for distributed speech recognition front-ends. Dans les actes de *Applied Voice Input Output Society Conference (AVIOS)*.

[Pelecanos et Sridharan, 2001] J. Pelecanos et S. Sridharan, 2001. Feature Warping for Robust Speaker Verification. Dans les actes de *2001 : A Speaker Odyssey - The Speaker Recognition Workshop*, 213–218.

[Petracca et Servetti, 2006] M. Petracca et J. Servetti, A. De Martin, 2006. Performance analysis of compressed-domain automatic speaker recognition as a function of speech coding technique and bit rate. Dans les actes de *ICME*.

[Poh et Bengio, 2005] N. Poh et S. Bengio, 2005. F-ratio client-dependent normalisation for biometric authentication tasks. Dans les actes de *ICASSP*, Volume 1, 721–724.

[Porter, 2000] J. Porter, 2000. *On the 30 Error Criterion*, 51–56. National Biometric Center Collected Works, J. Wayman, ed.

[Przybocki et Martin, 1998] M. A. Przybocki et A. F. Martin, 1998. NIST speaker recognition evaluation 1997. Dans les actes de *Workshop on speaker recognition and its commercial and forensic applications*, 120–123.

[Pujol et al., 2006] P. Pujol, D. Macho, et C. Nadeu, 2006. On real-time mean and variance normalization of speech recognition features. Dans les actes de *ICASSP*.

[Quatieri et al., 2000] T. Quatieri, R. Dunn, D. Reynolds, J. Campbell, et E. Singer, 2000. Speaker recognition using G.729 speech codec parameters. Dans les actes de *ICASSP*.

[Raj et al., 2001] B. Raj, J. Migdal, et R. Singh, 2001. Distributed speech recognition with codec parameters. Dans les actes de *ASRU*.

[Reynolds, 2003] D. A. Reynolds, 2003. Channel robust speaker verification via feature mapping. Dans les actes de *ICASSP*, 53–56.

[Reynolds et al., 2000] D. A. Reynolds, T. F. Quatieri, et R. B. Dunn, 2000. Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing 10*, 19–41.

[Reynolds et Rose, 1995] D. A. Reynolds et R. C. Rose, 1995. Robust text-independent speaker identification using Gaussian Mixture speaker Models. *Speech and Audio Processing, IEEE Transactions on 3*(1), 72–83.

[Rosenberg, 1992] A. Rosenberg, 1992. The use of cohort normalized scores for speaker verification. Dans les actes de *ISCLP*, 599–602.

[Rosenberg et al., 1992] A. Rosenberg, J. Delong, C.-H. Lee, B.-H. Juang, et K. Soong, 1992. The use of cohort normalized scores for speaker verification. Dans les actes de *ICSLP*.

[Rosenberg et Sambur, 1975] A. Rosenberg et M. Sambur, 1975. New techniques for speaker verification. *IEEE transactions on Acoustics, Speech and Signal Processing 23*(2), 169–176.

[Rosenberg et Soong, 1992] A. E. Rosenberg et F. K. Soong, 1992. *Advances in Speech Signal Processing*, Chapter Recent Research in Automatic Speaker Recognition, 701 – 738. Marcel Dekker.

[Scheffer, 2006] N. Scheffer, 2006. *Structuration de l'espace acoustique par le modèle générique pour la vérification du locuteur*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

[Scheffer et Bonastre, 2006] N. Scheffer et J.-F. Bonastre, 2006. Fusing generative and discriminative UBM-based systems for speaker verification. Dans les actes de *Proc. of the 2nd international workshop on MMUA (MultiModal User Authentification)*.

[Shriberg et al., 2005] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, et A. Stolcke, 2005. Modeling prosodic feature sequences for speaker recognition. *Speech communication 46*(3-4), 455–472.

[Shynk, 1992] J. Shynk, 1992. Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine 9*(1), 14–37.

[Solomonoff et al., 1998] A. Solomonoff, A. Mielke, M. Schmidt, et H. Gish, 1998. Clustering speakers by their voices. Dans les actes de *ICASSP*.

[Song et Rosenberg, 1986] F. Song et A. Rosenberg, 1986. On the use of instantaneous and transitional spectral information in speaker recognition. Dans les actes de *ICASSP*, 877–880.

[Soong et al., 1985] F. Soong, A. Rosenberg, L. Rabiner, et B. Juang, 1985. A vector quantization approach to speaker recognition. Dans les actes de *ICASSP*, Volume 10.

[Sturim et Reynolds, 2005] D. Sturim et D. Reynolds, 2005. Speaker adaptive cohort selection for Tnorm in text-independent speaker verification. Dans les actes de *ICASSP*.

[Supplee et al., 1997] L. Supplee, R. Cohn, J. Collura, et A. McCree, 1997. MELP : the new federal standard at 2400 bps. Dans les actes de *ICASSP*.

[Tierney, 1980] J. Tierney, 1980. A study of LPC analysis of speech in additive noise. *IEEE transactions on Acoustics Speech and Signal Processing 28*(4), 389– 397.

[Tsuge et al., 2002] S. Tsuge, S. Kuroiwa, M. Shishibori, F. Ren, et K. Kita, 2002. Robust feature extraction in a variety of input devices on the basis of ETSI standard DSR front-end. Dans les actes de *ICSLP*.

[Turk et Pentland, 1991] M. Turk et A. Pentland, 1991. Face recognition using eigenfaces. Dans les actes de *IEEE CVPR*, 586–591.

[Vair et al., 2006] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, et P. Laface, 2006. Channel factors compensation in model and feature domain for speaker recognition. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

[Vair et al., 2007] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, et P. Laface, 2007. Loquendo - politecnico di torino's 2006 NIST speaker recognition evaluation system. Dans les actes de *Interspeech*.

[Van Leeuwen, 2004] D. Van Leeuwen, 2004. Speaker adaptation in the NIST speaker recognition evaluation 2004. Dans les actes de *Interspeech*.

[van Leeuwen et Brummer, 2007] D. van Leeuwen et N. Brummer, 2007. An introduction to application independent evaluation of speaker recognition systems. Dans les actes de *Speaker Classification (1)*, 330–353.

[Viikki et Laurila, 1998] O. Viikki et K. Laurila, 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication 25*, 133–147.

[Vlaj et al., 2005] D. Vlaj, B. Kotnik, B. Horvat, et Z. Kacic, 2005. A computationnaly efficient Mel-filter bank algorithm for distributed speech recognition systems. *EURASIP Journal on Applied Signal Processing 29*(4), 487–497.

[Vogt et al., 2005] R. Vogt, B. Baker, et S. Sridharan, 2005. Modelling session variability in text-independent speaker verification. Dans les actes de *Eurospeech*.

[Vuuren, 1996] S. Vuuren, 1996. Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. Dans les actes de *ICSLP*, 1788–1791.

[Wan et Campbell, 2000] V. Wan et W. M. Campbell, 2000. Support vector machines for speaker verification and identification. Dans les actes de *Neural Networks for Signal Processing*, Volume 2, 775–784.

[Weddin et Winther, 2006] M. Weddin et O. Winther, 2006. Frame selection for speaker identification. Dans les actes de *ICASSP*.

[Wiener, 1949] N. Wiener, 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley.

[Woodland et Povey, 2002] P. Woodland et D. Povey, 2002. Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. *Computer Speech and Language 16*, 25–47.

[Wu et al., 2003] T. Wu, L. Lu, K. Chen, et H.-J. Zhang, 2003. UBM based incremental speaker adaptation. Dans les actes de *ICASSP*.

[Xiang et al., 2002] B. Xiang, U. Chaudhari, G. Ramaswamy, et R. Gopinath, 2002. Short-time Gaussianization for robust speaker verification. Dans les actes de *ICASSP*, 681–684.

[Xie et al., 2006] Y. Xie, M. Liu, Z. Yao, et B. Dai, 2006. Improved two-stage wiener filter for robust speaker identification. Dans les actes de *ICPR '06 : Proceedings of the 18th International Conference on Pattern Recognition*, 310–313. IEEE Computer Society.

[Yin et al., 2006] S.-C. Yin, P. P. Kenny, et R. Rose, 2006. Experiments in speaker adaptation for factor analysis based speaker verification. Dans les actes de *Odyssey - The Speaker Recognition Workshop*.

Publications Personnelles

[Bonastre et al., 2008] N. Bonastre, J-F. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, et J. Mason, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. Dans les actes de *Odyssey*.

[Preti, 2007] A. Preti, 2007. Utilisation de mesures de confiance pour l'adaptation non supervisée des modèles de locuteurs en vérification du locuteur. Dans les actes de *Rencontres Jeunes Chercheurs en Parole (RJCP)*, 116–119.

[Preti et Bonastre, 2006] A. Preti et J.-F. Bonastre, 2006. Unsupervised model adaptation for speaker verification. Dans les actes de *International Conference on Spoken Language Processing (Interspeech - ICSLP)*, 2090–2093.

[Preti et al., 2006] A. Preti, F. Capman, et J.-F. Bonastre, 2006. A continuous unsupervised adaptation method for speaker verification. Dans les actes de *International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE), Springer ISBN : 978-1-4020-261-2*.

[Preti et al., 2008a] A. Preti, F. Capman, B. Ravera, et J.-F. Bonastre, 2008a. An application constrained front-end for speaker verification. Dans les actes de *European Signal Processing Conference (EUSIPCO)*.

[Preti et al., 2008b] A. Preti, F. Capman, B. Ravera, et J.-F. Bonastre, 2008b. Surveillance vocale de réseaux de communications professionnels par la reconnaissance du locuteur. Dans les actes de *Journées d'Etudes sur la Parole (JEP)*.

[Preti et al., 2007] A. Preti, F. Matrouf, D.and Capman, B. Ravera, et J.-F. Bonastre, 2007. Confidence measure based unsupervised target model adaptation for speaker verification. Dans les actes de *International Conference on Spoken Language Processing (Interspeech - ICSLP)*, 754–757.

[Preti et al., 2006] A. Preti, N. Scheffer, et J.-F. Bonastre, 2006. Discriminant approaches for GMM-based speaker detection systems. Dans les actes de *Proceedings of the 2nd international workshop on MMUA (MultiModal User Authentification)*.