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

THÈSE

Structuration de l'espace acoustique par le modèle générique pour la vérification du locuteur

Présentée pour obtenir le grade de Docteur en Sciences
de l'Université d'Avignon et des Pays de Vaucluse

SPÉCIALITÉ : INFORMATIQUE

par

NICOLAS SCHEFFER

Soutenue publiquement le 18 décembre 2006 devant un jury composé de :

MM.	Frederic BIMBOT	CR1/CNRS, IRISA/INRIA, Rennes	Rapporteurs
	Patrick KENNY	Professeur, CRIM, Montreal, Canada	
MM.	Claude BARRAS	Maitre de conferences, LIMSI, Orsay	Examinateurs
	Samy BENGIO	Senior Researcher, IDIAP, Martini, Suisse	
	Edouard GEOFFROIS	Docteur, DGA, Arcueil	
	John MASON	Professeur, UWS, Swansea, Royaume-Uni	
M.	Jean-François BONASTRE	Maître de Conférences HdR, LIA, Avignon	Directeur de thèse

École Doctorale Science et Agronomie
Laboratoire Informatique d'Avignon

Résumé

La Vérification Automatique du Locuteur (VAL) consiste à confirmer ou infirmer par sa voix l'identité proclamée d'un individu. Pour cette tâche, il est nécessaire d'estimer la probabilité pour qu'un locuteur ne corresponde pas à l'identité qui a été proclamée. Dans le paradigme de modélisation par GMM, un modèle générique (ou modèle du monde) est employé à cet effet. Cependant, son utilisation va bien au delà de cette simple estimation. Il joue, en effet, un rôle structurant pour l'espace acoustique, à travers l'estimation des modèles de locuteur et les calculs de statistiques. Les travaux présentés dans cette thèse s'inscrivent dans le cadre de la VAL et sont orientés autour d'un axe principal : l'intégration du modèle générique au sein des nouveaux formalismes apparus ces dernières années. Parmi ceux-ci, deux catégories ont retenu notre attention, les systèmes s'appuyant sur une approche discriminante et les systèmes modélisant des caractéristiques du locuteur autres que celles issues de l'enveloppe spectrale à court terme (appelés systèmes « haut-niveau »). La première contribution de nos travaux consiste à représenter le signal par des événements acoustiques issus du modèle générique et à analyser la séquence de ces événements dont la dynamique est spécifique du locuteur. Ces événements acoustiques sont, de fait, indépendants de la structure de la langue et n'ont pas de signification linguistique propre. Ce système, nommé AES (Acoustic Event Sequences), présente des performances similaires aux systèmes haut niveau basés sur une analyse phonétique ou lexicale. La seconde contribution vise à l'élaboration d'un système employant une modélisation discriminante structurée par le modèle générique. Notre contribution se distingue par l'expression du problème de vérification dans une faible dimension et exploitant la capacité de modélisation du modèle générique. Les performances du système, nommé SVM-UBM, obtenues sont proches des systèmes standards. Les contributions proposées ouvrent de nombreuses perspectives attrayantes notamment l'unification des deux approches majeures présentées dans ce document, perspective qui représente une prochaine étape vers un paradigme riche en informations, dynamiques et discriminantes.

Abstract

Automatic Speaker Verification (ASV) consists in accepting or rejecting a person's claimed identity from their voice. For this task, it is usually required to estimate the probability that a speaker does not correspond to a given identity. In the case of a GMM modelling paradigm, a generic model (called world model) is employed for this estimate. However, the world model is used to a greater extent as it plays a structural role for speaker model estimation and statistics computation. The work presented in this thesis is centered on ASV and has one main axis : the integration of the generic model into new approaches that have appeared in the literature in recent years. Among these, there are two distinct categories of particular interest : first, a category of systems adopting a discriminant approach and second, a category of systems modeling speaker characteristics other than the shot-term spectral envelop (called "high-level" systems). The first contribution of the work described in this thesis aims at representing the signal by acoustic events which have been extracted from the generic model and at analysing event sequences of which the dynamic is speaker specific. These acoustic events are inherently independent from the language structure and do not have any linguistic significance. This system, called AES (Acoustic Event Sequences), achieves similar levels of performance compared to high-level systems based on phonetic or lexical analysis. The second contribution aims at implementing a system, structured by the generic model, with a discriminant approach. The novel aspect of this contribution lies in the expression of the verification problem in a small dimension while exploiting the modeling capacity of the generic model. The performance of this system, called SVM-UBM, is close to state-of-the-art GMM based systems. The proposed methods expose numerous interesting perspectives among which is the unification of the two main approaches detailled in this document. This perspective represents one step towards an innovative modeling approach, rich of with discriminant and dynamic information.

Remerciements

Plus qu'un document et un travail de recherche, une thèse est une expérience personnelle unique où les relations sociales prennent une place centrale dans sa réussite. Aussi, aux vues des innombrables rencontres que j'ai pu faire, et des personnes qui m'ont directement ou indirectement aidé ou apporter leur soutien, une page de document ne suffirait pas pour exprimer toute ma gratitude.

Je pense d'abord à mon environnemznt de travail au LIA, qui est sans commune mesure, un laboratoire chaleureux, ouvert et où les amitiés se lient facilement. Mes remerciements vont donc à tous mes collègues permanents, anciens et de passage du LIA.

J'exprime toute ma gratitude à mon directeur de thèse, Jean-François Bonastre, MdC/HdR au LIA, pour son soutien et sa confiance en mon travail et en mes idées. Sa direction de thèse m'a permis, au bout de trois années, d'acquérir la confiance nécessaire pour ma future carrière. Je le remercie de m'avoir permis de participer à de multiples conférences nationales et internationales, ce qui a fait de ma thèse une expérience riche et unique.

Je tiens sincérement à remercier les personnes qui m'ont fait confiance dès le départ, Jean-François Bonastre et Edouard Geoffrois, DGA. Merci à mes rapporteurs Patrick Kenny, Frédéric Bimbot, et membres du jury John Mason, Claude Barras, Samy Bengio. Je n'aurais pas pu espérer un jury d'une telle qualité.

Au sein du LIA, je tiens à remercier Renato De Mori et Marc El-Bèze, ex et actuel directeur. Remerciements particuliers à Sylvain Meigner et Téva Merlin, pour m'avoir guider au début de ma thèse, Benoit Favre (vive l'intranet), Driss Matrouf, tout spécialement pour les innombrables pauses cafés-cigarettes agrémentées de discussions scientifiques et d'idées nouvelles. Merci à Will, Alex, Corinne, Lolo, Laurianne, les Naths, les Christians, Jens, j'en oublis tellement !

Merci à ma famille, sans laquelle, je n'aurai jamais eu la possibilité d'aller jusqu'au bout d'un cursus universitaire sans leur soutien, tout au long de ces années. Merci à mon grand-père Auguste. Merci à Nicole pour son soutien et son amour au quotidien pendant ces quatre années à Avignon.

Merci enfin à tous ceux que j'ai omis de citer.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

Introduction Générale

Dans l'éventail des médias à sa portée, l'homme a toujours privilégié la parole, de par sa nature universelle et sa convivialité. La communication parlée constitue donc le moyen privilégié de communication entre les hommes. Elle n'en est pas moins très complexe et présente une richesse qui intéresse depuis toujours de nombreux domaines de recherche. En effet, la parole transporte le message linguistique lui-même mais aussi de nombreuses informations non verbales, transmises volontairement ou de manière automatique. La parole véhicule notamment des informations sur la personne à l'origine de ce signal, son identité en particulier, mais aussi ses origines géographiques et/ou son état émotionnel et pathologique. De nombreux intérêts scientifiques et industriels découlent de ces informations. Parmi ceux-ci, l'authentification de l'utilisateur par sa voix est actuellement un domaine en forte croissance. Il s'inscrit en effet dans la mouvance actuelle des techniques d'authentification biométrique, aptes à identifier une personne par des mesures morphologiques et comportementales (dont la voix fait partie). Une des explications de cet engouement pour l'authentification vocale tient à un facteur de disponibilité : la voix, outre sa convivialité déjà soulignée, est souvent la seule modalité biométrique disponible, dans les applications liées aux serveurs téléphoniques par exemple.

L'automatisation des systèmes d'authentification, en vue de leur déploiement à grande échelle, constitue par conséquent un enjeu important. Dans ce cadre, les systèmes de Reconnaissance Automatique du Locuteur (RAL) s'appuient sur les caractéristiques de la parole permettant de reconnaître les individus. Parmi les différentes tâches de RAL, la Vérification Automatique du Locuteur (VAL) consiste à vérifier l'identité proclamée par un individu. La vérification du locuteur, tâche privilégiée de cette thèse, a de nombreux défis technologiques à relever pour pouvoir être intégrée dans les applications de la vie courante. Les variations du canal téléphonique, par exemple, sont à l'origine de nombreuses perturbations dégradant les performances des systèmes automatiques. La variabilité intrinsèque à la parole est, elle aussi, une source de dégradation reconnue. Au delà de ces variabilités, une des limitations majeures des systèmes actuels tient à la modélisation du locuteur par des informations correspondant à l'enveloppe spectrale à court-terme. La recherche d'informations nouvelles constitue un des principaux thèmes de recherche actuels.

PROBLÉMATIQUE

Les systèmes de RAL actuels reposent majoritairement sur des approches probabilistes. Parmi ces approches, les systèmes « état de l'art » sont généralement basés sur une modélisation des locuteurs par des modèles génératifs, comme les modèles à mélange de gaussiennes (GMM), associée à une représentation du signal basée sur des paramètres cepstraux (enveloppe spectrale à court-terme). Les systèmes les plus performants utilisent classiquement un modèle générique, également appelé modèle du monde, ou *UBM* (*Universal Background Model*), pour représenter le modèle du non-locuteur. Depuis quelques années, afin de répondre aux défis présentés précédemment, les systèmes de VAL ont évolué selon deux tendances majeures :

- la première consiste à mieux modéliser les variabilités des locuteurs et du canal de transmission survenant au cours d'enregistrements successifs. Ces méthodes ont nécessité l'incorporation de grandes quantités de données ainsi que l'augmentation de la complexité des modèles, afin de modéliser et de normaliser ces variabilités ;
- la seconde tient à la nature des systèmes de VAL actuels qui associent généralement une multitude de systèmes différents, chacun traitant d'une source d'information spécifique ou apportant une nouvelle manière de modéliser les locuteurs par l'adoption de classifieurs de nature différente. Ainsi, la caractérisation du locuteur par des informations linguistiques ou syntaxiques (comme les phonèmes ou le lexique utilisé) et le développement d'approches discriminantes pour la modélisation sont les thèmes récurrents des travaux de recherche de la communauté. Le gain en performance est recherché par la fusion des informations nouvelles et complémentaires issues de ces différentes approches, au prix d'un accroissement notable de la complexité. Il est en effet nécessaire, d'une part, de mettre au point séparément chacun des systèmes, puis d'autre part, d'élaborer des méthodes robustes de combinaison de l'information.

Les approches présentées précédemment souffrent donc principalement d'une hétérogénéité dans leur formalisme qui complique la mise au point des systèmes et leur fusion. L'objectif principal de cette thèse est de pallier en partie ce problème en nous appuyant sur la capacité de modélisation du modèle générique utilisé dans les approches standards.

CONTRIBUTIONS

Dans ces travaux de thèse, nous proposons l'intégration du modèle générique dans les nouveaux formalismes décrits précédemment. La première contribution majeure consiste à représenter le signal par des événements acoustiques issus du modèle générique, et à analyser la séquence de ces événements dont nous pensons que la dynamique est spécifique du locuteur. Ces événements acoustiques sont, de fait, indépendants de la structure de la langue (sons voisés, phonèmes...) et n'ont pas de signification linguistique propre. Cette approche a été publiée dans [Scheffer et Bonastre, 2005] et [Scheffer et Bonastre, 2006a].

La deuxième contribution importante de cette thèse consiste à unifier les paradigmes des approches de modélisation générative et discriminante en utilisant le modèle générique comme élément central. Notre contribution se distingue d'autres travaux visant des objectifs similaires par la proposition d'un système exprimant le problème de vérification dans un espace de faible dimension, structuré par un seul modèle génératif. À l'instar du formalisme de représentation relative du locuteur (vu par exemple dans les approches de type « modèles d'ancrage »), nous proposons une

représentation relative du locuteur dans l'espace des scores du modèle génératif. Cette approche repose sur l'adaptation des techniques utilisées dans les systèmes « haut-niveau » pour laquelle une représentation du locuteur réduite mais pertinente pour la VAL est proposée. Cette approche a été publiée dans [Scheffer et Bonastre, 2006b].

ORGANISATION DU DOCUMENT

Le premier chapitre présente les principes des systèmes de reconnaissance automatique du locuteur et les techniques couramment employées. Les différentes tâches applicatives de la RAL ainsi que les techniques d'évaluation des systèmes sont détaillées. Les différentes paramétrisations du signal de parole ainsi que les techniques permettant de réduire l'influence du canal et la variabilité des locuteurs sont également présentées. Nous détaillons enfin les techniques visant à rendre plus robuste le réglage d'un seuil global pour le système de vérification.

La suite du document se divise en deux parties.

La première partie présente les différents types de systèmes de VAL utilisés à l'heure actuelle, en particulier dans le cadre des campagnes d'évaluations internationales. Dans le chapitre 2, nous abordons les techniques adoptant une modélisation générative du locuteur à base de modèles de mélange de gaussiennes (GMM). Nous insistons sur le rôle structurant du modèle générique tout au long du processus de VAL pour ce paradigme. Le chapitre 3 présente ensuite les méthodes discriminantes de modélisation du locuteur, basées sur les machines à vecteurs supports (SVM). Alors que le modèle générique semble intervenir peu dans ces méthodes, nous insistons sur le rôle des données de pseudo-imposteurs — utilisées classiquement pour apprendre le dit UBM — qui jouent quant à elles un rôle de normalisation non négligeable. Le chapitre 4 présente, de façon non-exhaustive, les systèmes dits « haut-niveau » apparus depuis quelques années. Nous présentons les différentes unités segmentales utilisées par ces systèmes, ainsi que les approches multilingues où plusieurs reconnaisseurs de langues différentes sont sollicités en parallèle. Ces approches sont divisées en deux catégories suivant leur manière d'analyser ces unités. Ainsi, nous définissons les systèmes à caractère « intra-segmental » comme ceux s'attachant à modéliser acoustiquement ces unités, alors que le caractère « inter-segmental » des systèmes de la deuxième classe traduit leur attachement à modéliser la dynamique de ces unités. Nous insistons sur le fait que ces systèmes, basés sur des caractéristiques linguistiques, s'attachent plus à générer une représentation du signal caractéristique du locuteur plutôt qu'à décoder le message linguistique proprement dit.

La seconde partie est dédiée au contexte expérimental de nos travaux et aux contributions apportées tout au long de la thèse. Le chapitre 5 présente le protocole expérimental utilisé pour les expériences. Celui-ci est basé sur des sous-ensembles des bases de données des corpus d'évaluations du NIST-SRE. Nous détaillons ensuite le système de référence du LIA, ALIZE/LIA_SpkDet, ainsi que son évolution au cours des années. Le chapitre 6 présente ensuite notre approche « haut-niveau » basée sur l'UBM et appelée *AES* pour *Acoustic Event Sequences*. Le chapitre 7 présente l'extension multi-classes de cette approche, nommée *C-AES* pour *Class-dependent AES*. Nous proposons, dans ce cadre, l'utilisation de plusieurs jeux d'événements à différentes résolutions, non pas temporelles, mais acoustiques. Le chapitre 8 présente l'élaboration d'un système de VAL adoptant une modélisation discriminante du locuteur structurée par le modèle générique. L'architecture adoptée constitue une réponse aux problèmes d'hétérogénéité et de complexité usuellement liés aux systèmes intégrant les paradigmes génératif et discriminant. Cette approche a été nommée

SVM-UBM. Nous concluons finalement en résumant nos contributions et nos principaux résultats ainsi qu'en ouvrant des perspectives pour des futurs travaux de recherche.

CADRE DE TRAVAIL

Ces travaux ont été réalisés dans le cadre d'un financement de thèse de la Délégation Générale de l'Armement DGA-CNRS[1]. Le laboratoire d'accueil a été le LIA[2], au sein de l'équipe « Traitement du Langage Naturel Oral ».

La dynamique de la thèse a été sans cesse soutenue par les campagnes d'évaluations internationales, ainsi que l'implication dans le projet ALIZE/LIA_SpkDet, réalisé dans le cadre du projet AGILE du programme *Technolangue*.

Une part non négligeable du travail présenté dans cette thèse découle directement des améliorations et des extensions apportées à la plate-forme ALIZE/LIA_SpkDet en vue de la participation annuelle aux campagnes d'évaluations du NIST-SRE (2004,2005 et 2006) mais aussi à la campagne ESTER pour la tâche de Suivi de Locuteur. Ces participations régulières ont permis d'évaluer dans ce cadre strict les différentes contributions au domaine présentées dans ce document. Au delà de ces aspects, ce travail de thèse s'est également inscrit dans une dynamique d'équipe pour le développement et la diffusion de la plate-forme ALIZE/LIA_SpkDet au sein de la communauté (notons par exemple, que notre système est le système de référence biométrie-voix pour le réseau d'excellence NOE-BIOSECURE).

[1] Délégation Générale pour l'Armement : http://www.recherche.dga.defense.gouv.fr/
[2] Laboratoire Informatique Avignon : http://www.lia.univ-avignon.fr/

CHAPITRE 1

Reconnaissance Automatique du Locuteur : Principes et Évaluations des systèmes

Sommaire

Ce chapitre est consacré aux principes de la reconnaissance automatique du locuteur et des approches dominantes utilisées dans ce domaine. Nous présentons dans un premier temps quelques notions concernant l'authentification biométrique ainsi que les applications dans lesquelles elle intervient. Ensuite, nous détaillons les différentes tâches de la reconnaissance du locuteur qui peuvent être définies selon l'application visée. Nous exposons les particularités de la biométrie « voix » en soulignant les différentes variabilités auxquelles celle-ci est soumise. Nous nous concentrons ensuite sur la structure générique d'un système de reconnaissance pour en présenter les méthodes d'évaluation des performances. Enfin, nous abordons les techniques standards utilisées pour traiter le signal audio, normaliser les variabilités et régler le seuil de décision pour accepter ou rejeter un test de vérification.

1.1 Généralités sur l'authentification biométrique

Biométrie : bios=vivant , metron=mesure

Pour identifier une personne, trois approches (éventuellement complémentaires) sont possibles :

- Utiliser un identifiant : ce que l'on possède (carte, badge, document) ;
- Utiliser une connaissance : ce que l'on sait (un mot de passe) ;
- Utiliser une biométrie : ce que l'on est.

Les deux premiers points diffèrent de la biométrie au sens où ils sont concernés par la possession d'un élément (mots de passe, ou clés) qui peut être utilisé par des fraudeurs pour usurper l'identité d'un tiers, tandis que la biométrie correspond à des critères morphologiques de l'être humain, qui doivent être uniques pour chaque personne.

1.1.1 Définition et catégorisation

Le terme biométrie est défini dans le *Petit Robert* par la *Science qui étudie à l'aide des mathématiques (statistiques, probabilités) les variations biologiques à l'intérieur d'un groupe déterminé*. Le terme « biométrie » se réfère étymologiquement à la mesure du vivant. Il serait plus naturel de faire référence au terme français exact d'« anthropométrie » pour *les techniques de mensurations du corps humain et de ses diverses parties*. L'anthropométrie judiciaire fait d'ailleurs référence *aux méthodes d'identification des criminels par ces mensurations*.

L'emploi du terme biométrie permet d'étendre le terme d'anthropométrie en englobant également les techniques visant à caractériser une personne par des traits comportementaux. C'est pourquoi les modalités biométriques sont subdivisées selon un critère de variation dans le temps. Les biométries invariantes au cours du temps pour un individu, tenant de la morphologie humaine, sont souvent nommées « morphologiques » ou « statiques ». Celles qui présentent une grande variation temporelle sont nommées « comportementales » ou « dynamiques ».

Quelques exemples sont répertoriés ci dessous :

- morphologiques : l'iris, les empreintes digitales, les empreintes palmaires, le visage ;

– comportementales : signature (image et dynamique), frappe au clavier, démarche et voix (cette dernière est souvent considérée comme appartenant aux deux catégories).

1.1.2 Du rôle de la multimodalité

L'authentification biométrique multi-modale constitue un enjeu important pour les prochaines années. L'intérêt croissant pour la multi-modalité tient à plusieurs facteurs. Premièrement, la combinaison naturelle de différentes sources d'informations permet d'augmenter les performances d'authentification. Ensuite, la « disponibilité » d'une biométrie parmi d'autres est accrue, *i.e.* le système peut changer de modalité dès lors qu'une modalité donnée devient indisponible.

De plus, la multi-modalité permet de diminuer les contraintes utilisateurs liées au processus d'authentification. L'acquisition de l'image d'un iris, par exemple, est souvent très contraignante. La voix peut être plus aisément acquise par des microphones. Finalement, pour des systèmes réels sous certaines conditions d'utilisation, il a été remarqué que le contrôle d'accès couplé à l'utilisation de l'iris (une biométrie très performante) est mal perçu par les personnes utilisant le système (problème d'acceptabilité de la modalité biométrique). Il y a donc un compromis à gérer entre les taux d'erreurs présentés pour une modalité et son taux d'acceptabilité. La voix et le visage sont des modalités dont les taux d'erreurs sont beaucoup plus élevés que l'iris, mais qui sont en revanche très bien acceptées par les utilisateurs.

1.1.3 La voix comme modalité biométrique

L'application évidente d'un système d'authentification par la voix est le contrôle d'accès au sein d'un bâtiment ou d'un système informatique. À cela s'ajoute aujourd'hui, les applications de vérification d'identité lors des transactions téléphoniques et les applications criminalistiques (et de renseignement), comme les écoutes sur les réseaux téléphoniques. Au delà des applications concernant l'authentification à des fins de sécurité, d'autres domaines se sont intéressés à l'identification du locuteur où les taux d'erreurs actuels sont suffisamment bas pour répondre à certaines applications commerciales. Par exemple, on pourra trouver des modules de reconnaissances du locuteur pour le chargement automatique d'un profil utilisateur. Dans le domaine du multimédia, l'augmentation drastique des quantités de données radiodiffusées et télévisuelles nécessite, au delà de l'archivage, une indexation permettant d'effectuer des recherches dans les flux. Avec l'identification vocale, il est alors possible de rechercher le discours d'une personne en envoyant une requête comportant son identité.

La mise en œuvre d'un système de reconnaissance automatique du locuteur (*RAL*) est une tâche difficile et qui, à l'heure actuelle, ne peut rivaliser avec des systèmes à base d'empreintes digitales ou génétiques. La communauté scientifique met en garde le système judiciaire sur l'utilisation abusive de la voix comme preuve tangible d'authentification de l'individu, via des publications [Bonastre et al., 2003], ou des associations (l'Association Française de la Communication Parlée)[1]. Le message principal consiste à insister sur un « devoir de précaution », sachant qu'il n'y a pas de preuve scientifique permettant d'affirmer l'existence d'une « empreinte vocale » qui aurait la même unicité qu'une empreinte digitale. Le terme « signature vocale » sera plus justement employé,

[1] AFCP : http://www.afcp-parole.org

permettant de souligner le caractère non reproductible d'une production d'un signal audio par l'appareil vocal humain.

1.2 Applications et tâches pour la RAL

La reconnaissance automatique du locuteur consiste à reconnaître l'identité d'un individu à partir de sa voix [Doddington, 1985; Rosenberg et Soong, 1991]. Les applications des systèmes de RAL se distinguent par leur contexte applicatif et leur niveau de sécurité. Ces contraintes peuvent être prises en compte pour la définition d'une tâche spécifique de la RAL. Il est communément admis de regrouper ces tâches dans des grandes catégories : identification, vérification et indexation. Les différences entre ces tâches amènent des stratégies de décision alternatives selon que le nombre de locuteurs soit connu, que l'identité du locuteur soit proclamée ou que le système soit de nature ouverte ou non à des utilisateurs inconnus. Dans la suite de ce paragraphe, nous présentons brièvement les principales tâches associées à la RAL.

1.2.1 Identification Automatique du Locuteur

A partir d'un ensemble de locuteurs référencés dans le système, la tâche d'Identification Automatique du Locuteur (*IAL*) consiste à déterminer l'identité du locuteur présent dans un signal vocal (signal de test) [Atal, 1976; O'Shaughnessy, 1986]. Deux conditions d'identification sont connues : milieu ouvert ou fermé. Dans le premier cas, seuls des locuteurs connus peuvent accéder au système quand, dans le second, le système peut émettre une réponse de type « rejet » correspondant à l'hypothèse où aucun des locuteurs référencés dans la base n'est présent dans le signal.

En milieu fermé, chaque accès de test est comparé à tous les modèles de locuteurs référencés dans le système. L'identité du locuteur possédant la référence la plus proche est émise en sortie du système. Les performances d'un système d'IAL se dégradent au fur et à mesure que la population de locuteurs concernés augmente.

1.2.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (*VAL*) consiste à vérifier l'identité proclamée par un individu par la comparaison d'un signal vocal et d'un modèle de référence du locuteur présumé, préalablement appris par le système. Un système de VAL a donc deux entrées : une identité et un accès de test. Le résultat de cette comparaison est considéré comme une mesure de similarité avant d'être comparé à un seuil d'acceptation. Lorsque la mesure de similarité est supérieure à ce seuil, l'individu est accepté, il est rejeté dans le cas contraire. Le tutoriel en vérification du locuteur proposé par [Bimbot et al., 2004] donne une bonne introduction aux approches courantes de la VAL.

1.2.3 Indexation en Locuteur

L'indexation en locuteur de document audio regroupe plusieurs tâches : segmentation, regroupement et suivi de locuteur. La segmentation en locuteur consiste à découper le signal audio en segments homogènes du point de vue du locuteur. Le regroupement en locuteur consiste à déterminer le nombre des locuteurs présents dans l'enregistrement et les interventions associées à chacun d'entre eux. Ces deux tâches réunies constituent l'indexation en locuteur proprement dite, nommée *Speaker diarization*. La difficulté de la tâche provient du manque d'information *a priori* : nombre de locuteurs, identités ou un échantillon de voix pour établir une première référence ne sont pas disponibles. Au niveau applicatif, cette tâche s'inscrit dans le cadre des applications d'indexation en locuteurs de bases de données multimédia (télévisuelles, radiophoniques, réunions...)[Johnson, 1999; Delacourt, 2000].

Suivi de Locuteur La tâche de Suivi de Locuteur (*SVL*) [Sonmez et al., 1999; Bonastre et al., 2000], consiste, à partir d'un ensemble de locuteurs référencés, à délimiter et à identifier les zones du signal où chacun des locuteurs est intervenu. Il est possible d'avoir un seul locuteur à suivre dans un document où de multiples locuteurs interviennent. Les applications correspondant à ce type de tâche intéressent les gouvernements et les services de renseignement dans le cadre d'écoutes téléphoniques mais aussi la recherche d'information dans les documents multimédia.

1.3 Facteurs limitant : les variabilités du signal vocal

Le signal de parole est un signal complexe qui présente de nombreuses variabilités [Rossi, 1989]. Un signal de parole est unique car il est considéré comme une « performance » du fait de sa non-reproductivité au sein de l'appareil vocalique. Dans ce signal se mélangent informations linguistiques, informations caractéristiques du locuteur, informations relatives à la transmission...

Toutes ces informations présentent une grande variabilité et les recherches actuelles se portent essentiellement sur la maîtrise (ou la normalisation) de ces variabilités pour ne garder que l'information concernant le locuteur. En effet, la capacité des systèmes de RAL à authentifier une personne repose essentiellement sur la modélisation de la variabilité inter-locuteur, *i.e.* la modélisation de la variation entre différents individus. Les approches modernes tentent d'augmenter cette modélisation par la prise en compte de la variabilité intra-locuteur et la modélisation du canal de transmission.

Il est à noter que la plupart des systèmes de RAL à l'heure actuelle reposent sur une coopération totale entre le système et l'individu, ce qui n'est pas forcément le reflet de la réalité. Un système de RAL doit être robuste à ce facteur [Homayounpour, 1995].

Nous présentons brièvement les variabilités existant dans le signal de parole afin de mieux comprendre les enjeux sur lesquels repose la recherche en RAL.

1.3.1 Variabilités intra-locuteur

Les systèmes de RAL se basent sur le principe de la variation du signal de parole entre les individus. Il est cependant clair qu'une variation du signal de parole est aussi présente pour un même individu. Cette variation peut être due à beaucoup de facteurs dont :

- la nature intrinsèquement variable de la communication parlée ;
- des facteurs pathologiques de type fatigue, rhume,... ou émotionnels. Ces facteurs provoquent des altérations momentanées de la voix ;
- à plus long terme, une altération de la voix due à l'âge est présente chez tous les individus.

1.3.2 Facteur de distorsion

Les facteurs de distorsion dans le signal de parole peuvent être de deux natures, la première provient des facteurs environnementaux, la seconde provient des changements des conditions d'acquisition et de transmission entre deux enregistrements.

Lorsque la chaîne de transmission est maîtrisée et stable d'un enregistrement à un autre, les variations de celle-ci ne posent pas de problèmes particuliers et toute normalisation par rapport à ces variations est peu utile. Dans un cadre applicatif, ce n'est souvent pas le cas et cette variation doit être modélisée.

Le protocole opératoire peut apporter, de par sa définition, des variabilités entre deux enregistrements. Celles-ci apparaissent par la quantité de données disponible, ou par l'utilisation d'un microphone différent entre les deux enregistrements. Par exemple, pour une application embarquée dans un véhicule, l'enregistrement peut être fait chez l'utilisateur dans un environnement calme à l'aide d'un microphone de qualité alors que l'accès aux données se fait dans la voiture, la fenêtre potentiellement ouverte (environnement) avec un téléphone de qualité réduite (acquisition et transmission).

1.3.3 Contenu linguistique

Le contenu linguistique du message est une des informations présentes dans le signal de parole. Tirer profit de cette information, déjà difficile à extraire automatiquement, pour la reconnaissance du locuteur est une tâche complexe. NDans [Ramaswamy et al., 2003], cette information permet de faciliter et de rendre plus robuste la tâche de vérification en utilsant des informations personnelles au locuteur [Ramaswamy et al., 2003]. Une autre approche, proposée par [Doddington, 2001], repose sur la caractérisation des locuteurs par leurs « tics de langage », en mode indépendant du texte.

La connaissance du contenu linguistique prononcé permet de contraindre la production de la parole pour estimer de façon robuste les paramètres caractéristiques du locuteur. Cette approche est souvent adoptée dans des applications où le signal disponible est de courte durée (type mot de passe). L'analyse pourra se porter sur les caractéristiques des mots et des phonèmes connus à l'avance. Une sécurité supplémentaire sera ajoutée par la validation du mot de passe.

1.4 Structure et Évaluation des systèmes de RAL

La structure et le déploiement d'un système de reconnaissance du locuteur contiennent des similitudes fortes quelque soit la tâche choisie. Dans ce paragraphe, nous présentons les étapes principales communes à tous les systèmes de RAL et l'utilisation du test d'hypothèse bayésien pour prendre la décision.

1.4.1 Structure d'un système de RAL

Un système de RAL possède deux modes de fonctionnement :
- un mode *apprentissage*, où un modèle est estimé pour chaque locuteur « client » du système puis servira de référence pour les tâches de reconnaissance à venir ;
- un mode *test*, où l'étape de reconnaissance (vérification, identification...) est effectuée. En sortie de ce module, le système émet une réponse : une identité pour la tâche d'identification, une décision accès/rejet pour la vérification.

1.4.2 Prise de décision

Au delà de la structure commune entre les différentes tâches d'un système de RAL, la stratégie de décision est différente selon la tâche choisie. Le plus souvent, cette stratégie se formalise par dans un cadre bayésien que nous présentons ci-dessous pour les tâches de vérification et d'identification.

1.4.2.1 Identification du Locuteur

Pour la tâche d'identification du locuteur en milieu fermé, tous les locuteurs possibles sont connus. Il s'agit de déterminer l'identité du locuteur dans la phrase $\mathbf{X}$. La stratégie de décision consiste, à partir d'un ensemble de locuteur $\overline{S} = \{S_1, \ldots, S_n\}$, à fournir l'identité correspondant au modèle qui a la plus forte vraisemblance sur ces données. Sans information *a priori* sur l'apparition des locuteurs ($p(S_1) = \cdots = p(S_n)$) et dans le cas d'un milieu fermé, l'identité S^* est acceptée si elle remplit la condition suivante :

$$S^* = \arg\max_{S_i} p(S_i|\mathbf{X}) = \arg\max_{S_i} p(\mathbf{X}|S_i). \tag{1.1}$$

1.4.2.2 Vérification du Locuteur

Dans le cadre d'une application de vérification du locuteur, la stratégie de décision est basée sur un test d'hypothèse. Étant donné un segment de parole $\mathbf{X}$ et une identité proclamée S, le test d'hypothèse est défini de la manière suivante :
- H_0 : $\mathbf{X}$ provient du locuteur dont l'identité S a été proclamée ;
- H_1 : $\mathbf{X}$ ne provient pas du locuteur dont l'identité S a été proclamée.

Le rapport de vraisemblance (*Likelihood Ratio* (LR)) entre ces deux hypothèses est estimé puis comparé à un seuil de décision θ :

$$LR(\mathbf{X}, H_0, H_1) = \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_1)} \lessgtr \theta \begin{cases} H_1 & \text{est accepté} \\ H_0 & \text{est accepté} \end{cases} . \tag{1.2}$$

H_0 correspond généralement à la vraisemblance du modèle de locuteur et H_1 à la vraisemblance du modèle de non-locuteur (*i.e.* tous les autres locuteurs possibles hormis S). La représentation de l'hypothèse inverse est difficile, elle est souvent remplacée par une approximation (*e.g.* un modèle générique, voir chapitre 2).

1.4.3 Évaluation d'un système VAL

La thèse étant centrée sur la tâche de vérification du locuteur (VAL), nous nous intéressons uniquement aux techniques d'évaluation pour cette tâche. L'évaluation de la qualité d'un système de VAL dépend de plusieurs facteurs. Ce sont les performances en termes de taux d'erreurs qui vont en déterminer la qualité.

Cependant, un système d'authentification en phase d'exploitation dépend aussi des échecs d'apprentissage, *i.e.* si pour des raisons de défauts matériels, ou parce que l'enregistrement est de trop mauvaise qualité pour servir à l'authentification, le système décide de rejeter le signal et de procéder à une nouvelle phase d'entraînement. Ces mesures d'échecs sont courantes dans les modalités comme l'iris ou les empreintes digitales, où les techniques d'acquisition jouent un rôle primordial. Elles sont en revanche peu prises en compte pour la parole, dans le cadre de travaux de laboratoire.

Un autre critère consiste à prendre en considération le corpus sur lesquels ces mesures ont été effectuées. En effet, un corpus comportant peu de variabilité ou un trop petit nombre de locuteurs peut guider à une mauvaise interprétation des résultats.

Dans la suite de ce paragraphe, nous nous concentrons sur l'évaluation d'un système de VAL par ses taux d'erreurs. Nous présentons les principales mesures ainsi que les divers corpus existants et les applications qui peuvent leur être associées.

1.4.3.1 Mesures des performances

Le test d'hypothèse de l'équation 1.2 possède quatre issues possibles. Les deux premières sont les authentifications réussies : en acceptant le bon locuteur ou en rejetant un imposteur. Les performances d'un systèmes de VAL s'évaluent cependant plutôt sur les deux types d'erreurs d'authentification : les fausses acceptations *FA*, correspondant à l'acceptation à tort d'un imposteur, et les faux rejets *FR*, correspondant au rejet à tort du locuteur correspondant à l'identité proclamée (locuteur client). C'est à partir des taux de rejet P_{FA} et P_{FR} que l'évaluation du système est réalisée.

$$P_{FA} = \frac{\text{Nombre d'imposteur acceptés}}{\text{Nombre d'accès imposteurs}}, \tag{1.3}$$

$$P_{FR} = \frac{\text{Nombre de client rejetés}}{\text{Nombre d'accès client}}. \tag{1.4}$$

1.4.3.2 Point de fonctionnement

L'évaluation d'un système en mode opérationnel doit être effectuée sur un point précis de fonctionnement, *i.e.* une valeur du seuil. Pour cela, une mesure pondérée des taux d'erreurs appelée fonction de coût de décision ou *Decision Cost Function (*DCF*)* est utilisée. Les pondérations relatives à cette mesure sont déterminées par l'application visée par le système. Les paramètres en jeu sont : les coûts associés à chaque taux d'erreur C_{FA}, C_{FR} et les probabilités *a priori* des populations imposteurs P_{Imp} et clients P_{Cl}. Cette fonction de coût s'exprime sous la forme :

$$DCF = C_{FR} \cdot P_{Cl} \cdot P_{FR} + C_{FA} \cdot P_{Imp} \cdot P_{FA}. \tag{1.5}$$

La DCF est souvent normalisée pour obtenir des valeurs entre 0 et 1. Cette mesure est utilisée pour évaluer les systèmes en mode opérationnel, *i.e.* lorsque le réglage du seuil de décision a été effectué. Dans la phase de développement, les taux d'erreurs sont fonction du seuil θ. Dans les expériences en laboratoire (et dans celles de ce document), la mesure de performance utilisée est celle qui minimise la fonction de coût $DCF_{\min} = \min_\theta DCF$. Nous assimilerons pour des notions de clarté la mesure $DCF_{\min}$ à la mesure DCF tout au long de ce document.

Lorsque les coûts associés aux erreurs sont identiques et qu'il n'existe pas d'*a priori* sur la population (i.e. $P_{Cl} = P_{Imp} = 0.5$) alors ce point de fonctionnement correspond à l'*HTER* (pour *Half Total Error Rate*). Une autre mesure largement utilisée pour mesurer les performances est l'*Equal Error Rate*, *EER* où les taux d'erreurs sont identiques $P_{FA} = P_{FR}$. Ces différentes mesures sont illustrées sur une courbe DET présentée en exemple à la figure 1.1.

1.4.3.3 Courbes caractéristiques de l'évaluation

La courbe DET (*Detection Error Tradeoff curve*) a été présentée par [Martin et Przybocki, 1997] comme une variante des courbes ROC (Receiving Operating Characteristic) [Oglesby, 1995], dont les axes sont les taux d'erreurs P_{FA} et P_{FR}. Ces courbes permettent la comparaison rapide des performances entre des systèmes sur tous les points de fonctionnement (*i.e.* toutes les valeurs possibles du seuil de décision θ). L'échelle des axes suit une loi normale et par conséquence, sous l'hypothèse de gaussianité des scores, la courbe DET est une droite [Auckenthaler et al., 2000]. Cette courbe illustrera la plupart des résultats expérimentaux présentés dans ce document[2].

[2]Il existe des alternatives à la courbe DET, comme l'EPC (*Expected Performance Curve*) [Bengio et Mariethoz, 2004] partant de l'hypothèse que la comparaison globale de deux systèmes uniquement par leur courbe DET présente un biais.

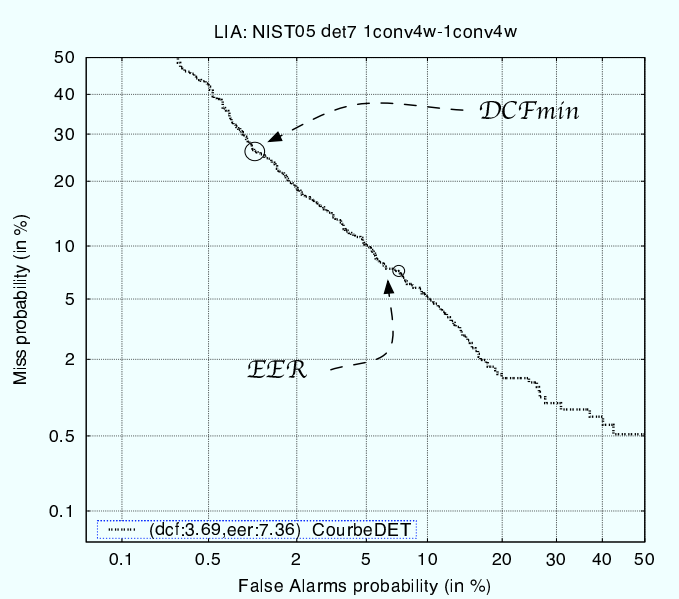


FIG. 1.1: *Exemple d'une courbe DET ainsi que les mesures de performances usuelles d'un système de VAL. La DCFmin représente la valeur du seuil où la fonction de coût est minimale, l'EER représente le point où les taux de faux-rejet et de fausses-acceptations sont égaux.*

1.4.3.4 Calibration des scores

Une technique nouvelle proposée par [Brümmer et du Preez, 2006] a été adoptée par le NIST[3] pour les prochaines évaluations. Cette méthode consiste à calibrer les scores d'un système de vérification afin d'obtenir des scores de confiance valables pour toute application (pour tout point de fonctionnement). Elle assure aussi une robustesse dans le réglage du seuil. Pour obtenir ce calibrage, l'auteur adopte une fonction de coût basée sur les rapports de vraisemblances, nommée C_{llr}, à minimiser. La mesure qui en découle représente les performances globales d'un système pour toutes les valeurs du seuil, et non sur un point précis de fonctionnement comme la mesure DCF.

1.4.3.5 Corpus et Campagnes d'évaluation

L'évaluation des systèmes tient aussi au corpus utilisé et au protocole expérimental adopté. Depuis 1996, l'institut américain NIST organise annuellement des campagnes d'évaluation des systèmes de VAL, en mode indépendant du texte [Przybocki et Martin, 1998; Martin et Przybocki, 1997]. Ces évaluations sont réalisées dans un contexte conversationnel en milieu téléphonique. A la suite de ces campagnes, un atelier réunissant tous les participants est organisé afin de partager les connaissances, de souligner les difficultés rencontrées et de préparer les prochaines éditions.

La tâche principale de ces campagnes est la détection du locuteur. A cette tâche il faut ajouter les différentes variantes existantes permettant de s'intéresser à des problèmes spécifiques, tels que :

[3]NIST : National Institue of Standards and Technologies, `www.nist.gov/speech`

– des variations dans les durées d'apprentissage et de test ;
– des variations de conditions d'enregistrement entre l'apprentissage et le test ;
– récemment, l'utilisation des microphones « exotiques », dits auxiliaires, pouvant atteindre des rapports signal sur bruit (SNR) très faibles ;
– des variations dans la langue utilisée par le même locuteur (locuteurs bilingues).

Afin d'entraîner les systèmes automatiques, les laboratoires sont libres d'utiliser toute base de données à leur disposition (généralement celles des campagnes précédentes). La phase d'évaluation est limitée dans le temps (environ 3 semaines). Durant cette phase, les participants reçoivent les données d'apprentissage pour de nouveaux locuteurs ainsi qu'une large série de tests à réaliser en aveugle. Chaque participant se doit de répondre à chaque test par un score de confiance et une décision de rejet ou d'acceptation (*true* ou *false*).

Il existe d'autres bases de données pour évaluer les qualités d'un système de VAL (*Polyvar* [Chollet et al., 1997], *Timit*[4],...), mais les campagnes NIST restent un outil majeur et robuste pour l'évaluation des systèmes de VAL.

1.5 Techniques communes aux systèmes de VAL par approche probabiliste

La thèse étant centrée sur la tâche de vérification du locuteur, nous présentons par la suite les techniques courantes utilisées par les systèmes actuels. Le lecteur pourra se référer à [Bimbot et al., 2004] pour une introduction plus détaillée aux techniques de la vérification du locuteur. La plupart de ces techniques sont cependant utilisables pour les autres tâches de reconnaissance.

Trois modules principaux composent un système générique de VAL :
– le module d'analyse acoustique (étape de paramétrisation) : ce module permet d'extraire une représentation du signal de parole appropriée pour la tâche de reconnaissance ;
– le module de modélisation se charge d'estimer les modèles des locuteurs. Ceux-ci seront stockés dans le système et serviront de référence pour les mesures de comparaison avec un signal de test ;
– le module de décision permet de comparer un signal de parole donné à un ou plusieurs modèles de locuteurs (selon l'application) en produisant une mesure de similarité pour la décision.

1.5.1 Paramétrisation du signal de parole

De par la complexité intrinsèque du signal de parole et la quantité d'informations présente dans ce signal, les techniques de reconnaissance automatique du locuteur n'utilisent pas ce signal sous sa forme brute. La littérature abonde de nombreux types de codage de la parole, dédiés aux tâches de reconnaissance permettant d'extraire au mieux l'information contenu dans ce signal[Furui, 1981; Reynolds, 1994; Hermansky, 1990].

[4] http://www.mpi.nl/world/tg/corpora/timit/timit.html

De par la propriété pseudo-stationnaire de la parole, ce codage est généralement réalisé périodiquement (*e.g.* toutes les 10 à 30 ms). Le flux d'information résultant est une suite de vecteurs de paramètres acoustiques dont le calcul nécessite généralement une fenêtre temporelle (type Hamming) de 20 à 30 ms. Nous présentons dans ce paragraphe les techniques de base d'extraction d'information utile à la reconnaissance et particulièrement les paramètres cepstraux. Nous évoquons finalement les techniques de normalisation de ces paramètres.

1.5.1.1 Paramètres de l'analyse spectrale

L'analyse spectrale permet d'extraire des paramètres représentatifs des caractéristiques de l'appareil phonatoire des individus. Trois classes de paramètres se distinguent :

- les analyses par bancs de filtres : il s'agit d'une analyse assez simple du système auditif de l'homme qui consiste à calculer l'énergie du signal vocal dans différentes bandes de fréquences ;
- Les analyses à base de transformée de Fourier : les coefficients issus de la transformée de Fourier peuvent être utilisés pour une analyse en bancs de filtres ainsi que pour les calculs de coefficients cepstraux. Parmi les plus connus se trouvent les coefficients LFCC (*Linear Frequency Cepstrum Coefficient*) ;
- Les analyses par prédiction linéaire [Grenier, 1977] : les coefficients issus d'un modèle autorégressif (de l'analyse LPC) peuvent être utilisés comme paramètres caractéristiques. Parmi les plus connus se trouvent les coefficients LPCC (*Linear Predictive Cepstrum Coefficient*).

Le lecteur pourra se reporter à [Oppenheim et Schafer, 1989] pour une présentation des divers codages de la parole.

Formalisme des coefficients cepstraux La représentation du signal par coefficients cepstraux est très utilisée dans les tâches de reconnaissance associées à la parole. Ces coefficients caractérisent bien la forme du spectre et permettent de séparer l'influence de la source du signal vocal de celle du conduit vocal. Cette séparation est rendue possible grâce à un filtre déconvolutif. Le cepstre est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale de puissance. Soit $s(t)$ le signal de parole, défini comme la convolution entre le signal glottique $e(t)$ et la réponse impulsionnelle du conduit vocal $h(t)$:

$$s(t) = e(t) * h(t). \tag{1.6}$$

Le logarithme de la densité spectrale de puissance est défini comme :

$$\log|S(f)| = \log|E(f)| + \log|H(f)|. \tag{1.7}$$

Par l'application transformée inverse, le domaine de cette représentation du signal est appelé quéfrentiel[5].

$$s'(q) = e'(q) + h'(q). \tag{1.8}$$

Les coefficients cepstraux ont la propriété d'être faiblement corrélés entre eux, ce qui permet d'éviter l'analyse de corrélation inter-coefficients lors de la modélisation (*e.g.* la modélisation fait généralement intervenir des matrices de covariance diagonales).

[5] Toutes les opérations de traitements du signal dans ce domaine sont définies en inversant les premières lettres de chaque mot (liftrage, cepstre, quéfrence, *etc*)

Echelle de Mel L'échelle de Mel est utilisée pour la représentation d'un signal audio basé sur la perception auditive de l'être humain. Les bandes de fréquences sont positionnées logarithmiquement et permettent de mieux approximer le système auditif humain que dans le cas d'un positionnement linéaire. Ce rôle perceptif est bien approprié pour les tâches de reconnaissance.

L'application de l'échelle de Mel sur des coefficients LFCC permet d'obtenir les coefficients MFCC, très employés en reconnaissance du locuteur. Il est à noter que l'échelle de Mel n'est pas utile dans la bande téléphonique. Les coefficients PLP sont les coefficients perceptifs dérivés des LPC. La littérature [Hermansky, 1990] donne des formules perceptives précises dont les coefficients ont été déterminés empiriquement. En pratique, l'application de l'échelle de Mel (ou Bark) permet d'obtenir des coefficients PLP similaires en performance pour les différentes tâches de reconnaissance. De manière étonnante, les paramètres les plus usités en reconnaissance du locuteur sont très proches des paramètres utilisés en reconnaissance de la parole (où paradoxalement l'objectif est d'être indépendant du locuteur).

Dérivées des coefficients cepstraux L'approche la plus répandue pour la prise en compte de la dynamique à court-terme du signal de parole est d'utiliser les dérivées des coefficients cepstraux. Cette méthode a été proposée par [Furui, 1981]. Les informations apportées par les dérivées des coefficients (d'ordre 1, 2, et même 3) sont couramment utilisées pour améliorer les performances dans les tâches de reconnaissance en général.

1.5.1.2 Paramètres prosodiques

Les paramètres prosodiques caractérisent en grande partie le style d'élocution d'un locuteur. Ces paramètres peuvent être la vitesse d'élocution (débit), durée et fréquence des pauses, ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement...). Ces paramètres sont généralement associés aux paramètres de l'analyse spectrale pour améliorer les performances des systèmes. Des travaux comme ceux de [Atal, 1976] ont en effet montré que ces paramètres ne sont pas suffisamment discriminants pour être utilisés seuls. Une méthode alternative et performante de l'utilisation de ces paramètres a été développée par [Adami et al., 2003] et est présentée au chapitre 4.

1.5.1.3 Robustesse aux variabilités par normalisation dans l'espace cepstral

Les techniques de normalisation acoustique ont pour principal objectif d'enlever les distorsions dues à l'environnement acoustique et à la transmission, facteur limitant des performances des systèmes. Dans ce domaine, deux classes de méthodes prévalent.

La première est une normalisation par moyenne et/ou variance. La soustraction de la moyenne cepstrale [Furui, 1981] (CMS) part de l'hypothèse que les distorsions apportées par l'environnement acoustique peuvent être modélisées par une constante dans le domaine cepstral. L'estimation et la soustraction de cette moyenne sur un fichier audio permet d'atténuer cette distorsion. Il est maintenant courant de normaliser le fichier audio par la variance des coefficients cepstraux avant de procéder à la modélisation des locuteurs. La variance peut être estimée soit sur la totalité d'un fichier soit sur une fenêtre glissante [Openshaw et Mason, 1994].

La seconde est nommée « Gaussianisation ». Cette technique apparue en 2001 dans [Pelecanos et Sridharan, 2001], appelée *feature warping*, part du principe que les distorsions dues au canal n'affectent pas seulement la moyenne et la variance estimées sur le signal entier mais aussi la distribution marginale de chacun des coefficients. En pratique, cette normalisation consiste à modifier la répartition des coefficients cepstraux sur une fenêtre glissante en la faisant correspondre à une gaussienne de moyenne nulle et de variance unité. Cette technique est actuellement beaucoup utilisée, particulièrement avec les techniques de compensation de canal.

1.5.2 Détection de parole

Le procédé de sélection des trames utiles joue un rôle très important sur les performances des systèmes (voir chapitre 5 pour une expérience sur la sensibilité du système LIA_SpkDet à cette sélection). Cette étape est souvent assimilée à un processus de segmentation « silence/parole », ou plus exactement « parole/non-parole ». Une des techniques employées pour cette segmentation consiste à modéliser par une bi-gaussienne l'énergie des trames du signal audio. Les trames de paroles utiles sont supposées appartenir à la gaussienne de haute énergie, les autres trames considérées comme de la non-parole (silence, bruit...) appartiennent à la gaussienne de basse énergie. Un seuil dépendant des paramètres des gaussiennes est estimé pour affecter les trames à une classe [Bonastre et al., 2004]. Des procédés plus complexes, basés sur une modélisation HMM deux états (parole/non-parole) donnent aussi de bons résultats.

1.5.3 Prise de décision, réglage du seuil et normalisation des scores

La prise de décision en VAL est l'étape finale du processus d'authentification. L'enjeu est d'obtenir un système performant pour la fonction de coût associée à la décision. Nous présentons les techniques de normalisation des scores de confiance, visant à rendre plus robustes les décisions prises par le système de VAL.

1.5.3.1 Réglage du seuil

Le test d'hypothèse est un score constituant le degré de confiance avec lequel nous affirmons l'hypothèse H_1, *i.e.* $\mathbf{X}$ provient du locuteur dont l'identité S a été proclamée. Dans un contexte opérationnel, il est nécessaire d'estimer un seuil global pour le système de vérification afin d'accepter ou de rejeter le test de vérification. Ce seuil est généralement réglé sur un corpus de développement pour un point de fonctionnement précis (*e.g.* : minimum de la fonction de coût de décision lors des campagnes d'évaluation NIST-SRE).

[Bengio et al., 2001] proposent de modifier la fonction de coût de décision, en affirmant que l'équation 1.2 est valide lorsque les deux hypothèses sont bien modélisées (ce qui n'est généralement pas le cas). Les auteurs proposent une formulation plus générique de la fonction de décision afin d'en apprendre les coefficients par un classifieur SVM.

1.5.3.2 Normalisation des scores

La normalisation des scores est une étape visant à rendre robuste la détermination d'un seuil global pour la fonction de décision. En effet, lorsqu'un seuil global est fixé, l'hypothèse sous-jacente est de considérer que la variance des tests clients et imposteurs est constante quelque soit le locuteur. Or de nombreuses distorsions apparaissent selon le modèle ou le segment de test présenté. Ainsi, dans [Li et Porter, 1988], une grande variation des vraisemblances inter et intra-locuteurs a été constatée.

Les normalisations sur la distribution des scores intra-locuteurs étant difficiles à réaliser (vu la faible quantité de données disponibles), les techniques de normalisation dans l'espace des scores se concentrent sur la normalisation de la distribution imposteurs (variance inter-locuteurs). Considérons un test de vérification constitué de l'accès de test $\mathbf{X}$ et du modèle de locuteur S ; la formulation générique de normalisation du score de vérification $\mathcal{Y}_S(\mathbf{X})$ est la suivante :

$$\hat{\mathcal{Y}}_S(\mathbf{X}) = \frac{\mathcal{Y}_S(\mathbf{X}) - \mu}{\sigma} \tag{1.9}$$

où μ et σ sont la moyenne et la variance de la distribution représentant la variabilité inter-locuteur à normaliser. Dans tous les cas, la distribution des scores imposteurs doit se rapprocher le plus possible d'une distribution de moyenne 0 et de variance unité après normalisation.

Parmi les techniques de normalisation les plus connues, se trouvent la *Z-normalisation* (Znorm) [Reynolds, 1997] et la *T-normalisation* (Tnorm) [Auckenthaler et al., 2000]

- *Znorm* : La distribution des scores est normalisée comme à l'équation 1.7 par la distribution des scores issus de la réponse du modèle du locuteur sur des segments imposteurs $\mathbf{X}_{imp}$. Pour chaque modèle de locuteur S, la distribution de variabilité inter-locuteurs est estimée par $\mu = \mu_S = \mathcal{E}_{\mathbf{X}_{imp}}[\mathcal{Y}_S(\mathbf{X}_{imp})]$ et $\sigma^2 = \sigma_S^2 = \mathrm{Var}_{\mathbf{X}_{imp}}[\mathcal{Y}_S(\mathbf{X}_{imp})]$;
- *Tnorm* : La distribution des scores est normalisée comme à l'équation 1.7 par la distribution des scores issus de la réponse de modèles imposteurs I sur le segment de test. Pour chaque accès de test $\mathbf{X}$, la distribution de variabilité inter-locuteurs est estimée par $\mu = \mu_{\mathbf{X}} = \mathcal{E}_I[\mathcal{Y}_I(\mathbf{X})]$ et $\sigma^2 = \sigma_{\mathbf{X}}^2 = \mathrm{Var}_I[\mathcal{Y}_I(\mathbf{X})]$.

En plus du gain en performance généralement observé par ce procédé, ces techniques permettent de projeter les scores de différents systèmes dans le même espace, ce qui rend efficace les fusions de scores de type arithmétiques. Il existe bien d'autres normalisations prenant en compte le canal (*H-Norm*) ou dépendante d'une classe de locuteur (*C-Norm*) mais nous ne les expliciterons pas ici.

Les techniques de normalisation décrites ci-dessus produisent des scores dans un espace non borné. De ce fait, l'interprétation et le choix de la valeur du seuil de décision est malaisé. La normalisation *WMAP* [Fredouille et al., 1999] (ou « World + MAP) transforme l'espace des scores vers un espace probabiliste. Cette transformation produit des probabilités *a posteriori* en prenant en compte les probabilités *a priori* des P_{cl} et P_{imp} de la fonction de coût 1.3.

1.6 Conclusion

Tout au long de ce chapitre, nous avons présenté les principes de la reconnaissance automatique du locuteur (RAL) ainsi que les différentes méthodes d'évaluation des performances. Différentes taĉhes (section 1.2) sont associées à la reconnaissance du locuteur et permettent de définir le contexte applicatif associé au système. La voix est une modalité biométrique qui possède de grandes variabilités temporelles et qualitatives que nous avons présenté en section 1.3. Nous avons ensuité détaillés la structure et l'évaluation des systèmes automatiques de reconnaissances du locuteur en section 1.4. Enfin, les techniques standards de VAL ont été présentées en 1.5. Parmi ces techniques, celles consistant à normaliser les variabilités dans l'espace des scores ont un rôle déterminant pour l'obtention d'un seuil unique de décision.

La vérification du locuteur est basée sur un test d'hypothèse bayésien et la modélisation de l'hypothèse inverse est un enjeu crucial. Celle-ci est généralement représentée par un modèle générique modélisant le non-locuteur. De notre point de vue, ce modèle joue un rôle central et nous présentons son implication tout au long du processus de reconnaissance dans la suite du document.

Première partie

Vérification du locuteur : Approches principales et rôle du modèle générique

CHAPITRE 2

Structure de l'espace acoustique dans la modélisation générative

Sommaire

2.1 Introduction

L'approche de modélisation du locuteur par mélange de gaussiennes (GMM : Gaussian Mixture Model) constitue l'état de l'art depuis son introduction par [Reynolds, 1995]. Ce système est généralement appelé GMM-UBM car cette modélisation nécessite l'utilisation d'un modèle générique appelé modèle du monde, ou UBM (pour *Universal Background Model*).

L'objectif d'une telle approche est d'aboutir à une modélisation générative, *i.e.* l'estimation de la distribution qui a pu générer les vecteurs cepstraux du signal d'apprentissage (modèle de production). En terme statistique, l'apprentissage consiste à estimer les paramètres du GMM maximisant la vraisemblance des données d'apprentissage. Cette approche générative est à comparer à une approche discriminante où l'objectif est de minimiser les erreurs de classification à l'apprentissage grâce à des contre-exemples (modèle de perception).

Dans ce chapitre, nous présentons les différentes composantes d'un système GMM-UBM (section 2.2). Nous abordons ensuite la problématique de la modélisation du non-locuteur et particulièrement l'apprentissage du modèle du monde. Puis nous présentons les différentes techniques utilisées pour la modélisation du locuteur en explicitant les contraintes apportées aux paramètres des modèles en section 2.3. Le modèle générique est au coeur de ces travaux de thèse et nous soulignons tout au long de ce document son rôle structurel pour la VAL. Dans le cas du système GMM-UBM, ce rôle est décrit en section 2.4.

2.2 Approche état de l'art : le système GMM-UBM

La structure générale d'un système à base de modèles de mixture de gaussiennes est présentée dans la suite de cette section. Le schéma 2.1 illustre les différents composants de ce système. Nous présentons tout d'abord le modèle GMM et les statistiques associées. Ensuite, l'apprentissage du modèle de non-locuteur, nécessaire au test d'hypothèse, sera explicité. Puis nous abordons le principe de la modélisation du locuteur. Nous finissons enfin par présenter les techniques de production d'un score de vérification pour le système GMM-UBM.

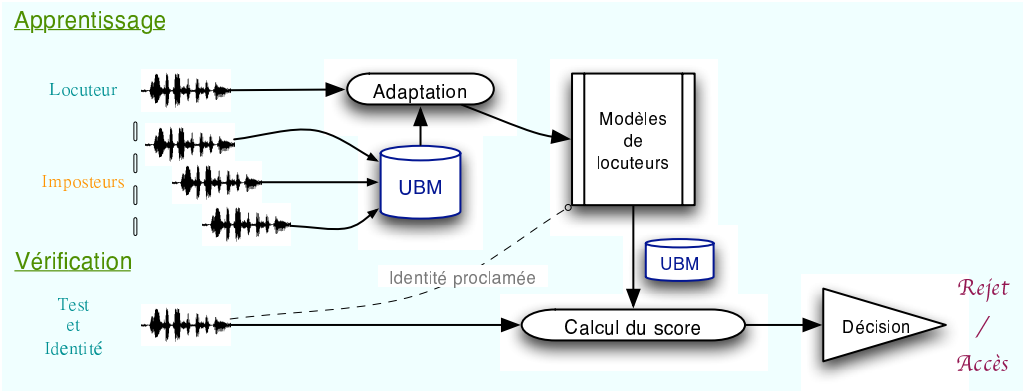


FIG. 2.1: *Structure du système GMM-UBM en VAL. L'apprentissage nécessite la construction préalable d'un modèle générique UBM. L'UBM est adapté sur les données d'apprentissage d'un locuteur pour estimer les paramètres du modèle spécifique à ce locuteur. Lors du test de vérification, le calcul de score fait intervenir l'UBM et le modèle correspondant à l'identité proclamée (correspondant au segment de test). La décision rejet/accès est prise par rapport à ce score.*

2.2.1 Modélisation par mixture de gaussiennes

La densité de probabilité d'une mixture de gaussiennes à N composantes pour une variable aléatoire $\mathbf{x}$ s'exprime sous la forme suivante :

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \gamma_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{2.1}$$

sous la contrainte $\sum_i \gamma_i = 1$ et $\forall i : \gamma_i \geq 0$. $\boldsymbol{\gamma}$ est le vecteur de poids de la mixture, $\mathcal{N}(x; \mu, \Sigma)$ est la loi gaussienne de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$, $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}]^T$ est le vecteur de paramètre global du GMM. Si $\mathbf{x}$ est de dimension d alors, une mixture de gaussienne est paramétrée par $N \times d$ paramètres de moyennes, $N \times d^2$ paramètres de variance, et N paramètres de poids. La densité d'une distribution normale à d dimensions est exprimée par :

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]. \tag{2.2}$$

Dans le cas de la reconnaissance du locuteur, les matrices de covariances sont généralement estimées sous forme diagonale. Dans le cas d'un problème de séparation bi-classe, chaque classe étant représentée par une gaussienne de l'équation 2.2, le type de matrices de covariances $\boldsymbol{\Sigma}$ associées à la gaussienne permet de changer la forme de la fonction de décision résultante. En particulier, si $\boldsymbol{\Sigma} = \sigma^2 \cdot I$ alors, géométriquement, les distributions sont des hyper-sphères de dimension d et de même rayon. La frontière de décision peut alors se résumer à un hyperplan qui sépare les régions au minimum d'erreur de classification. De plus, si les probabilités *a priori* des deux classes sont identiques alors la frontière se situe à mi-chemin entre les moyennes.

Pour calculer la vraisemblance d'une séquence de trame $\mathbf{X} = \{x_1 \ \ldots x_T\}$, pour un modèle paramétrée par $\boldsymbol{\theta}$, le logarithme est généralement utilisé en considérant l'indépendance des réalisations de la séquence d'apprentissage. Posons la notation $\log(p(.)) = \ell(.)$, alors

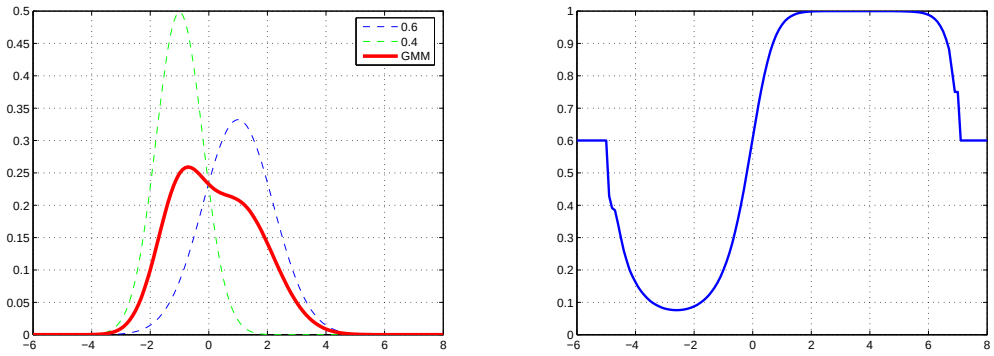
$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \ell(\mathbf{X}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \sum_{i=1}^{N} \gamma_i \mathcal{N}(x_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{2.3}$$

L'apprentissage d'un GMM est généralement réalisé avec l'algorithme EM [Dempster et al., 1977]. En effet, la maximisation de $p(\mathbf{X}|\boldsymbol{\theta})$ nécessite l'introduction de variables cachées dont la connaissance permet de trouver une forme analytique au problème (statistiques suffisantes). Ces variables cachées sont représentées dans le vecteur $\boldsymbol{\gamma} = \{\gamma_i\}_{i=1\ldots N}$.

La modélisation à l'aide de GMM peut être utilisée de deux façons :

- Pour une estimation de n'importe quelle distribution, le mélange de distributions gaussiennes étant reconnue comme un approximateur universel d'un large éventail de distributions (voir figure 2.2(a)).
- Pour du *soft clustering*, où chaque gaussienne est considérée comme un représentant (ou un état), voir figure 2.2(b). Contrairement à la méthode du *K-moyennes* où l'appartenance à une classe est quantifiée de façon binaire (1 ou 0), pour un modèle à mixture de gaussiennes, l'appartenance est la probabilité *a posteriori* de la classe connaissant les données. C'est par exemple le cas d'un détecteur parole/non-parole (voir 1.5.2), où l'objectif est de séparer les trames de silence de celles de parole.

En VAL, l'objectif original est de modéliser au mieux la distribution ayant généré les vecteurs cepstraux. En pratique il existe une contradiction car les méthodes basées sur les GMM ont un *a priori* très fort sur une correspondance gaussienne à gaussienne entre les différents modèles GMM (voir 2.4).



(a) (b)

FIG. 2.2: *a) Le GMM comme un estimateur de densité de probabilité (courbe pleine : combinaison linéaire des deux gaussiennes). b) Le GMM comme un classifieur souple, la courbe représentant la probabilité de l'appartenance à la classe de droite grâce à une décision bayésienne.*

2.2.2 Modélisation du non-locuteur

Pour la modélisation de l'hypothèse inverse du test bayésien de l'équation 1.2, deux classes de méthodes sont couramment usitées, celles modélisant le non-locuteur par une cohorte de locuteurs et celles utilisant un modèle généraliste représentant le modèle du non-locuteur pour n'importe quel locuteur.

2.2.2.1 Le test d'hypothèse bayésien pour le modèle UBM

Le paradigme de vérification du locuteur basé sur les GMM et l'utilisation d'un UBM (Universal Background Model) a été introduit par [Reynolds, 1995][Carey et Parris, 1992]. A l'époque, ce paradigme donnait des performances bien supérieures aux méthodes classiques (quantification vectorielle par exemple).

Le rôle de l'UBM tient à la modélisation de l'hypothèse inverse dans la stratégie de décision de l'équation 1.2. La modélisation de l'hypothèse inverse se fait grâce à la construction d'un modèle universel appelé modèle du monde, ou UBM, et dénoté W. Précisément, si S et $\overline{S}$ représentent respectivement le modèle du locuteur et celui du non-locuteur et $\mathbf{X}$, un segment de test dont l'identité proclamée correspond à S, alors le rapport de vraisemblance est donné par :

$$LR(\mathbf{X}, H_0, H_1) = LR(\mathbf{X}, S, W) = \frac{p(\mathbf{X}|S)}{p(\mathbf{X}|\overline{S})} \simeq \frac{p(\mathbf{X}|S)}{p(\mathbf{X}|W)}. \tag{2.4}$$

Il est clair qu'une modélisation précise du non-locuteur (dans le cas de modèles génératifs) n'est pas réalisable, *i.e.* l'approximation faite par l'UBM est qu'une distribution générique des vecteurs cepstraux représente tout les autres locuteurs hormis le locuteur concerné (et ceci *quelque soit le locuteur*). Nous mettons en évidence, dans la suite de ce chapitre, que l'UBM est surtout utilisé pour son rôle de structuration de l'espace acoustique propre à la vérification du locuteur.

Apprentissage de l'UBM Ce modèle généraliste est appris en utilisant des centaines d'heures de signal audio provenant de multiples locuteurs. Il est à noter que durant l'apprentissage du modèle du monde, l'utilisation d'un seuillage des paramètres de variance (*variance flooring*) pour l'estimation des variances des gaussiennes a permis d'améliorer les performances. Cette technique consiste à définir une borne inférieure pour les variances du modèle GMM. Ce paramètre est généralement élevé (50% de la variance globale pour le système de référence du LIA), ce qui modifie la solution optimale de l'algorithme et ne permet pas d'atteindre le même maximum de vraisemblance que lors d'un apprentissage non-contraint.

Dans le chapitre 5, nous montrons une expérience illustrant l'importance de la modélisation de la variance. La justification la plus probable de ce seuillage tient à l'élimination de singularités dans le modèle de mixture. Il faut pour cela prendre un exemple extrême : si un vecteur d'entrée x_t a la même valeur que la moyenne d'une composante alors la vraisemblance de ce vecteur pour cette composante se réduit à :

$$\mathcal{N}(x_t|x_t = \boldsymbol{\mu}, \sigma) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}. \tag{2.5}$$

Pour cette composante, la maximisation de la vraisemblance résulte en une diminution de la variance correspondante. Le *variance flooring* permet de contourner ce problème en évitant à une gaussienne de modéliser peu de données, soit d'éviter le sur-apprentissage (ou *overfitting*). Ce seuillage est donc utilisé afin d'augmenter la capacité de généralisation du modèle. L'utilisation de cette technique permet aux modèles génératifs des systèmes état de l'art d'avoir un grand nombre de composantes (de l'ordre de 2000).

2.2.2.2 Modélisation par cohorte de locuteurs imposteurs

Une seconde approche dans la modélisation du non-locuteur consiste à sélectionner une cohorte de locuteur $\overline{I} = \{I_i\}_{i=1,\ldots,N}$. Cette approche a été suggérée dans [Higgins et al., 1991] puis par [Rosenberg et al., 1992]. Selon que cette cohorte est dépendante ou non du locuteur, l'hypothèse inverse peut s'exprimer soit comme la vraisemblance d'un modèle unique appris sur la cohorte dépendante du modèle, soit comme la moyenne des vraisemblances sur chacun des locuteurs de la cohorte.

$$LR(\mathbf{X}, H_0, H_1) = \frac{p(\mathbf{X}|S)}{p(\mathbf{X}|\overline{S})} \simeq \frac{p(\mathbf{X}|S)}{\mathcal{E}_i[p(\mathbf{X}|I_i)]} \text{ ou } \frac{p(\mathbf{X}|S)}{p(\mathbf{X}|\overline{I})}, \tag{2.6}$$

où $\mathcal{E}_i[p(\mathbf{X}|I_i)]$ représente la moyenne des vraisemblances des modèles de la cohorte, et $\overline{I}$ le modèle GMM appris sur les données d'apprentissages des locuteurs de la cohorte.

2.2.3 Modélisation du locuteur

La modélisation du locuteur en VAL diffère de l'estimation du modèle UBM car les données disponibles ne sont pas en quantité suffisante pour estimer les paramètres du modèle si le nombre de composantes est élevé. Les méthodes dites d'*adaptation* permettent d'estimer de manière robuste des modèles spécifiques au locuteur en ajoutant de l'information *a priori* sur la distribution des paramètres. A la section 2.3, nous essayons de répertorier les techniques essentielles. Celles les plus usitées en VAL sont largement tirées de l'adaptation bayésienne, particulièrement celle du MAP (*maximum a posteriori*). Il est d'ailleurs intéressant de remarquer que cette approche tend généralement vers l'estimation ML lorsque la quantité de données disponible pour un locuteur est infinie.

2.2.4 Calcul du score de vérification

Le score de vérification correspondant à la vraisemblance d'une séquence de données de test $\mathbf{X} = \{x_1 \dots x_t\}$ sur un modèle de locuteur S est exprimé sous la forme de l'espérance du logarithme du rapport de vraisemblance sur toutes les trames du segment de test présenté $\mathcal{Y}_S(X)$ (ou *Expected Log-Likelihood Ratio*, généralement appelé LLR). Précisément :

$$\mathcal{Y}_S(\mathbf{X}) = \mathcal{E}_{\mathbf{X}}[\mathrm{LLR}(\mathbf{X}|S,W)] = \frac{1}{T}\sum_t \log\Big(\frac{p(x_t|S)}{p(x_t|W)}\Big), \tag{2.7}$$

où $\mathcal{E}_{\mathbf{X}}[\mathrm{LLR}(\mathbf{X}|S,W)]$ est l'espérance mathématique du LLR sur le segment de test $\mathbf{X}$, $p(x_t|S)$ et $p(x_t|W)$ sont les vraisemblances du vecteur cepstral x_t respectivement sur le modèle du locuteur S et sur le modèle du monde W.

En considérant la vraisemblance de la séquence de trames par rapport au modèle du locuteur, la normalisation par rapport au nombre de trames n'a pas lieu d'être. En effet, le score devrait être strictement la somme des LLR. Ce terme de normalisation a néanmoins démontré son efficacité car il permet de traiter des séquences de trames de longueurs variables (ce qui est souvent le cas en parole).

Une technique permettant de capter d'autres informations, à l'origine adoptée pour son côté pratique, consiste à inverser le rôle du signal d'apprentissage et de test lorsque la durée du segment de test est plus longue que celle de l'apprentissage. Néanmoins, la performance peut être amélioré même lorsque les signaux ont la même longueur. Le système GMM-UBM n'est donc pas symétrique dans le traitement des deux signaux formant le test de vérification.

2.3 Estimation des paramètres des modèles de locuteurs

Nous présentons dans cette section les techniques d'adaptation utilisées dans les systèmes de VAL actuels. Ces techniques sont présentées sous la forme de contraintes imposées au modèle génératif et à l'espace acoustique, afin de n'estimer que les paramètres nécessaires à la caractérisation d'un locuteur. Il est en effet raisonnable de penser que la différence entre deux locuteurs

peut s'exprimer dans un espace acoustique contraint. Il existe en effet des invariants comme le système de production de parole, les caractéristiques phonétiques ou coarticulatoires dans une même langue. Ceci laisse à penser que tous les paramètres du modèle génératif ne sont pas nécessaires pour modéliser un locuteur spécifique.

Trois types de méthodes attirent notre attention :
- La première est l'adaptation *maximum a posteriori* (MAP), qui est la plus usitée dans les systèmes GMM de VAL à l'heure actuelle.
- La seconde est basée sur une homogénéisation de l'adaptation des modèles GMM en vue d'obtenir des scores normalisés. Cette technique assure une meilleur robustesse lors de l'estimation du seuil global pour la prise de décision.
- La dernière se focalise sur la structuration de l'espace acoustique pour les modèles du locuteur, soit par une représentation relative du locuteur dans l'espace des scores, soit par l'estimation d'un sous-espace géométrique ne contenant que les variabilités intrinsèques aux locuteurs.

2.3.1 Adaptation MAP des paramètres de moyenne du GMM

La méthode d'adaptation la plus usitée en VAL est celle du *maximum a posteriori*. Elle consiste à définir des distributions *a priori* $p(\theta)$ pour les paramètres du modèle et à maximiser leurs probabilités *a posteriori* $p(\theta|\mathbf{X})$ sur un signal d'apprentissage $\mathbf{X}$. Le critère d'adaptation pour l'estimation des nouveaux paramètres $\hat{\theta}$ s'écrit comme suit :

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{X}) = \arg\max_{\theta} p(\mathbf{X}|\theta)p(\theta). \tag{2.8}$$

Des formules adaptées à la modélisation GMM ont été développées par [Gauvain et Lee, 1994] en proposant un choix spécifique des densités *a priori* sur les paramètres. Ce choix s'oriente vers les distributions *a priori* conjuguées (*conjugate prior*) permettant aux distributions *a posteriori* d'appartenir à la même famille qu'aux distributions *a priori*. L'adoption de ces distributions permet de conserver l'utilisation de l'algorithme EM pour l'implémentation du MAP. Dans le cas des GMMs, ce choix s'oriente vers une distribution gaussienne comme *a priori* pour les paramètres moyenne/variance et une distribution de Dirichlet pour les paramètres de poids.

En pratique, dans un système de VAL indépendant du texte, seuls les paramètres de moyenne sont modifiés. Les moyennes du modèle du monde sont les *a priori* pour celles du locuteur. Dans ce cas, l'estimation de la moyenne $\hat{\boldsymbol{\mu}}_k$ pour une composante est obtenue par une combinaison linéaire des moyennes *a priori* $\boldsymbol{\mu}_k$ et empiriques $\bar{y}_k$, issues des données d'apprentissage.

$$\hat{\boldsymbol{\mu}}_{kk} = \frac{n_k}{n_k + \tau_k}\bar{y}_k + \frac{\tau_k}{n_k + \tau_k}\mu_k \text{ avec } n_k = N * \gamma_k, \tag{2.9}$$

où γ_k est le vecteur des variables cachées d'EM et N le nombre de trames d'apprentissage. Le facteur τ, appelé *relevance factor*, permet de contrôler l'adaptation du modèle aux données en modifiant la confiance sur la distribution *a priori* des paramètres de moyenne. Cette formule d'adaptation pose la distribution *a priori* sur les moyennes comme une gaussienne de moyenne μ_k et de variance $\frac{\sigma_k^2}{\tau_k}$. Ce facteur peut donc régler la confiance sur l'estimation de la moyenne. La propriété intéressante de cette formule d'adaptation est qu'elle tend vers l'estimation au maximum de vraisemblance en présence d'une infnité de données disponibles.

2.3.2 Adaptation homogène par divergence KL : D-MAP

Dans [Ben et al., 2002] et [Ben et Bimbot, 2003], l'auteur introduit deux techniques de modélisation du locuteur et de normalisation des scores basées sur la divergence de *Kullback–Leibler* (KL) : *D-Norm* et *D-Map*. La divergence KL peut s'avérer intéressante, car la distance qu'elle mesure entre deux modèles n'est plus géométrique mais informative. Elle représente en effet l'entropie relative entre deux modèles.

L'adaptation D-MAP a été construite afin d'effectuer cette normalisation de façon implicite au niveau de la modélisation et non des scores de vérification. Le but de l'adaptation D-MAP est d'homogénéiser les modèles de locuteurs vis-à-vis de leur distance KL par rapport au modèle du monde. Tous les locuteurs sont donc placés dans une « hypersphère entropique » dont le centre est le modèle UBM et le rayon D_{ref} à définir (représentant la distance de référence). Le facteur d'adaptation s'exprime alors en fonction de cette distance de référence et la distance D_{ML} correspondant au modèle du locuteur appris avec l'algorithme EM.

Ces travaux s'appuient sur un calcul itératif de la divergence KL symétrique entre un modèle client et le modèle du monde (utilisation d'une méthode de Monte-Carlo). Cette distance, dénotée $KL2$, est définie par la somme de deux divergences KL :

$$KL2 = \mathcal{E}_{p_S}[log\frac{p_S}{p_W}] + \mathcal{E}_{p_W}[log\frac{p_W}{p_S}], \tag{2.10}$$

où p_S est la densité de probabilité du modèle du locuteur S et $\mathcal{E}_{p_S}(.)$ est l'espérance estimée sur la densité de p_S. L'auteur prouve empiriquement qu'il existe une relation linéaire entre la moyenne des scores imposteurs pour un modèle de locuteur et la distance KL2, *i.e.* $KL2 = \alpha\mathcal{E}_I\{\mathcal{Y}_I(S)\}$. La D-Norm permet de normaliser les scores suivant cette relation.

2.3.3 Ancrage des modèles de locuteurs dans l'espace des scores imposteurs

Les méthodes actuelles de modélisation du locuteur par GMM sont gourmandes en temps de calcul et en espace mémoire. Pour tenter de remédier à ce problème, une perspective intéressante est de modéliser le locuteur par une représentation relative par rapport à un ensemble de modèles de références bien appris. Les techniques de modèles d'ancrage s'inscrivent dans cette représentation relative. Cette démarche permet de contraindre l'espace acoustique (précisément l'espace des scores) par des locuteurs imposteurs en espérant diminuer le nombre de paramètres à estimer.

Les premiers travaux sur cette représentation peuvent être trouvés dans [Merlin et al., 1999], le formalisme des *modèles d'ancrages* en reconnaissance du locuteur dans [Mami et Charlet, 2006] et en indexation audio [Sturim et al., 2001]. De nombreux points communs existent entre ces techniques et celles connues sous le nom d'*eigenvoices* ou voix propres [Kuhn et al., 2000]. Cependant, la représentation relative des locuteurs a généralement lieu dans l'espace des scores alors que les techniques des voix propres travaillent sur les paramètres des modèles.

Dans l'espace des modèles d'ancrages, un locuteur est caractérisé par un vecteur unique $\boldsymbol{S}$ dont la dimension est égale au nombre de locuteurs de référence (potentiellement virtuels) $\boldsymbol{I} = \{I_i\}_{i=\{1...E\}}$. Le vecteur $\boldsymbol{S}$ contient les logarithmes de rapport de vraisemblance (scores) $\mathcal{Y}$ du signal

d'apprentissage **X** du locuteur par rapport à chacun des modèles de référence :

$$\boldsymbol{S} = [\mathcal{Y}_{I_1}(X) \ \ldots \ \mathcal{Y}_{I_E}(X)]^T. \tag{2.11}$$

Pour calculer la distance entre deux locuteurs dans l'espace des modèles d'ancrage, il suffit de calculer une similarité entre les vecteurs. De nombreuses similarités peuvent être trouvées dans [Collet, 2006]. Le choix de l'espace de référence est capital et plusieurs méthodes sont répertoriées dans [Mami, 2003] : méthodes de regroupement ascendant des locuteurs, sélection d'un sous-ensemble optimal de locuteurs, ou analyse ACP afin de trouver les axes de variation les plus représentatifs et réduire la dimensionnalité.

2.3.4 Sous-espaces de variation des paramètres des modèles de locuteur

Dans la suite du paragraphe, nous appelons *super-vecteur* de moyenne le vecteur contenant touts les paramètres de moyenne d'un modèle génératif. Précisément, soit un modèle GMM M à N composantes de dimension d, chaque composante k possède un vecteur de moyenne de dimension $d : \mu^k = [\mu_1^k \ldots \mu_d^k]$. Le super-vecteur de moyenne $\mu(M)$ de ce modèle GMM M est de dimension Nxd et à la forme :

$$\boldsymbol{\mu}(M) = [\mu^1 \ldots \mu^N] = [\mu_1^1 \ldots \mu_d^1, \ \ldots \ , \mu_1^N \ldots \mu_d^N]^T. \tag{2.12}$$

2.3.4.1 Espace des voix propres

Les techniques de modélisation cherchant à exprimer les variations inter-locuteurs dans un espace de dimensionnalité réduite sont regroupées sous le terme *eigenvoices*. Ces approches ont été développées initialement dans le cadre de l'adaptation au locuteur [Kuhn et al., 2000] en s'inspirant d'une des techniques les plus usitées en reconnaissance du visage : *eigenfaces* [Turk et Pentland, 1991]. Les voix propres peuvent être générées à partir d'algorithmes de réduction de dimensionnalité (ACP). Considérons la formulation équivalente de l'adaptation MAP (équation 2.6) :

$$\boldsymbol{\mu}(S) = \boldsymbol{\mu}(W) + D \cdot z, \tag{2.13}$$

où $\boldsymbol{\mu}(S)$ et $\boldsymbol{\mu}(W)$ sont les super-vecteurs du locuteur et du monde, z est un vecteur contenant les « speaker factors » et D correspond à l'*a priori* sur la distribution des moyennes. Dans le cas de MAP, D est diagonale. Les méthodes essayant de tirer parti des corrélations inter-gaussiennes sont nommées *Extended MAP* [Zavaliagkos, 1995], où D est une matrice pleine. Le nombre de paramètres à estimer est très grand, les techniques des voix propres permettent de répondre à ce problème en partant du principe qu'on peut trouver une matrice V, de rang faible, caractérisant les variations du locuteur.

$$\boldsymbol{\mu}(S) = \boldsymbol{\mu}(W) + V \cdot y. \tag{2.14}$$

Le locuteur sera ainsi estimé par un nombre réduit de paramètres correspondant à la dimension de V, ces paramètres, y dans l'équation, sont appelées « facteurs locuteurs » (les colonnes de V sont les « voix propres »). Dans [Thyes et al., 2000][Kenny et al., 2005a], les auteurs proposent des méthodes d'estimation des vecteurs propres du sous-espace vectoriel engendré par V par maximum de vraisemblance, réduisant ainsi le nombre de paramètres à estimer.

2.3.4.2 Dissocier l'effet du canal et du locuteur

Les travaux de [Kenny et al., 2005a] apportent une alternative à l'adaptation MAP classique, en utilisant les bases des modèles *eigenvoices*. Les auteurs introduisent une méthode robuste permettant d'estimer les paramètres de l'équation **??** avec peu de données d'apprentissage grâce au formalisme du *factor analysis*.

L'apport principal de cette technique tient à son application dans l'espace du canal, donnant naissance aux techniques *eigenchannel* permettant de répondre au problème de variation de canal entre l'apprentissage et le test. La formulation du problème suppose que le super-vecteur d'un locuteur peut être décomposé en une somme de deux vecteurs, l'un dépendant du locuteur, l'autre dépendant du canal. A chacun est associé un sous-espace de variation de faible dimension. Tout comme les variations pour un même locuteur, les variations dues au canal peuvent être contraintes dans un sous-espace de faible dimension. Le corollaire est que la matrice de covariance du canal est de rang faible. Si la matrice de variation du canal était de rang plein, alors l'influence du canal se porterait dans tout l'espace acoustique, ce qui se traduirait par le fait qu'un locuteur pourrait être transformé en un autre locuteur par un changement de canal.

L'originalité de cette approche est de considérer l'espace du canal comme un espace continu, alors que les méthodes telles que le *feature mapping* considèrent l'espace du canal comme étant discret (cette technique nécessite en effet une étiquetage des données en un nombre fini de canaux différents). Dans [Kenny et Dumouchel, 2004a], le formalisme du *joint factor analysis* permet d'estimer conjointement un sous-espace propre au locuteur et un sous-espace propre au canal. Ainsi durant la phase d'apprentissage, le vecteur du locuteur sera appris en ayant retiré l'influence du canal, considéré comme du bruit. La formulation du problème de modélisation devient :

$$\begin{aligned} \boldsymbol{\mu}\hat{S} &= \boldsymbol{\mu}S + \boldsymbol{\mu}C & (2.15)\\ &= (\boldsymbol{\mu}(W) + V \cdot y) + (U \cdot x), & (2.16) \end{aligned}$$

où U est une matrice de rang faible, dont les vecteurs colonnes sont les « canaux propres », générant le sous-espace vectoriel associé aux variations du canal et x contient un nombre faible de paramètres à estimer, nommés les « facteurs de canal ».

Pour décider si l'identité proclamée dans un test appartient à un locuteur, [Vogt et al., 2005] estime *a priori* que le locuteur est présent dans le signal et remplace l'influence du canal dans l'apprentissage par celui du segment de test. Ces techniques sont coûteuses et difficiles à mettre en oeuvre, [Kenny et al., 2005b] proposent de nombreuses simplifications théoriques et une implémentation plus rapide. Plusieurs techniques visant des objectifs similaires ont été introduites ces dernières années. Parmi celles-ci, les plus connues sont *Nuisance Attribute Projection (NAP)* dans [Solomonoff et al., 2005] et *Within-Class Covariance Normalization (WCCN)* dans [Hatch et al., 2006].

2.4 Du rôle structurant de l'UBM

L'UBM a un rôle central dans le paradigme GMM-UBM. Il est en effet la modélisation unique, pour tous les locuteurs, de l'hypothèse inverse dans le test bayésien. Il représente ainsi le modèle

du non-locuteur pour tous les locuteurs, c'est en ce sens qu'il est qualifié d'« universel ». Dans les paragraphes suivants, nous approfondissons son rôle dans le processus en affirmant qu'il permet de structurer l'espace acoustique pour tous les modèles. Tout au long du document, ce rôle de structuration sera illustré pour tous les systèmes de VAL qui sont présentés.

2.4.1 Rôle dans la normalisation par rapport au canal

Nous avons déjà abordé la normalisation des vecteurs cepstraux pour retirer l'influence du canal au chapitre 1. Cette normalisation est un enjeu important dans le processus de vérification car cette étape est souvent dédiée au retrait de l'influence du canal pour la reconnaissance. Nous présentons le rôle du modèle générique dans deux méthodes ayant cet objectif, le *feature mapping* et la compensation du canal par le retrait des « canaux propres ».

2.4.1.1 Feature mapping pour la robustesse au canal

La technique du *feature mapping* a été introduite par [Reynolds, 2003]. Elle fait parti des premiers efforts de la communauté pour résoudre le problème de variation de canal entre les données d'apprentissage et de test.

Le principe, tiré du *stochastic matching*[Sankar et Lee, 1996], est basé sur une modélisation supervisée des différents canaux (en pratique entre 2 et 8 lors des campagnes NIST-SRE) par l'adaptation d'un modèle générique sur des données au canal bien défini.

La méthode du *feature mapping* peut se décomposer en trois étapes :
- Estimation de l'UBM général sur un grand nombre de données ;
- Adaptation de l'UBM à des canaux spécifiques (cellular, landline,...) grâce à des données étiquetées ;
- Estimation de la classe d'appartenance de chaque fichier audio par calcul de vraisemblance (identification) ;
- Normalisation des trames par projection sur une gaussienne indépendante du canal.

Cette normalisation consiste à projeter les trames d'un espace dépendant du canal vers un espace indépendant du canal en se basant sur l'indice de la gaussienne la plus vraisemblable dans le modèle indépendant du canal. Ainsi, si G_{CI} représente la gaussienne dans le modèle indépendant du canal et G_{CD} la gaussienne correspondante (avec le même indice) dans le modèle adapté, la normalisation d'une trame x_t s'effectue de la façon suivante :

$$\hat{x}_t = \frac{\sigma_{G_{CI}}}{\sigma_{G_{CD}}}(x_t - \mu_{G_{CD}}) + \mu_{G_{CI}}, \tag{2.17}$$

où μ et σ sont les moyennes et variances correspondant aux gaussiennes G_{CI} et G_{CD}. Cette normalisation correspond à une projection d'un espace dépendant à indépendant du canal.

Pour cette méthode, la structuration de l'espace acoustique par le modèle générique intervient au niveau de la correspondance entre les gaussiennes. Pour que cette méthode fonctionne, il faut que l'estimation soit faite de façon à garder une association entre les gaussiennes de même indice dans le modèle générique et dans le modèle du locuteur. C'est en effet l'UBM qui permet d'estimer l'indice de la gaussienne la plus vraisemblable pour une trame et c'est ce même indice qui est utilisé dans le modèle dépendant du canal.

2.4.1.2 Compensation du canal par retrait de l'influence des « eigenchannels »

Les dernières techniques de robustesses au canal utilisent le formalisme du *Factor Analysis* introduit par [Kenny et Dumouchel, 2004b]. [Vair et al., 2006] proposent une méthode de normalisation des trames pour compenser le canal. Le principe est de soustraire à chaque trame une quantité estimée dans un sous-espace représentant les variabilités dues au canal. Une méthode similaire a été appliquée avec succès par [Kenny et al., 2006] pour la reconnaissance de la parole.

Précisément, si μ_k^S, μ_k, U_k, c_k^S représentent respectivement, pour la gaussienne d'indice k, le vecteur de moyenne du locuteur, le vecteur de moyenne de l'UBM, la sous-matrice des vecteurs propres de l'espace de variation du canal associés à cette gaussienne et les facteurs canal du signal audio alors la normalisation du vecteur d'observation $\mathbf{x}_t$ s'effectue de la façon suivante :

$$\hat{\mathbf{x}}_t = \mathbf{x}_t - \sum_{i=1}^{K} \gamma_k(t) U_k \cdot c_k. \tag{2.18}$$

Là encore, l'UBM joue un rôle déterminant dans l'obtention des statistiques nécessaires à la normalisation puisque les occupations $\gamma_k(t)$ sont calculées sur la gaussienne k de l'UBM, les modèles dépendant du locuteur n'étant pas encore estimés.

2.4.2 Intervention au niveau des modèles

L'UBM joue un rôle important dans l'estimation des paramètres des modèles dépendant du locuteur. La technique d'adaptation du *maximum a posteriori* illustre bien son rôle.

En pratique, dans un système de VAL, l'*a priori* sur les paramètres de variance et poids est une distribution de *Dirac*, ainsi tous les locuteurs modélisés partagent les mêmes paramètres de variance et de poids que le modèle du monde. Le nombre de paramètres pour caractériser un locuteur est donc de $N \times D$.

Cet *a priori* assez surprenant ne trouve pas ou peu de justification théorique, il apporte cependant des avantages :

- il permet de garder la proportion de trames affectée à chaque gaussienne et définit une relation inter-gaussienne tout au long du processus.
- il définit une fonction, qui à une séquence de trames $\mathbf{X}$, affecte un vecteur $\Phi = \mu_i^j$ avec $i \in [1 \ldots D], j \in [1 \ldots N]$ donnant naissance aux techniques basées sur des super-vecteurs de moyennes.
- il permet de façon approximative d'introduire une contrainte sur les degrés de liberté du GMM, partant du principe qu'il faut peu de paramètres pour différencier deux locuteurs, comparé à la modélisation de l'espace acoustique entier, dédiée à l'UBM.

Il est clair au vu des points précédent que l'UBM joue un rôle structurant pour l'espace acoustique. En effet, les contraintes amenées par la modélisation du locuteur font intervenir les paramètres du modèle du monde (les paramètres de variance et de poids sont les mêmes pour tous les locuteurs). Par exemple dans la méthode D-MAP, l'espace acoustique est contenu dans une hyper-sphère de rayon fixe et de centre le modèle du monde.

2.4.3 Intervention au niveau des scores

Nous avons déjà abordé la normalisation des scores de vérification au chapitre 1. Cette normalisation est un enjeu important dans le processus de vérification car elle facilite le réglage d'un seuil global pour le système de vérification, tout en normalisant les variabilités inter-locuteurs dans l'espace des scores. Dans la suite, nous présentons le rôle du modèle générique à la fois pour la normalisation des scores mais aussi pour le calcul du LLR.

2.4.3.1 Calcul rapide du rapport de vraisemblances

L'UBM a été introduit dans le calcul du LLR à des fins d'économie de temps de calcul. Cette économie est possible grâce à une propriété intéressante qui relie toujours l'UBM et un modèle de locuteur, l'association gaussienne à gaussienne.

Le calcul du rapport de vraisemblance (LR) pour une trame x_t nécessite le calcul de deux vraisemblances de GMM, celle du locuteur S et celle de l'UBM W. Or, les systèmes de VAL utilisent la technique du *top-K* gaussiennes, consistant à utiliser l'indice des K meilleurs gaussiennes de l'UBM pour calculer la vraisemblance sur le locuteur client. L'approximation est formulée ci-dessous :

$$LR(x_t|S,W) = \frac{p(x_t|S)}{p(x_t|W)} = \frac{\sum_{k=1}^{K} p(x_t|S^k) + \sum_{k=K+1}^{N} p(x_t|W^k))}{p(x_t|W)}, \tag{2.19}$$

où S^i représente la composante k du GMM du locuteur. Le calcul du score se résoud donc par $N+mK$ calcul de vraisemblance au lieu de $(m+1)N$, où m est le nombre de locuteurs à tester. En pratique, le rapport entre ces quantités est de $100 < \frac{N+mK}{N+mN} \simeq N/K < 500$, résultant en un gain de temps de calcul très important. Cette technique s'avère très efficace dans le cadre des évaluations NIST-SRE où le calcul de plusieurs tests peut être mutualisé (dans un système réel de VAL, cette mutualisation n'est possible que si une *T-normalisation* est appliquée).

Dans le cas d'une modélisation purement générative, l'association gaussienne à gaussienne n'a pas de sens puisque l'objectif est de modéliser au mieux la distribution des données. Toutes les techniques citées dans cette section, *feature mapping*, *maximum a posteriori* utilisent cette propriété qui laisse à penser que le paradigme GMM-UBM est utilisé comme une mixture de classifieurs correspondant aux gaussiennes [Mariethoz, 2006].

2.4.3.2 Normalisation des scores *T-normalisation*

La *T-normalisation* introduite par [Auckenthaler et al., 2000] est une méthode performante, simple à mettre en oeuvre mais très coûteuse. Son apport est tellement significatif qu'elle est maintenant indissociable d'un système de VAL dans le cadre des évaluations NIST-SRE. Nous montrons dans la suite que l'influence de l'UBM se restreint à son rôle structurel lors de l'application de cette technique.

La formulation de la *T-normalisation* de l'équation 1.7 est rappelée ci-dessous. Notons le logarithme de la vraisemblance comme $log(p(.)) = \ell(.)$:

$$\hat{y}_S(X) = \frac{y_S(X) - \mu_X}{\sigma_X} = \frac{LLR(X|S,W) - \mathcal{E}_I[LLR(X|I,W)]}{\sqrt{\text{Var}_I[LLR(X|I,W)]]}}. \tag{2.20}$$

Au numérateur, l'influence de l'UBM peut être enlevée puisque :

$$LLR(X|S,W) - \mathcal{E}_I[LLR(X|I,W)] = \ell(X|S) - \ell(X|S) + \ell(X|W) + \mathcal{E}_I[\ell(X|I)]. \tag{2.21}$$

Au dénominateur, la variance d'une variable aléatoire étant invariante à une translation, $\sigma_I[\ell(X|I) - \ell(Y|W)] = \sigma_I[\ell(X|I)]$ et donc :

$$\hat{y}_S(X) = \frac{\ell(X|S) - \mathcal{E}_I[\ell(Y|I)]}{\sqrt{\text{Var}_I[\ell(X|I)]}}. \tag{2.22}$$

Le modèle du monde, à l'origine utilisé pour approximer l'hypothèse inverse dans le test d'hypothèses bayésien, voit son influence totalement disparaître de la formule du score final après une Tnorm. A un facteur de variance près, le score final est celui d'une cohorte imposteur présentée comme dans l'équation 2.4. Ceci renforce notre idée que le modèle du monde sert plus à structurer l'espace acoustique pour le paradigme GMM-UBM plutôt qu'à modéliser l'hypothèse inverse. Le rôle de l'UBM n'est donc pas direct dans le processus de vérification mais prédominant puisque essentiel au calcul des top-K gaussiennes (association gaussienne à gaussienne) et à la modélisation des locuteurs (distribution *a priori* pour les paramètres des modèles).

2.5 Conclusion

Tout au long de ce chapitre, nous avons essayé de parcourir les techniques liées à l'utilisation des GMM en vérification du locuteur. Après avoir présenté les différentes composantes du système GMM-UBM en 2.2, ainsi que l'apprentissage du modèle UBM, nous avons présenté les différentes techniques de modélisation du locuteur (2.3). Nous avons insisté sur les contraintes imposées par ces approches afin de modéliser le locuteur avec moins de paramètres que ceux du modèle générique. Enfin, nous soulignons le rôle structurel du modèle du monde tout au long du processus (2.4). A tous les niveaux, trames, modèles, scores, normalisation, le modèle UBM intervient pour le calcul des statistiques, pour le calcul d'indices des gaussiennes importantes, *etc*.

CHAPITRE 3

Machines à vecteurs supports pour les approches discriminantes en VAL

Sommaire

3.1 Introduction

La majorité des systèmes de détection du locuteur sont basés sur une modélisation générative des vecteurs cepstraux issus du signal vocal d'un locuteur. L'utilisation du paradigme GMM-UBM présenté au chapitre 2 est dorénavant une étape indispensable pour obtenir des performances proches de l'état de l'art dans des campagnes d'évaluations internationales telles que les campagnes NIST-SRE. Ces dernières années ont vu l'apparition d'approches discriminantes présentant des performances proches des méthodes génératives. Dans le cadre des campagnes NIST-SRE, ces

approches pour la reconnaissance du locuteur ont été appliquées avec succès dans [Wan et Campbell, 2000] et l'intérêt pour ces méthodes est encore en pleine croissance.

Les méthodes à base de machines à vecteurs supports (*SVM*) présentent des intérêts particuliers : leur capacité à traiter des problèmes de grande dimension et la bonne réalisation du compromis complexité/généralisation. De plus, leur mise en oeuvre est aisée au vu des multiples logiciels de grande qualité disponibles en licence libre pour la communauté des chercheurs.

Dans ce chapitre, nous présentons le formalisme des machines à vecteurs supports en insistant sur leurs caractéristiques. Le formalisme des modèles génératifs, détaillé au chapitre précédent, est particulièrement adapté lorsque les données à classifier sont des séquences de vecteurs de longueurs variables. Un des principaux défis pour appliquer une approche basée sur les SVM à la vérification du locuteur, particulièrement en présence de grandes quantités de données, est d'adapter les techniques existantes au traitement de données séquentielles. Nous abordons ensuite les solutions apportées en prenant pour exemples les systèmes qui ont prouvé leur efficacité dans les campagnes NIST-SRE. Enfin, nous présentons la structure générique d'un système de VAL basé sur les SVM, en soulignant l'utilisation des données imposteurs dans la structuration de l'espace acoustique. Cette structuration apparaît comme une normalisation des valeurs dans l'espace des caractéristiques (*feature space*) afin de calculer des distances appropriées à la VAL.

3.2 Machines à vecteurs supports

Les algorithmes de classification de type « machines à vecteurs supports » (SVM : *Support Vector Machines*) sont aujourd'hui considérés comme les méthodes parmi les plus performantes pour de nombreux problèmes réels de classification et de régression. A l'origine, ils ont été conçus pour construire une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes. Ils constituent par conséquent une alternative intéressante aux approches génératives classiques pour la vérification du locuteur, où les deux classes peuvent être représentées par celles du test d'hypothèse bayésien de l'équation 1.2

3.2.1 Classification binaire linéaire

Nous présentons brièvement le problème de la classification linéaire à 2 classes, relié étroitement avec le formalisme des SVM. Un SVM peut être exprimé comme un classifieur bi-classe [Cristianini et Shawe-Taylor, 2000], les stratégies d'extensions multi-classes étant souvent exprimées comme des extensions du modèle binaire.

Considérons un jeu de données d'apprentissage $(x_t, y_t)_{t=1...T} \in \mathbb{R}^d \times \{-1, +1\}$ correspondant à un problème de classification, *i.e.* associer chaque x_t à une classe y_t. Plutôt que de construire une fonction qui associe directement l'espace d'entrée à la classe, la classification linéaire consiste à trouver une fonction $f(x)$ dont le signe donne l'appartenance à la classe. Cette fonction est un séparateur linéaire et s'exprime sous la forme :

$$\begin{cases} f(x) = \boldsymbol{w} \cdot x + \boldsymbol{b} \\ y = sgn(f(x)) \end{cases}, \tag{3.1}$$

où $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

3.2.2 Maximisation de la marge

Parmi tous les hyperplans de séparation possibles, le SVM considère l'hyperplan optimal comme étant celui qui maximise la marge (voir figure 3.1. La marge du séparateur f et d'un point (x_t, y_t) est définie par $yf(x)$ (une marge négative correspond à une erreur de classification). Pour un ensemble d'exemples, la marge maximale est la distance aux points les plus proches. La distance de l'hyperplan défini précédemment à un point est donnée par :

$$d(x_t) = \frac{|\boldsymbol{w}x_t + b|}{||\boldsymbol{w}||}, \tag{3.2}$$

où $||\boldsymbol{w}||$ est la norme euclidienne du vecteur caractéristique $\boldsymbol{w}$, et b l'ordonnée à l'origine de l'hyperplan. Il est ainsi aisé de formuler la distance entre les deux hyperplans correspondant aux classes $\{-1, +1\}$:

$$d(\boldsymbol{w}x + b = +1, \boldsymbol{w}x + b = -1) = \frac{2}{||\boldsymbol{w}||}. \tag{3.3}$$

Par conséquent, la maximisation de la marge se traduit par la minimisation de $||\boldsymbol{w}||$.

Dans le cas d'un problème linéairement séparable, les exemples d'apprentissage se trouvent en dehors ou sur la marge. Cette propriété est ajoutée au problème d'optimisation sous forme de contraintes (inégalités) et s'exprime alors comme :

$$\begin{cases} \min \frac{||\boldsymbol{w}||^2}{2} \\ y_t(\boldsymbol{w}x_t + b) \geq 1 \ , \ \forall t \end{cases} . \tag{3.4}$$

3.2.3 Résolution du problème de minimisation

La solution à un problème de minimisation est généralement obtenue en annulant la dérivée de la fonction étudiée. En présence de contraintes exprimées sous la forme d'inégalités, le problème de minimisation peut se résoudre dans l'espace dual. Cette formulation du problème équivaut à injecter les contraintes dans la fonction objective. La formulation du problème dans l'espace dual peut être trouvée dans la littérature (*e.g.* [Burges, 1998]), nous n'allons pas l'expliciter ici. La fonction de décision peut se réécrire ainsi :

$$f(x) = \sum_t \alpha_t y_t x_t \cdot x + b, \tag{3.5}$$

avec $\sum_t \alpha_t y_t = 0$ et $\forall t, \ \boldsymbol{\alpha_t} \geq 0$.

La solution de ce problème se traduit par une somme pondérée de produits scalaires entre les exemples d'apprentissage. Les α_t sont les variables ajoutées pour l'expression du problème dans l'espace dual (multiplicateurs de Lagrange).

Les *vecteurs supports* sont les exemples d'apprentissage pour lesquels $\boldsymbol{\alpha_t} \neq 0$. La maximisation de la marge peut être vue comme une gestion de la complexité du problème, puisque le SVM ne mémorise que ces exemples pour discriminer les deux classes. Le paramètre α est nul pour les

autres exemples. La figure 3.1 représente un séparateur linéaire et ses vecteurs supports associés et illustre le principe de la marge maximale.

Ce principe suit celui du *Minimum Description Length* postulant que la meilleure représentation des données, en terme de généralisation, est celle qui a nécessité la plus faible quantité d'information pour la décrire. Des algorithmes efficaces de résolution de ce problème peuvent être trouvés dans [Platt, 1999], le plus connu étant le *Sequential Minimisation Optimisation* (SMO).

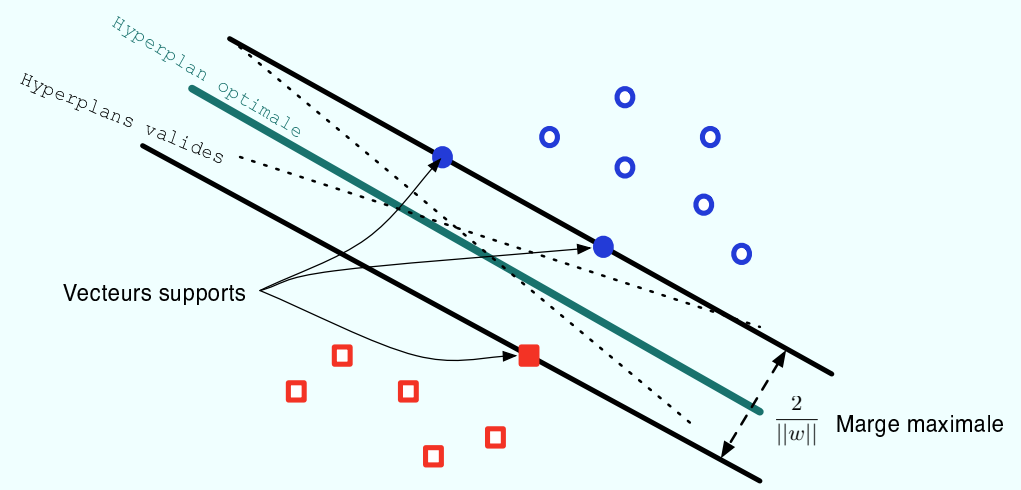


FIG. 3.1: *Maximisation de la marge et vecteurs supports. La distance à maximiser est donnée par* $\frac{2}{||w||}$. *Les vecteurs supports sont représentés par les points sur la marge.*

3.2.4 Traitement des erreurs et passage à un espace de grande dimension

Si nous prenons maintenant le cas d'un problème de classification non-linéairement séparable (le cas généralement rencontré dans les problèmes réels), deux techniques utilisées dans les SVM permettent de traiter ce cas. La première est l'introduction d'une marge souple, la seconde est le passage à un espace de dimension suffisamment grande, appelé espace des caractéristiques (ou *feature space*), pour que la frontière de séparation soit linéaire. Ces deux techniques sont utilisées conjointement.

3.2.4.1 SVM à marge souple

Dans le cas d'un problème non-linéairement séparable, il est nécessaire d'introduire des *variables ressorts* dans le problème de minimisation, permettant de pénaliser les erreurs de classification réalisées sur l'ensemble d'apprentissage. Ces variables représentent la distance qui sépare un exemple mal classifié à l'hyperplan de la classe de cette exemple. Le problème de minimisation s'exprime alors comme :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{t=1}^{T}\xi_t \\ y_t(\boldsymbol{w}\cdot x_t + b) \geq 1 - \xi_t),\ \forall t \\ \xi_t \geq 0, \forall t, \end{cases} \tag{3.6}$$

où les variables ressorts ξ_t sont associées à un paramètre de coût C. La minimisation de la fonction objective prend alors en compte la minimisation des erreurs de classification. L'avantage est que le problème garde la même forme que dans le cas séparable à l'exception d'une borne supérieure sur les α : $\forall i, 0 \leq \alpha_i \leq C$. Le paramètre C permet de fixer le coût associé aux erreurs, c'est un hyperparamètre souvent obtenu à l'aide d'une validation croisée.

Le schéma 3.2 résume les différentes valeurs possibles des paramètres α_t pour les vecteurs supports et les exemples en dehors de la marge (dans les cas de bonne et de mauvaise classification).

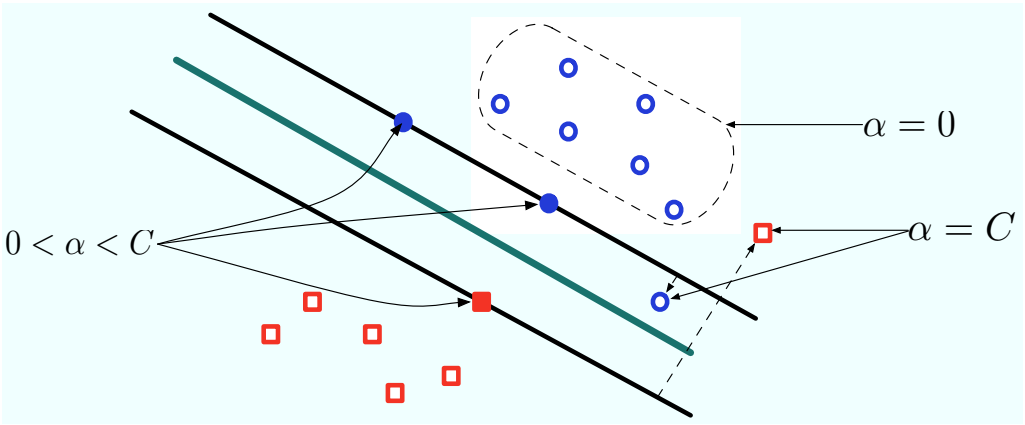


FIG. 3.2: *SVM à marge souple. Les paramètres α sont non-nuls pour les vecteurs supports et les exemples mal classifiés. La marge souple permet d'amener la prise en compte de la minimisation des erreurs de classification*

De plus, cette introduction de *variables ressorts* induit l'utilisation d'une fonction de perte particulière pour les SVM [Burges, 1998] : le *hinge loss*. En effet, l'introduction des contraintes de l'équation 3.4 dans le problème primal transforme la fonction objective comme :

$$\min \frac{1}{2}\|w\|^2 + C\sum_{t=1}^{T}[1 - Yf(X)]_+, \tag{3.7}$$

$$\text{avec } z_+ = max(z, 0). \tag{3.8}$$

D'après cette expression, si la marge est supérieure à 1 (ξ_i=0) le coût d'une erreur de classification est nul. Si la marge est inférieure à 1, alors le coût est linéaire et vaut $C\xi_i$.

3.2.4.2 Projection dans un espace de grande dimension

L'attrait pour les classifieurs SVM tient à leur capacité à traiter des problèmes non-linéaires en gardant le formalisme décrit précédemment. Le principe sous jacent est qu'un problème non séparable linéairement peut être transformé en un problème séparable linéairement dans un espace de dimension suffisamment grande. Une fonction $\Phi : \mathbb{R}^d \rightarrow F$, appelée *feature mapping*, est définie. Elle transforme les données de l'espace d'entrée, *input space*, vers un espace de plus grande dimension, l'espace des caractéristiques ou *feature space*. Le problème garde toujours la même

forme et la solution ne dépend que des produits scalaires entre les exemples d'apprentissage, après projection dans le *feature space*.

Plutôt que de choisir une fonction Φ, une fonction *noyau* K est utilisée. Cette fonction a la propriété suivante : $K(x,x') =< \Phi(x), \Phi(x') >$. Pour un noyau K bien choisi (voir paragraphe suivant), le calcul explicite de la projection dans l'espace des caractéristiques n'est pas nécessaire. Ainsi la fonction de décision s'exprime sous la forme suivante :

$$f(x) = \sum_t \alpha_t y_t < \Phi(x_t), \Phi(x) > + b = \sum_t \alpha_t y_i K(x_t, x) + b. \tag{3.9}$$

L'équation ci-dessus illustre « l'astuce du noyau » ou *kernel trick*. La connaissance explicite de Φ n'est pas nécessaire pour effectuer le produit scalaire. La fonction noyau permet d'effectuer implicitement le produit scalaire dans le *feature space* et réduit considérablement le coût de calcul dans le cas de problèmes de grande dimension. Il est ainsi aisé d'inférer des calculs de distance dans le *feature space* à l'aide des fonctions noyaux pour des espaces de dimensions potentiellement infinies.

Au vu des caractéristiques énoncées, l'attrait pour ce classifieur est évident. Le coût pour effectuer une classification d'un problème non-linéaire peut être assimilé au coût d'une classification linéaire binaire.

3.2.4.3 Noyaux généraux

La classe de fonction respectant les conditions nécessaires pour former un noyau est bien définie. N'importe quelle fonction peut être utilisée pour peu qu'elle représente un produit scalaire dans un certain espace. Une liste de noyaux les plus courants peut être trouvée dans la littérature, les principaux étant : linéaire, polynomial et gaussien. Plus généralement, un noyau valide doit satisfaire la condition de Mercer [Vapnik, 1998]. Une fonction symétrique $K(x,y)$ peut ainsi s'exprimer sous la forme d'un produit scalaire $K(x,y) =< \Phi(x), \Phi(y) >$ pour un certain Φ si et seulement si :

$$\int K(x,y)g(x)g(y)dxdy \geq 0, \tag{3.10}$$

quelque soit g satisfaisant $\int g^2(x)dx$ finie. Dans la plupart des cas, il n'est pas facile de vérifier la condition de Mercer. Un des corollaires est que les matrices générées par le noyau (matrices de Gram) sont semi-définies positives, et donc inversibles. Le problème de classification est entièrement défini une fois ces matrices noyaux calculées. L'avantage est de transformer un problème de grande dimension en un problème de dimension égale à celle du nombre d'exemples.

3.2.4.4 Un compromis optimal entre complexité et généralisation

Les méthodes discriminantes ont généralement l'objectif de minimiser les erreurs de classification à l'apprentissage. Cet objectif se traduit par l'introduction du *risque empirique* dans la fonction objective. Le risque empirique est une fonction représentant les erreurs et le coût associé à une mauvaise classification, qu'il faut minimiser.

Lorsque qu'une méthode d'apprentissage a pour objectif de minimiser le risque empirique et la complexité du classifieur, cette méthode minimise un risque appelé « risque régularisé ». Le concept de risque régularisé permet, par sa minimisation, de résoudre le compromis complexité/généralisation. Ceci est illustré à la figure **??**. Si la complexité d'une machine est trop faible, le classifieur ne peut atteindre ses performances maximales et se trouve dans le cas de *sur-généralisation*. A l'inverse, un trop grand nombre de paramètres conduit au sur-apprentissage et donc à la modélisation d'événements non informatifs tels que le bruit.

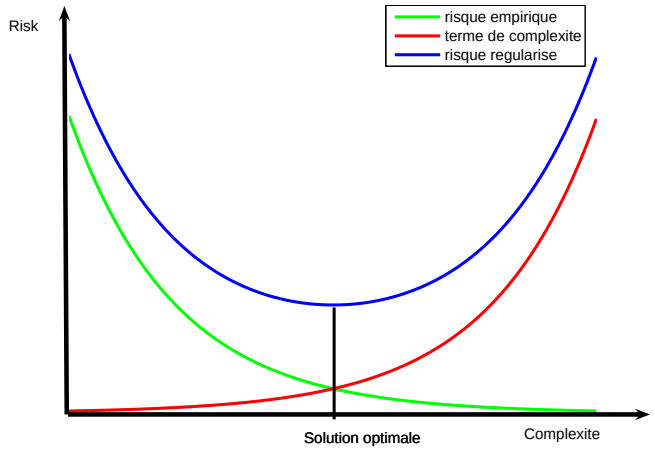


FIG. 3.3: *La minimisation du risque régularisé permet de trouver une solution optimale au compromis sur/sous apprentissage.*

Dans les approches de modélisation à base de GMM, comme nous l'avons abordé au paragraphe précédent, une des façons de contrôler la capacité de généralisation consiste à fixer une borne inférieure pour les variances des gaussiennes et de procéder à la maximisation de vraisemblance.

Dans le cas des SVM, la fonction objective de l'équation 3.4 implique la minimisation du risque régularisé. Le premier terme représente la complexité du modèle reliée au nombre de vecteurs supports. Le second représente le risque empirique. L'intégration du risque régularisé dans la fonction objective constitue l'originalité majeure des classifieurs SVM.

3.3 Les SVM appliqués à la VAL, noyaux de séquences

Une des manières intuitives d'appliquer un SVM à la vérification du locuteur est d'utiliser la même démarche que dans celle de la modélisation générative, *i.e.* apprendre des modèles discriminants dans l'espace des vecteurs acoustiques et combiner les scores en phase de test pour décider de la classe d'une séquence. Le score de vérification $\mathcal{Y}_S(X)$ d'une séquence $X = \{x_1 \dots x_T\}$ sur un modèle de locuteur S est alors défini ainsi :

$$\mathcal{Y}_S(X) = \frac{1}{T}\sum_T \mathcal{Y}_S(x_t). \tag{3.11}$$

Le modèle S contient toutes les trames d'apprentissage du locuteur et un nombre important de trames de locuteurs imposteurs comme exemples négatifs. Cependant, la quantité importante de données d'apprentissage en VAL empêche les SVM de traiter ce problème de cette manière. Cette méthode est en effet très coûteuse (elle nécéssite T^2 opérations) et donne de faibles performances. Les méthodes de réduction des vecteurs d'entrée sont une alternative mais ne donnent pas de résultats satisfaisant.

Pour pallier à cette difficulté, le paradigme généralement adopté ces dernières années est de traiter la séquence de vecteurs dans son ensemble via l'utilisation de *noyau de séquences* (terme potentiellement dangereux car l'ordre des trames dans le signal n'a pas d'importance pour ces noyaux). Cette méthode permet de traiter des grandes quantités de données à un coût réduit (linéaire par rapport au nombre de trames).

3.3.1 Moyennage de noyaux vectoriels sur la séquence

Les premiers noyaux de séquences appliqués dans le cadre des évaluations NIST-SRE proviennent des travaux de [Campbell et al., 1999] pour les classifieurs polynomiaux. Nous présentons deux approches basées sur des moyennes d'expansion des trames de la séquence. Le GLDS, très utilisé dans les systèmes de VAL, s'appuie sur une expansion polynomiale. L'autre permet de pallier les limitations du GLDS en généralisant l'approche pour n'importe quel noyau.

3.3.1.1 Generalized Linear Discriminant Sequence Kernel : GLDS

Un noyau efficace, présentant, des performances proches des méthodes GMM-UBM aux évaluations NIST-SRE, est le noyau *GLDS* [Campbell et al., 2006]. Il consiste en une projection explicite des séquences dans un espace de dimension fixe en utilisant une moyenne des expansions polynomiales des vecteurs cepstraux, suivi d'un produit scalaire linéaire. Dans ce cas, l'astuce du noyau n'est pas réellement utilisée.

La forme originale du noyau GLDS fait intervenir une expansion polynomiale Φ_p, composée de monômes jusqu'à un degré donné p. Par exemple, si $p = 2$ alors l'expansion polynomiale est un vecteur à deux dimensions,

$$x = [x_1, x_2]^T \Rightarrow \Phi_2(x) = [1, x_1, x_2, x_1^2, x_1 \cdot x_2, x_2^2]^T.$$

Le noyau entre deux séquences de vecteurs $X = x_1 \dots x_{T_x}$ et $Y = y_1 \dots y_{T_y}$ est exprimé par un produit scalaire normalisé entre les moyennes d'expansion Φ_p :

$$K_{\text{GLDS}}(X,Y) = \overline{\Phi_p}(X) R^{-1} \overline{\Phi_p}(Y), \tag{3.12}$$

avec $\overline{\Phi_p}(X) = \frac{1}{T_X} \sum_{t=1}^{T_X} \Phi_p(x_t)^T$.

R est la matrice des moments d'ordre 2 des expansions $\overline{\Phi_p}_B = [\overline{\Phi_p}(B_1) \dots \overline{\Phi_p}(B_N)]$ estimée sur une population de développement $B = B_1, \dots, B_N$. Pour des raisons d'efficacité, la matrice R est généralement diagonale et l'ordre p de l'expansion ne dépasse pas 3.

Dans le cas du GLDS, ce sont donc les données imposteurs qui structurent l'espace des caractéristiques (*feature space*) grâce à cette matrice de normalisation. Cette matrice est en effet calculée

sur une population correspondant aux pseudo-imposteurs, généralement les données d'apprentissage du modèle du monde.

3.3.1.2 Généralisation du GLDS à toutes les expansions

[Louradour et Daoudi, 2005] définit une classe de noyaux permettant d'implémenter le GLDS pour des expansions autre que polynomiales et pour n'importe quelle p. Cette approche permet de pallier aux limitations théoriques du noyau GLDS. Il n'est en effet pas possible de généraliser l'approche en l'état à des expansions infinies (*e.g.* comme dans le cas des noyaux gaussiens).

La méthode employée s'inspire de l'*empirical kernel map* [Schölkopf et Smola, 2002]. La stratégie consiste à effectuer des approximations en utilisant un dictionnaire d'exemples en nombre fini (*e.g.* des trames représentatives dans les données pseudo-imposteur). Le noyaux de séquence s'exprime comme une moyenne de valeurs de noyaux entre des éléments inter-séquences (entre la séquence des données d'apprentissage du locuteur et les séquences du dictionnaire). Posons $B = \{b_1, \ldots, b_n\}$ comme le dictionnaire d'exemples, et k, le noyau à approximer, la projection s'exprime alors comme :

$$\overline{\Phi}(x_1 \ldots x_t) = \frac{1}{T} \sum_{i=1}^{T} \Phi(x_t), \tag{3.13}$$

$$\Phi(x) = [k(x, b_1) \ldots k(x, b_n)]^T. \tag{3.14}$$

Le principal défi est de définir une méthode pour construire un dictionnaire pertinent. Dans [Louradour et al., 2006], l'utilisation de la factorisation de Cholesky incomplète constitue une solution à la réduction de la dimensionnalité en minimisant l'erreur dans l'approximation faite.

Pour ces deux méthodes, appliquées avec succès aux campagnes NIST-SRE, la structuration de l'espace des caractéristiques est guidée par une matrice de normalisation estimée sur les données des pseudo-imposteurs. Le modèle générique décrit au chapitre 2 n'intervient donc pas directement, mais ses données d'apprentissage sont utilisées.

3.3.2 Exploitation des modèles génératifs dans les classifieurs discriminants

L'intérêt des modèles GMM et l'approche GMM-UBM pour la VAL ont été évoqués au chapitre précédent. Nous venons d'aborder l'intéret des méthodes discriminantes à base de SVM dont certaines ont été appliquées avec succès dans les campagnes d'évaluations du NIST-SRE. Dans ce cadre, la combinaison des méthodes discriminantes et génératives est particulièrement intéressante. Cette approche n'est pas nouvelle et consiste à projeter une séquence de données d'entrée de longueur variable (comme un signal de parole) sur un vecteur de dimension fixe, en utilisant les paramètres des modèles génératifs. Nous présentons deux types de méthodes, celles travaillant dans l'espace des paramètres et celles travaillant dans l'espace des scores.

3.3.2.1 Noyaux dans l'espace des paramètres

Une méthode permettant de travailler dans l'espace des modèles et fortement reliée à D-MAP (2.3.2) consiste à adapter la formulation de la divergence de *Kullback-Liebler* (KL) pour son utilisation dans un SVM. La divergence KL entre deux distributions, $f(x)$ et $g(x)$ s'exprime de la façon suivante :

$$KL(f|g) = \int f(x) log\Big(\frac{f(x)}{g(x)}\Big) dx. \tag{3.15}$$

De par sa nature anti-symétrique, la divergence KL ne satisfait pas les conditions de Mercer et n'est donc pas utilisable directement comme noyau. Les premières solutions apparaissent dans [Moreno et al., 2004] par une symétrisation et une exponentiation de la distance.

Une formulation performante, parue dans [Campbell et al., 2006], consiste à trouver une borne supérieure à la divergence KL et à utiliser directement les super-vecteurs de moyennes dans le calcul du noyau (espace des paramètres). Précisément, dans le cas où seuls les paramètres de moyenne des modèles GMM des locuteurs ont été adaptées, la divergence KL entre deux modèles de locuteurs S et S' est bornée par :

$$KL(S|S') \leq \sum_{i=1}^{N} \alpha_i KL(\mathcal{N}(-;\boldsymbol{\mu}_i^S, \boldsymbol{\Sigma}_i) \| \mathcal{N}(-;\boldsymbol{\mu}_i^{S'}, \boldsymbol{\Sigma}_i)), \tag{3.16}$$

où μ_i^S et $\mu_i^{S'}$ sont les vecteurs de moyenne correspondant à la composante i des modèles S et S', et Σ_i est la matrice de covariance de la composante i partagée par les deux modèles. Cette reformulation considère que la divergence KL est bornée par la combinaison linéaire des divergences entre chaque gaussienne. Le noyau approprié qui satisfait les conditions de Mercer est obtenu par linéarisation (en supposant les matrices de covariances diagonales).

$$K(X_S, X_{S'}) = \sum_{i=1}^{N} \Big(\sqrt{\alpha_i} \Sigma^{-\frac{1}{2}} \mu_i^S \Big)^t \Big(\sqrt{\alpha_i} \Sigma^{-\frac{1}{2}} \mu_i^{S'} \Big). \tag{3.17}$$

Cette méthode obtient de très bons résultats aux évaluations NIST-SRE et permet l'application conjointe des techniques de compensation de canal pour les SVM.

3.3.2.2 Noyaux dans l'espace des scores

La combinaison des méthodes discriminantes et génératives dans l'espace des scores a été traitée avec succès dans [Wan et Renals, 2002; Quan et Bengio, 2002; Fine et al., 2001]. La méthode originelle est connue sous le nom de *Noyau de Fisher*, développée par [Jaakkola et Haussler, 1998]. Elle fut ensuite généralisée par [Smith et al., 2001] sous le nom de méthodes de l'espace des scores ou *Score spaces*. Dans [Schölkopf et Smola, 2002], ces méthodes sont regroupées sous le nom de *Noyau naturels*. Nous présentons d'abord la formulation générique de ces techniques pour ensuite examiner un cas spécifique : l'espace des scores de vraisemblance.

[Wan, 2003] définit l'espace des scores de la façon suivante : étant donné un ensemble de k modèles génératifs, M_k et leurs paramètres θ_k, la formulation générique de la projection d'une séquence $\mathbf{X} = \{x_1, \ldots, x_n\}$ dans l'espace des scores est donnée par :

$$\Psi_{\hat{F}}^f(\mathbf{X}) = \Psi_{\hat{F}} f(p_k(\mathbf{X}|M_k, \boldsymbol{\theta}_k)), \tag{3.18}$$

où $f(p_k(\mathbf{X}|M_k,\boldsymbol{\theta}_k))$ est une fonction des scores d'un ensemble de modèles génératifs, appelée « argument du score ». $\Psi_{\hat{F}}$ est la projection de cette fonction par l'opérateur de score $\hat{F}$. Les propriétés de l'espace résultant dépendent du choix de l'opérateur et de l'argument utilisé. Différentes options ont été proposées par [Smith et al., 2001]. Nous en présentons un cas particulier : le noyau de Fisher.

Espace des scores de vraisemblances Prenons comme argument de score f, la log-vraisemblance (on notera cette quantité $\ell(.)$ pour $\log(p(.))$) d'un seul modèle génératif, M, paramétrisé par $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}]$, et choisissons la dérivée première comme opérateur de score $\hat{F}$, nous obtenons la projection dans l'espace des scores de vraisemblance :

$$\Psi^{\ell}_{\nabla}(\mathbf{X}) = \nabla_{\theta}\ell(\mathbf{X}|M,\boldsymbol{\theta}) = [\nabla_{\boldsymbol{\mu}}\ell(\mathbf{X}|M,\boldsymbol{\theta}) \ , \ \nabla_{\boldsymbol{\Sigma}}\ell(\mathbf{X}|M,\boldsymbol{\theta}) \ , \ \nabla_{\boldsymbol{\gamma}}\ell(\mathbf{X}|M,\boldsymbol{\theta})]. \tag{3.19}$$

En statistique, la *fonction score* est définie comme la dérivée première du log-vraisemblance. Cette projection est connue sous le nom de *Fisher mapping*, Ψ_{Fisher}, et a été appliquée avec succès à l'analyse de séquences biologiques par [Jaakkola et Haussler, 1998]. Chaque composante du vecteur correspond à la dérivée de la log-vraisemblance par rapport à un des paramètres du modèle. Dans [Wan, 2003], la projection est augmentée de la valeur du log-vraisemblance. L'intuition générale du noyau de Fisher est de discriminer deux jeux de données en mesurant leur influence sur les paramètres du modèle, *i.e.* définir en quelle mesure les paramètres du modèle doivent être modifiés pour s'adapter aux données.

Noyaux de Fisher Pour une projection dans l'espace des scores, la métrique définie par les modèles génératifs n'est généralement pas euclidienne. Pour calculer correctement un noyau, il faut définir une matrice de normalisation. Dans le cas du *mapping* de Fisher, cette matrice est la matrice d'information de Fisher, $\mathcal{I}(\boldsymbol{\theta})$. Elle est calculée comme la matrice de covariance des projections de l'espace des scores en utilisant les données de pseudo-imposteurs (comme la matrice R du noyau GLDS). Le noyau de Fisher est donc défini comme :

$$K_{Fisher}(\mathbf{X},\mathbf{Y}) = \Psi_{Fisher}(\mathbf{X})^{T} \cdot \mathcal{I}(\boldsymbol{\theta})^{-1} \cdot \Psi_{Fisher}(\mathbf{Y}). \tag{3.20}$$

Dans [Wan, 2003], la normalisation par la matrice d'information de Fisher peut être interprétée comme une normalisation vers une moyenne 0 et une variance unité des composantes de l'espace des scores. De par les tailles souvent importantes des problèmes, la matrice est souvent considérée diagonale ou remplacée par l'identité. Il faut noter la forme similaire du noyau de Fisher et celle du GLDS présentée au paragraphe 3.3.1.

Méthodes de représentation relative des locuteurs : modèles d'ancrage Les méthodes de représentation relative des locuteurs vues au chapitre 2 peuvent être formalisées dans l'espace des scores. Ces méthodes utilisent en effet les scores de vraisemblance d'une cohorte de modèles de locuteur bien appris afin de représenter le locuteur cible de façon relative dans l'espace des scores.

Il serait intéressant de normaliser les exemples d'apprentissage à l'entrée du SVM par une matrice de covariance pour cette approche, ce qui intégrerait le formalisme des modèles d'ancrage

aux méthodes à noyaux[1]. Il est à noter qu'aucune méthode à base de SVM n'a été proposée dans ce cadre.

3.4 Structure applicative d'un système de VAL basé sur les SVMs

Nous présentons dans cette section la structure générique des systèmes de VAL basés sur les SVM. Ainsi, nous détaillons les étapes d'apprentissage, de calcul de scores et de normalisation pour le classifieur SVM et les noyaux de séquences associés.

3.4.1 Projection des séquences de trames vers des vecteurs

La mise en oeuvre d'un noyau est difficile pour des séquences de taille variable. Ainsi, les systèmes de VAL basés sur les SVM s'attachent à trouver une expansion de la séquence (ou *feature mapping*) correspondant à la fonction $\phi(.)$ pour ensuite appliquer un produit scalaire et former un noyau $K(.,.) = <\phi,\phi>$ (voir le schéma 3.3 de la structure générique d'un système SVM). L'astuce du noyau (*kernel trick*) n'est pas réellement utilisée puisqu'un noyau linéaire est employé après le passage dans un espace de grande dimension.

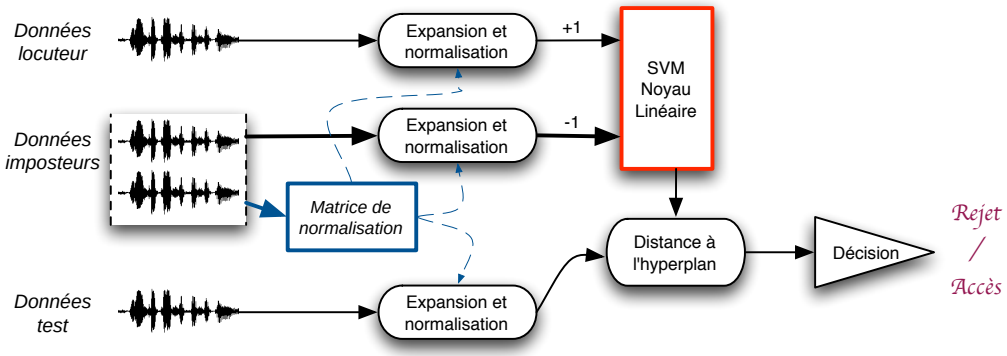


FIG. 3.4: *Structure d'un système de vérification du locuteur basé sur un SVM. Les données du locuteur (étiquetées +1) et les données imposteurs (étiquetées -1) subissent une expansion vers l'espace des caractéristiques. Une matrice de normalisation est utilisée. Les données de test sont projetées de la même manière et le score de vérification correspond à la distance à l'hyperplan du modèle SVM de l'identité proclamée.*

3.4.2 Mise en oeuvre des systèmes basés sur les SVM

L'utilisation des noyaux de séquences décrit précédemment permet de répondre au problème des données de longueurs variables en projetant une séquence de longueur variable vers un vecteur de dimension fixe. Pour un système de VAL, la mise en oeuvre d'un système à base de SVM utilise toujours la procédure suivante :

[1] La technique VZ-Norm permet d'obtenir un effet similaire [Collet, 2006]

– les exemples positifs sont représentés par la projection des données d'apprentissage du locuteur. Pour un locuteur donné, il y a autant d'exemples positifs que de sessions d'apprentissage. Les stratégies visant à couper le signal d'apprentissage en plusieurs segments pour augmenter le nombre d'exemples positifs se sont révélées sans utilité ;
– les exemples négatifs sont représentés par un nombre maximum de locuteurs provenant du corpus de données de pseudo-imposteurs ;
– un modèle de coût lors de l'apprentissage est souvent adopté afin de compenser la grande disproportion entre les données du locuteur et les imposteurs. Ainsi, une erreur de classification sur les exemples positifs est N fois plus coûteuse que sur les exemples négatifs.

En phase de test, les données sont projetées de la même manière que les exemples en phase d'apprentissage. Le score de vérification correspond à la distance entre les données de test et l'hyperplan du modèle correspondant à l'identité proclamée. Cette procédure est illustrée sur le schéma 3.3.

3.4.3 Normalisation des exemples

Une étape de normalisation des vecteurs d'entrée avant d'apprendre l'hyperplan séparateur est généralement employée. Il existe deux classes principales de méthodes utilisées à cet effet.

La première est la normalisation par moyenne/variance : sur chaque dimension du problème de classification (taille du vecteur dans le *feature space*), les données sont centrées et réduites. La moyenne et la variance de chacune de ces dimensions sont estimées sur les données imposteurs.

La seconde est la normalisation par rang [Shriberg et al., 2004] permettant d'obtenir des gains en performance notables. La méthode consiste à projeter les exemples sur une distribution uniforme afin d'obtenir des comportements dynamiques équivalents pour chaque dimension. Ceci est particulièrement utile lorsque aucune connaissance *a priori* n'est disponible. Dans chaque dimension, chaque valeur est remplacée par son rang. Celui-ci est calculé à partir du nombre d'instances négatives dont la valeur est inférieure à celle analysée. Ce rang est ensuite divisé par le nombre total d'instances négatives pour obtenir un nombre entre 0 et 1. Le rang est très intuitif puisque la différence entre deux valeurs ainsi normalisées correspond aux nombres d'instances négatives dans le jeu de données imposteurs, présentes entre ces valeurs.

3.5 Conclusion

Les techniques de modélisation discriminantes du locuteur ont pu être appliquées avec succès dans les évaluations de grandes ampleurs grâce à l'utilisation du formalisme des machines à vecteurs supports (SVM). Pour des problèmes complexes de classification comme celui de la VAL, la capacité des SVM à traiter des problèmes de grande dimension est particulièrement intéressante. De plus, là où la capacité de généralisation des modèles GMM est contrôlée grâce à des seuillages des variances des gaussiennes, la complexité du classifieur SVM est directement minimisée dans sa fonction objective. Afin de traiter de grandes quantités de données, les noyaux de séquences permettent de réduire le coût prohibitif des méthodes s'appuyant sur des calculs de noyaux à la trame. Dans ce cadre, la combinaison des méthodes génératives et discriminantes peut s'avérer intéressante. Le noyau de Fisher et plus généralement les méthodes de l'espace des scores sont

des formalismes adaptées à l'exploitation des GMM pour les SVM. Un de leur désavantage est la grande taille du problème résultant.

Pour toutes ces méthodes, les données imposteurs jouent un rôle de normalisation des distances dans l'espace caractéristique du SVM (*feature space*). Ainsi le modèle générique n'est pas utilisé, mais son influence n'est qu'indirecte, puisque ses données d'apprentissage sont utilisées dans la normalisation des exemples pour les noyaux de séquences.

CHAPITRE 4

Caractérisation du locuteur par des informations « haut-niveau »

Sommaire

4.1 Introduction

Les systèmes état-de-l'art, dits « bas-niveau » ou « acoustique », s'attachent à modéliser les distributions des coefficients cepstraux (extraits sur une fenêtre très courte) de manière non-ordonnée.

En effet, les systèmes à base de modèles à mélange de gaussiennes ou de machines à vecteur support modélisent la séquence de trames dans son ensemble sans prendre en compte l'ordre temporel ou une structure spécifique du signal de parole.

Deux problèmes fondamentaux sont liés à cette approche. Le premier est d'ignorer la structure sous-jacente de la parole, où les spécificités du locuteur sont liées à des fenêtres d'analyse plus grandes, comme peuvent l'être les phonèmes, syllabes, mots ou les événements acoustiques. Des approches émergentes s'attachent d'ailleurs à combiner les différentes sources en développant des systèmes dits « haut-niveau » sur des fenêtres plus grandes pour extraire de l'information orthogonal aux systèmes « bas-niveau ». Le second problème tient à la nature même du signal de parole qui, comme nous l'avons vu dans le chapitre 1, contient aussi de l'information linguistique et des distorsions dues au canal. Or les stratégies « bas-niveau » prenant en compte le signal de manière globale, ne permettent pas de normaliser complètement ces informations non caractéristiques du locuteur, surtout en mode indépendant du texte. Certaines techniques vues au chapitre 2 permettent de normaliser l'effet du canal, cependant le problème du contenu linguistique dans le signal reste entier.

La segmentation du signal de parole permet une analyse du signal de parole sur des unités prédéfinies. Cette segmentation utilise des *a priori* acoustiques sur la structure même de la parole. Parmi ces techniques, les plus courantes sont les analyses lexicales, phonétiques, prosodiques, *etc*. Tous ces systèmes sont regroupés sous le terme « haut-niveau » ou encore « stylistiques » car ils s'attachent à analyser les distinctions entre locuteurs par des caractéristiques reliées à la structure de la langue. Les performances attribuées aux systèmes « haut-niveau » ne rivalisent généralement pas avec les systèmes cepstraux mais ces systèmes permettent d'extraire de l'information nouvelle et complémentaire. Bien que le développement de ces méthodes ait eu lieu grâce à l'augmentation de la quantité de données (comme le montre les travaux de l'atelier *SuperSID* [Reynolds et al., 2002] pour la tâche *NIST-SRE Extended Data Task*), ces systèmes montrent des performances satisfaisantes dès lors que la durée d'apprentissage est raisonnable.

Nous présentons en premier lieu les différents niveaux de segmentation possibles pour la construction de systèmes haut-niveau. Nous distinguerons ensuite les deux types principaux de méthodes utilisées après l'étape de segmentation par la façon dont elles traitent le signal ainsi segmenté :

- Les premières s'attachent à la modélisation de chaque type de segment du signal de parole. Un locuteur sera par exemple modélisé par plusieurs modèles acoustiques correspondant aux phonèmes. Nous appellerons cette stratégie de modélisation « intra-segmental » ;
- Les secondes utilisent les segments comme unités (ou *token*) et ont pour objectif d'analyser la dynamique entre les unités. Nous appelons cette stratégie de modélisation « inter-segmental ».

Nous montrons comment ces techniques permettent de normaliser l'influence du contenu linguistique dans la suite de ce chapitre. Nous soulignons de plus que l'objectif de ces techniques est de générer une représentation du signal spécifique au locuteur et non une représentation du message, car l'analyse se porte sur la façon dont le message a été prononcé et non sur le contenu linguistique.

4.2 Segmentation du signal de parole pour la VAL

La segmentation en unités acoustiques du signal de parole pour la vérification du locuteur a surtout été utilisée dans le cadre d'applications dépendantes du texte. Pour les applications à base de mots de passe, il est nécessaire de pouvoir modéliser le maximum d'information pour une très courte durée du signal. En effet, l'enveloppe spectrale d'un locuteur est difficile à estimer sur peu de données.

Nous présentons dans la suite les différentes unités acoustiques possibles pour la segmentation du signal de parole. Nous examinons ensuite les stratégies de production de segmentations multiples grâce à une approche multilingue couramment utilisée en identification automatique de la langue (IAL).

4.2.1 Unités segmentales du signal de parole

Formellement, un segmenteur en unités peut s'exprimer comme la transformation d'une séquence de trames $X = \{x_1, \ldots, x_T\}$ en une séquence de symboles d'un dictionnaire $D = \{d_1, \ldots, d_N\}$ avec $N \leq D$. A chacun des symboles produits est généralement associé une unité temporelle ou une frontière dans le signal. C'est le problème général de la reconnaissance de la parole (RAP) avec un dictionnaire phonétique ou lexical.

Les différentes unités segmentales peuvent être regroupées selon l'unité temporelle qui les caractérisent. De façon non-exhaustive, ces différentes unités sont :

- Les unités prosodiques : généralement une quantification des caractéristiques prosodiques sur une fenêtre glissante (valeurs du pitch (f0), de l'énergie, ...), ou de la durée des phonèmes comme dans [Ferrer et al., 2003] ;
- Les unités phonétiques : générées par un décodage acoustico-phonétique (DAP) du signal de parole. Ces unités sont dépendantes de la langue ;
- Les unités de classes phonétiques : les classes phonétiques sont des regroupements de phonèmes ou de triphones, permettant de caractériser des traits caractéristiques (nasales, fricatives, voisées/non-voisées, ...). Ces unités sont très utilisées en RAP pour l'apprentissage des modèles et plus récemment en VAL ;
- Les unités syllabiques : ces unités correspondent à une unité temporelle plus grande que les phonèmes. Elles peuvent correspondre à des séquences de pseudo-phonèmes indépendants de la langue, produits par un reconnaisseur phonétique multilingue (Broad Phone Recogniser dans [Martin et al., 2006]), ou plus classiquement par des unités syllabiques provenant d'un dictionnaire de règles expertes, comme pour le logiciel tsylb2 [1] ;
- Les unités lexicales : ce sont les unités les plus structurées. Elles sont souvent obtenues par un décodage automatique de parole, à défaut de posséder des données transcrites ;
- Les unités non-supervisées : elles représentent des événements acoustiques indépendants de la structure de la langue et possèdent une structure implicite propre, *i.e.* elles n'ont pas de signification linguistique propre (*e.g.* zones de stabilité fréquentielle).

[1] Développé par Bill Fisher, disponible sur http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm

4.2.2 Approche multilingue pour la segmentation multiple du signal

Les approches multilingues pour la génération de multiples segmentations du signal de parole ont d'abord été utilisées dans les systèmes d'identification automatique de la langue à base de modèles phonétiques. Ces approches trouvent leur fondement à travers des adaptations des travaux de [Zissman, 1996]. Dans ces travaux, l'auteur présente les quatre approches principales pour la reconnaissance de la langue. Parmi ces quatre approches, PRLM (*Phone Recognition followed by Language Modeling*) et son extension multilingue PPRLM (*Parallel PRLM*) sont les plus couramment utilisées. Les systèmes « haut-niveau » basés sur des unités phonétiques en VAL utilisent aussi ce paradigme.

L'approche PRLM est basée sur un décodage phonétique puis une modélisation N-gramme pour la caractérisation du langage. L'extension multilingue de PRLM, PPRLM, utilise plusieurs décodeurs de différentes langues en parallèle. Le schéma 4.1 résume l'approche PPRLM de [Zissman, 1996]. Plusieurs reconnaisseurs phonétiques de langages différents sont utilisés en parallèle pour segmenter le signal de parole en unités phonétiques. Les langues cibles ne sont pas forcément celles du décodeur phonétique. Un classifieur permet de prendre la décision.

Les dernières améliorations ont été apportées par [Gauvain et al., 2004] en proposant une solution plus robuste pour l'estimation des vraisemblances des N-grammes à travers l'utilisation des treillis de phonèmes générés par le décodeur. L'application de cette technique en VAL peut être trouvée dans [Hatch et al., 2005].

Normalisation du texte et émergence d'une représentation caractéristique du locuteur Considérons un système utilisant les phonèmes pour la vérification du locuteur. La problématique est de distinguer, à l'extrême, deux locuteurs prononçant la même phrase. La sortie d'un reconnaisseur phonétique est alors strictement identique quelque soit le locuteur et n'apporte aucune information permettant de distinguer les locuteurs. Deux techniques permettent de répondre à ce problème, elles ont le même objectif : faire émerger des informations caractéristiques du locuteur plutôt que de décoder le contenu linguistique.

La première consiste à relâcher les contraintes linguistiques sur les modèles de langage phonétique, *i.e.* ne pas optimiser le moteur de reconnaissance afin d'être cohérent avec une structure linguistique. Généralement, un reconnaisseur phonétique en boucle ouverte, *i.e.* sans mémoire, est utilisé à cet effet[Hatch et al., 2005][Campbell et al., 2003]. Le principe du mode en boucle ouverte est d'utiliser un modèle de langage phonétique à distribution uniforme (*i.e.* dont les implications pour la reconnaissance du locuteur ont été étudiées dans [Charlet, 1997]). La stratégie est donc différente du cas de la reconnaissance de la parole : dans les approches inspirées de PRLM, le modèle phono-tactique n'est pas utilisé pour produire une transcription exacte du message mais il est construit *a posteriori* pour caractériser le locuteur. Ainsi, une même phrase pour deux locuteurs différents possède deux décodages phonétiques différents. Ceci produit ainsi des représentations caractéristiques de la façon dont le locuteur a prononcé le message, plutôt que du message prononcé.

La seconde approche consiste à utiliser de multiples reconnaisseurs appris sur des langages différents afin d'obtenir de multiples segmentations phonétiques. L'idée étant de générer différentes représentations du signal pour faire émerger des différences caractéristiques entre les locuteurs.

La richesse du dictionnaire est implicitement augmentée par les nuances de phonèmes provenant des différentes langues. Deux phonèmes « a » provenant de deux locuteurs différents pourront être différenciés par un jeu de phonèmes d'une autre langue. Comme sur la figure 4.1, il est tout à fait possible de ne pas utiliser la langue parlée par les locuteurs pour le reconnaisseur phonétique car ce qui est recherché est une représentation du signal caractéristique du locuteur et non le décodage phonétique réel.

Cette façon de procéder peut se généraliser à toutes les unités acoustiques possibles (c'est ce que nous ferons pour l'AES présenté au chapitre 6) en cherchant à faire émerger une structure de la parole indépendante de la langue et du locuteur. Cette structure produit alors les séquences d'unités acoustiques pour un locuteur donné et les informations caractéristiques du locuteur sont tirées de cette représentation.

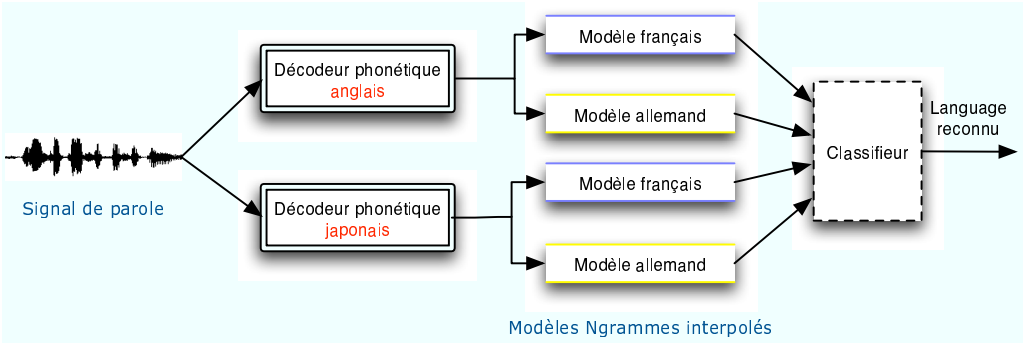


FIG. 4.1: *L'approche PPRLM. Plusieurs reconnaisseurs phonétiques de langages différents sont utilisés en parallèle afin de segmenter le signal de parole en unités phonétiques. Les langues cibles ne sont pas forcément celles du décodeur phonétique. Un classifieur permet de prendre la décision.*

4.3 Modélisation acoustique des unités segmentales du signal

Nous présentons dans la suite les méthodes visant à modéliser le locuteur par plusieurs modèles correspondant aux unités segmentales du signal de parole présentées dans le paragraphe 4.2. Cette approche est qualifiée d'*intra-segmentale*. Cette modélisation intervient à différents niveaux et nous présentons de façon non-exhaustive les systèmes de VAL appartenant à cette catégorie.

4.3.1 Modélisation des unités phonétiques

[Andrews et al., 2001] proposent l'utilisation d'un système basé sur les phonèmes pour la VAL. La nouveauté de ce système est de se concentrer sur la prononciation des phonèmes en enrichissant le système par des reconnaisseurs phonétiques multilingues. Les auteurs montrent l'amélioration apportée par cette intégration. Le locuteur est donc modélisé par une cohorte de modèles GMM, chacun étant spécifique à un phonème et à une langue précise. Les auteurs valident l'hypothèse selon laquelle le reconnaisseur utilisant la langue des locuteurs n'est pas nécessaire, puisque son retrait du système n'amène qu'une dégradation légère des performances.

4.3.2 Modélisation syllabique

[Martin et al., 2006] proposent une représentation du signal indépendante de la langue en se basant sur des pseudo-syllabes comme unité de segmentation. Une pseudo-syllabe est représentée par une séquence d'unités phonétiques indépendantes de la langue. Quatre classes d'unités sont définies : voyelles/diphtongues, nasales/semi-voyelles(glides), fricatives et enfin stops/silence. Des approches similaires à base de multi–grammes peuvent être trouvées dans [Farinas, 2002].

[Baker et al., 2005] étendent ce paradigme à la VAL. La modélisation de ces 4 classes permet d'obtenir des performances satisfaisantes. La modélisation d'événements pseudo-syllabiques consiste à modéliser les 4^3 triplés de syllabes possibles. Un modèle GMM à 32 composantes ne modélisant que les 16 triplets les plus fréquents suffit à obtenir un système robuste.

Des travaux sur la modélisation d'unités syllabiques provenant de *tsylb2* utilisant une modélisation prosodique de chacun des événements syllabiques ont été développés dans le cadre du système *SNERF* de [Shriberg et al., 2004].

4.3.3 Modélisation lexicale contrainte

Les travaux de [Sturim et al., 2002] proposent une méthode de modélisation du locuteur par l'apprentissage de GMM sur des mots spécifiques. Cette méthode nécessite l'utilisation d'un reconnaisseur de parole ou de données transcrites. Deux méthodes de sélection des mots clés sont proposées montrant des taux de performances équivalents :

- Sélection des mots les plus fréquents (environ 50)
- Sélection des mots selon des *a priori* linguistiques propres à capturer des informations spécifiques au locuteur.

Cette méthode correspond aussi à une normalisation du texte pour la modélisation du locuteur. Les mots sélectionnés ont une grande probabilité d'apparaître dans les signaux de parole conversationnelle. Cette sélection permet de réduire l'influence du message prononcé. Une extension de ces travaux par une modélisation HMM peut être trouvée dans [Boakye et Peskin, 2004].

4.3.4 Modélisation de zones de stabilité

La modélisation des zones de stabilité fait partie des unités de segmentation non-supervisées car elle est totalement décorrélée de la structure de la langue. Un méthodologie d'obtention d'unités acoustiques indépendantes du contenu linguistique est proposée dans [Chollet et al., 1999]. Ces unités sont appelées ALISP pour *Automatic Language Independent Speech Processing*.

4.3.5 Modélisation des classes phonétiques

Les approches précédentes ont pour désavantage de fragmenter les données et demandent une certaine quantité de données pour être performantes, *i.e.* il existe une bijection entre un modèle et ses données. [Stolcke et al., 2005] proposent une méthode pour éviter cette fragmentation en utilisant les transformations MLLR (*Maximum Likelihood Linear Regression*) utilisées dans un

décodeur de parole afin de modéliser des classes phonétiques. Ces transformations étant partagées par plusieurs classes phonétiques, cette méthode évite la fragmentation des données. Cette approche est coûteuse car elle nécessite l'utilisation d'un décodeur de parole.

Le système utilise plusieurs transformations MLLR tirées du décodeur de parole et estimées à deux niveaux : le premier niveau utilise les transformations caractérisant les phonèmes « obstruants » et « non-obstruants », le second définit huit transformations supplémentaires (stops et fricatives voisées/non-voisées, voyelles basses et hautes, phonèmes retroflexes, nasales).

Le système, devenu très prisé par la communauté, utilise les matrices de transformations MLLR comme une expansion des données vers un vecteur caractérisant le locuteur, utilisé ensuite comme entrée dans un SVM. Ces transformations sont celles des modèles HMM phonétiques utilisés dans un reconnaisseur de parole.

Dans le cadre d'un modèle GMM, la transformation MLLR simple consiste à trouve une transformation affine $< A, b >$ à appliquer aux vecteurs de moyenne du modèle pour passer d'un modèle indépendant à dépendant du locuteur. Si $\boldsymbol{\mu}_S^k$ et $\boldsymbol{\mu}_W^k$ sont respectivement les vecteurs de moyennes de dimension d de la composante d'indice k du locuteur cible et du modèle du monde, alors la régression s'effectue comme suit :

$$\boldsymbol{\mu}_S^k = A \cdot \boldsymbol{\mu}_W^k + b. \tag{4.1}$$

A l'inverse de MAP, où la transformation est appliquée par gaussienne, la transformation MLLR est la même pour toutes les gaussiennes. Elle consiste à effectuer une régression dans l'espace des paramètres cepstraux. Les matrices A et b ont donc pour dimension $d \times d$ et $d \times 1$. Après concaténation des vecteurs, le vecteur résultant aura une taille de $K \times d(d+1)$ pour les K transformations. Un système de ce type montre des performances proches des systèmes GMM-UBM. De plus, l'information capturée est complémentaire car la combinaison des deux systèmes permet de réduire significativement les taux d'erreur.

4.4 Modélisation de la dynamique des unités segmentales

Nous abordons ici de manière générale les systèmes de vérification du locuteur basés sur une analyse dynamique des unités segmentales. Cette approche est qualifiée d'*inter-segmental*. Ces systèmes utilisent les représentations du signal telles qu'elles ont été abordées au paragraphe 4.2. La mise en oeuvre de ces systèmes est abordée à la section suivante. Nous présentons de façon non exhaustive les systèmes de ce type utilisés en VAL.

4.4.1 Analyse dynamique des unités prosodiques

[Adami et al., 2003] proposent une méthode pour segmenter le signal de parole par des unités prosodiques. Une analyse du contour prosodique permet de définir deux unités, **+** et **-**, selon que le pitch et l'énergie aient respectivement une dérivée positive ou négative. Les régions non voisées sont modélisées par le symbole **uv**. Une extension consiste à modéliser le pitch et l'énergie par des symboles modélisant les états joints de ces deux tokens, précisément **++**, **--**, **+-**, **-+**.

L'ajout d'information de durée se fait par quantification : **L** (long), **M** (medium) et **S** (short). Le contexte phonétique peut être ajouté grâce à un alignement préalable. D'après les résultats, le système utilisant toutes ces unités (Slope-Phone-Duration) présente des performances proches d'un système GMM-UBM pour un protocole utilisant 8 conversations par locuteur pour l'apprentissage (environ 20 minutes de parole).

4.4.2 Analyse dynamique des unités phonétiques

Un système de VAL analysant la dynamique d'unités phonétiques a été proposé par [Andrews et al., 2001] et [Campbell et al., 2004a] en s'inspirant des travaux de [Zissman, 1996] sur l'approche PPRLM. Il est à noter que le décodage est non-contraint en utilisant une reconnaisseur phonétique en boucle ouverte[Campbell et al., 2003][Hatch et al., 2005]. Le système est de plus composé de multiples langages générant ainsi de multiples séquences de phonèmes dans des langages différents.

L'analyse spécifique du locuteur s'effectue par une vectorisation du flux en calculant la fréquence et la nature des N-grammes de phonèmes. Chaque vecteur ainsi produit est concaténé avec ceux des autres langages afin d'être utilisé en entrée d'un SVM. Dans ce cas, l'utilisation de multiples langages permet de générer de multiples représentations du signal ne correspondant pas forcément au message prononcé.

Dimension inter-flux [Jin et al., 2003] choisissent l'analyse dans la dimension inter-langage comparée à l'analyse dans la dimension temporelle. Les séquences d'unités sont de longueur égale au nombre de langages et la corrélation mesurée est inter-flux. Cette étape nécessite un alignement dynamique entre les flux de phonèmes. L'analyse en N-grammes ne prend pas en compte l'ordre des tokens qui n'a pas de sens dans cette dimension. Les auteurs montrent que la dimension inter-flux et la dimension temporelle sont complémentaires puisque la combinaison des deux permet de réduire les taux d'erreurs.

4.4.3 Analyse des différences idiolectales

Les différences idiolectales, *i.e.* ou approximativement « tics de langage », sont des systèmes se concentrant sur l'analyse lexicale du signal de parole pour discriminer les locuteurs. Leur implémentation n'a été possible qu'avec l'apparition de tâches de vérification comportant une quantité importante de données. L'étude de [Doddington, 2001] fait référence en la matière, en proposant une méthode de construction d'un modèle de langage sur les mots prononcés. L'étude se base sur les transcriptions (automatiques ou non) d'une base de données audio. Une modélisation N-grammes d'ordre 2 permet d'obtenir un système de vérification avec des performances satisfaisantes lorsqu'il est combiné avec un système standard GMM-UBM.

Le désavantage certain d'un tel système est son caractère monolingue et indépendant du texte. En effet, un tel système ne marche pas sur une autre langue que celle traitée par le décodeur de parole. Parmi les systèmes présentés, ce système est le seul à s'appuyer sur la structure de la langue en analysant le message prononcé.

4.4.4 Analyse dynamique des unités segmentales non-supervisées

Il est possible d'utiliser des segmenteurs entièrement non-supervisés *i.e.* totalement indépendants de la structure de la langue. L'analyse dynamique des unités ALISP a, par exemple, donné de bons résultats [El Hannani et Petrovska-Delacretaz, 2005].

Dans [Torres-Carrasquillo et al., 2002], la généralisation du PPRLM pour l'identification de la langue en utilisant des GMM par langue comme segmenteur est appliquée avec succès. Les unités utilisées dans le flux de sortie sont les indices des gaussiennes du GMM. Pour chaque trame, l'unité produite est l'indice de la gaussienne à plus forte vraisemblance. C'est une unité non-supervisée. L'avantage de cette méthode est qu'il est très facile d'augmenter le nombre de langues puisque il n'est pas nécessaire d'apprendre des décodeurs phonétiques spécifiques.

Le système AES présenté dans le chapitre 6 trouve son fondement dans cette approche mais pour la reconnaissance du locuteur. Dans notre cas, c'est le modèle du monde qui permet de générer une représentation du signal adaptée à la vérification du locuteur.

4.5 Mise en oeuvre des systèmes « inter-segmental »

Nous formalisons le problème de reconnaissance du locuteur lorsque un système analyse la dynamique du signal, représenté comme une séquence d'unités segmentales. Les systèmes à unités phonétiques, prosodiques ou idiolectales ont été présentés au paragraphe précédent.

La structure générale de ces systèmes est présentée dans la suite de ce chapitre. La méthode généralement employée consiste à modéliser les données d'un locuteur par une approche N-grammes puis d'obtenir les scores de vérification par une formulation du rapport de vraisemblance appropriée pour ce type de modèle.

4.5.1 Modélisation du locuteur

A partir de la segmentation produite pour le signal d'apprentissage d'un locuteur, la modélisation des données du locuteur est effectuée par un modèle N-gramme. Les spécificités du locuteur exploitées par une telle modélisation prennent en compte la nature des événements (représentée par les N-grammes) et leur apparition (représentée par les probabilités associées).

Le type de modèle utilisé est appelé N-gramme d'unités (*token-Ngram*), ou approche par « sac de N-gramme ». Cette approche est effectuée à la manière des travaux de [Doddington, 2001]. Le principe est de transformer une séquence de symboles, généralement dans un ordre temporel (mots, prosodie, phonèmes, ...) en unités k et leur probabilité associée $p(k)$.

Soit une séquence d'événements acoustiques $\mathbf{X}_S = e_1,\ldots,e_T$ des données d'apprentissage du locuteur S. Pour une analyse en 2-gramme, chaque séquence du sac de N-grammes $B = \{e_1\_e_2,\ldots,e_{T-1}\_e_T\}$ est considérée comme une unité k dont la probabilité est calculée. Considérons le sac de N-grammes B, contenant tous les 2-gramme $k = \{e_i\_e_j\}$ uniques, alors le modèle contient la probabilité de ces séquences donnée par la fréquence d'apparition dans les données X_S

pour le locuteur S.

$$p(k|\mathbf{X}_S) = p(e_i\_e_j|X_S) = \frac{C(k|\mathbf{X}_S)}{\sum_k C(k|\mathbf{X}_S)}, \forall k \in B, \tag{4.2}$$

où $C(.)$ est l'opérateur de compte d'unités. Le schéma **??** illustre l'approche en utilisant un dictionnaire de 2 tokens et une analyse en 2-gramme.

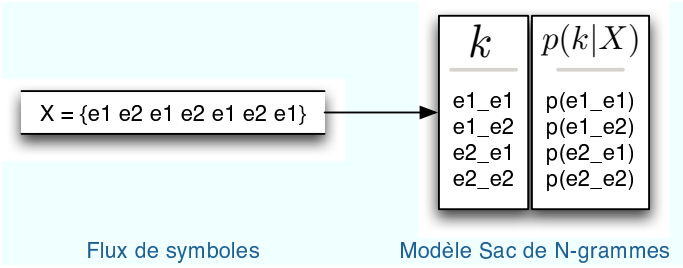


FIG. 4.2: *Modélisation par sac de* N*-grammes. Le dictionnaire d'événements acoustiques contient 2 symboles. Une analyse en* N*-grammes (appelé token-Ngram) calcule la probabilité de chacune des séquences par leur probabilité d'apparition dans le flux de symboles.*

Pour chaque fichier d'apprentissage correspondant à un locuteur, un modèle de ce type est construit. Il est possible de regrouper des séquences de longueurs différentes dans un même modèle afin d'augmenter la capacité de modélisation. Un modèle N-gramme du monde est également nécessaire pour calculer la probabilité d'apparition dans les données imposteurs.

4.5.2 Calcul du score de décision

Après avoir construit des modèles N-grammes, représentant les statistiques sur les événements acoustiques, nous rappelons les deux techniques principales pour calculer un score de vérification.

4.5.2.1 Détecteur Rapport de Vraisemblance : LLR

La première méthode consiste à estimer le score de vérification par le logarithme du rapport de vraisemblance (LLR) entre un segment de test $\mathbf{X}$, les données d'un locuteur $\mathbf{X}_S$ et celles du modèle du monde $\mathbf{X}_W$. Le score de ce test est exprimé ci-dessous :

$$\mathcal{Y}_S(\mathbf{X}) = \text{LLR}(\mathbf{X}|\mathbf{X}_S, \mathbf{X}_W) = \frac{1}{\sum_k C(k|\mathbf{X})} \cdot \sum_k C(k|\mathbf{X}) \log\Big(\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}\Big), \tag{4.3}$$

où $C(k|\mathbf{X})$, $p(k|\mathbf{X}_S)$ et $p(k|\mathbf{X}_W)$ sont le compte de l'unité k dans le segment de test et leurs probabilités dans les modèles N-gramme.

Cette modélisation est généralement augmentée par une estimation par *maximum a posteriori*. En effet, les stratégies les plus connues pour résoudre le problème de couverture est l'utilisation des techniques dénommées *back-off*. Elles permettent d'estimer la probabilité d'une unité manquante en effectuant des combinaisons des probabilités des N-grammes d'ordre inférieur. Dans le cas de la vérification du locuteur, l'utilisation de MAP permet de revenir à la probabilité dans le modèle

du monde. Dans [Baker et al., 2004], une solution analytique est proposée et exprimée comme suit :

$$\widetilde{C}(k|\mathbf{X}_S) = C(k|\mathbf{X}_S) + \alpha C(k|\mathbf{X}_W), \tag{4.4}$$

où $C(k|\mathbf{X}_S)$ représente le compte du token k dans les données d'apprentissage du modèle S et α est le poids d'adaptation dans l'intervalle $(0,1)$. La vraisemblance devient :

$$\widetilde{p}(k|\mathbf{X}_S) = \frac{\widetilde{C}(k|\mathbf{X}_S)}{\sum_k \widetilde{C}(k|\mathbf{X}_S)}. \tag{4.5}$$

Le schéma **??** résume la méthode de calcul du score par le logarithme du rapport de vraisemblance.

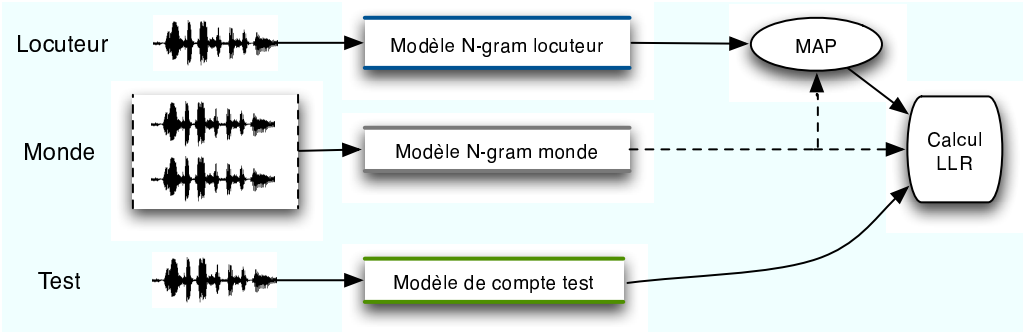


FIG. 4.3: *Calcul du score comme le logarithme du rapport de vraisemblance pour les systèmes « haut-niveau ». Le modèle* N*-gramme du monde est adapté à partir des données du locuteur. Un modèle de compte sur le test est construit pour calculer le LLR.*

4.5.2.2 Méthode basée sur les SVM

La deuxième méthode est basée sur la construction d'un noyau en vue d'une utilisation dans un classifieur SVM. Comme illustrent les expériences présentées en 6.3, la méthode SVM obtient les meilleurs performances. Le noyau TFLLR (*Term Frequency Log-Likelihood Ratio*), introduit par [Campbell et al., 2004b], est une alternative à la méthode TF-IDF (*Term Frequency - Inverse Document Frequency* [Salton et al., 1975]) afin de projeter des N-grammes dans un espace approprié pour l'utilisation d'un SVM. Cette méthode a prouvé qu'elle permettait d'améliorer les performances des systèmes, comparée à un calcul de LLR classique (équation 4.3).

La méthodologie de construction d'un noyau pour un SVM tient à sa formulation présentée dans le chapitre 3. Il s'agit d'exprimer le noyau comme une distance dans un certain espace entre deux signaux d'apprentissage de deux locuteurs. Cette distance est construire en reformulant l'équation 4.3 du LLR sous forme d'un noyau approprié. Le SVM calculera tous les noyaux entre toutes les paires d'exemples. Prenons deux locuteurs S et S', nous exprimons ci-dessous le noyau TFLLR entre leur données respectives : $K_{TFLLR}(\mathbf{X}_S, \mathbf{X}_{S'})$.

Considérons l'unité k appartenant au sac de N-grammes B. En reprenant, l'équation du LLR standard exprimée avec des unités, en posant $\sum_k C(k|\mathbf{X})$ constant, et en estimant $C(k|\mathbf{X})$ sur les données du locuteur S', il est possible de construire un noyau approprié. La vraisemblance de ce

token k sur une séquence de données X est notée $p(k|\mathbf{X})$. La construction du noyau TFLLR est donnée dans la suite :

$$score = \sum_k \frac{C(k|\mathbf{X}_{S'})}{\sum_k C(k|\mathbf{X}_{S'})} \cdot \log\Big(\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}\Big), \tag{4.6}$$

$$= \sum_k p(k|\mathbf{X}_{S'}) \cdot log\Big(\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}\Big). \tag{4.7}$$

$$\tag{4.8}$$

Après linéarisation de la fonction logarithme grâce à son expansion en série de Taylor d'ordre 1 où $\log(x) \sim x - 1$, le noyau s'exprime comme :

$$score \approx \sum_k p(k|\mathbf{X}_{S'}) \cdot \frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)} - \sum_k p(k|\mathbf{X}_{S'}), \tag{4.9}$$

$$= \sum_k \frac{p(k|\mathbf{X}_S)}{\sqrt{p(k|\mathbf{X}_W)}} \frac{p(k|\mathbf{X}_{S'})}{\sqrt{p(k|\mathbf{X}_W)}} - 1. \tag{4.10}$$

$$\tag{4.11}$$

Pour que la matrice du noyau soit semi-définie positive, le noyau est donnée par :

$$K_{TFLLR}(\mathbf{X}_S, \mathbf{X}_{S'}) = \sum_k \frac{p(k|\mathbf{X}_S)}{\sqrt{p(k|\mathbf{X}_W)}} \frac{p(k|\mathbf{X}_{S'})}{\sqrt{p(k|\mathbf{X}_W)}}. \tag{4.12}$$

La construction du noyau réside dans la pondération des vraisemblances des locuteurs par la vraisemblance de l'unité sur les données d'apprentissage du modèle du monde.

En pratique, pour chaque accès, il est nécessaire de calculer la statistique $\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}$. Par conséquent, cette méthode demande la construction d'un modèle N-gramme comme au paragraphe précédent pour chaque accès de test. Le score de vérification correspond finalement à la distance entre le vecteur de test et l'hyperplan appris par le SVM utilisant le noyau TFLLR.

Relation avec la construction du noyau GLDS Il est possible de définir plus formellement le noyau TFLLR en reprenant le formalisme du noyau GLDS défini en 3.8.

$$K_{\text{TFLLR}}(\mathbf{X}_S, \mathbf{X}_{S'}) = \overline{\Phi_p}(\mathbf{X}_S) \cdot R^{-1} \cdot \overline{\Phi_p}(\mathbf{X}_{S'}). \tag{4.13}$$

Il suffit alors de définir Φ_p et R selon le formalisme du TFLLR, avec le sac de N-gramme $B = \{k_1, \ldots, k_b\}$ regroupant toutes les unités.

$$\Phi_p(X, B) = [C(k_1|\mathbf{X}) \ \ldots \ C(k_B|\mathbf{X})], \tag{4.14}$$

$$\overline{\Phi_p}(\mathbf{X}, B) = \frac{1}{\sum_B \Phi_p(\mathbf{X}, B)} \Phi_p(\mathbf{X}, B), \tag{4.15}$$

$$R = \overline{\Phi_p}(\mathbf{X}_W, B)^T \cdot \mathbf{I}_{\dim(B)} \tag{4.16}$$

$$= [p(k_1|\mathbf{X}_W) \ \ldots \ p(k_B|\mathbf{X}_W)]^T \cdot \mathbf{I}_{\dim(B)}. \tag{4.17}$$

La modélisation N-gramme devient alors intégrée au formalisme de l'expansion, où p correspond à l'ordre considéré. La projection $\Phi_p(X,B)$ étant définie comme l'opérateur de compte d'unités, moyenné dans $\overline{\Phi_p}$ pour obtenir le calcul de N-grammes pour les données d'un locuteur. La matrice de normalisation est définie comme une matrice diagonale correspondant aux probabilités N-grammes pour les données pseudo-imposteurs.

Le schéma **??** résume la stratégie de modélisation et de calcul du score pour cette méthode basée sur les SVM en reprenant la structure générique d'un système SVM.

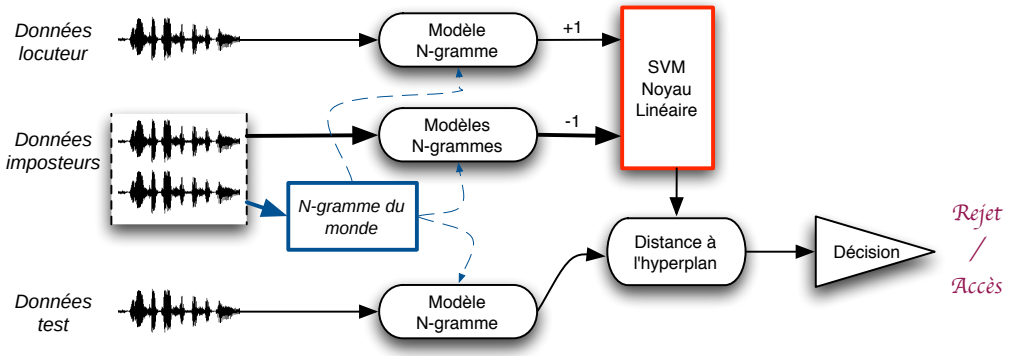


FIG. 4.4: *Mise en oeuvre de la modélisation des locuteurs par la méthode utilisant le noyau TFLLR. Pour chaque accès, un modèle* N*-gramme est estimé. L'expansion TFLLR consiste à pondérer les probabilités par celles estimées sur les données du monde. Chacun de ces modèles est ensuite normalisé par la matrice des probabilités des unités sur les données du monde. Le score de vérification correspond à la distance à l'hyperplan entre le vecteur de test et le modèle SVM issu des vecteurs d'apprentissage (locuteur considéré et pseudo-imposteurs).*

4.6 Conclusion

Ce chapitre a pour objectif de présenter les systèmes dits de « haut-niveau » apparus dernièrement dans les campagnes d'évaluations NIST-SRE. Ces systèmes ont pour particularité de segmenter le signal de parole en unités afin d'effectuer des analyses locales dans le signal. Ces unités ont été présentées en 4.2. Nous avons montré aussi comment ces systèmes utilisent une approche multilingue pour générer plusieurs représentations du signal. Au paragraphe 4.3, nous présentons les différents systèmes utilisant une modélisation par unités segmentales générées alors qu'en 4.4 nous nous focalisons sur les systèmes analysant la dynamique de ces unités pour caractériser le locuteur. Enfin, la mise en oeuvre de ces systèmes à base d'analyse N-gramme a été présentée en 4.5

Nous soulignons que ces techniques ont un objectif sous-jacent visant à extraire l'information spécifique au locuteur plutôt que le décodage du contenu linguistique du message (il n'est par exemple pas possible de distinguer deux locuteurs prononçant une même phrase si un décodage phonétique parfait est produit). Nous assimilons cette approche à une normalisation du texte en générant de multiples représentations du signal, pour faire émerger des différences caractéristiques entre les locuteurs.

Deuxième partie

Contributions : Structuration de l'espace acoustique par le modèle générique dans les approches séquentielles et discriminantes

CHAPITRE 5

Présentation du système GMM-UBM du LIA et du contexte expérimental

Sommaire

Les travaux de thèse présentés dans ce document ont été ponctués par les campagnes internationales *Speaker Recognition Evaluation* organisées annuellement par le NIST [1]. Le système GMM/UBM du LIA[2], *LIA_SpkDet*, a été modifié et amélioré au cours des années en vue de la participation à ces campagnes.

[1] www.nist.gov/speech
[2] Laboratoire Informatique Avignon, www.lia.univ-avignon.fr

Dans ce chapitre, nous présentons dans un premier temps le système et la plate-forme LIA_SpkDet ainsi que les modifications qui ont été apportées au cours des années, dans un second temps, le protocole expérimental utilisé dans le cadre de ces travaux de thèse. Ainsi, la base de données utilisée ainsi que les tâches visées sont présentées. Les participations régulières aux campagnes NIST-SRE ont constitué un travail intensif sur ce système ; nous insistons sur certains développements en présentant des résultats de références ainsi que des expériences comparatives.

5.1 Le système ALIZE/LIA_SpkDet du LIA

Ces dernières années, le LIA a engagé un projet de développement d'outils pour la reconnaissance du locuteur dont les objectifs principaux sont : le développement selon un modèle libre (ou open-source), la qualité du code pour l'utilisation à des fins de recherche, et la simplicité d'utilisation pour leurs diffusions à des fins pédagogiques.

5.1.1 Le toolkit ALIZE

Le Laboratoire Informatique d'Avignon développe une plate-forme open-source (licence LGPL) pour la reconnaissance du locuteur baptisé *ALIZE* [3] [Bonastre et al., 2005]. Ce projet a été réalisé dans le cadre de l'appel à projet *Technolangue*[4]. Cette plate-forme a pour objectif de faciliter le développement d'applications en vérification, identification et segmentation du locuteur. Elle est principalement centrée sur des problèmes de classification à base de GMM et HMM.

Des projets concernant l'intégration de différentes biométries au sein de la plate-forme sont en cours. Cette intégration consiste à étendre les fonctionnalités de la plate-forme pour permettre de traiter plusieurs biométries avec la même philosophie de programmation. Ainsi unifié, le développement d'applications multimodales est facilité (c'est l'un des objectifs d'un projet de recherche en cours[5] au LIA).

5.1.2 Le système LIA_SpkDet

ALIZE est une bibliothèque de fonctions et d'objets (API). Conjointement à celle-ci, le LIA développe une sur-couche contenant les exécutables nécessaires à un système opérationnel dans le domaine de la RAL. Cette plate-forme, LIA_RAL [6], est aussi distribuée en licence libre (GPL) [Bonastre et al., 2005].

LIA_RAL est subdivisé en sous parties dont LIA_Seg pour la segmentation en locuteur et LIA_SpkDet pour la vérification du locuteur. Cette dernière constitue le système de base que le LIA présente aux évaluations NIST-SRE chaque année. Ce « package » est diffusé à la communauté chaque année, permettant aux autres laboratoires de reproduire nos expériences, mais aussi

[3] http ://www.lia.univ-avignon.fr/heberges/ALIZE/
[4] www.technolangue.net
[5] Projet ANR-RNTL 2006 Mistral
[6] http ://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

de faciliter la tâche aux nouveaux venus. Ainsi, les efforts de recherche se portent sur un autre domaine que l'implémentation d'un système GMM-UBM état de l'art.

5.1.3 Les systèmes GMM-UBM, ALIZE/LIA_SpkDet

Les particularités composant les systèmes GMM-UBM du LIA utilisés tout au long de ces années peuvent être trouvées en annexe A. Ces trois systèmes - *Sys04*, *Sys05* et *Sys06* - correspondent respectivement au système pour l'évaluation 2004, 2005 et 2006.

5.2 Protocole expérimental

Le protocole expérimental employé au cours de ces travaux de thèse est basé sur les évaluations NIST-SRE (décrites au paragraphe 1.4.3.5). Ces évaluations sont réalisées de façon rigoureuse afin de permettre une comparaison de multiples systèmes de différents laboratoires.

5.2.1 Généralités sur le protocole

Le protocole NIST-SRE [NIST, 2006] offre de multiples tâches de vérification qui diffèrent de par la durée des enregistrements (apprentissage et test). Parmi celles-ci, une tâche dénommée *core test* est requise lors de la participation à l'évaluation. La plupart des expériences présentées au cours de ces travaux sont réalisées sur cette tâche. Le NIST sélectionne ensuite un sous-ensemble de tests de vérification sur lesquels les systèmes vont être évalués. Ce sous-ensemble est nommé *primary test*. Pour les années 2004 à 2006, il est constitué de tous les accès où la langue est l'anglais uniquement.

Pour cette condition, les segments d'apprentissage et de test contiennent deux minutes et demie (2,5min) de parole en moyenne et proviennent d'un fichier audio originel contenant 2 locuteurs (conversations d'un durée de 5 minutes). La segmentation en locuteur n'est pas nécessaire car les deux voix sont séparées sur les deux canaux. Depuis 2004, ces données proviennent de la base de données MIXER[7] dont les enregistrements sont constitués de conversations téléphoniques de personnes participant volontairement à cette campagne d'enregistrement (un système de paiement des participant par rapport au nombre de conversations est au coeur de la stratégie de collecte).

Le protocole de vérification n'a pas évolué durant la période couverte par ces travaux. Chaque test de vérification (identité + segment de test) est pris indépendemment, *i.e.* l'utilisation d'informations d'autres tests pour prendre une décision est interdite. Les résultats de l'évaluation sont donnés par les mesures statistiques DCF (*Detection Cost Function*) et EER *Equal Error Rate*) , présentées au chapitre 1. Pour toutes les expériences, c'est la mesure *DCFmin* correspondant au minimum de la fonction de coût qui est présentée. Celle-ci correspond à un système dont le seuil de décision aurait été déterminé de façon optimal[8].Une courbe DET [Martin et Przybocki, 1997]

[7] http ://mixer.ldc.upenn.edu/

[8] Bien qu'*a posteriori*, cette mesure est généralement adoptée pour présenter les résultats d'un système de vérification. En annexe de ce document, nous présentons quelques résultats officiels des campagnes NIST-SRE en annexe correspondant à la DCF actuelle, *i.e.* le seuil a été réglé sur un ensemble de développement.

accompagne généralement les résultats pour illustrer plus précisément le comportement d'un système sur plusieurs points de fonctionnement.

5.2.2 Bases de données utilisées dans le cadre de cette thèse

Les différents corpus d'évaluation utilisés dans le cadre de ces travaux proviennent d'un sous-ensemble des bases de données NIST-SRE correspondant respectivement aux bases utilisées pour l'évaluation 2004, 2005 et 2006. Ces sous-ensembles sont définis par les tests de vérification correspondant uniquement à :

- la partie *female* pour NIST-SRE-2004, notée *NIST04*,
- la partie *male* pour NIST-SRE-2005, notée *NIST05*,
- la partie *male* pour NIST-SRE-2006, notée *NIST06*.

À chaque résultat, nous mentionnons le protocole utilisé sous la forme d'un numéro associé à la courbe DET (*e.g.* det 7). Le détail de ces protocoles peut être trouvé en annexe A. Enfin, nous mentionnons les durées d'enregistrement d'un test de vérification sous la forme *Xconv4w-Yconv4w* ou X et Y représentent respectivement le nombre de conversations utilisées pour l'apprentissage et pour le test.

5.2.3 Référencement des résultats d'expérience

Les résultats des expériences dans la suite du document font référence à un des systèmes *Sys04*, *Sys05* ou *Sys06*. Ceci permet de caractériser les bases de données utilisées pour le modèle du monde et pour les modèles pseudo-imposteurs et plus généralement la paramétrisation des signaux, les spécificités pour la détection de la parole...

Pour résumer, un résultat sera donc défini par quatre champs :

- le corpus d'évaluation : NIST*XX*,
- le protocole de vérification : det*X*,
- les durées d'apprentissage et de test : *X*conv4w-*X*conv4w,
- les traitements acoustiques et les bases de données utilisées du système GMM-UBM correspondant : Sys*XX*.

5.3 Résultats de référence du système ALIZE/LIA_SpkDet

Nous présentons dans cette partie quelques résultats sur les protocoles décrits précédemment afin d'illustrer le comportement du système GMM-UBM du LIA.

5.3.1 Systèmes de référence

Nous présentons les performances des systèmes décrits dans le paragraphe précédent sur les protocoles définis en 5.2.2. Ces trois systèmes correspondent au système de référence GMM-UBM du LIA présenté aux campagnes d'évaluation NIST-SRE.

TAB. 5.1: *Systèmes de référence présentés au cours des années 2004 à 2006. Les résultats sont donnés sur les bases de données NIST-SRE-2004 pour Sys04 et NIST-SRE-2005 pour Sys05 et Sys06.*

Système / Protocole	DCFmin (x100)	EER (%)
Sys04 / NIST04	4.51	11.08
Sys05 / NIST05	3.49	8.73
Sys06 / NIST05	3.11	7.99

Les gains observés au cours de ces trois années proviennent de plusieurs facteurs, les principaux étant :
- la sélection plus rigoureuse des trames informatives pour le gain entre *Sys04* et *Sys05* ;
- l'augmentation du nombre de composantes des vecteurs acoustiques et l'application du *feature mapping* pour le gain entre *Sys05* et *Sys06*.

5.3.2 Influence de la détection parole/non-parole

La sélection des trames présentant de l'information utile se révèle être un enjeu important pour les performances des systèmes de VAL. En utilisant la méthodologie de sélection présentée en annexe A, nous réalisons une expérience présentant la sensibilité du classifieur à la sélection des trames utiles. La figure **??** illustre ces différences en présentant les performances des systèmes où la sélection des trames est très sélective (ED0.0, environ 30% des trames sélectionnées) jusqu'à peu sélective (ED0.9, environ 60% des trames sélectionnées).

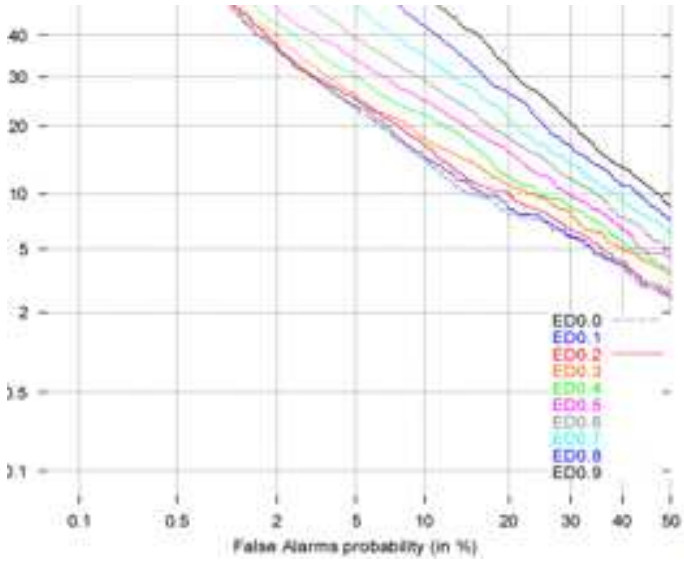


FIG. 5.1: *Sensibilité du système ALIZE/LIA_SpkDet à la sélection des trames utiles. Le cas le plus favorable se présente lorsque la détection de parole est la plus sélective (ED0.0). Le protocole utilise des modèles de 128 gaussiennes pour la tâche 1conv4w-1conv4w*

Les résultats montrent clairement une grande sensibilité à la sélection de trames. L'améliora-

tion entre le système *Sys04* et *Sys05* est principalement due à ce réglage. Le cas le plus favorable est lorsqu'aucune trame appartenant à la gaussienne de moyenne énergie n'est sélectionnée (ED0.0). Lorsque 90% des trames de la gaussienne de moyenne énergie sont sélectionnées, la perte observée est d'environ 10% (absolu) en termes d'EER.

5.3.3 Influence de la taille des vecteurs cepstraux

Nous expliquons les gains de performance du système de référence (*Sys05* à *Sys06*) par l'augmentation de la dimension des vecteurs acoustiques. Dans l'expérience présentée, le nombre de paramètres cepstraux passe de 16 à 19 coefficients. L'ajout des dérivées secondes ainsi que le paramètre de log-énergie aux vecteurs d'entrée est également analysé. La figure **??** montre l'amélioration du système *Sys05* en utilisant 32 coefficients puis 50 coefficients. Ce gain est de 7.8% à l'EER et de 3.7% à la DCFmin en relatif

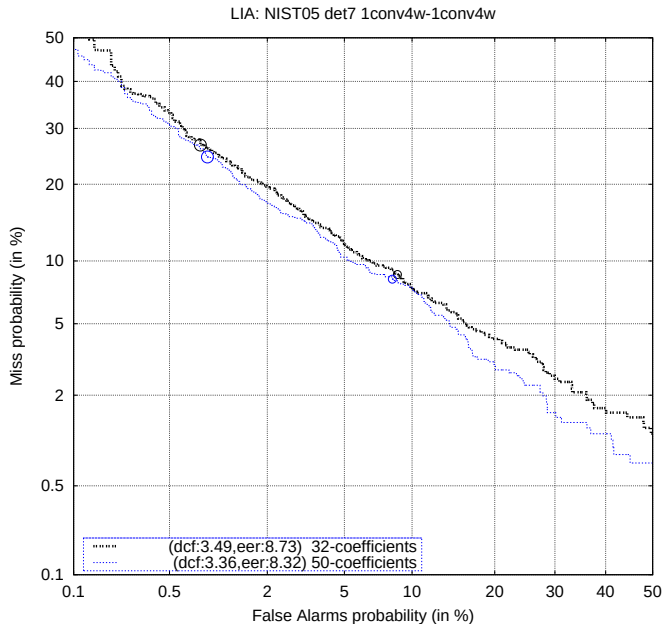


FIG. 5.2: *Effet du nombre de coefficients cepstraux sur une expérience de VAL sur NIST05, det7, système* Sys05. *Le vecteur de paramètres cepstraux augmente de 16 à 19 coefficients. L'ajout des dérivées secondes au vecteur ainsi que celui du log-énergie au vecteur son également considérés.*

5.3.4 Influence du genre du locuteur

La détection du genre du locuteur n'est pas une étape requise par les participants à la campagne NIST-SRE. Les signaux sont en effet étiquetés par le genre du locuteur (*male* ou *female*) et la correspondance du genre du locuteur entre l'apprentissage et le test est respectée.

Les systèmes du LIA utilisent un modèle du monde par genre. L'utilisation d'un modèle du monde indépendant du genre ne nous a pas permis d'atteindre des performances équivalentes.

L'expérience présentée à la figure 5.1 illustre les différences de comportement du système selon le genre du locuteur.

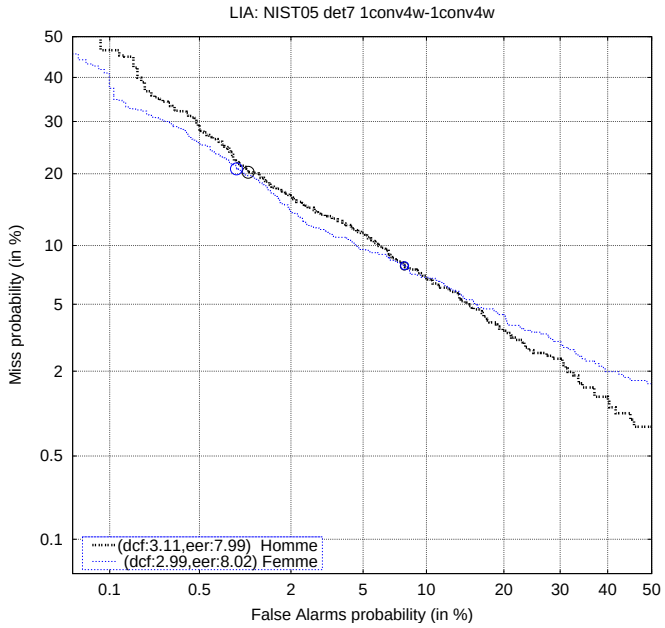


FIG. 5.3: *Effet du genre du locuteur sur une expérience de VAL sur NIST05, système* Sys06*.*

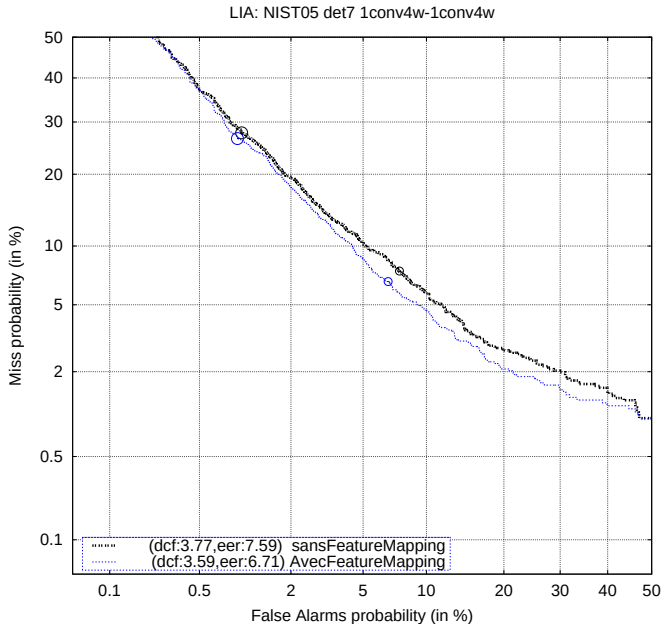


FIG. 5.4: *Amélioration due au feature mapping en utilisant des modèles dépendant du canal dont les paramètres de moyenne et de variance ont été adaptés.*

5.3.5 Amélioration apportée par le *feature Mapping* pour la normalisation de canal

La figure 5.2 illustre l'amélioration apportée par la technique du *feature mapping* [Reynolds, 2003] pour la normalisation de canal sur le système de référence *Sys05* du LIA.

L'effet du *feature mapping* est global sur la courbe DET présentée. En effet, la min est réduite de 3.77% à 3.59% et l'EER de 7.59% à 6.71%.

5.3.6 Influence de la quantité de données imposteurs

L'acquisition du corpus Fisher a permis d'augmenter considérablement la durée d'apprentissage du modèle, de 1.3 millions de trames à environ 10 millions, l'objectif étant de permettre au modèle du monde d'estimer de façon plus robuste des statistiques indépendantes du locuteur. Cependant, aucune amélioration significative due à cet ajout de données n'a été remarquée (voir figure **??**).

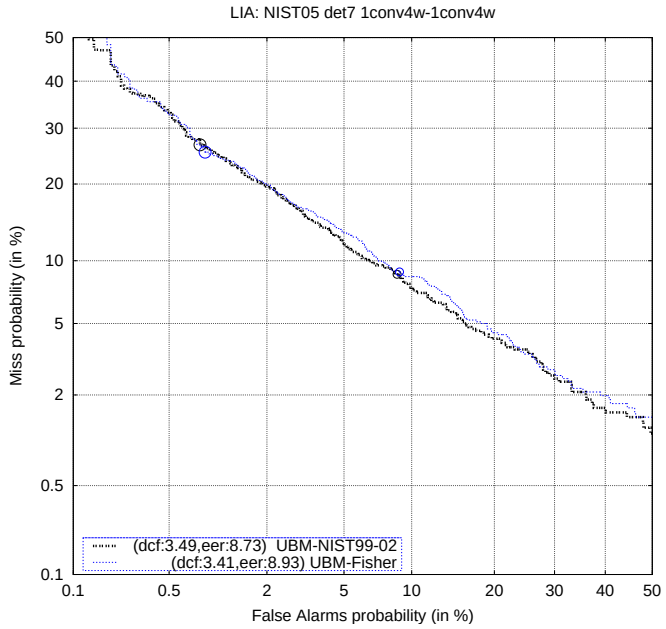


FIG. 5.5: *Influence de la quantité de données de pseudo-imposteurs pour l'apprentissage du modèle du monde. Performance du système* Sys05 *utilisant 1.3 millions de trames (base NIST) ou environ 10 millions de trames (base Fisher de* Sys06*). Corpus d'évaluation NIST05*

Nous verrons cependant par la suite (chapitre 8) que ce nouveau modèle UBM est différent, les données étant réparties plus équitablement sur la mixture en utilisant une quantité de données plus importante.

5.3.7 Influence de la variance

Dans le chapitre 2, nous avons évoqué le *variance flooring* comme une technique permettant d'éviter le sur-apprentissage pour les méthodes à base de maximum de vraisemblance. Cette technique influence beaucoup la modélisation du modèle puisque environ 50% des paramètres de variances sont seuillés (dans le cas du *Sys06*, environ 42000 variances seuillées pour un total de 102400 à estimer).

Si les paramètres de variances sont peu importants, alors les matrices de covariances du modèle peuvent être modélisées par $\mathbf{\Sigma}_k = \sigma_k^2 \cdot \mathbf{I}$, où k est la composante d'indice k et $\mathbf{I}$ la matrice identité. Nous réalisons deux expériences en modélisant les matrices de covariance de cette façon afin d'illustrer l'importance de la modélisation de la variance.

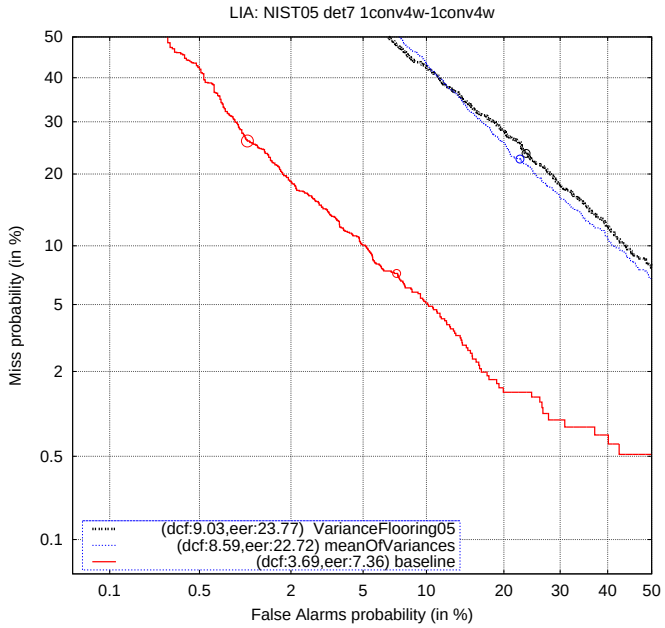


FIG. 5.6: *Influence de la modélisation de la variance sur une expérience de VAL, configuration* Sys06*. Environ 50% des paramètres de variance sont seuillés lors de l'apprentissage classique. Les résultats montrent que les paramètres non-seuillés jouent un rôle crucial dans la modélisation générative. Ainsi, le variance flooring est utilisé comme facteur de régularisation afin de prévenir le sur-apprentissage.*

Dans la première expérience, toutes les variances sont seuillées à 0.5 (cette valeur correspond au paramètre de seuillage du système LIA_SpkDet, sur *Sys04*,*Sys05* et *Sys06*). Les matrices de covariance sont de la forme : $0.5 \cdot I$;

Pour la seconde expérience, la matrice de covariance est estimée à partir de la moyenne des variances de la matrice de covariance dans l'estimation ML. Les matrices de covariance sont de la forme $\overline{\sigma_k^2} \cdot I$ où $\overline{\sigma_k^2}$ est la moyenne des variances de chaque dimension.

Les résultats sont donnés à la figure 5.3 d'où il ressort de ces expériences que la variance joue un rôle crucial dans la modélisation : le *variance flooring* est utilisé comme facteur de régularisa-

tion du modèle, afin d'éviter le sur-apprentissage en augmentant la capacité de généralisation de la mixture.

5.4 Conclusion

Tout au long de ce chapitre, nous avons présenté le protocole expérimental ainsi que les différents systèmes qui structurent nos expériences et résultats dans la suite du document. Trois corpus d'évaluation représentant des sous-ensembles des bases de données NIST-SRE sont utilisés : *NIST04*, *NIST05* et *NIST06*. Ils correspondent à la partie *male* des corpus entiers. Trois systèmes *Sys04*, *Sys05* et *Sys06* sont implémentés, faisant chacun référence à une variante de LIA_SpkDet utilisée par le LIA pour l'évaluation correspondante. Ces systèmes de référence sont utilisés par la suite à des fins de comparaison et de fusion des scores. Ils seront également indiqués dans les expériences pour les autres systèmes afin de référer aux techniques de paramétrisation, détection de la parole,... Nous avons présenté quelques expériences comparatives permettant d'illustrer la sensibilité des systèmes à certaines techniques et variations (sélection des trames utiles, genre du locuteur,...). Ces résultats sont le reflet du travail effectué tout au long de ces années pour proposer un système de référence GMM-UBM correspondant à l'état de l'art.

CHAPITRE 6

Analyse dynamique du signal par une segmentation basée sur l'UBM

Sommaire

6.1 Introduction

Dans les chapitres 2 et 3, nous avons abordé l'importance des données pseudo-imposteurs et comment elles jouaient un rôle structurel dans les approches générative à travers la construction d'un modèle générique. Au chapitre 4, nous avons ensuite présenté les systèmes dits de « haut niveau » permettant de générer des représentations du signal de parole en utilisant des unités reliées à la structure de la langue (unités phonétiques, lexicales, syllabiques, *etc*). Ces systèmes permettent implicitement de faire émerger une structure appropriée pour la VAL en relâchant les contraintes linguistiques dans la génération des représentations du signal (reconnaisseur phonétique en boucle-ouverte, approche multilingue...).

Ce chapitre présente un nouveau système de vérification du locuteur utilisant les techniques de modélisation des systèmes « haut-niveau » et le modèle UBM pour structurer le signal de parole. Nous avons souligné le fait que ces systèmes « haut niveau » s'intéressent plus à générer des représentations du signal qu'à décoder le message proprement dit. Nous montrons comment l'UBM peut permettre de faire émerger une structure sous-jacente pertinente pour la VAL.

Au lieu d'utiliser des décodeurs reliés fortement à la nature du matériel « parole », nous présentons une technique utilisant un décodeur basé sur un modèle générique indépendant du locuteur et de la langue : l'UBM. Ces travaux rejoignent ceux de [Torres-Carrasquillo et al., 2002] pour l'IAL, dans lesquels le décodeur est appelé *GMM tokeniser*. Cette stratégie est motivée par le constat de rupture entre la modélisation « bas-niveau » (basée sur la modélisation cepstrale) et « haut-niveau ». Les systèmes haut-niveau n'essaient pas de tirer profit de la capacité de modélisation des systèmes à base de GMM et particulièrement, du rôle structurel de l'UBM pour l'espace acoustique.

Notre méthode consiste à utiliser le modèle générique UBM pour la conception d'un système « haut-niveau » en présentant une méthodologie de construction d'événements acoustiques dont la dynamique sera employée pour caractériser les locuteurs. Le système que nous introduisons est appelé AES pour *Acoustic Event Sequences* [Scheffer et Bonastre, 2005]. Dans ce système, nous distinguons une étape indépendante du locuteur et de la langue incluant : la construction du dictionnaire et la génération d'événements acoustiques (présentée en section 6.2.1), et une étape dépendante du locuteur : la modélisation et le test de vérification (présentée en section 6.2.2). De nombreux résultats expérimentaux, en section 6.3, viennent illustrer notre propos et permettent une analyse du comportement de ce système.

6.2 Le système AES : « Acoustic Event Sequences »

Cette section détaille le principe du système AES. Il repose sur deux parties distinctes : une première partie est indépendante du locuteur et consiste à construire les événements acoustiques, une deuxième partie est dépendante du locuteur et est dédiée à la tâche de vérification et de modélisation proprement dite.

A la manière des systèmes aperçus au chapitre 4, le système AES a pour objectif de générer une représentation du signal appropriée pour la VAL. Parmi les systèmes haut-niveau, l'AES se place dans la catégorie « inter-segmental ». Cette catégorie regroupe les systèmes s'intéressant à la dynamique des unités segmentales pour caractériser le locuteur.

6.2.1 Processus indépendant de la langue et du locuteur

Nous présentons dans ce paragraphe, la construction du dictionnaire d'événements acoustiques permettant de générer la segmentation du signal à partir de l'UBM.

6.2.1.1 Dictionnaire acoustique indépendant du locuteur

Les symboles du dictionnaire AES sont générés en utilisant l'UBM du système GMM-UBM présenté en 5.3. La quantité de données d'apprentissage est volontairement élevée afin que l'information contenue dans le GMM soit maximale. Contrairement aux systèmes par segmentation GMM en identification de la langue qui utilisent des mixtures avec peu de gaussiennes, la stratégie adoptée ici consiste à utiliser une dimensionnalité obtenant des performances maximales dans les systèmes état-de-l'art.

L'objectif poursuivi dans ce paragraphe est de construire un dictionnaire de taille réduite dont les éléments correspondent à des regroupements de gaussiennes issues du modèle générique. Ces classes issues du regroupement seront appelées « événements acoustiques ». Nous décrivons dans la suite la stratégie adoptée qui se décompose en trois étapes :

- la définition du dictionnaire de taille maximale ;
- la construction d'une matrice renseignant les critères de confusion entre les gaussiennes ;
- un processus de réduction de dimensionnalité de cette matrice afin de définir les événements acoustiques.

Le dictionnaire final sera composé d'éléments chacun constitué d'un ou plusieurs indices de gaussiennes.

Définition du dictionnaire de taille maximale La première étape consiste à considérer le dictionnaire de taille maximale comme celui dont les éléments correspondent aux indices des gaussiennes du modèle générique. De cette première étape résulte un dictionnaire dont la taille est égale au nombre de composantes dans l'UBM (2048 dans ces travaux, cette taille étant généralement celle qui donne les meilleurs taux de performance dans un système état de l'art). Il est assez évident que cette taille est trop élevée pour effectuer une analyse en séquence car sans *a priori*, le nombre de séquences possibles peut atteindre 2048^n où n est la taille de la fenêtre d'analyse.

Cette limitation tient au principe de la couverture en N-grammes qui diminue lorsque la taille de la fenêtre d'analyse augmente. Prenons deux jeux de données différents X et Y et leurs ensembles respectifs E_X et E_Y contenant tous les N-grammes présents dans ces données, alors la couverture (ou rappel) de l'ensemble E_X par rapport à l'ensemble E_Y, $R(E_X|E_Y)$ est une mesure non-symétrique reliée à l'intersection des deux ensembles soit :

$$R(E_X|E_Y) = \frac{\#(E_X \cap E_Y)}{\#E_Y}, \tag{6.1}$$

où # renseigne sur le cardinal d'un ensemble. Cette quantité permet d'évaluer la capacité d'un système à modéliser un problème par une analyse N-gramme. Dans les expériences, la couverture en 2-grammes est d'environ 22% pour un dictionnaire de taille 128 et 5% pour un dictionnaire de taille 2048. Pour pallier ce problème de couverture, il est donc nécessaire de réduire le nombre de symboles.

Construction de la matrice de confusion inter-gaussiennes L'étape suivante consiste à construire une matrice de confusion permettant d'envisager par la suite un regroupement entre les gaussiennes. A cet effet, des indicateurs de confusion inter-gaussiennes sont générés dans une matrice

carrée M de taille N^2 (N étant le nombre de composantes du modèle UBM). Les indices des lignes (ou colonnes) de la matrice correspondent aux indices de gaussiennes dans le modèle UBM. Cette matrice est obtenue en utilisant toutes les données d'apprentissage du modèle du monde.

La méthode de génération de cette matrice prend en compte à chaque trame, les K gaussiennes à la plus forte vraisemblance. Notons Ω_K l'ensemble de ces gaussiennes et $\overline{\Omega_K}$ les $N-K$ restantes. Posons ensuite G_i comme la composante d'indice i du modèle génératif de moyenne $\boldsymbol{\mu}_i$, de variance Σ_b et de poids γ_i. Alors pour une trame x_t, les K gaussiennes de plus forte vraisemblance sont données par :

$$p(G_i|x_t) > p(G_j|x_t)\ ,\ \forall i \in \Omega_K\ ,\ \forall j \in \overline{\Omega_K}, \tag{6.2}$$

avec $p(G_i|x_t)$ la probabilité d'appartenance de la trame à la gaussienne G_i définie comme :

$$p(G_i|x_t) = \frac{\gamma_i p(x_t|G_i)}{p(x_t)} \simeq \gamma_i\ p(x_t|\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_b)). \tag{6.3}$$

Si $idx_1, \ldots, idx_K$ correspondent aux indices des *top-K* gaussiennes, alors pour chaque trame des données d'apprentissage du modèle du monde, l'étape de mise à jour est donnée par :

$$M(idx_1, idx_k) += 1, \forall k \in [1, K] \tag{6.4}$$

La matrice résultante contient sur chaque ligne i :

- l'élément sur la diagonale correspond au nombre de fois où la gaussienne i était en première position ;
- les autres éléments indiquent la présence des autres gaussiennes parmi les *top-K* lorsque G_i apparaît en premier.

De nombreuses expériences consistant à changer l'étape de mise à jour en fonction de l'appartenance ont été réalisées, mais les résultats ont montré que cette fonction de mise à jour est la plus efficace.

Réduction de la taille du dictionnaire L'étape suivante consiste à regrouper les gaussiennes pour former les événements acoustiques proprement dits afin d'augmenter la couverture entre les données d'apprentissage et de test. Nous procédons à une réduction de cette matrice en définissant un critère de regroupement des gaussiennes. Ce critère est celui du minimum de confusion. A cet effet, le taux de confusion entre les classes C_i et C_j, noté $\tau_{i,j}$, est calculé en utilisant la matrice M de la façon suivante :

$$\tau_{i,j} = M(i,j) + M(j,i) \tag{6.5}$$

En utilisant ces indicateurs, la méthode de regroupement en classes s'effectue séquentiellement en cherchant le couple (C_i, C_j) où τ est maximum, puis :

- dans le cas d'un taux de confusion identique, la priorité est donnée à la classe de cardinal le plus faible évitant ainsi un comportement trop agrégatif de l'algorithme ;
- les classes (C_i, C_j) sont alors regroupées dans une nouvelle classe $C'_i = (C_i, C_j)$.

L'algorithme peut être qualifié de regroupement *bottom-up* puisque le regroupement s'effectue de manière agrégative. Le nombre optimal de classes n'est pas défini *a priori*, il doit être déterminé empiriquement sur un corpus de développement.

Le résultat de cette réduction est un dictionnaire qui associe chaque gaussienne de l'UBM à une classe de plus haut-niveau correspondant à un regroupement de gaussiennes de l'UBM originel. Le schéma **??** illustre la procédure de réduction d'un dictionnaire de trois éléments vers un dictionnaire de deux éléments.

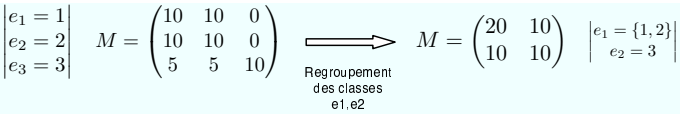


Fig. 6.1: *Exemple de réduction d'un dictionnaire de trois éléments vers un dictionnaire de deux éléments pour l'approche AES. Les taux de confusions entre les deux premières classes sont plus les plus élevées, le regroupement s'effectuent donc en créeant une nouvelle classe contenant les éléments e_1 et e_2.*

Classe Out of Vocabulary Afin de lisser la distribution des symboles et de ne pas prendre en compte les classes peu informatives, une classe « poubelle » est crée, généralement appelée *OOV* pour *Out of Vocabulary*. Dans cette classe se trouvent les événements acoustiques qui sont apparus moins de n fois en première position, n étant déterminé empiriquement.

La procédure de construction du dictionnaire d'événements acoustiques est résumée sur le schéma **??**.

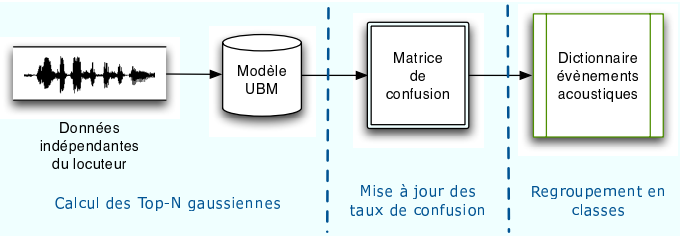


Fig. 6.2: *Le système Acoustic Event Sequence : Partie indépendante du locuteur et de la langue. Un processus de regroupement des gaussiennes permet de créer un dictionnaire d'événements acoustiques de taille réduite.*

6.2.1.2 Génération de la représentation du signal par un flux d'événements acoustiques

Ayant construit le dictionnaire de taille réduite, il reste à générer la représentation du signal de parole associée à ce dictionnaire. La stratégie adoptée consiste à transformer tous les signaux paramétrisés en des séquences de symboles. Chaque fichier appartenant au corpus d'apprentissage, de test, et imposteurs est soumis au même processus de génération.

Étant donné chaque trame active d'un signal, l'indice de la gaussienne la plus probable est produit. Cet indice est obtenu par un calcul d'appartenance de cette trame à chacune des gaussiennes. Pour une trame x_t le symbole généré s correspond à l'indice de la gaussienne à probabilité maximale pour cette trame, soit :

$$s = \arg\max_i p(G_i|x_t), \tag{6.6}$$

où $p(G_i|x_t)$ a été défini à l'équation 6.3. L'indice de cette composante est ensuite remplacé par le symbole correspondant dans le dictionnaire construit précédemment.

Le flux de symboles résultant contient autant de symboles qu'il y a de trames dans le fichier. La procédure de génération de la représentation du signal par un flux d'événements acoustiques est résumée sur le schéma 6.1.

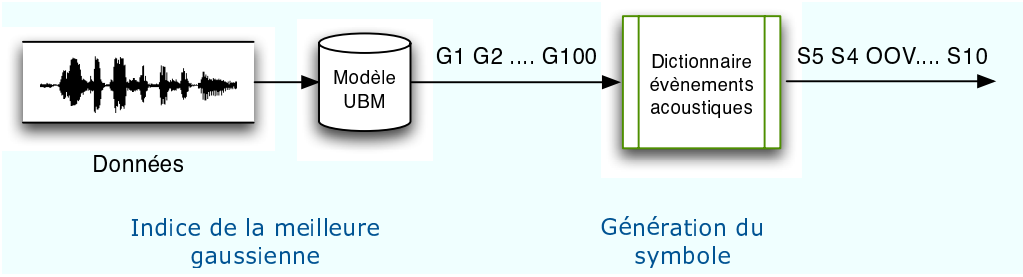


FIG. 6.3: *L'approche Acoustic Event Sequence : Partie indépendante du locuteur et de la langue. Le modèle générique est utilisé pour décoder le signal de parole. La représentation générée est un flux de symboles tirés issus de ce dictionnaire.*

6.2.2 Processus dépendant du locuteur

Ayant produit le dictionnaire indépendant du locuteur et de la langue, nous nous intéressons maintenant à la méthode de construction de modèles spécifiques aux locuteurs basée sur ce dictionnaire. L'objectif est de modéliser les séquences dont les statistiques sont spécifiques au locuteur. En effet, comme pour les phonèmes [Andrews et al., 2001] et pour l'idiolecte [Doddington, 2001] où l'analyse de séquence s'est montrée efficace pour la discrimination entre locuteurs, l'idée principale de cette méthode est de généraliser l'approche à des informations non linguistiques, comme les événements acoustiques définis par les symboles du dictionnaire.

Dans la suite, la méthode de modélisation adoptée pour construire des modèles spécifiques au locuteur est présentée. La première partie présente la technique adoptée à base de N-grammes, la seconde précise les différentes méthodologies de calcul du score de vérification. Cette partie est largement inspirée de la partie 4.5. Nous en rappelons les principaux résultats et nous en reprenons les schémas, dans un souci de clarté.

6.2.2.1 Modélisation du locuteur

A partir de la segmentation produite pour le signal d'apprentissage d'un locuteur, la modélisation des données du locuteur est effectuée par un modèle N-gramme. Les spécificités du locuteur par la construction d'un tel modèle prennent en compte la nature des événements (représentés par les N-grammes) et leur apparition (représentée par les probabilités associées).

Le type de modèle utilisé est appelé N-gramme d'unités (*token-Ngram*), ou approche par « sac de N-gramme ». Ce modèle est utilisé dans les systèmes d'analyse N-gramme comme dans les travaux de [Doddington, 2001]. Le principe est de transformer une séquence de symboles en unités k et leur probabilité associée $p(k)$.

Soit la représentation en flux d'événements acoustiques $\mathbf{X}_S = e_1, \ldots, e_T$ des données d'apprentissage du locuteur S. Pour une analyse en 2-gramme, chaque séquence $\{[e_1\_e_2], [e_2\_e_3], \ldots, [e_{T-1}\_e_T]\}$ est considérée comme une unité dont la probabilité est calculée. Considérons le sac de N-grammes B, contenant tous les 2-gramme k uniques, alors le modèle contient la probabilité de ces séquences donnée par la fréquence d'apparition dans les données $\mathbf{X}_S$ pour le locuteur S.

$$p(k|\mathbf{X}_S) = \frac{C(k|\mathbf{X}_S)}{\sum_k C(k|\mathbf{X}_S)}, \tag{6.7}$$

où $C(k|\mathbf{X})$ est l'opérateur de compte de l'unité k sur les données $\mathbf{X}$. Le schéma 6.2 illustre l'approche en utilisant un dictionnaire de deux unités et une analyse en 2-gramme.

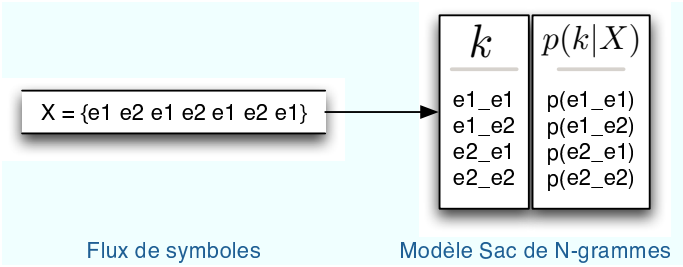


Fig. 6.4: *Exemple d'une modélisation par sac de* N*-grammes pour l'approche AES. Le dictionnaire d'événements acoustiques contient 2 symboles. Une analyse en* N*-grammes (appelé token-Ngram) calcule la probabilité de chacune des séquences par leur probabilité d'apparition dans le flux de symboles.*

Pour chaque fichier d'apprentissage correspondant à un locuteur, un modèle de ce type est construit. Il est possible de regrouper des séquences de longueurs différentes dans un même modèle afin d'augmenter la capacité de modélisation (voir paragraphe 6.3.2). Un modèle N-gramme du monde est généralement nécessaire pour pouvoir calculer les scores correspondant aux tests de vérification.

Construction d'un codebook Afin de limiter le nombre de séquences possibles, les unités composants le sac de N-gramme sont déterminées sur les données d'apprentissage du modèle du monde. La génération de ce « codebook » vise à éliminer les séquences qui ont été rencontrées rarement dans les données d'apprentissage du modèle du monde. Cette méthode permet de garantir une robustesse dans les statistiques estimées ainsi qu'une réduction de la complexité du problème.

La mesure de la couverture dans les expériences réalisées est mesurée par rapport à ce codebook. Soit toutes les représentations des signaux en événements acoustiques $\mathbf{X} = \{\mathbf{X}_i\}_{i=[1\ldots N]}$ d'une expérience de vérification, le codebook B et leurs ensembles de N-gramme respectifs $E_{\mathbf{X}_i}$ et E_B, alors la couverture R_{AES} (eq.6.1) en N-grammes pour le système AES est calculée comme la moyenne de la couverture sur tous les fichiers soit :

$$R_{AES} = \frac{1}{N} \sum_{i=1}^{N} R(E_{\mathbf{X}_i}|E_B). \tag{6.8}$$

6.2.2.2 Calcul du score de décision

Après avoir construit des modèles N-grammes, représentant les statistiques sur les événements acoustiques, nous rappelons les deux techniques principales pour calculer un score de vérification, qui ont déjà été vues au chapitre 4.

Détecteur Rapport de Log-Vraisemblance LLR La première méthode consiste à estimer le score de vérification par le logarithme du rapport de vraisemblance (LLR) entre un segment de test $\mathbf{X}$, les données d'un locuteur $\mathbf{X}_S$ et celles du modèle du monde $\mathbf{X}_W$. Le score $\mathcal{Y}_S(\mathbf{X})$ de ce test est exprimé ci-dessous :

$$\mathcal{Y}_S(\mathbf{X}) = \text{LLR}(\mathbf{X}|\mathbf{X}_S, \mathbf{X}_W) = \frac{1}{\sum_k C(k|\mathbf{X})} \cdot \sum_k C(k|\mathbf{X}) \log\Big(\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}\Big), \tag{6.9}$$

où $C(k|\mathbf{X})$, $p(k|\mathbf{X}_S)$ et $p(k|\mathbf{X}_W)$ sont respectivement le compte de l'unité k sur le segment de test $\mathbf{X}$, et les probabilités de l'unité dans les modèles N-grammes S et W.

Cette modélisation est généralement augmentée par une estimation par *maximum a posteriori* décrite au chapitre 4. Le schéma **??** résume la méthode de calcul du score par le logarithme du rapport de vraisemblance.

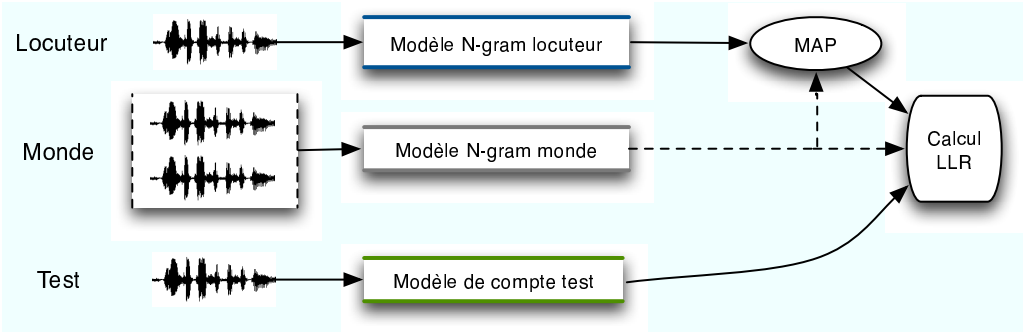


FIG. 6.5: *Calcul du score par le logarithme du rapport de vraisemblance pour le système AES. Le modèle* N*-gramme du monde est adapté à partir des données du locuteur. Le compte des unités sur le test est nécessaire pour calculer le LLR.*

Méthode basée sur les SVM La deuxième méthode est basée sur la construction d'un noyau en vue d'une utilisation dans un classifieur SVM. Le noyau TFLLR (*Term Frequency Log-Likelihood Ratio*), introduit par [Campbell et al., 2004b], est une alternative à la méthode *TF-IDF*[1] afin de projeter des N-grammes dans un espace approprié pour l'utilisation d'un SVM.

Considérons l'unité k appartenant au sac de N-grammes B. La vraisemblance d'une séquence de données $\mathbf{X}$ pour le paramètre k est notée $p(k|\mathbf{X})$, la formulation générale du noyau TFLLR est la suivante :

$$\sum_k \frac{p(k|\mathbf{X}_S)}{\sqrt{p(k|\mathbf{X}_W)}} \frac{p(k|\mathbf{X}_{S'})}{\sqrt{p(k|\mathbf{X}_W)}} - 1, \tag{6.10}$$

[1] *Term Frequency - Inverse Document Frequency*

En pratique, pour chaque accès, il est nécessaire de calculer la statistique $\frac{p(k|\mathbf{X}_S)}{p(k|\mathbf{X}_W)}$. Par conséquent, cette méthode demande la construction d'un modèle N-gramme pour chaque accès de test et pour le modèle du monde. La construction du noyau réside dans la pondération des vraisemblances des locuteurs par la vraisemblance de l'unité sur les données d'apprentissage des modèles du monde.

Le score de vérification correspond finalement à la distance entre le vecteur de test et l'hyperplan appris par le SVM utilisant le noyau TFLLR. Le schéma **??** résume la stratégie de modélisation et de calcul du score pour cette méthode basée sur les SVM.

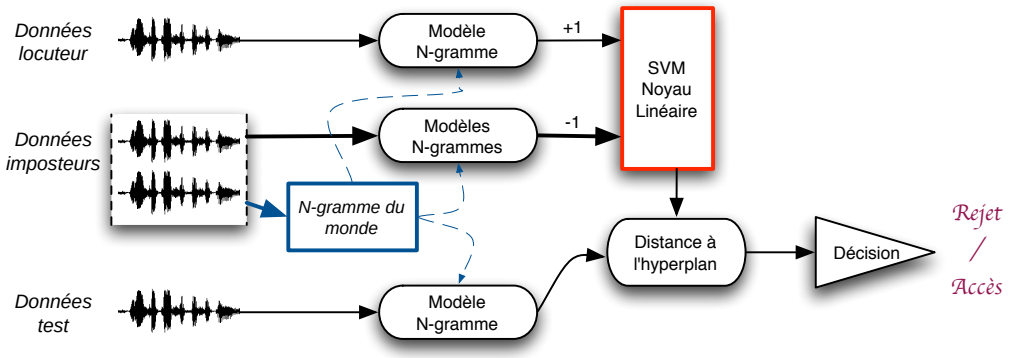


FIG. 6.6: *Mise en oeuvre de la modélisation des locuteurs par la méthode utilisant le noyau TFLLR. Pour chaque accès, un modèle* N*-gramme est estimé. L'expansion TFLLR consiste à pondérer les probabilités par celles estimées sur les données du monde. Le score de vérification correspond à la distance à l'hyperplan entre le vecteur de test et le modèle SVM issu des données d'apprentissage (locuteur considéré et peusdo-imposteurs).*

6.3 Résultats et Expériences

Les expériences décrites par la suite ont pour objectif de présenter le comportement général du système AES en utilisant les résultats sur les protocoles définis au chapitre 5.

Dans un premier temps, le système AES utilisant une longueur fixe de séquence est étudié, *i.e.* le N des modèles N-grammes est fixé. La comparaison entre les deux méthodes de vérification vues au paragraphe précédent est effectuée. Des expériences sur la taille du dictionnaire d'événements acoustiques permettent de valider la stratégie adoptée.

Nous présentons ensuite un système AES intégrant plusieurs longueurs de séquences et examinons la façon de les intégrer dans le modèle. Enfin, la combinaison avec un système GMM/UBM état de l'art est présentée afin d'étudier la complémentarité de l'information apportée par le système AES par rapport au système GMM-UBM.

6.3.1 Système AES utilisant une longueur fixe de séquence

Les expériences présentées ci-dessous sont réalisées avec un système AES où la longueur des séquences d'unités k est fixe.

6.3.1.1 Méthode du LLR standard

Nous présentons les résultats d'un système AES pour la méthode du LLR standard utilisant un dictionnaire de 256 événements acoustiques. La figure 6.3 présente les courbes DET, sur *NIST04*, lorsque la modélisation est effectuée en utilisant des modèles 2 et 3-grammes sans adaptation MAP. Les résultats sont assez faibles puisque les systèmes présentent 36% et 37% d'EER pour des longueurs de 2 et 3-gramme respectivement.

Nous considérons ensuite l'effet de l'adaptation MAP sur la performance du système d'analyse séquentielle d'événements acoustiques. Les courbes DET (figure 6.3) illustrent l'efficacité de la méthode MAP pour les systèmes 2 et 3-gramme (avec $\alpha = 0.01$ de l'équation 4.5). Ces résultats sont encourageant puisque une réduction de l'EER de 36 à 26% et de 37 à 32% pour les systèmes AES 2 et 3-gramme respectivement est observée.

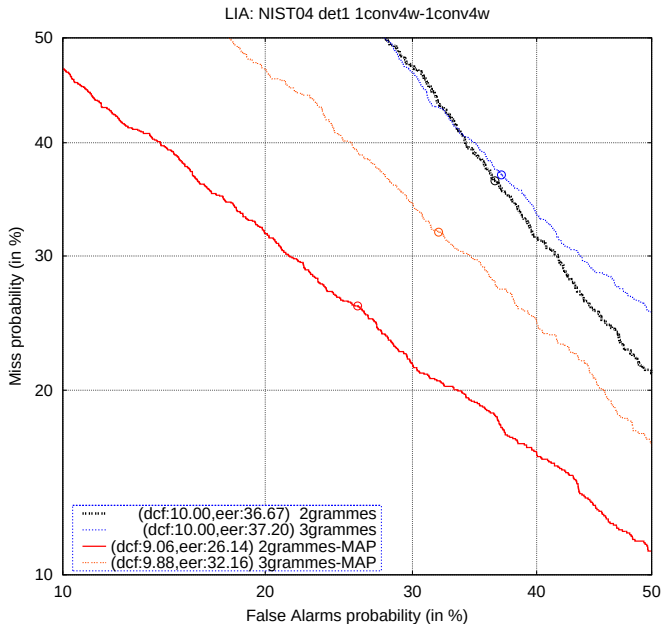


FIG. 6.7: *Performance d'un système AES utilisant pour la méthode LLR avec et sans adaptation MAP. Dictionnaire de 256 symboles pour des séquences de 2,3-grammes. Configuration* Sys04.

6.3.1.2 Méthode basée sur les SVM : Noyau TFLLR

Nous présentons ensuite l'intégration de la méthode TFLLR pour le système AES. Cette méthode utilise un classifieur SVM et nécessite par conséquent la génération d'exemples positifs et négatifs.

Afin de représenter les données de la classe négative dans le SVM, des pseudo-imposteurs utilisés pour apprendre le modèle du monde sont employés. L'entrée du classifieur est constitué de l'accès d'apprentissage du locuteur et de tous les imposteurs. La décision à marge maximale est trouvée en passant cette entrée à un noyau linéaire[2]

[2]Pour le corpus NIST04, l'outil *SVMTorch* de [Collobert et Bengio, 2001] a été utilisé pour calculer les SVM et pour

Afin de compenser la grande disproportion entre les tests client et imposteurs, un modèle de coût a été adopté lors de l'apprentissage. Ainsi, une erreur de classification sur les exemples positifs sera 200 fois plus coûteuse que sur les exemples négatifs (une valeur trouvée empiriquement).

La figure **??** présente les courbes DET comparant les deux techniques de vérification (LLR standard et TFLLR) pour un système AES utilisant un dictionnaire de taille 64. Sur cette figure, nous pouvons observer deux choses :

– En premier, le système AES avec un dictionnaire de 64 symboles obtient de meilleurs performances que celui à 256 symboles pour des séquences de 3-gramme. En effet, ce système obtient un EER proche pour les 2-gramme (28% au lieu de 26%) mais meilleur pour les 3-gramme (27% au lieu de 32%) ;
– Un gain significatif peut être observé lorsque la méthode du TFLLR est utilisée. Le gain (absolu) est de 5% et 3% en termes d'EER pour les modèles 2 et 3-gramme respectivement.

Il ressort de ces expériences qu'une taille de dictionnaire réduite permet d'être plus stable pour de plus grandes longueurs de séquences. Il semble aussi que la méthode basée sur les SVM permet d'obtenir un gain significatif. Dans toute la suite des expériences, nous utiliserons cette méthode pour le système de référence AES.

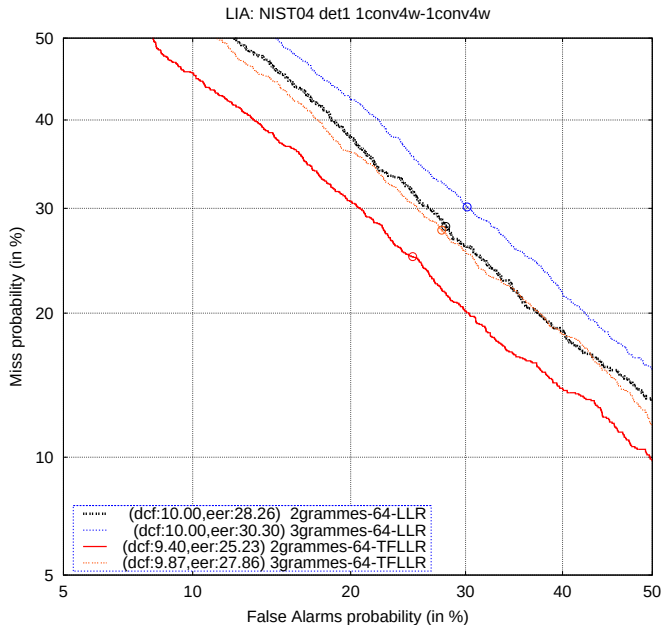


FIG. 6.8: *Approche AES utilisant un dictionnaire de 64 symboles et des longueurs de 2 et 3-gramme. Cette taille de dictionnaire permet d'obtenir de bonnes performances en 3-gramme. L'utilisation de la technique TFLLR permet d'obtenir un gain significatif comparé au LLR standard. Configuration* Sys04

classifier les instances. Pour les autres protocoles, c'est l'outil SVM-Light de Thorsten Joachims [Joachims, 1999] qui a été utilisé.

6.3.1.3 Construction des événements acoustiques

Nous présentons ici une justification à la construction des événements acoustiques en comparant un système AES avec l'utilisation simple des indices de gaussiennes de l'UBM originel. La figure **??** présente les résultats en terme d'EER du système AES en utilisant un dictionnaire de 2048 pour des longueurs de séquences allant de 1 à 3, et de 128 pour des longueurs de séquences allant de 1 à 5.

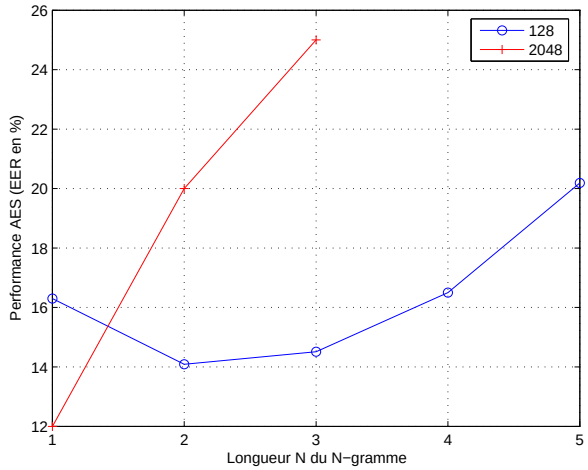


FIG. 6.9: *Différents systèmes AES avec des longueurs de* N*-gramme allant de 1 à 5 et un dictionnaire de taille 128, et de 1 à 3 pour une taille de 2048 (pas de normalisation des scores, configuration* Sys05*). Un dictionnaire de 128 permet de capturer de l'information dynamique puisque les performances sont meilleures en 2,3-grammes qu'en 1-gramme. Un dictionnaire de 2048 voit ses performances décroître rapidement.*

Lorsqu'un dictionnaire de grande taille est utilisé, 2048 ici, les performances se dégradent avec l'augmentation de la longueur de séquence. En revanche, l'utilisation d'un dictionnaire d'événements acoustiques de taille réduite, 128 ici, permet d'obtenir des performances plus stables en augmentant la longueur de séquence. De plus, sur la figure **??**, il est à noter que les systèmes 2 et 3-gramme, avec un dictionnaire de 128, obtiennent de meilleures performances que l'uni-gramme seul.

6.3.1.4 Déstructuration de la séquence temporelle du signal de parole

La performance observée du système AES peut être attribuée à d'autres facteurs que la modélisation des séquences d'événements acoustiques ou à l'augmentation des paramètres du problème. Nous présentons deux expériences visant à tirer aléatoirement les trames du signal d'entrée afin de détruire la séquence temporelle du signal de parole. L'objectif est de prouver que le gain observé est du à la modélisation de la dynamique du signal (les expériences ont été réalisées en utilisant des modèles 2-grammes, un dictionnaire de 64, avec la méthode du TFLLR). Les résultats de ces expériences, présentés sur la figure 6.4, tendent à prouver que :

– la performance du système n'est pas due à l'augmentation du nombre de paramètres pour modéliser les locuteurs. Ceci est mis en évidence par l'application d'un processus de tirage aléatoire des trames de test et d'apprentissage (courbe *Random-Train-Test*) ;
– l'information discriminante entre les locuteurs provient de la modélisation des séquences d'événements acoustiques. Ceci est mis en évidence par l'application d'un processus de tirage aléatoire des trames de test uniquement (courbe *Random-Test*).

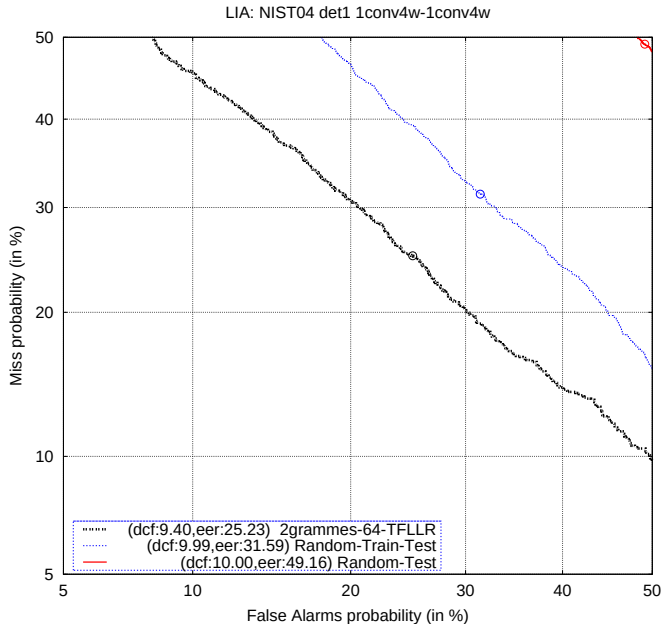


FIG. 6.10: *Expérience par destruction de l'ordre aléatoire des trames sur les fichiers d'apprentissage et de test.*

6.3.1.5 Prise en compte de la durée des événements acoustiques dans l'AES

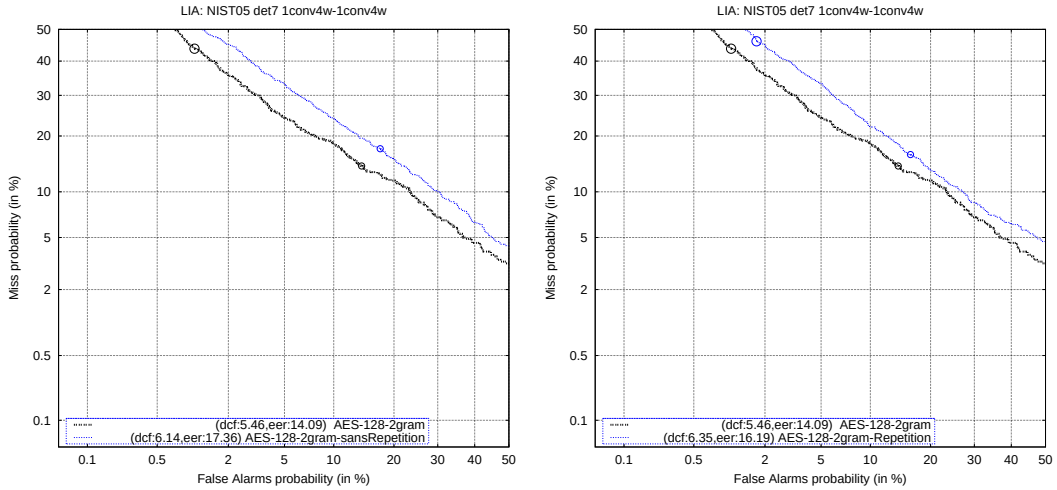
L'objectif de l'AES est d'analyser les séquences d'événements acoustiques spécifiques au locuteur, *i.e.* permettant de les discriminer les uns par rapport aux autres.

Ici, nous nous intéressons à l'information apportée par la répétition des symboles (que nous pouvons appeler durée). En effet, si les événements acoustiques sont considérés comme des zones de stabilité, une démarche naturelle serait d'analyser les transitions entre ces zones et non la durée de celles-ci.

Nous analysons l'impact des statistiques de durée sur la performance du système AES en comparant :
– les performances du système AES lorsque les répétitions de symboles dans les signaux sont remplacées par le symbole lui même ;
– l'information apportée par les séquences « hétérogènes », *i.e.* sans aucune répétition de symboles.

Il est démontré à la figure 6.5(a) que la durée des événements acoustiques porte de l'information puisque le système se dégrade dans cette configuration. D'autre part la figure 6.5(b) illustre la

perte en performance d'un système où seulement les séquences contenants le même symbole sont présentes dans le codebook. Ces expériences tendent à prouver que les deux types d'informations, séquences de durée ou séquences « hétérogènes », sont importantes et qu'il faut conserver les deux types.



(a) (b)

FIG. 6.11: *a) Analyse de l'approche AES lorsque les répétitions de symboles sont remplacées par le symbole seul. Taille du dictionnaire 128, 2-grammes. b) Analyse de l'approche AES lorsque les séquences de répétition de symboles sont les seules présentes dans le codebook. Taille du dictionnaire 128, 2-grammes.* Sys05

6.3.1.6 Augmentation de la durée d'entraînement

Le protocole NIST offre plusieurs durées d'apprentissage pour les locuteurs. Nous présentons dans le tableau **??** le comportement du système AES en utilisant trois fois plus de données, soit la tâche 3conv4w-1conv4w.

TAB. 6.1: *Comportement du système AES pour différentes durées d'apprentissage.*

Protocole	EER(%)	DCFmin(x100)
NIST05 det7 : 1conv4w-1conv4w	14.51	5.67
NIST05 det3 : 3conv4w-1conv4w	9.41	4.37

Cette expérience illustre le bon comportement du système AES lorsque la quantité de données pour l'apprentissage est plus importante. En effet, le système arrive à tirer profit de l'augmentation de données à l'instar des systèmes cepstraux classiques puisqu'un gain (absolu) de 5% à l'EER et de 1% à la DCFmin est observé.

Nota : En général, il y autant d'exemples positifs pour l'apprentissage du SVM qu'il y a de sessions d'apprentissage (ici 3 sessions). Dans l'expérience précédente, pour rester consistant avec le

protocole utilisé par le GMM-UBM, les statistiques pour le locuteur ont, au contraire, été moyennées dans un seul vecteur.

6.3.2 Concaténation de l'information provenant de différentes longueurs de *N*-grammes

Ce paragraphe se penche sur l'intégration de l'information des différentes longueurs de *N*-grammes pour le système AES. L'intégration de ces multiples informations peut être réalisée de deux façons :

- La première est d'appliquer plusieurs systèmes AES et d'effectuer une fusion des scores *a posteriori*.
- La seconde est d'intégrer les différentes longueurs de séquences en concaténant les vecteurs d'expansion TFLLR pour chaque longueur de *N*-gramme à l'entrée du SVM.

Le tableau 6.1 montre les performances des systèmes AES pris indépendemment lorsque chacun utilise une longueur de *N*-gramme fixe. Ce sont les systèmes de référence qui vont être fusionnés.

TAB. 6.2: *Systèmes AES avec des longueurs de* N*-gramme allant de 1 à 6. Pas de normalisation des scores. Les meilleurs systèmes sont les 2,3-gramme. NIST05, det7,* Sys05.

N du *N*-gramme	1	2	3	4	5	6
EER (%)	16.30	14.09	14.51	16.5	20.19	43.74
DCFmin(x100)	5.75	5.46	5.67	6.35	7.38	9.29

Il est clair que les performances dépendent de la longueur de séquence et qu'il y a une limite à l'ordre des *N*-grammes que le système AES peut gérer. A partir des modèles 6-gramme, la couverture n'étant plus suffisante, le système ne présente plus des performances satisfaisantes. Dans le tableau **??**, nous présentons la moyenne de la couverture des signaux par le le codebook (calculée comme à l'équation 6.1).

TAB. 6.3: *Couverture en* N*-grammes entre le codebook et les données pour un dictionnaire de taille 128. NIST05, det7,* Sys05.

N	1	2	3	4	5	6
Couverture (%)	100	22.5	13.75	9.8	6.5	2.5

Les deux prochaines expériences visent à capturer l'information présente dans les différentes longueurs d'analyse afin de tirer parti des événements acoustiques de taille variable.

6.3.2.1 Combinaison de systèmes AES indépendants

Cette expérience vise à combiner les scores des systèmes *a posteriori* grâce à une moyenne arithmétique non-pondérée des scores. Le tableau 6.2 présente les résultats obtenus.

TAB. 6.4: *Fusions dans l'espace des scores pour l'intégration des différentes longueurs de séquences dans l'approche AES. Le score final est obtenu par une moyenne arithmétique non-pondérée. La fusion des systèmes de 1 à 6 grammes dégrade fortement les performances, et celle de 1 à 4 grammes ne permet pas d'amélioration par rapport au meilleur système AES. NIST05, det7,* Sys05.

Type de fusion	DCFmin	EER
Moyenne 1 à 6-gramme	5.62	15.11
Moyenne 1 à 4-gramme	5.18	14.62

Une fusion de tous les systèmes a été essayée sans succès ; une perte significative en performance est observée. Ceci est du à l'intégration des systèmes dont l'ordre des N-grammes est supérieur à 5, dont les taux d'erreurs sont comparativement élevés. La fusion des systèmes uniquement d'ordre 1 à 4 est présentée dont les résultats tendent à prouver qu'une fusion simple n'est pas souhaitable pour l'intégration des différentes informations provenant de séquences de différentes longueurs puisque une perte en performance est observée par rapport au meilleur système (2-gramme).

6.3.2.2 Intégration des différentes longueurs de séquences dans le sac de Ngram

Pour cette expérience, toutes les séquences de différentes longueurs ont été intégrées dans un seul vecteur qui constitue l'entrée du classifieur SVM. La construction de ce vecteur résulte de la simple concaténation de tous les vecteurs contenant la projection TFLLR pour chaque longueur de séquence. Les séquences qui n'ont pas été vues au moins 10 fois dans les données du modèle du monde sont éliminées du codebook.

Le tableau 6.3 présente les résultats des deux expériences.

- toutes les séquences d'ordre 1 à 6 ont été fusionnées, ce qui résulte en une sélection de 15000 N-grammes dans le codebook. Cette configuration présente une légère amélioration par rapport au système de référence, mais le gain n'est pas significatif. En revanche, le SVM n'est pas perturbé dans sa sélection de l'information discriminante lorsqu'il traite un grand nombre de séquences, dont certaines apportent peu d'informations (6-gramme) ;
- la couverture en 5 et 6-gramme étant très faible (autour de 2%), le second résultat montre la fusion des séquences de 1 à 4, résultant à la sélection de 10000 N-grammes.

Les résultats de ces expériences montrent que le système réagit de la même façon indépendemment de l'ordre des N-grammes en entrée. Alors que dans les expériences précédentes, les systèmes à faible performance handicapent les meilleurs, cette méthode permet de conserver une performance équivalente au meilleur système.

TAB. 6.5: *Intégration de l'information des différentes longueurs de séquences dans l'approche AES par concaténation des statistiques TFLLR dans un seul vecteur. NIST05, det7,* Sys05.

Longueurs de N-grammes sélectionnées	DCFmin(x100)	EER(%)
1 à 6-grammes	5.46	14.09
1 à 4-grammes	5.43	14.09

6.3.3 Influence de la normalisation de canal par *feature mapping*

La technique du *feature mapping* décrite en 5.3.5 est largement utilisée pour les systèmes GMM-UBM car elle permet d'améliorer les résultats en normalisant les variations de canal entre les données d'apprentissage et de tests. Cette technique a été intégrée à LIA_SpkDet en 2006 (*Sys06*). La figure 6.6 montre l'effet de la normalisation de canal sur la performance du système AES. Dans cette configuration, l'AES ne profite pas de cette technique et subit même une dégradation. Il semble que dans ce cas, l'algorithme de regroupement est moins performant pour trouver des événements acoustiques pertinents.

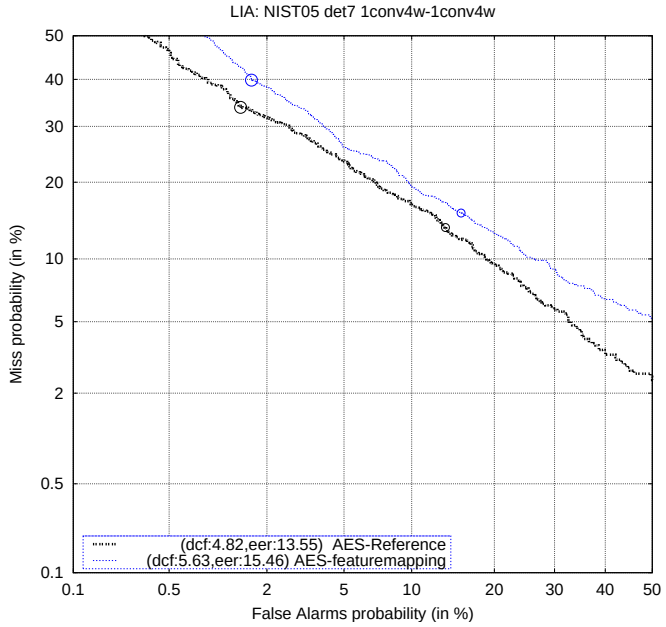


FIG. 6.12: *La technique du feature mapping appliquée au le système AES résulte en une dégradation des performances du système AES. L'algorithme de regroupement est moins performant pour trouver des événements acoustiques pertinents. NIST05, det7,* Sys06*.*

6.3.4 Combinaison avec un système état de l'art GMM-UBM

La plupart des systèmes dits « haut-niveau » n'ont pas pour objectif de rivaliser avec les performances des systèmes modélisant la distribution des vecteurs cepstraux. En revanche, ils essaient d'extraire de l'information nouvelle et complémentaire. Dans le cas de l'AES, l'information apportée provient de la dynamique du signal. Cette propriété n'est pas traitée par le GMM puisque l'ordre temporel des trames du signal n'est pas pris en compte par le modèle.

La figure 6.7 illustre la performance des deux systèmes de référence AES et GMM/UBM à l'aide d'une courbe DET et des mesures associées (minimum à la DCF et EER). La combinaison des deux systèmes à l'aide d'une fusion arithmétique (dont les poids ont été déterminés sur le jeu de développement) montre un gain faible mais significatif comparé au système GMM/UBM. En effet, l'EER descend de 8.73 à 7.99% et de 3.49 à 3.26 à la DCF.

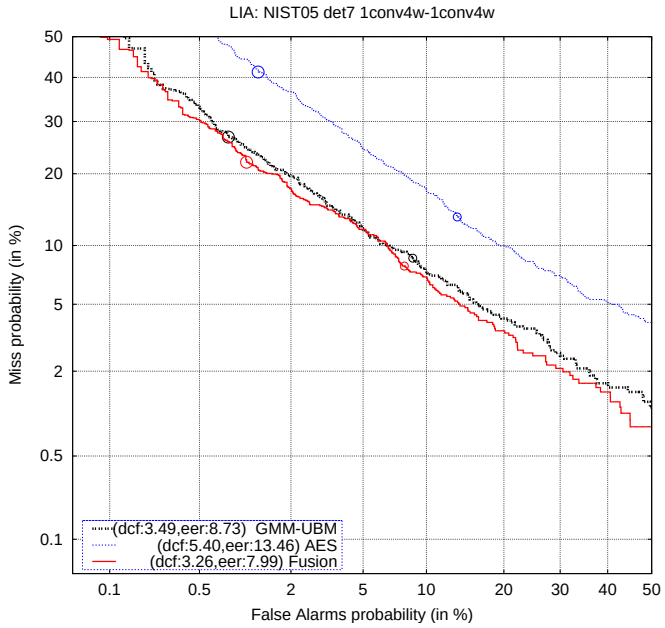


FIG. 6.13: *DET, minDCF et EER des systèmes GMM/UBM* Sys05*, AES et d'une fusion arithmétique pondérée (0.8 GMM-UBM/0.2 AES).*

6.4 Conclusion

Une approche innovante pour la modélisation des locuteurs a été présentée dans cette partie. En construisant un dictionnaire indépendant du locuteur par une procédure non-supervisée basée sur l'UBM et en analysant la dynamique des symboles générés, il est possible d'obtenir un système performant de vérification. Même si les performances n'atteignent pas le niveau des systèmes GMM-UBM, le système AES présente des performances à l'état de l'art comparés aux systèmes « haut-niveau » de type phonétiques ou lexicaux. Cette approche permet de conforter l'idée qu'une représentation appropriée pour la VAL peut être générée par des événements acoustiques dont la construction n'a nécessité que l'UBM.

La section 6.2 explicite le principe du systèmes AES : la partie indépendante (construction du dictionnaire) et la partie dépendante du locuteur (modélisation et vérification). Les expériences de la section 6.3 illustrent le comportement du système sous diverses conditions. Nous avons d'abord montré que la méthode TFLLR à base de SVM était la plus adaptée. La nécessité de travailler avec un dictionnaire issu de l'UBM mais présentant une taille réduite a ensuite été montrée expérimentalement. Ensuite, nous avons tenté d'intégrer la prise en compte de différentes tailles de N-grammes au système AES. Nous avons montré que la concaténation des expansions TFLLR à l'entrée du SVM permettait de garantir une certaine stabilité dans les performances. Finalement, la combinaison de l'AES avec un système GMM-UBM montre un gain notable confirmant l'apport d'information provenant de la dynamique du signal.

CHAPITRE 7

Extension du système AES à une analyse multi-résolution : C-AES

Sommaire

7.1 Introduction

La construction d'événements acoustiques et leur analyse séquentielle est une méthode dont nous avons prouvé l'efficacité pour la VAL dans le chapitre précédent. Cette méthodologie repose sur le rôle structurant de l'UBM pour l'espace acoustique en générant une représentation du signal à partir d'événements acoustiques reliés à ce modèle. L'analyse séquentielle de ces événements permet de capturer de l'information dynamique caractéristique du locuteur.

Ce système permet de combiner l'efficacité de la modélisation acoustique reposant sur l'utilisation d'un modèle du monde (UBM) avec une analyse séquentielle issues des outils utilisés dans les systèmes appelés « haut-niveau ». Ce chapitre propose l'extension de cette méthodologie à une analyse de multiples systèmes AES sur différentes classes et à différentes résolutions : ce système est nommé C-AES pour *Class-dependant Acoustic Event Sequences*, l'approche a été publiée dans [Scheffer et Bonastre, 2006a].

L'originalité de cette approche est d'appliquer une analyse séquentielle à l'intérieur de classes d'événements acoustiques en partant de l'hypothèse qu'une analyse locale peut apporter de l'information complémentaire à une analyse globale sur tout le signal de parole.

Dans la littérature, les méthodes de modélisation acoustique par classes sont largement utilisées en reconnaissance automatique de la parole ; notamment pour l'apprentissage des modèles. En vérification du locuteur, des approches comme [Andrews et al., 2001] utilisent les caractéristiques phonétiques pour discriminer les locuteurs. Dans [Stolcke et al., 2005], la combinaison d'une approche multi-classe fondée sur les différentes transformations MLLR (*Maximum Likelihood Linear Regression*) résultant d'un décodage automatique de la parole et d'un système génératif a montré de bonnes performances.

Le système C-AES utilise plusieurs jeux d'événements acoustiques dont les tailles de dictionnaires diffèrent. Nous parlerons de faible *résolution* lorsque la taille du dictionnaire est faible et inversement. Cette résolution n'est pas temporelle mais acoustique. En effet, un dictionnaire de taille élevée permet une représentation plus riche de l'événement acoustique pour une trame donnée et donc une plus grande résolution dans l'analyse. L'idée principale du système C-AES consiste à appliquer un système AES sur des classes à une faible résolution avant de combiner toutes les informations pour effectuer l'analyse multi-classes. Ces classes sont définies comme des événements acoustiques et sont par conséquent de même nature que les événements du système AES. Elles sont dénotées *Class Events*, à cause de leur plus faible résolution comparativement aux événements standards de l'AES (qui seront dénotées *Feature Events*).

Cette méthodologie par classe a pour objectif de capturer de nouvelles informations par rapport aux systèmes s'appuyant sur des statistiques estimées sur le signal dans son ensemble, comme les systèmes GMM-UBM qui considèrent une seule transformation entre le modèle UBM et le modèle d'un locuteur (e.g. : adaptation MAP).

Ce chapitre est organisé de la façon suivante : nous détaillons d'abord le système C-AES en 7.2, notamment la génération des *Class Events* à faible résolution. Pour pouvoir effectuer de multiples analyses, l'information *a priori* sur chaque classe a son importance. De ce fait, une modification du noyau TFLLR est proposée, ainsi que la méthodologie de concaténation des informations inter-classes. Enfin, des expériences illustrant la méthodologie sur les bases de données NIST-SRE sont présentées en 7.3.

7.2 C-AES : Une extension multi-classes pour l'AES

Nous détaillons ici la méthodologie de construction d'un système C-AES, extension du système AES (voir chapitre 6) à une analyse multi-classe.

7.2.1 Principe

L'idée principale du C-AES est d'effectuer une analyse séquentielle à l'intérieur de classes acoustiques du signal de parole. De telles classes peuvent être de deux types :

- Classes basées sur des unités phonétiques, e.g. : voyelles, fricatives, nasales... ;
- Classes acoustiques obtenues de manière non supervisée.

L'approche choisie pour ce travail relève de la deuxième catégorie : les classes, dénommées *Class Event*, sont considérées comme un autre type d'événement acoustique et sont générées de la même façon qu'en 6.2.1. Ces *Class Events* sont cependant à une résolution beaucoup plus basse. Cette diminution de résolution s'applique à l'analyse acoustique et non à l'axe temporel.

L'implémentation d'un système C-AES comprend les étapes suivantes, illustrées par la figure 7.1 :

1. génération des *Feature Event* (FE) à la manière d'un système AES classique ;
2. génération des *Class Event* (CE) (voir 7.2.2) ;
3. application d'un système AES indépendant pour chacun des CE ;
4. combinaison de l'information provenant des multiples systèmes (voir 7.2.3).

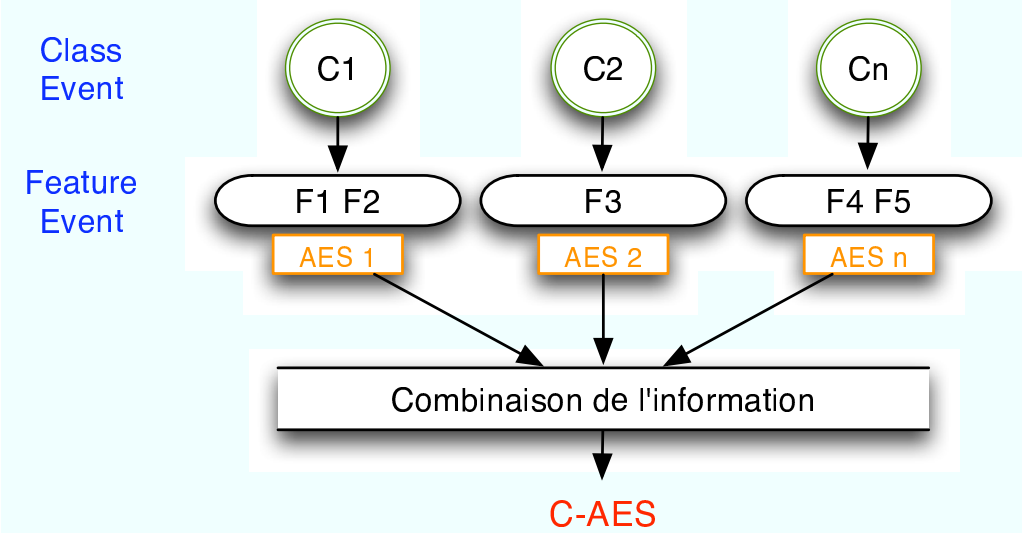


FIG. 7.1: *Méthode de combinaison de l'information multi-classes pour l'approche C-AES : le système AES est appliqué indépendamment sur chacun des n Class Events, chacun avec son propre dictionnaire de Feature Events. Le procédé de combinaison est sur les vecteurs produits par les systèmes AES.*

7.2.2 Génération des Class Event

Les *Feature Event (FE)* sont générés comme détaillé dans la méthodologie de construction du dictionnaire de l'AES en 6.2.1. Pour les *Class event (CE)*, le procédé de regroupement hiérarchique des événements est prolongé jusqu'à la résolution souhaitée.

Dans ce travail, la taille du dictionnaire est fixée à 128 pour les FE, et à 8 pour les CE. Ces valeurs ont été trouvées de façon empirique. Toutes les combinaisons de taille sont imaginables pour peu que $dim(CE) << dim(FE)$. Le processus de génération de symboles produit, pour chaque signal, deux représentations différentes du signal (en fait deux flux de symboles).

De par la nature même des CE — un regroupement de plusieurs FE — il est intéressant de remarquer que :

– le système C-AES maximise de façon implicite la couverture pour l'analyse séquentielle. Chaque *Class Event* possède un dictionnaire réduit comparé à un système AES standard. La couverture de toutes les séquences possibles à l'intérieur des classes est de fait maximisée. Afin d'illustrer ce dernier point, la table 7.1 présente la répartition de la taille du dictionnaire de FE parmi les différents CE ;
– cette méthode peut être interprétée comme un procédé de sélection de séquences dans un système AES standard, où les séquences de FE n'appartenant pas à la même classe sont éliminées. La figure 7.2 résume cette sélection en présentant les séquences éliminées dans un système C-AES en comparaison de celles présente dans le système AES.

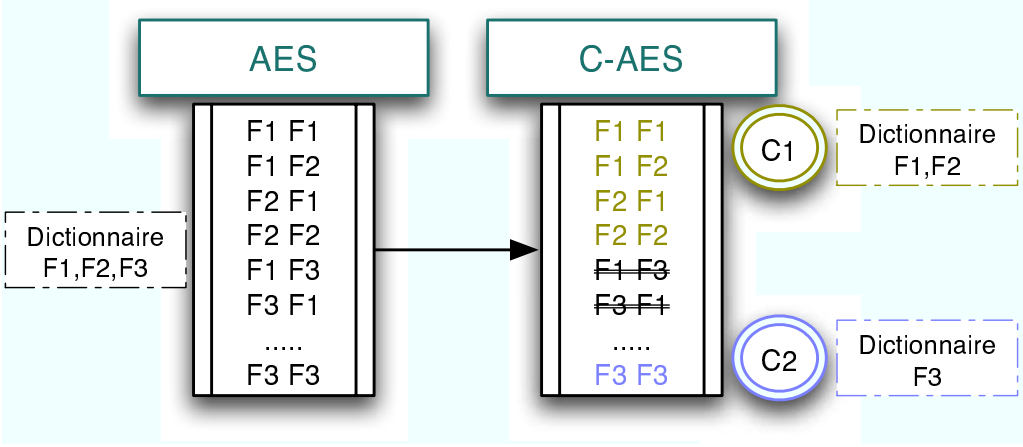


FIG. 7.2: *Comparaison entre les séquences 2-gramme prises en compte par un système AES par rapport au système C-AES. Les séquences inter-classes sont éliminées de l'analyse.*

TAB. 7.1: *Taille du dictionnaire des Feature Event pour le système C-AES à 8 CE.*

Class Event k	1	2	3	4	5	6	7	8
Taille du dictionnaire de FE correspondant	37	10	11	29	9	9	13	10

7.2.3 Construction du noyau

Le système AES est fondé sur les machines à vecteur support (SVM) comme méthode de classification, en utilisant le noyau TFLLR (*Term-Frequency Log-Likelihood Ratio kernel*, voir le paragraphe 4.5). Celui-ci permet de produire une expansion adéquate pour des systèmes de détection du locuteur reposant sur une modélisation N-gramme.

Dans notre cas, le noyau TFLLR ne répond pas à tous nos besoins, particulièrement pour l'intégration de l'information des *Class Event*. En l'état, le système nécessite l'information *a priori* sur les CE pour pouvoir modéliser l'influence de la classe sur l'analyse AES. Certaines techniques de normalisation (moyenne/variance, rang) pourraient nous aider à normaliser l'influence des classes, mais cela n'est pas notre objectif. En effet, ces techniques ne sont efficaces que quand aucune information *a priori* n'est disponible. La section 7.3 de ce chapitre tend à prouver, que cette information est cruciale, tout comme la façon de l'estimer.

Nous proposons de modifier le noyau TFLLR afin qu'il puisse produire des scores de vérification sur de multiples classes en utilisant l'influence de chaque CE. Considérons l'unité k appartenant au sac de N-gramme B et la vraisemblance de cette unité sur une séquence de donnée $\mathbf{X}$, $p(k|\mathbf{X})$, le score produit par le noyau TFLLR est défini par :

$$\sum_k \frac{p(k|\mathbf{X}_S)}{\sqrt{p(k|\mathbf{X}_W)}} \frac{p(k|\mathbf{X}_{S'})}{\sqrt{p(k|\mathbf{X}_W)}} - 1, \tag{7.1}$$

où $\mathbf{X}_S$,$\mathbf{X}_{S'}$,$\mathbf{X}_W$ sont les données d'apprentissage respectives de deux locuteurs, S et S' et du modèle du monde W. Le noyau s'exprime comme la pondération des vraisemblances sur les données des locuteurs par celle sur les données du monde.

Dans le cas du C-AES, chaque unité k^j est un N-gramme dont les symboles appartiennent au dictionnaire du *Class Event* correspondant (notée C_j). Le calcul du LLR entre un segment de test $\mathbf{X}$, les données d'un locuteur $\mathbf{X}_S$ et celles du modèle du monde $\mathbf{X}_W$ adapté pour une analyse multi-classe peut être exprimé sous la forme :

$$\mathcal{Y}_S(\mathbf{X}) = \sum_j \sum_{k^j \in C_j} p(C_j|\mathbf{X}) \mathcal{Y}_S(\mathbf{X}, C_j), \tag{7.2}$$

où $p(C_j|\mathbf{X})$ est la probabilité *a posteriori* du *Class Event* C_j sur les données et $\mathcal{Y}_S(\mathbf{X}, C_j)$ est le logarithme du rapport de vraisemblance (défini par l'équation 6.8) calculé pour les unités k^j appartenant au *Class Event* C_j.

Afin de prendre en compte l'information non uniformément répartie entre les classes, nous définissons dans la suite l'expansion des données $\mathbf{X}_S$ pour les N *Class Event* $C = \{C_j\}_{j=[1 \ldots N]}$.

Pour une classe C_j donnée, de taille M_j, et contenant toutes les unités correspondantes telles que $C_j = [k_1^j, \ldots k_{M_j}^j]$, l'expansion du noyau modifié a la forme suivante :

$$\Phi_{C_j}(\mathbf{X}_S) = [\phi(\mathbf{X}_S|k_1^j), \ldots, \phi(\mathbf{X}_S|k_{M_j}^j)] \tag{7.3}$$

$$\phi(\mathbf{X}_S|k_i^j) = p(C_j|\mathbf{X}_S) \frac{p(k_i^j|\mathbf{X}_S)}{\sqrt{p(k_i^j|\mathbf{X}_W)}}. \tag{7.4}$$

L'expansion pour une classe consiste à calculer la vraisemblance de l'unité sur les données du locuteur pondérée par la probabilité du modèle du monde et la probabilité du *Class Event* lui correspondant.

Afin de combiner l'information provenant des différentes classes, le vecteur résultat contient la concaténation de toutes les expansions des données pour les N *Class Event*.

$$\overline{\Phi}(\mathbf{X}_S) = [\Phi_{C_1}(\mathbf{X}_S), \ldots, \Phi_{C_N}(\mathbf{X}_S)]. \tag{7.5}$$

En pratique, pour chaque *Class Event*, un système AES indépendant est appliquée sur les séquences du dictionnaire correspondant à la classe. Ceci produit 8 vecteurs dans notre cas. Pour une instance (un accès), tous les vecteurs correspondant à chaque *Class Event* sont alors concaténés après avoir été pondérés par la probabilité de la classe décrite au paragraphe précédent.

Procédure de vérification Comme dans l'AES au paragraphe 6.3.1.2, afin de représenter les données de la classe négative dans le SVM, des pseudo-imposteurs utilisés pour apprendre le modèle du monde sont employés. L'entrée du classifieur est constitué de l'accès d'apprentissage du locuteur et de tous les imposteurs. La décision à marge maximale est trouvée en passant cette entrée à un noyau linéaire.

L'outil *SVM-Light* de Thorsten Joachims [Joachims, 1999] a été utilisé pour calculer les SVM et pour classifier les instances. Afin de compenser la grande disproportion entre les tests clients et imposteurs, un modèle de coût lors de l'apprentissage a été adopté. Ainsi, une erreur de classification sur les exemples positifs sera 200 fois plus coûteuse que sur les exemples négatifs (une valeur trouvée empiriquement). Les scores obtenus de cette manière sont normalisés par une *T-normalisation* .

7.3 Expériences et Résultats

Dans ce paragraphe, nous présentons les différentes expériences réalisées sur le système C-AES. Le protocole d'évaluation ainsi que le système de référence ont été définis au chapitre 5. Tout au long des expériences, la base de données utilisée est NIST05, protocole det7 et la configuration du système *Sys05*. Les résultats généraux du système C-AES sont présentés ainsi que les différentes méthodes pour estimer la probabilité des classes CE vu en 7.5. Nous concluons en comparant les performances d'un système GMM-UBM et d'un système AES, ainsi que les possibilités de combinaison.

7.3.1 Estimation des probabilités de classes CE

Nous présentons deux approches différentes pour l'estimation de la probabilité de la classe. La première consiste à prendre la probabilité *a priori* de chaque *Class Event*, la seconde fait appel à l'estimée utilisant le *maximum a posteriori* (MAP).

7.3.1.1 Probabilité *a priori* comme facteur de pondération

Pour estimer la probabilité *a priori* de chaque *Class Event*, toutes les données du modèle du monde $\mathbf{X}_W$ générées à cette résolution ont été utilisées. La probabilité *a priori* de la classe $P(C_j)$ est estimée comme :

$$P(C_j) = P(C_j|\mathbf{X}_W) \tag{7.6}$$

où $P(C_j|\mathbf{X}_W)$ est estimée comme la fréquence d'apparition de la classe sur $\mathbf{X}_W$. Dans notre cas, le tableau 7.2 donne les facteurs de pondérations utilisés pour les expériences.

TAB. 7.2: *Proabilité a priori des Class Event pour le modèle 8 classes du système C-AES.*

C_j	1	2	3	4	5	6	7	8
$P(C_j)$	0,41	0,02	0,11	0,26	0,07	0,03	0,06	0,03

7.3.1.2 Utilisation de MAP pour l'estimée

Une approche *maximum a posteriori* (MAP) peut être utilisée pour estimer la probabilité de la classe. L'estimation de la vraisemblance uniquement sur les données d'un locuteur n'étant pas assez précise, la technique du MAP permet d'utiliser la probabilité *a priori* si la classe y est peu présente. Précisément, si $\tilde{p}(C_j|\mathbf{X})$ est la nouvelle estimée de la probabilité, alors :

$$\tilde{p}(C_j|\mathbf{X}) = \alpha p(C_j|\mathbf{X}) + (1-\alpha)p(C_j|\mathbf{X}_W), \text{ avec } \alpha = \frac{C(k)}{C(k)+\tau}, \tag{7.7}$$

où $C(.)$ est l'opérateur de compte d'unités, et τ le *relevance factor* (déterminé empiriquement) de l'équation 2.6 (avec 8 classes, τ a été fixé à la valeur 1000).

7.3.1.3 Résultats

Le tableau 7.3 illustre l'effet de l'estimation de $p(C_k|\mathbf{X})$ sur la performance du système C-AES. Les résultats tendent à prouver que l'intégration de cette information est nécessaire. En effet, le système sans aucun facteur de pondération présente un taux d'erreur deux fois plus important que celui qui utilise l'estimation de la probabilité par son *a priori*. Le second résultat montre que la méthode d'estimation de cette probabilité joue également un rôle important. Un gain absolu de 1%, à la fois à l'EER et à la DCFmin est observé lorsque l'estimation MAP est utilisée.

TAB. 7.3: *Performance du système C-AES en comparant les différentes méthodes d'estimation des probabilité des Class Event : pas de pondération, estimée de la probabilité du Class Event par son a priori et par MAP. NIST05, det7,* Sys05

Système	DCFmin (x100)	EER
Sans pondération	9.80	28.5
a priori	6.46	14.87
maximum a posteriori	5.87	13.89

7.3.2 Combinaison des scores des systèmes

L'objectif du système C-AES est d'apporter de nouvelles informations complémentaires aux systèmes standards, lesquels caractérisent le signal vocal par une transformation globale. Nous présentons une expérience de combinaison des sorties des systèmes afin de valider notre approche.

Afin de fusionner les systèmes, tous les scores des systèmes ont été normalisés (Tnorm) et une moyenne arithmétique entre les scores des systèmes est appliquée. Le système GMM-UBM correspond au système de référence *Sys05* présenté dans 5.3. Le tableau 7.4 présente différentes combinaisons entre les systèmes de référence et le C-AES, alors que la figure **??** illustre les performances sous forme d'une courbe DET (les poids de la fusion ont été déterminés sur le jeu de développement afin d'optimiser la DCFmin).

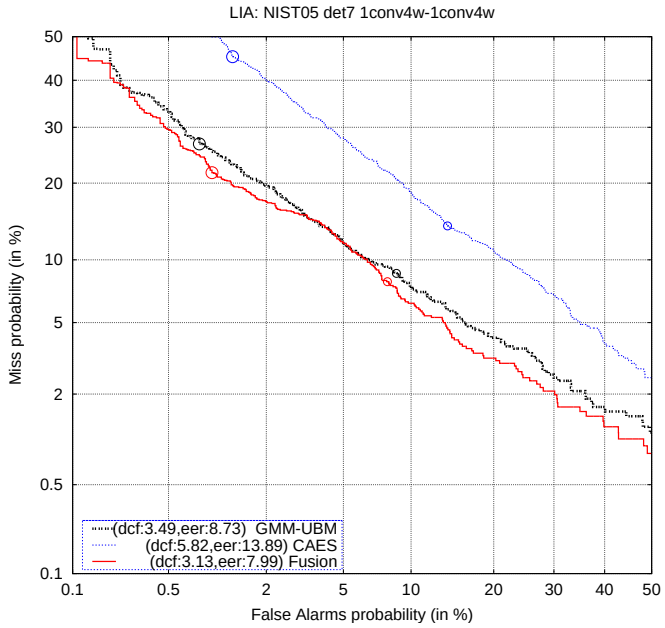


FIG. 7.3: *Courbes DET pour le système de référence GMM-UBM, le CAES et leur combinaison. La fusion des systèmes permet d'améliorer les performances du système GMM-UBM de référence. NIST05, det7,* Sys05

TAB. 7.4: *Performance des systèmes de référence GMM-UBM, AES, C-AES et combinaison des différents systèmes. Les poids de fusion sont indiqués dans le tableau. La fusion entre GMM-UBM et C-AES obtient de meilleurs taux d'erreurs qu'avec le système AES standard. NIST05, det7,* Sys05

Système	DCFmin (x100)	EER (en %)
GMM-UBM *Sys05*	3.58	8.54
AES	5.37	13.33
CAES	5.87	13.89
GMM-UBM *Sys05* (0.8) + AES (0.2)	3.31	8.04
GMM-UBM *Sys05* (0.7) + CAES (0.3)	3.16	7.96

Alors que les performances du système C-AES sont plus faibles par rapport au système AES classique, il est intéressant de remarquer le gain en performance après fusion. En effet, un gain (relatif) de 12% et 7% est observé à la DCFmin et à l'EER respectivement, à comparer à 7% et 6% pour le système standard GMM-UBM fusionné avec l'AES.

Ces expériences tendent à prouver qu'une partie importante de l'information réside à l'intérieur de classes acoustiques du signal de parole. Les relations interclasses peuvent être omises pour une telle analyse dynamique. Il est cependant possible d'envisager l'analyse séquentielle des *Class Event* mais ceci ne fait pas partie de notre étude.

7.4 Conclusion

En proposant une méthodologie pour appliquer de multiples analyses séquentielles (C-AES) sur des classes à plus faible résolution, nous avons montré, à travers les expériences, que l'information intra-classe est caractéristique du locuteur et qu'elle apporte une autre information que celle véhiculée par le signal dans son ensemble (à la façon de MAP).

Pour tirer ces conclusions, nous avons étendu la définition du noyau TFLLR à son application multi-classe en intégrant l'information *a priori* sur les classes. Nous avons montré que son estimation devait être effectuée avec précaution (7.3). En effet, les résultats sans le noyau TFLLR modifié présentent des taux d'erreurs deux fois plus important. De plus, l'estimation MAP donne un gain de 1% absolu à la DCFmin et à l'EER. La combinaison de ce système avec un système GMM-UBM montre un gain relatif de 12% à la DCFmin et de 7% à l'EER, ce qui donne de meilleurs résultats qu'avec le système AES.

Dans la construction du système C-AES, nous n'avons pas étudié la possibilité d'une analyse séquentielle à l'échelle des CE. Les CE ont un rôle différent comparé aux FE dans les expériences réalisées, mais ceci pourra faire l'objet de travaux futurs.

Les perspectives de ce travail portent aussi sur l'extension de ces expériences en générant des jeux *Class Event* et de *Feature Event* à de multiples résolutions pour les combiner dans un seul vecteur. Au delà de la simple performance du système, l'intérêt principal de ce travail est d'avoir proposé une méthodologie d'intégration de l'AES dans un contexte multi-résolution.

CHAPITRE 8

Structuration de l'espace acoustique par l'UBM pour une méthode discriminante : l'approche SVM-UBM pour la VAL

Sommaire

8.1 Introduction

L'approche majoritairement utilisée en VAL est basée sur les modèles génératifs pour représenter le locuteur (chapitre 2. L'utilisation du paradigme GMM-UBM [Reynolds, 1995] apparaît maintenant comme une étape indispensable pour obtenir des performances proches de l'état de l'art dans des campagnes d'évaluation internationales telles que les campagnes NIST-SRE. Ces dernières années ont vu l'apparition d'approches discriminantes basées sur l'utilisation des machines à vecteurs supports (SVM). Certaines de ces approches ont été appliquées avec succès dans [Wan et Campbell, 2000] et présentent des performances proches de celles de la modélisation générative. Ainsi, ces méthodes font toujours l'objet d'un intérêt croissant.

Dans ce cadre, la combinaison des méthodes discriminantes et génératives est particulièrement intéressante. De par leur capacité à bien représenter les données, les mixtures de modèles (GMM) sont souvent employées comme modèle génératif pour ces techniques. Cependant, ces méthodes sont complexes à mettre en oeuvre. Les méthodes basées sur la distance de *Kullback* (voir en 3.3.2) interviennent seulement en aval des méthodes GMM-UBM. Elles nécessitent le processus coûteux d'adaptation du modèle génératif aux locuteurs. D'autre part et malgré leur attrait théorique, les techniques basées sur les noyaux de *Fisher* ne sont pas employées durant les campagnes NIST-SRE car elles se prêtent mal aux évaluations de grande ampleur. La mise en oeuvre de ces techniques est en effet complexe de par la grande taille du problème de classification qui en résulte.

Ce chapitre présente une méthode simple et peu coûteuse permettant de combiner les approches génératives et discriminantes en utilisant le modèle UBM comme unique modèle génératif. Alors que les méthodes à noyaux de séquences n'utilisent les données pseudo-imposteurs que pour normaliser les expansions dans l'espace des caractéristiques (voir chapitre 3), notre approche fait intervenir le modèle générique UBM dans le processus afin de structurer l'espace acoustique à l'instar des méthodes de modélisation générative du chapitre 2. Comme tout au long de ces travaux, nous pensons que le modèle générique permet d'extraire les statistiques nécessaires pour la modélisation du locuteur. A cet effet, nous proposons une dérivation du noyau TFLLR (*Term Frequency Log-Likelihood Ratio kernel*) proposée dans [Campbell et al., 2004b] pour la prise en compte des composantes du modèle UBM. Nous utilisons le modèle générique UBM pour structurer la représentation des locuteurs dans l'espace des scores.

La dérivation du noyau adapté à notre problème est présentée dans le paragraphe 8.2. Ce système est appelé SVM-UBM et a été publié dans [Scheffer et Bonastre, 2006b]. Le paragraphe 8.3 détaille le protocole expérimental et les résultats des expériences. Nous montrons que l'utilisation du noyau TFLLR, très proche d'une formulation réduite du noyau de Fisher, donne des performances proches du système de détection de locuteur standard : le GMM-UBM. Nous abordons aussi le problème de la normalisation des exemples à l'entrée du SVM. Nous concluons en analysant l'apport d'un tel système dans le cas de sa combinaison avec un système GMM-UBM.

8.2 Exploitation du modèle générique pour la modélisation discriminante des locuteurs

Ce paragraphe présente la méthodologie adoptée afin de construire un système de détection du locuteur à caractère discriminant en utilisant le modèle générique. Ce système doit apporter de l'information complémentaire à la modélisation purement générative du paradigme GMM-UBM. Le modèle générique UBM constitue la fondation de ce système.

La première partie rappelle brièvement les principes des SVM et leur application à la VAL. Ensuite, les techniques exploitant les modèles génératifs dans un SVM sont rappelées afin d'introduire une nouvelle manière d'interpréter le noyau TFLLR lorsqu'il est utilisé conjointement avec le modèle UBM.

8.2.1 Les machines à vecteurs supports en vérification du locuteur

Les machines à vecteurs supports (SVM) décrites par [Vapnik, 1998] sont généralement utilisées comme un classifieur bi-classes en VAL (client/imposteur). Pour répondre à un problème linéairement séparable, le SVM construit l'hyperplan de décision optimal qui maximise la marge entre deux classes.

Un sous-ensemble des données d'apprentissage, les *vecteurs supports*, est mémorisé. En VAL, cela implique que le locuteur est modélisé par ses données d'apprentissages et par un sous-ensemble de données imposteurs, les plus proches de l'hyperplan. La dimension du problème est réduite considérablement tout en gardant une bonne capacité de généralisation. La plupart du temps, les données ne sont pas linéairement séparables, et l'introduction de *variables ressorts* pour la prise en compte des coûts de mauvaise classification se révèle nécessaire.

La discrimination de séquences de longueur variable, comme les trames d'un signal de parole, est une tâche difficile. Cependant, des techniques visant à projeter une séquence de trames vers un vecteur de longueur fixe existent (section 3.3) et sont bien adaptées à la vérification du locuteur. Dans le cadre des campagnes d'évaluations NIST-SRE, ces méthodes ont été appliquées dans [Wan et Campbell, 2000], avec des noyaux polynomiaux d'ordre 3 maximum et dans [Louradour et Daoudi, 2005] avec des ordres supérieurs en montrant des performances satisfaisantes sur les évaluations internationales. Le schéma 8.1 (déjà présenté au chapitre 3) rappelle la structure d'un système de VAL utilisant les SVM, en prenant pour exemple le noyau GLDS (*Generalized Linear Discriminant Sequence kernel*). Le noyau de séquence est construit par une moyenne d'expansions polynomiales des vecteurs cepstraux et une normalisation par une matrice de corrélation. Nous adaptons cette structure pour y faire intervenir le modèle génératif.

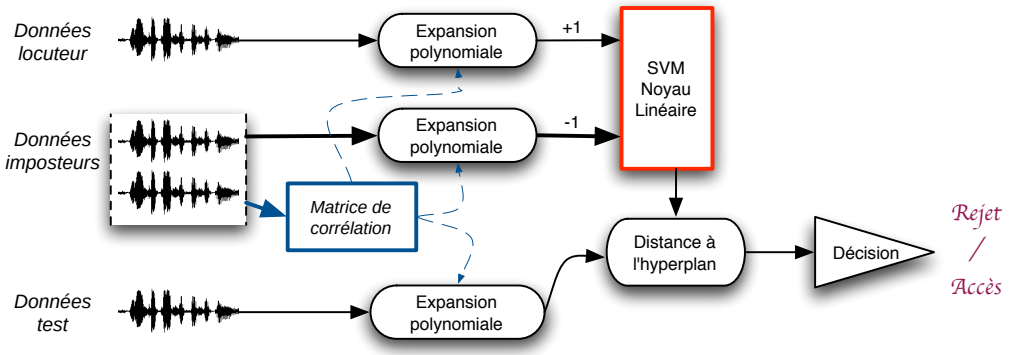


FIG. 8.1: *Exemple d'utilisation des SVM en VAL : le système GLDS. L'expansion polynomiale normalisée des vecteurs cepstraux permet de construire un noyau approprié pour la VAL.*

8.2.2 Noyaux de séquences exploitant les modèles génératifs

Les noyaux exploitant les capacités des modèles génératifs ont été introduits par [Jaakkola et Haussler, 1998] sous le nom de *noyaux de Fisher* et généralisés ensuite dans un ensemble de méthodes de l'espace des scores ou *score-space* [Smith et al., 2001]

Comme abordé au chapitre 3, la mise en oeuvre d'un noyau est difficile pour des séquences de taille variable. Ainsi, les systèmes de VAL basés sur les SVM s'attachent à trouver une expansion de la séquence (ou *feature mapping*) correspondant à la fonction $\phi(.)$ pour ensuite appliquer un produit scalaire pour former un noyau $K(.,.) =< \phi, \phi >$. La méthode que nous employons s'inscrit dans cette catégorie. Notre objectif consiste dans un premier temps à trouver un *feature mapping* approprié, puis d'appliquer un produit scalaire entre les paires d'exemples dans le *feature space* dans un deuxième temps.

Dans ces travaux, un intérêt particulier est porté à l'espace des scores de vraisemblance, le *likelihood score space*. Le lecteur est invité à se référer à [Wan et Renals, 2002] pour une dérivation plus détaillée des autres espaces des scores. Soit M un modèle GMM, paramétré par θ. Le *Fisher mapping* Ψ_{Fisher} d'une séquence de trames $\mathbf{X}$ est donné comme la dérivée première de la fonction de log-vraisemblance. Soit plus précisément :

$$\Psi_{Fisher}(\mathbf{X}) = \nabla_{\boldsymbol{\theta}} \log(p(\mathbf{X}|M,\boldsymbol{\theta})). \tag{8.1}$$

Chaque valeur du vecteur résultant contient la dérivée du log-vraisemblance par rapport à chacun des paramètres de $\boldsymbol{\theta}$. Considérons la dérivation de cette expression par rapport au poids α_j de la composante G_j du GMM M :

$$\frac{\partial}{\partial \alpha_j} \log(p(\mathbf{X}|M,\boldsymbol{\theta})) = \frac{1}{T} \sum_{t=1}^{T} \frac{\frac{\partial}{\partial \alpha_j} p(x_t|M,\boldsymbol{\theta})}{p(x_t|M,\boldsymbol{\theta})} \tag{8.2}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \frac{p(x_t|G_j)}{\sum_{n=1}^{N} \alpha_n p(x_t|G_n)}. \tag{8.3}$$

Cette dérivée partielle correspond à une accumulation pour chaque trame de la vraisemblance non-pondérée pour une composante donnée, divisée par la vraisemblance totale pour cette trame. La

dimension de ce vecteur est égale au nombre de composantes dans le GMM. Nous introduisons cette formulation spécifique car nous allons retrouver son expression dans la suite des travaux.

8.2.3 De l'utilisation de l'UBM et d'un SVM pour la vérification du locuteur

L'approche présentée dans ce chapitre repose sur l'information donnée par un seul GMM, précisément l'UBM. Au lieu d'apprendre les modèles clients par une adaptation MAP (ou un critère ML) et d'effectuer la tâche de vérification avec un SVM, nous proposons dans la suite une méthode utilisant seulement les composantes de l'UBM pour l'apprentissage discriminant.

Les exemples à l'entrée du SVM — lesquels ont été extraits à partir des paramètres de l'UBM — doivent représenter le comportement de ce modèle sur les données d'apprentissage d'un locuteur. Le noyau TFLLR présenté dans [Campbell, 2006] est utilisé afin de produire des projections appropriées pour des approches à base de N-grammes. Sa formulation nous sert à dériver un noyau adapté à notre problème.

Considérons des unités k appartenant à un sac de N-gramme B. Soit la vraisemblance de cette unité sur une séquence de données $\mathbf{X}$, notée $p(k|\mathbf{X})$, le noyau TFLLR est calculé comme suit :

$$K_{TFLLR}(\mathbf{X}_S, \mathbf{X}_{S'}) = \sum_{k \in B} \frac{p(k|\mathbf{X}_S)}{\sqrt{p(k|\mathbf{X}_W)}} \frac{p(k|\mathbf{X}_{S'})}{\sqrt{p(k|\mathbf{X}_W)}}, \tag{8.4}$$

où $\mathbf{X}_S$,$\mathbf{X}_{S'}$,$\mathbf{X}_W$ sont les données d'apprentissage respectives de deux locuteurs et du modèle du monde. La construction du noyau réside finalement dans la pondération des vraisemblances des locuteurs par la vraisemblance du modèle du monde.

Posons maintenant l'unité comme la composante gaussienne de l'UBM W (définie comme W_j pour la composante d'indice j) et considérons sa probabilité comme son occupation sur les données d'apprentissage. Alors, pour une séquence donnée $\mathbf{X}$ et une unité définie comme W_k, l'expansion $\Phi_{W_k}(\mathbf{X})$ associée au noyau TFLLR est donnée par :

$$\frac{p(W_j|\mathbf{X})}{\sqrt{p(W_j)}} = \frac{1}{T} \sum_{i=1}^{T} p(W_j)^{-1/2} \frac{p(x_t|W_j)p(W_j)}{p(x_t)} \tag{8.5}$$

$$= \frac{1}{T} p(W_j)^{1/2} \sum_{t=1}^{T} \frac{p(x_t|W_j)}{\sum_n p(x_t|W_n)p(W_n)} \tag{8.6}$$

$$= \sqrt{p(W_j)} \nabla_{\alpha_j} \log(p(\mathbf{X}|W, \boldsymbol{\theta}_W)), \tag{8.7}$$

où $p(W_j|\mathbf{X})$ correspond à l'occupation de la gaussienne W_j pour la séquence $\mathbf{X}$. Vues d'une autre manière, ces statistiques correspondent au poids des composantes du modèle résultant de la première itération de l'algorithme EM.

Ce noyau apparaît très clairement relié au *Fisher mapping* décrit dans l'équation 8.2. La racine carrée additionnelle de la composante gaussienne peut être vue comme une normalisation lissant la dynamique des features. Ce lissage est utile dans le cas de données creuses, cette fonction peut cependant être remplacée par une fonction plus agressive (*e.g.* logarithme) comme dans [Campbell, 2006]. Pour notre taille de problème et d'après les expériences, son influence se révèle restreinte.

Pour faire correspondre ce noyau à celui de Fisher, une normalisation par la matrice d'information de Fisher est nécessaire. Cette matrice est souvent approximée par une matrice de covariance diagonale ou par une matrice identité, ce qui est le cas dans nos travaux. Aucune expérience utilisant cette normalisation n'a été réalisée mais de telles expériences peuvent être l'objet de travaux futurs (le terme sous la racine de l'expansion TFLLR peut être vu comme une forme de normalisation).

La structure du système SVM-UBM est résumée à la figure 8.2, où l'UBM intervient dans l'expansion des trames du signal en utilisant le noyau TFLLR. La matrice de normalisation est considérée comme l'identité dans ces travaux.

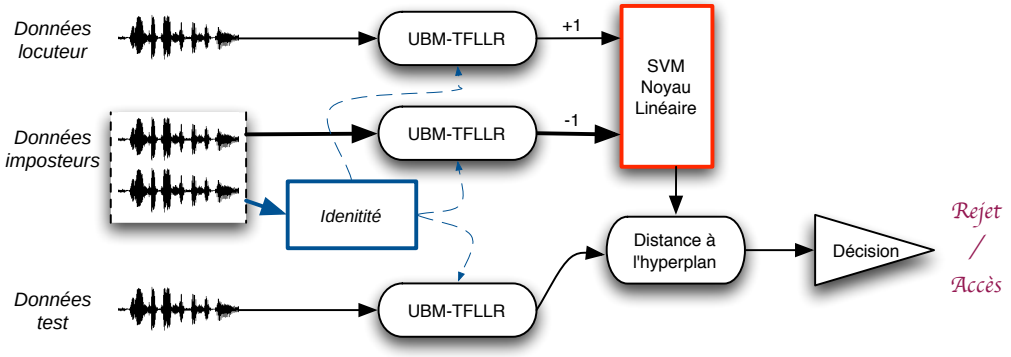


FIG. 8.2: *Schéma du système SVM-UBM. La projection des données dans le feature space est représentée par l'expansion TFLLR appliquée à l'UBM. Ceci revient à utiliser une forme réduite du Fisher mapping*

8.3 Développement Expérimental

Les expériences présentées par la suite illustrent les résultats principaux du système SVM-UBM. Nous analysons aussi les comportements du système par rapport à la normalisation des exemples, des scores, et à la quantité de données disponibles pour l'apprentissage.

8.3.1 Expériences et protocoles

Les expériences de vérification du locuteur, présentées en 8.3, sont effectuées sur la base NIST05, protocole det7, sauf si cela est clairement mentionné. Le système GMM-UBM correspond à la version *Sys05* de LIA_SpkDet.

Afin de construire les modèles imposteurs, des locuteurs provenant du modèle du monde ont été utilisés, ici 161 locuteurs. Durant l'apprentissage, l'entrée du classifieur comporte tous les vecteurs imposteurs étiquetés négativement (issus des données du monde) et le vecteur du locuteur étiqueté positivement (issu de ses données d'apprentissage). La marge de décision maximale est obtenue en passant cette entrée à un noyau linéaire. Lors du processus de vérification, le vecteur de test est donné comme entrée au SVM. Pour chaque test de vérification, le score correspond

à la distance du noyau TFLLR entre le segment de test et l'hyperplan correspondant à l'identité proclamée.

Nous avons utilisé l'outil *SVM-Light* de [Joachims, 1999] pour l'apprentissage des SVM et la classification des instances. Afin de compenser la grande disproportion entre les tests client et imposteur, un modèle de coût lors de l'apprentissage a été adopté. Ainsi, une erreur de classification sur les exemples positifs à l'apprentissage sera 200 fois plus coûteuse que sur les exemples négatifs (une valeur trouvée empiriquement). Les scores obtenus de cette manière ont ensuite été normalisés avec une *T-normalisation* (sauf si cela est clairement mentionné).

8.3.2 Influence de la taille du modèle UBM

Pour les expériences, deux tailles différentes de modèles UBM ont été utilisées, 128 et 2048 composantes. Cette variété va nous permettre d'analyser les performances du systèmes en fonction de la complexité du modèle UBM.

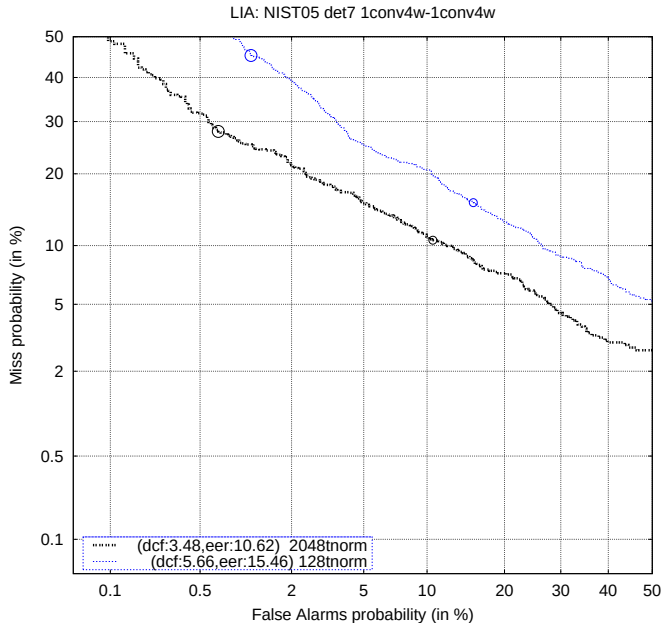


FIG. 8.3: *Comparaison entre deux systèmes SVM-UBM utilisant un modèle générique de taille différente : 128 (fin) et 2048 (épais)*

Les résultats de la figure 8.3 montrent clairement que le système SVM-UBM utilisant un modèle UBM à 2048 composantes surpasse celui basé sur le modèle à 128 composantes en termes de performance. En effet, un gain absolu de 5% peut être observé à l'EER. Comme dans un système GMM-UBM standard, le nombre de composantes est donc crucial et les taux d'erreurs diminuent quand ce nombre augmente. Cependant les expériences au delà de cette dimension n'ont pas apporté d'améliorations significatives, ce qui est vraisemblablement dû aux difficultés inhérentes à l'apprentissage de tels modèles.

8.3.3 Normalisation de la dynamique des exemples

Dans ce paragraphe, la technique de *normalisation par rang* (voir paragraphe 3.4.3) est appliquée aux vecteurs d'entrée du SVM. La méthode consiste à projeter les exemples sur une distribution uniforme afin d'obtenir des comportements dynamiques équivalents pour chaque dimension.

Cette normalisation est particulièrement utile lorsque aucune connaissance *a priori* n'est disponible. Dans chaque dimension, chaque valeur est remplacée par son rang. Celui-ci est calculé à partir du nombre d'instances négatives dont la valeur est inférieure à celle analysée. Ce rang est ensuite divisé par le nombre total d'instances négatives. Cette technique a été appliquée avec succès dans plusieurs systèmes comme dans [Shriberg et al., 2004] ou [Stolcke et al., 2005]. Le tableau 8.1 montre l'influence de cette normalisation sur le système SVM-UBM.

TAB. 8.1: *Normalisation par rang pour l'approche SVM-UBM utilisant différentes tailles de modèles UBM : 128 et 2048 sans T-normalisation .*

Taille du modèle/normalisation	DCFmin(x100)	EER(%)
128 / -	6.30	16.29
128 / rang	5.35	13.86
2048 / -	4.07	11.46
2048 / rang	5.22	15.04

Les résultats montrent clairement un gain significatif en performance pour le système SVM-UBM à 128 composantes lorsque la normalisation par rang est appliquée (15% relatif à l'EER et 5,5% relatif à la DCFmin), démontrant l'efficacité de la méthode. En revanche, une perte absolue de 3.8% est observée lorsque la normalisation par rang est appliquée au système UBM à 2048 composantes. Le prochain paragraphe tente d'expliquer cette perte surprenante.

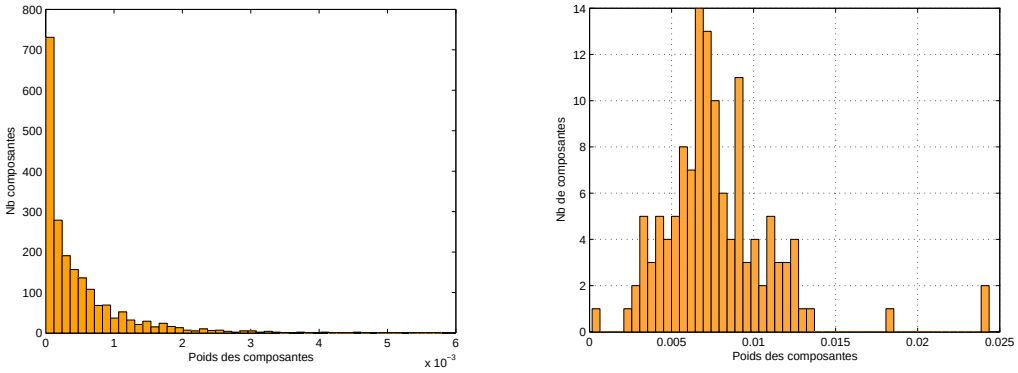
8.3.3.1 Réduction de la dimensionnalité du problème

Le classifieur SVM a une propriété intéressante, il sélectionne de façon optimale les exemples d'apprentissage qui maximisent la marge : les vecteurs supports. En revanche, il ne sélectionne pas les dimensions les plus intéressantes lorsqu'il est utilisé dans sa forme linéaire (où la matrice du noyau est constituée des produits scalaires de paires d'exemples).

La normalisation par rang a pour but d'homogénéiser les dynamiques de chaque dimension des vecteurs d'entrée, supposant ainsi que toutes les dimensions ont le même pouvoir discriminant. La perte importante en performance, subie au paragraphe précédant pour le modèle 2048, est reliée à cette propriété.

La figure 8.4 montre les distributions des poids de l'UBM permettant d'illustrer la dynamique très différente entre les deux tailles de modèles. En se référant aux expériences sur la normalisation par rang dans le tableau 8.1, il semble que pour une taille de 128, le pouvoir discriminant soit équitablement réparti entre les dimensions, ce qui n'est pas le cas pour un modèle de taille 2048. En parallèle, la figure 8.4(a) illustre une bonne répartition des poids de la mixture dans le modèle de taille 128, ce qui n'est pas le cas pour le modèle de taille 2048. Le pouvoir discriminant semble

donc être relié à cette répartition dans l'UBM. Nous proposons une méthode permettant de sélectionner les dimensions utiles en fonction de ce critère. Cette méthode vise à faire correspondre la dynamique d'un modèle 2048 à celle d'un modèle 128, puisque la normalisation par rang est performante pour ce dernier.



(a) (b)

FIG. 8.4: *Distribution des poids des gaussiennes pour différentes tailles de modèles de l'UBM : 2048 (a) et 128 (b)*

Un filtrage des gaussiennes non-nécessaires est proposé pour résoudre le problème. La figure 8.5 montre la distribution des poids du modèle résultant. Le tableau 8.2 montre les résultats suivant les différentes manières de filtrer les vecteurs d'entrée. La première enlève les 700 composantes de plus faible poids, la seconde retire de plus les 20 gaussiennes de plus fort poids. Cette sélection repose uniquement sur le modèle UBM et peut donc être faite *a priori*.

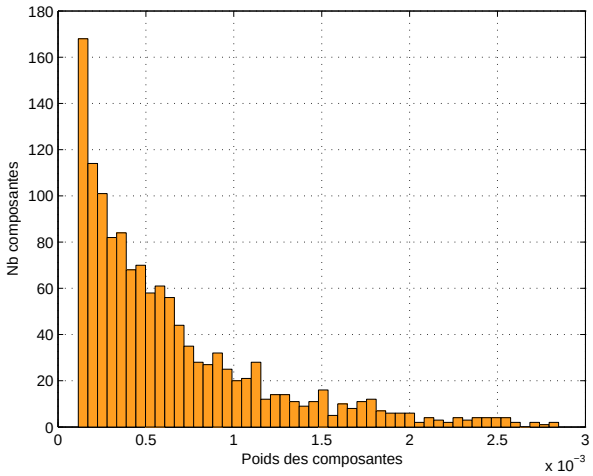


FIG. 8.5: *Distribution des poids des gaussiennes de l'UBM où les gaussiennes de faibles et forts poids ont été retirées (700 plus faibles, 20 plus fortes)*

TAB. 8.2: *Performance de l'approche SVM-UBM sans normalisation par rang. Les résultats varient peu en filtrant environ 30% des gaussiennes (pas de T-normalisation)*

Système SVM-UBM	DCFmin(x100)	EER (%)
Référence	4.07	11.46
700 gaussiennes de poids faible retirées	4.00	10.64
700 gaussiennes de poids faible et 20 de poids fort retirées	4.00	10.94

Ces résultats montrent clairement que la performance du système n'est pas altérée par la réduction de la taille du vecteur d'entrée, en utilisant le critère proposé. En effet, le retrait d'environ 30% des gaussiennes dans l'UBM 2048 n'affecte pas la performance globale du système, apportant même un faible gain. Ceci renforce l'idée que le pouvoir discriminant des dimensions réside dans le poids des gaussiennes du modèle générique UBM.

De ce fait, la normalisation par rang peut être utilisée à des fins de réduction de la dimensionnalité à l'entrée du classifieur SVM. Le tableau 8.3 présente l'impact de la normalisation par rang lorsque les vecteurs d'entrée sont filtrés.

TAB. 8.3: *Performance du système SVM-UBM normalisé par rang utilisant le modèle UBM filtré. Le système obtient des performances similaires au système de référence en comparaison à la perte importante obtenue sans filtrage (pas de Tnorm).*

Système SVM-UBM normalisé par rang	DCFmin (x100)	EER (%)
Référence	5.22	15.04
700 gaussiennes de poids faibles et 20 de poids forts retirées	4.19	11.26

Le système obtient des performances similaires au système de référence en comparaison à la perte importante obtenue sans filtrage. Il est décevant de remarquer qu'aucun gain en performance n'a été atteint en filtrant le problème, cependant ceci constitue une approche solide pour la réduction de la dimensionnalité du problème. Ainsi, pour la même performance, la dimension des exemples a été réduite aux environs de la moitié.

8.3.3.2 Normalisation par rang pour la performance du système

Nous avons montré au paragraphe précédent comment le filtrage des données d'entrée permet d'égaler le même comportement pour les modèles UBM de tailles différentes. Cependant, cette étape de normalisation permet aussi de capturer d'autres informations que celles du système non-normalisé.

A cet effet, nous présentons une expérience combinant deux systèmes SVM-UBM par une moyenne arithmétique non-pondérée. Le premier étant le système de référence, le second voit ses données normalisées par rang. Le modèle du monde utilisé est celui de *Sys06*, appris avec beaucoup plus de données. Il n'a pas été nécessaire de filtrer les gaussiennes pour ce modèle car la répartition des poids était déjà équilibrée. Les résultats sont illustrés à la figure 8.6).

Nota : La construction de ce nouveau modèle a eu peu d'influence sur les performances du système GMM-UBM. On remarque dans ce cas que l'augmentation de la quantité de données a

permis d'obtenir un modèle UBM avec une meilleure répartition des données (moins de gaussienne au poids presque nuls).

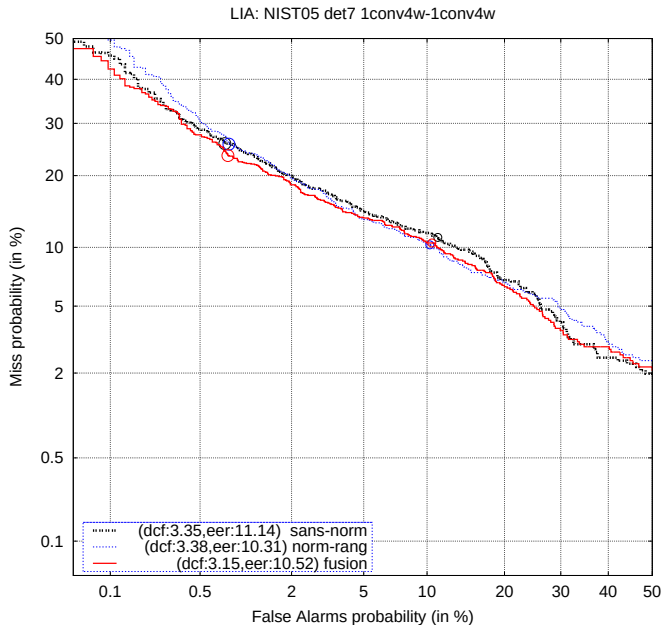


FIG. 8.6: *La normalisation par rang peut être utilisée pour améliorer les performances par la combinaison d'un système SVM-UBM de référence et d'un autre où les données ont été normalisées par rang. Configuration NIST05, det7 et système* Sys06

Il est intéressant de remarquer que la combinaison permet d'améliorer les performances du système, particulièrement à la DCFmin. L'information supplémentaire apportée par ce traitement provient probablement de l'ajout d'information inter-coefficients. En effet, la normalisation permet implicitement d'étudier la corrélation entre les exemples.

8.3.4 Influence de la T-normalisation

La *T-normalisation* 1.5.3 est largement utilisée dans les systèmes de VAL, nous examinons son influence sur les performances du système. Les locuteurs imposteurs sont les mêmes que ceux utilisés comme exemples négatifs, *i.e.* les locuteurs issus des données imposteurs. Pour certaines méthodes basées sur les SVM, la technique de *T-normalisation* est effectuée implicitement et n'apporte pas de gain. Dans notre cas, elle apporte deux avantages :

- un gain significatif, particulièrement à la DCFmin ;
- une projection des scores dans le même espace que ceux d'un autre système, facilitant le processus de fusion.

La figure **??** montre l'effet de la *T-normalisation* des scores sur les performances des systèmes.

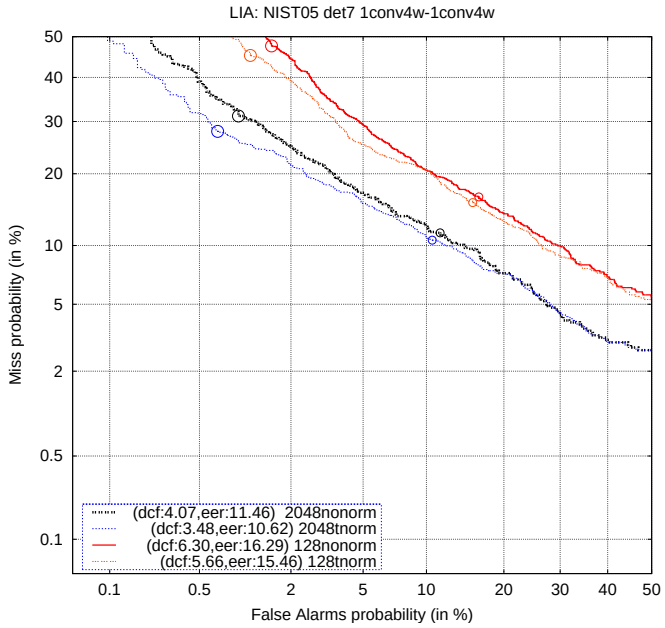


FIG. 8.7: *Effet de la T-normalisation sur les performances de l'approche SVM-UBM utilisant deux tailles différentes de modèles UBM : 2048 et 128. Courbe DET sans T-normalisation (large), avec (fines)*

8.3.5 Influence du feature mapping pour la normalisation de canal

Comme dans le cas du système AES, nous examinons l'effet du *feature mapping*, à des fins de normalisation de canal (paragraphe 2.4), sur la performance du système.

Nous remarquons, au tableau 8.4, qu'aucun gain n'est observé par l'application de cette technique. Une première justification peut être contenue dans les travaux de [Kenny et al., 2005b], décrits en 2.3.4.2. Le sous-espace contenant les variabilités dues au canal est contenu dans l'espace des paramètres de moyenne. Le système SVM-UBM étant fortement relié aux paramètres de poids des modèles, il semble que l'espace des paramètres de poids ne soit pas approprié pour la normalisation de canal.

TAB. 8.4: *Influence du feature mapping sur les performances du système de référence SVM-UBM. Configuration* Sys06.

Système	DCFmin(x100)	EER (%)
Référence	3.35	10.94
Feature mapping	3.35	11.14

8.3.6 Influence de la quantité de données

Le protocole NIST offre plusieurs durées d'apprentissage pour les locuteurs, nous présentons en figure 8.7 le comportement du système AES en utilisant trois fois plus de données, soit la tâche

3conv4w-1conv4w.

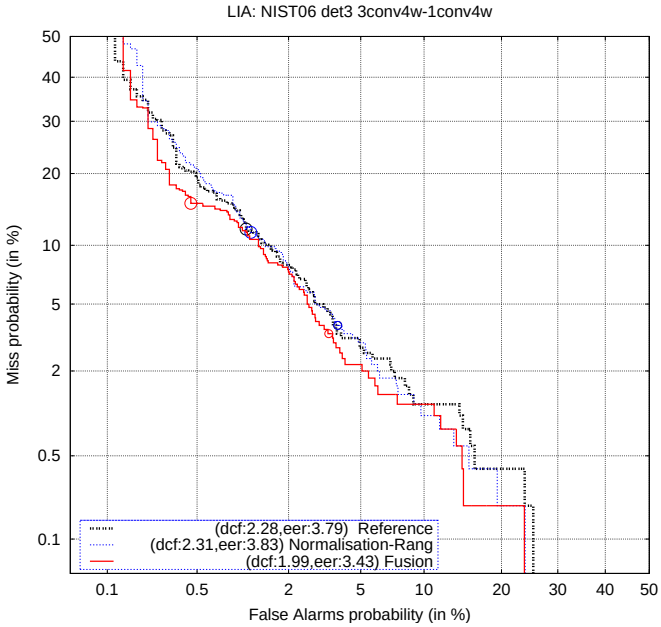


FIG. 8.8: *Performance du système SVM-UBM pour une quantité de données d'apprentissage trois fois supérieure. NIST06, det3, configuration* Sys06.

Cette expérience illustre le bon comportement du système SVM-UBM lorsque la quantité de données pour l'apprentissage est plus importante. En effet, le système arrive à tirer profit de l'augmentation de données à l'instar des systèmes cepstraux classiques. La gain (absolu) est environ de 7% à l'EER et de 1% à la DCFmin.

Nota : En général, il y a autant d'exemples positifs pour l'apprentissage du SVM qu'il y a de sessions d'apprentissage (ici 3 sessions). A des fins comparatives pour rester consistant avec le protocole utilisé par le GMM-UBM, les statistiques pour le locuteur ont été moyennées dans un seul vecteur.

8.3.7 Combinaison des scores avec un système GMM-UBM

En considérant le système de référence GMM-UBM présenté au chapitre 5, le système SVM-UBM a des performances similaires en termes de DCFmin, bien que sa complexité soit plus faible. Un gain significatif est attendu en combinant les deux, puisque ce système doit apporter l'information discriminante.

Deux combinaisons de systèmes sont présentées, consistant chacune en une moyenne arithmétique des scores des deux systèmes. La première est optimisée pour la DCFmin, la seconde est optimisée pour l'EER (ces paramètres ont été trouvés sur le jeu de développement). Le tableau 8.5 et la figure 8.8 résument les résultats et renseignent sur les poids alloués à la fusion.

TAB. 8.5: *Fusion arithmétique entre les systèmes SVM-UBM et GMM-UBM. Les poids de fusion sont indiqués en %.*

Système	DCFmin(x100)	EER(%)
1 :UBM-GMM	3.49	8.73
2 :UBM-SVM	3.48	10.62
Fusion 1 :50% / 2 :50%	3.06	8.41
Fusion 1 :70% / 2 :30%	3.17	8.20

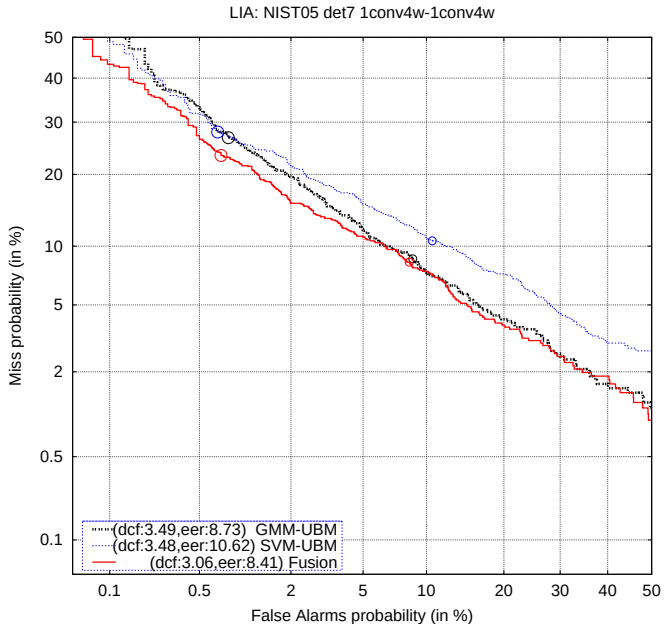


FIG. 8.9: *Les systèmes GMM-UBM et SVM-GMM de référence ainsi que leur combinaison pondérée. Cette expérience souligne le caractère complémentaire de l'information provenant du système SVM-UBM.*

Une fusion simple montre que la combinaison des deux systèmes apporte un gain significatif. Selon les poids de fusion choisis, le gain peut être observé à des points de fonctionnement différents. La fusion non-pondérée améliore la DCFmin par un gain relatif autour de 12%, l'autre combinaison améliore à la fois la DCFmin (gain de 9% relatif) et l'EER (gain de 6% relatif).

8.4 Discussion

Pour conclure sur notre approche, cette section souligne deux interprétations possibles du système SVM-UBM. La première est de remarquer que ce système est performant pour caractériser le locuteur par les vecteurs de poids des gaussiennes. La seconde est de valider notre stratégie de décodage du signal de parole pour le système AES en comparant le système SVM-UBM à un système AES où les modèles N-grammes sont d'ordre 1.

8.4.1 Analogie avec les systèmes basés sur les *super-vecteurs*

Les systèmes à base de « super-vecteurs » sont apparus récemment et présentent de bonnes performances. Un super-vecteur est défini comme un vecteur concaténant les paramètres d'un modèle statistique (voir paragraphe 2.3.4). Les plus utilisés sont les super-vecteurs de moyenne des gaussiennes d'un GMM. Ils sont utilisés pour la normalisation du canal (Factor Analysis [Kenny et al., 2005b]) ou dans les SVM comme fonction de projection (distance de Kullback-Leiber). Dans ce cadre, le système SVM-UBM peut être interprété comme un système SVM caractérisant le locuteur par les super-vecteurs des paramètres de poids, où :

- les vecteurs de poids sont appris par une itération de l'algorithme EM avec le modèle UBM comme initialisation ;
- ils sont ensuite normalisés par la probabilité *a priori* de leur gaussienne respective, soit le poids dans l'UBM originel.

Nous présentons une expérience comparant l'utilisation des paramètres poids dans le paradigme GMM-UBM à notre système SVM-UBM où ces paramètres sont utilisés dans un super-vecteur. Dans le cas d'un système GMM-UBM, l'adaptation des paramètres n'est appliquée que sur les moyennes et les systèmes travaillant sur les poids donnent de mauvais résultats. La figure 8.9 illustre cette faible performance en présentant les résultats d'un système GMM-UBM n'utilisant que les poids pour modéliser les locuteurs. Ces derniers sont estimés de deux façons :

- une estimation ML par une itération de l'algorithme EM : Courbe *Estimation-ml-poids* ;
- une estimation MAP, censée être plus robuste, en utilisant un *relevance factor* de 14 : Courbe *Adaptation-map-poids*.

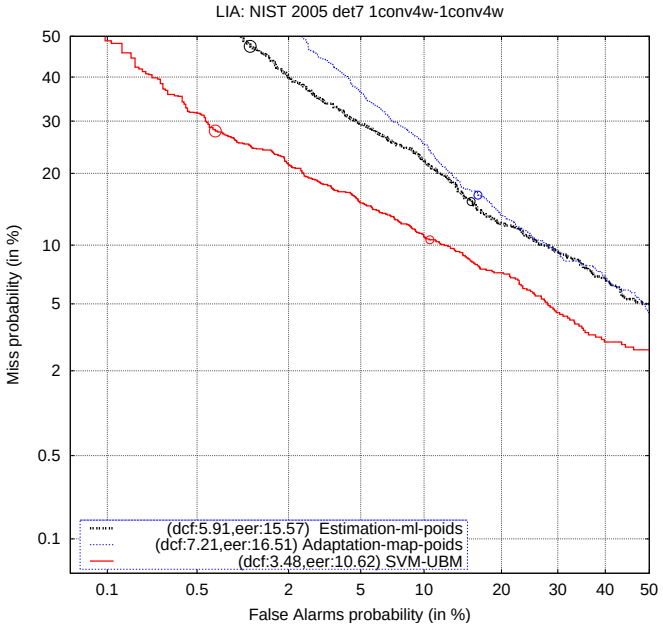


FIG. 8.10: *Comparaison du système GMM-UBM LIA_SpkDet en adaptant les poids des modèles de locuteur uniquement (par ML et MAP) et le système SVM-UBM utilisant les poids comme un super-vecteur.*

Les résultats sont en faveur du système SVM-UBM puisqu'un gain d'environ 6% à l'EER et de 3% à la DCFmin en absolu est observé. Ceci renforce l'idée que notre méthodologie permet de discriminer les locuteurs uniquement par le super-vecteur de poids, ce qui est plus difficile à mettre en oeuvre avec un système GMM-UBM classique.

8.4.2 Analogie entre les systèmes SVM-UBM et AES 1–gramme

Il existe un parallèle fort entre le système SVM-UBM et le système AES lorsque ce dernier utilise uniquement les 1–gramme. Dans ce cas, le système AES n'analyse plus la dynamique des événements mais juste leurs fréquences. Lorsque nous avons considéré l'unité k du noyau TFLLR comme la gaussienne du modèle générique, nous avons implicitement contraint la séquence du N-gramme à une longueur 1. Vu d'une autre manière, le système SVM-UBM correspond à un cas particulier du système AES.

Partant de cette constatation, nous pouvons émettre deux remarques importantes. La première concerne la taille des modèles utilisés. Dans le cas du système SVM-UBM, nous avons vu en 8.3.2, que la taille optimale était celle du modèle générique. Il semble donc que, pour une analyse non-dynamique des événements acoustiques, l'étape de réduction ne soit plus nécessaire. Alors que nous avons montré (paragraphe 6.3.1.3) qu'une réduction de la taille du dictionnaire permettait de capturer de l'information dynamique pour le système AES.

La seconde remarque nécessite la réalisation d'une expérience. Puisque l'AES et le SVM-UBM sont reliés, nous pouvons valider notre stratégie de décodage et de réduction de symbole. L'expérience consiste à comparer un système SVM-UBM et un système AES 1–gramme utilisant une taille de dictionnaire équivalente à celle de l'UBM utilisé pour le SVM-UBM. La figure 8.10 présente les résultats pour les tailles 128 et 2048.

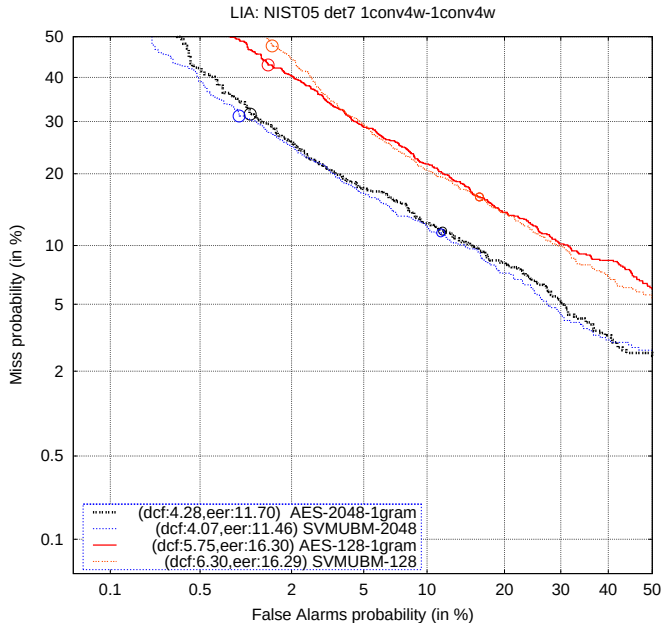


FIG. 8.11: *Système SVM-UBM utilisant un modèle de taille réduite (128) et un système AES 1–gramme utilisant un dictionnaire de taille équivalente. Les systèmes obtiennent les mêmes performances, ce qui valide notre procédé de réduction présenté au chapitre 6)*

Les performances des deux configurations étant similaires, cette expérience permet de valider notre stratégie de réduction de dimensionnalité et de décodage du signal de parole présentée au chapitre 6. En ce qui concerne le décodage du signal, il semble que l'hypothèse, consistant à décoder le signal de parole en utilisant la gaussienne la plus vraisemblable comme unité, soit validée. En effet, la probabilité *a posteriori* d'une composante peut être efficacement estimée par sa fréquence d'apparition comme gaussienne de plus forte vraisemblance, ou plus formellement pour une séquence de T trames :

$$p(W_k|\mathbf{X}) \simeq \frac{1}{T}\#[W_k = \arg\max_{W_l} p(W_l|\mathbf{X})], \tag{8.8}$$

avec $\#[W_k = \arg\max_{W_l} p(W_l|\mathbf{X})]$ représentant le nombre de fois où la gaussienne W_k a été la plus vraisemblable sur la séquence $\mathbf{X}$.

En ce qui concerne le regroupement par un critère de minimum de confusion des gaussiennes, celui-ci n'a pas amené de pertes d'informations puisque les systèmes utilisant une taille de modèle/dictionnaire de 128 présentent des performances similaires. Il est possible d'établir un parallèle entre cette remarque et les méthodes basées sur l'analyse dynamique d'unités phonétiques. Pour ces dernières, l'utilisation de treillis permet de réduire les taux d'erreurs, en gardant le maximum d'information avant de prendre une décision quant au phonème prononcé. Pour notre approche, la probabilité de la meilleur gaussienne étant très élevée, l'utilisation de treillis pour un décodage du signal ne sera que peu utile.

8.5 Conclusion

Le problème de projection d'une séquence de données de parole vers un vecteur de dimension fixe a été adressé précédemment dans la littérature, sous forme de méthodes complexes à mettre en oeuvre pour des quantités de données importantes. Le travail présenté ici propose cependant une méthodologie simple et performante pour la prise en compte des modèles génératifs et discriminants en utilisant le noyau TFLLR.

Il a été montré en section 8.2 que les vecteurs d'entrée calculés avec le noyau TFLLR sont étroitement reliés au le *mapping* de Fisher, lorsque ce dernier est réduit à la dérivation de la vraisemblance par rapport aux poids. Ce système est peu coûteux en terme de calcul si les statistiques nécessaires ont été calculés auparavant à l'aide d'un système GMM-UBM disponible dans le système global de vérification du locuteur.

L'originalité de cette approche est de démontrer que le modèle UBM à lui seul est suffisant pour calculer les statistiques nécessaires pour discriminer les locuteurs, renforçant ainsi l'idée que ce dernier joue un rôle structurant et que ce rôle est dominant. Alors que d'autres méthodes doivent construire des modèles spécifiques aux locuteurs (par le critère ML ou MAP), nous nous affranchissons de cette étape pour n'utiliser que le modèle générique.

Il a aussi été prouvé au paragraphe 8.3.3, que la normalisation par rang pouvait être utilisée pour réduire la dimensionnalité du problème mais aussi que cette normalisation apportait de l'information supplémentaire en combinant les deux systèmes. Nous montrons finalement, que notre approche SVM-UBM combinée avec un GMM-UBM classique peut apporter un gain relatif autour de 12% à la DCFmin en comparaison au GMM-UBM seul.

Nous soulignons dans la section 8.4 la bonne performance du système SVM-UBM, lorsque l'approche est assimilée à une caractérisation du locuteur par les vecteurs de poids des modèles. Ce point est particulièrement remarquable car cette stratégie donne de faibles performances dans le cas du paradigme GMM-UBM. Les dernières expériences permettent de souligner dans quelle mesure le système SVM-UBM est un cas particulier du système AES présenté au chapitre 6. Cette démonstration a permis de valider notre approche de décodage du signal et réduction de dimensionnalité utilisée pour l'AES.

Conclusion et Perspectives

La Vérification Automatique du Locuteur (VAL) consiste à confirmer ou infirmer l'identité proclamée d'un individu par sa voix. Les travaux présentés dans cette thèse s'inscrivent dans le cadre de cette tâche et sont orientés autour d'un axe principal : l'intégration du modèle générique utilisé dans la modélisation générative au sein des nouveaux formalismes apparus ces dernières années.

Parmi ceux-ci, deux catégories d'approches ont particulièrement retenu notre attention :

- la première regroupe les systèmes dits « haut-niveau » où la représentation du signal de parole s'appuie sur des caractéristiques autres que celles issues de l'enveloppe spectrale à court terme (en général, ces approches utilisent des informations plus proches de la linguistique que de l'acoustique) ;
- la seconde regroupe les systèmes basés sur une modélisation discriminante des locuteurs. Les systèmes de cette catégorie utilisent généralement le formalisme des machines à vecteurs supports (SVM).

Les systèmes récents de reconnaissance du locuteur issus des approches précédemment citées associent en général un reconnaisseur génératif de type GMM-UBM et plusieurs nouveaux classifieurs et/ou sources d'information.

Les contributions apportées dans ce document s'inscrivent dans cette démarche, mais en essayant d'unifier les différents formalismes et de simplifier la structure globale du système. Le principe fondamental à l'origine de cette thèse tient en une constatation simple : alors que le modèle générique (modèle du monde) joue un rôle primordial dans les systèmes génératifs (GMM-UBM), il n'est utilisé que par l'intermédiaire de ses données d'apprentissage dans le cadre des systèmes discriminants et, en général, pas du tout dans les systèmes « haut niveau ». En intégrant le modèle générique à ces nouveaux systèmes, nous tirons profit non seulement de la robustesse de son estimation — ce modèle est appris sur des corpus de très grande taille, à l'aide d'algorithmes spécifiquement optimisés — mais également des formalismes des nouveaux types de systèmes.

REPRÉSENTATION ET ANALYSE DU SIGNAL DE PAROLE PAR DES ÉVÉNEMENTS ACOUSTIQUES

La première partie de cette thèse a été consacrée à l'élaboration d'un système analysant la dynamique du signal de parole. Cette analyse utilise une représentation en unités acoustiques générées par le modèle générique (UBM).

Ce système appartient à la catégorie des systèmes dits « haut-niveaux » qui utilisent classiquement une segmentation basée sur des *a priori* linguistiques pour normaliser l'influence du texte. Le relâchement des contraintes linguistiques pour les approches haut-niveau a influencé nos travaux nous amenant à développer un système analysant la dynamique d'unités qui ne sont plus en rapport avec la structure de la langue. Dans ce cadre, nous avons proposé un système basé sur le modèle générique (UBM), pour générer la segmentation en unités. Une des originalités de notre approche se situe dans l'utilisation du même modèle du monde que celui employé dans les systèmes à base de GMM. De par la taille importante de ce modèle, nous avons proposé un algorithme de réduction efficace permettant de regrouper les gaussiennes afin de former des événements acoustiques. Ce système est appelé AES pour *Acoustic Event Sequences* et a été présenté au chapitre 6,.

La modélisation des locuteurs consiste à analyser la dynamique entre les unités issues de l'étape de segmentation. Cette étape est réalisée par une approche, dite « sac de N-grammes », couramment utilisée dans les systèmes d'identification de la langue, ainsi que dans les systèmes « haut-niveau » en VAL. De nombreuses expériences viennent valider notre approche. Nous avons montré comment une méthode reposant sur les SVM (noyau *TFLLR*) pouvait se montrer plus performante que les approches classiques fondées sur le calcul du rapport de vraisemblance. Nous montrons aussi que, pour une taille donnée, l'analyse de la dynamique apporte plus d'informations que l'unité seule. Ensuite, la destruction aléatoire de l'ordre temporel des trames permet de s'assurer que la dynamique des vecteurs d'entrée du système est bien prise en compte dans notre approche. Puis, nous montrons comment intégrer différentes longueurs de N-grammes dans le système. Enfin, la combinaison de l'AES avec un système GMM-UBM amène une amélioration notable des performances.

Nous proposons ensuite l'extension de ce système à une analyse multi-classes, C-AES, présentée au chapitre 7. Nous montrons qu'il est possible de tirer profit de plusieurs représentations du signal, chacune utilisant un jeu d'événements acoustiques différent. Le système C-AES repose sur la construction d'événements acoustiques à une plus faible résolution d'analyse (en terme spectrale et non temporelle), nommés *Class Events*. Nous avons défini une méthodologie permettant d'intégrer l'information de ces classes, en appliquant un système AES indépendant sur chacun des dictionnaires d'événements acoustiques correspondant. Nous avons pour cela proposé une modification du noyau TFLLR et montré que l'estimation de la probabilité des classes est importante. Le système C-AES, bien que présentant des performances similaires à l'AES, surpasse celui-ci lorsqu'il est combiné avec le GMM-UBM.

INTÉGRATION DU MODÈLE GÉNÉRIQUE DANS UN SYSTÈME DISCRIMINANT

A l'instar de l'intégration du modèle générique pour la construction d'un système « haut-niveau », la seconde partie de cette thèse a été consacrée à l'élaboration d'un système discriminant structuré par le modèle générique. Comme pour le système AES, le seul modèle génératif utilisé pour ce système est le modèle UBM (ce système, appelé SVM-UBM est présenté au chapitre 8).

Les techniques visant à combiner les méthodes génératives et discriminantes induisent des problèmes de très grande taille et difficiles à mettre en oeuvre. Notre contribution se distingue par la proposition d'un système exprimant le problème de vérification dans une faible dimension, ne nécessitant qu'un seul modèle génératif. Nous avons appliqué avec succès cette méthode, puisque les performances obtenues sont proches des systèmes GMM-UBM. En outre, lorsque le système SVM-UBM est combiné avec un système GMM-UBM, nous montrons que celui-ci apporte de l'information complémentaire. En effet, une réduction significative des taux d'erreurs est observée.

L'originalité principale de notre approche tient aux interprétations nouvelles qui naissent du formalisme proposé :

- la première est l'interprétation du noyau TFLLR lorsqu'il est utilisé pour un modèle génératif. Le noyau TFLLR est en effet une forme réduite et appropriée du noyau de *Fisher*, lorsqu'il utilise les statistiques des composantes de l'UBM. Le noyau de Fisher a déjà été appliqué en VAL, mais des problèmes de dimensionnalités limitent généralement son déploiement, pour des modèles génératifs de grande taille. La forme réduite que nous avons proposée présente de bonnes performances, associées à un coût calculatoire très réduit ;
- la seconde est d'interpréter le système SVM-UBM comme un système de VAL travaillant sur les super-vecteurs de poids des modèles génératifs. En effet, la forme réduite proposée du noyau de Fisher ne considère que les paramètres de poids (probabilités *a priori* et *a posteriori* des composantes de l'UBM). Alors que dans la modélisation générative, l'adaptation des poids des gaussiennes donnent de faibles résultats, le système proposé offre une méthodologie satisfaisante pour tirer profit de ces informations la plupart du temps délaissées.

PERSPECTIVES

Les perspectives de ce travail sont multiples et concernent chacune des contributions détaillées dans ce document.

Le système AES, en adoptant le formalisme N-gramme pour la modélisation des séquences, présente une limitation forte quant à la longueur maximale de la fenêtre d'analyse (due majoritairement aux problèmes de couverture). Deux approches peuvent faire l'objet de travaux futurs, toutes deux visent à augmenter la couverture entre les données d'apprentissage et de test.

La première est la modélisation d'événements acoustiques de longueurs variables. Dans le cadre du système AES, les événements acoustiques sont de longueur fixe, correspondant à la trame. Il serait intéressant de concevoir une méthodologie de construction automatique d'événements de longueur variable. Une des solutions possibles peut être trouvée dans des critères d'information mutuelle pour le regroupement des symboles du dictionnaire. Des techniques similaires sont utilisées en reconnaissance de la parole pour la construction optimale du lexique.

La seconde approche se concentre sur la structure des séquences d'événements. En effet, une structure sous jacente, guidée par le modèle du monde, contraint l'espace des séquences générées. Dans notre système, nous avons utilisé un dictionnaire de toutes les séquences possibles afin de faire émerger cette structure. La recherche d'une forme plus précise de cette structure permettrait de réduire la complexité du problème et de concentrer les efforts de modélisation sur les aspects spécifiques du locuteur. Dans [Navratil et al., 2003], une approche basée sur un formalisme

d'arbres binaires pour la prédiction des séquences et construite sur des critères entropique semble prometteuse, en autorisant des longueurs des séquences d'ordre supérieur.

Au delà de ces problèmes de couverture, une voie intéressante serait d'enrichir la modélisation par d'autres informations. Par exemple, à l'instar des techniques basée sur la divergence de *Kullback-Liebler* dans l'espace des paramètres, il serait possible de prendre en compte les distances entre les séquences, à travers des distances d'édition. Une autre solution pour rajouter de l'information peut être d'enrichir le modèle par l'utilisation d'une multitude de modèles génériques dont la nature reste à trouver (modèles de différentes langues par exemple).

Nous avons proposé dans le cadre du système C-AES, un formalisme permettant d'intégrer différents jeux d'événements acoustiques à différentes résolutions. Les prochaines expériences doivent consister à utiliser de multiples jeux à de multiples résolutions, puis à combiner ces informations dans un vecteur unique. Une perspective immédiate est de s'intéresser à l'analyse séquentielle des classes à basse résolution (Class Event) et de trouver une manière d'intégrer cette analyse dans le formalisme. D'autre part, le système C-AES pourrait lui aussi profiter d'un enrichissement du nombre de modèles génériques. Il serait par exemple intéressant d'effectuer une analyse croisée entre les dictionnaires d'événements des différents modèles (les classes proviennent d'un modèle, les événements acoustiques d'un autre).

Le système SVM-UBM pourrait lui aussi profiter de l'utilisation de multiples modèles génériques. Un parallèle intéressant pourrait être effectué avec les modèles d'ancrage dans le cas d'une cohorte de modèles génériques. La dimension de ce système est suffisamment faible pour qu'une augmentation de la complexité ne soit pas un problème majeur. Une autre perspective intéressante serait d'analyser la complémentarité entre l'espace des scores et celui des paramètres, en combinant par exemple ce système avec un système s'appuyant sur les super-vecteurs de moyenne.

Toutes les contributions apportées dans ce document peuvent aussi profiter des techniques de normalisation de canal (*Joint Factor analysis*, *Nuisance Attribute Projection*...) De futurs travaux de recherche doivent analyser la possibilité d'appliquer ces techniques aux manières de caractériser le locuteur dans ces systèmes. Nous avons montré que la technique du *feature mapping* se révélait sans effet pour nos méthodes, il serait nécessaire de vérifier si l'espace des scores est approprié pour représenter les variations du canal.

La dernière perspective consiste à étudier un modèle unifiant toutes les contributions présentées dans ce document. Nous avons déjà abordé ce point en soulignant que le système AES d'ordre 1 était une bonne approximation du système SVM-UBM. Le système AES, quant à lui, peut être relié au système C-AES, lorsqu'une seule classe est utilisée. Le formalisme du C-AES semble assez puissant pour constituer la base d'une approche unifiée. Il reste cependant à définir les différents jeux d'événements acoustiques associés aux différentes classes et utilisant des longueurs de N-gramme variables. Cette approche unifiant finalement tous les systèmes basés sur le rôle structurant du modèle générique constitue la perspective principale de ces travaux.

Troisième partie

Annexes

ANNEXE A

Description des protocoles et des systèmes utilisés

A.1 Protocoles utilisés

Les résultats des expériences réalisées au cours de ces travaux sont donnés sur les protocoles détaillés par la suite.

A.1.1 NIST04, det1, 1conv4w-1conv4w

Ce protocole est constitué d'un sous-ensemble de la base de donnée NIST-SRE de l'année 2004. Précisément, uniquement la partie Femme de la tâche requise *core test*, *1conv4w-1conv4w* (originellement nommée *1side-1side*) est considérée. Les expériences présentées comportent les tests de vérification défini par le NIST comme *all tests* (det 1) [NIST, 2004]. Il en résulte un protocole avec 370 locuteurs clients à référencer dans le système. L'expérience de vérification entière tient en 14373 tests (dont 1320 sont des tests clients).

A.1.2 NIST05, det7

A.1.2.1 1conv4w-1conv4w

Ce protocole est constitué d'un sous-ensemble de la base de donnée NIST-SRE de l'année 2005. Précisément, uniquement la partie Homme de la tâche requise *core test, 1conv4w-1conv4w*. Les expériences présentées comportent les tests de vérification défini par le NIST comme *primary test* (det 7) [NIST, 2005]. Il en résulte un protocole avec 274 locuteurs clients à référencer dans le système. L'expérience de vérification entière tient en 9012 tests (dont 951 sont des tests clients).

A.1.2.2 3conv4w-1conv4w

Certaines expériences sont présentées sur un sous-ensemble de la base de donnée NIST-SRE de l'année 2005 contenant trois fois plus de données d'apprentissage pour les locuteurs. Précisément, uniquement la partie Homme de la tâche *3conv4w-1conv4w*. Les expériences présentées comportent les tests de vérification défini par le NIST comme *primary test* (det 7). Il en résulte un protocole avec 221 locuteurs clients à référencer dans le système. L'expérience de vérification entière tient en 8309 tests (dont 829 sont des tests clients).

A.1.3 NIST06, det3

A.1.3.1 1conv4w-1conv4w

Ce protocole est constitué d'un sous-ensemble de la base de donnée NIST-SRE de l'année 2006. Précisément, uniquement la partie Homme de la tâche requise *core test, 1conv4w-1conv4w*. Les expériences présentées comportent les tests de vérification défini par le NIST comme *english trials* (det 3) [NIST, 2006]. Il en résulte un protocole avec 354 locuteurs clients à référencer dans le système. L'expérience de vérification entière tient en 9720 tests (dont 741 sont des tests clients).

A.1.3.2 3conv4w-1conv4w

Certaines expériences sont présentées sur un sous-ensemble de la base de donnée NIST-SRE de l'année 2006 contenant trois fois plus de données d'apprentissage pour les locuteurs. Précisément, uniquement la partie Homme de la tâche *3conv4w-1conv4w*. Les expériences présentées comportent les tests de vérification défini par le NIST comme *english trials* (det 3). Il en résulte un protocole avec 258 locuteurs clients à référencer dans le système. L'expérience de vérification entière tient en 5507 tests (dont 496 sont des tests clients).

A.2 Description des systèmes

A.2.1 Généralités

Ces coefficients sont obtenus de la façon suivante : un banc de filtre de 24 coefficients est calculé sur des fenêtres de Hamming de 20 ms toutes les 10ms. La bande passante est limitée à la bande téléphonique soit 300-3400Hz. Une transformation en cosinus discrète permet d'obtenir les coefficients cepstraux. Le toolkit utilisé est SPRO[1].

La sélection des trames de parole est basée sur l'analyse du coefficient d'énergie du signal. La distribution de ce coefficient est tout d'abord normalisé sur une distribution de moyenne 0 et de variance unité. Un GMM à 3 composantes est ensuite appris visant à sélectionner les trames informatives.

[1] développé par Guillaume Gravier, http://gforge.inria.fr/projects/spro

L'apprentissage du modèle du monde est effectué grâce à l'algorithme EM du toolkit ALIZE/LIA_SpkDet. Le modèle de mixture du monde est composée de 2048 gaussiennes. La borne inférieure des valeurs de variance est 0.5 (*variance flooring c.f.* chapitre 2). Les modèles de locuteurs sont dérivés par adaptation bayésienne MAP, en appliquant la transformation aux moyennes de l'UBM contraint par un *relevance factor* d'une valeur de 14 (une seule itération de l'algorithme EM est utilisé).

Un processus de normalisation des moyennes et des variances du modèle est appliquée pour tous les modèles GMM après les itérations EM ou MAP. Cette normalisation est une transformation des paramètres pour que la moyenne et la variance globale du modèle soit exactement de moyenne nulle et de variance unité.

A.2.2 *Sys04*

Pour le système *Sys04*, le signal est paramétrisé par 32 coefficients incluant 16 coefficients LFCC et leur dérivées premières (générés par SPROv3). Un processus de normalisation des trames est appliquée afin que la distribution de chaque coefficient cepstral soit de moyenne 0 et de variance 1, où les paramètres de moyenne et de variance ont été estimées fichier par fichier.

La base de pseudo-imposteurs Imp_{NIST} est composée de :
- Les données « landline » uniquement du corpus NIST SRE 1999.
- Les données « cellular, GSM, CDMA » uniquement du corpus NIST SRE 2001.
- La même composition que le point précédent pour le corpus NIST SRE 2002.

Les données pour le modèle du monde sont celles des corpus NIST-SRE de 1999 à 2002 où 1.3 millions de trames de paroles a été utilisé (3.5 heures). Les modèles du monde sont dépendant du genre est sont appris sur des données équilibrée entre les types de canaux téléphoniques « landline, GSM et CDMA ». Les moyennes du modèle sont initialisés en utilisant 0.004% des trames, sélectionnées aléatoirement. Ensuite, 24 itérations de l'algorithme EM sont effectuées sur 10% des trames, sélectionnées aléatoirement à chaque itération. Enfin, deux itérations de l'algorithme sont appliquées sur toutes les données d'apprentissage.

La *T-normalisation* est effectuée en utilisant 80 modèles imposteurs dépendant du genre, composé de 27 segments CDMA (2002), 26 segments GSM (2001), et 27 segments landline (1999). Une *Z-normalisation* est ensuite appliquer en utilisant des locuteurs selon la même proportion.

Les exemples étiquetés négatifs pour les systèmes basés sur un SVM et utilisant cette configuration, sont formés de 200 locuteurs dont les données ont été divisés en 5 sous-parties formant ainsi 1000 exemples négatifs.

A.2.3 *Sys05*

La paramétrisation et le modèle du monde sont les mêmes que pour *Sys04*. Le toolkit utilisé est en revanche SPRO version 4. Un processus de normalisation des trames est appliquée afin que la distribution de chaque coefficient cepstral soit de moyenne 0 et de variance 1, où les paramètres de moyenne et de variance ont été estimées fichier par fichier. La détection d'énergie est plus

sélective. Une fois les segments de parole dans le signal sélectionné, un processus de raffinement de la segmentation est appliqué dans lequel :

- les recouvrements entre les segments de parole des deux côtés de la conversation sont retirés ;
- des règles morphologiques, consistant à ajouter ou à retirer des trames d'un segment, sont appliquées pour éviter les segments trop court.

La base de données de développement utilisée est la même que celles de *Sys04*. Une *T-normalisation* dépendant du genre est effectuée en prenant des locuteurs de la population des données du monde (81 locuteurs équitablement répartis entre les types de téléphone). 80 autres locuteurs extraits du corpus NIST-SRE-04 sont ajoutés.

Les modèles imposteurs pour la *T-normalisation* sont des locuteurs provenant des données d'apprentissage du modèle du monde et des locuteurs pris sur NIST-SRE-2004, soit 161 locuteurs en tout.

Les exemples étiquetés négatifs pour les systèmes basés sur un SVM et utilisant cette configuration, sont les données des modèles de la *T-normalisation* , soit 161 exemples par genre.

A.2.4 *Sys06*

Les trames sont composées de 19 coefficients LFCC, de leurs dérivées premières, des 11 premières dérivées secondes, et du log-énergie (50 coefficients). La technique du *feature mapping* est utilisée et la méthodologie d'implémentation de la méthode a été la suivante :

- Construction d'un modèle du monde générique sur le corpus *Fisher* ;
- Adaptation des modèles dépendant du canal avec des données étiquetées. Nous avons utilisés trois étiquettes : landline, cordless, cellular. Les moyennes et les variances des modèles ont été adaptés par MAP (facteur de régulation à 14) ;
- Association des données du corpus NIST-SRE aux modèles dépendants du canal par un calcul de vraisemblance ;
- Pour chaque trame, application de la projection définie à l'équation 2.13 du modèle indépendant du canal vers le modèle dépendant du canal.

L'apprentissage du modèle du monde s'effectue sur le corpus *Fisher* du LDC [2] en utilisant 735 conversations de locuteurs homme différents (5 minutes de parole en moyenne par conversation, soit environ 10 millions de trames). Une proportion égal entre les canaux *cordless, cellular et landline* a été respecté pour la sélection de ces locuteurs.

Les locuteurs formant la cohorte imposteur pour la *T-normalisation* sont au nombre de 180 et respectent aussi cette proportion.

Les exemples étiquetés négatifs pour les systèmes basés sur un SVM et utilisant cette configuration, sont les données des modèles de la *T-normalisation* , soit 180 exemples par genre.

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

BIBLIOGRAPHIE

[Adami et al., 2003] A. Adami, R. Mihaescu, Reynolds, D. A., et J. Godfrey, 2003. Modeling prosodic dynamics for speaker recognition. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China*, pp. 788–791. 17, 57

[Andrews et al., 2001] W. D. Andrews, M. A. Kohler, J. P. Campbell, et J. J. Godfrey, 2001. Phonetic, idiolectal, and acoustic speaker recognition. Dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Greece*. 55, 57, 83, 98

[Atal, 1976] B. S. Atal, 1976. Automatic recognition of speakers from their voices. Dans *IEEE transactions*, Volume 644, pp. 460–475. 8, 17

[Auckenthaler et al., 2000] R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. Score normalization for text-independent speaker verification system. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop* *10*(1-3), 42–54. 13, 19, 33

[Baker et al., 2004] B. Baker, R. Vogt, M. Mason, et S. Sridharan, 2004. Improved phonetic and lexical speaker recognition through map adaptation. Dans *Odyssey'04, the Speaker Recognition Workshop, Toledo, Spain*. 60

[Baker et al., 2005] B. Baker, R. Vogt, et S. Sridharan, 2005. Gaussian mixture modelling of broad phonetic and syllabic events for textindependent speaker verification. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa, Portugal*, pp. 2429–2432. 55

[Ben et Bimbot, 2003] M. Ben et F. Bimbot, 2003. D-MAP : a distance-normalized MAP estimation of speaker models for automatic speaker verification. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China*, Volume 2. 28

[Ben et al., 2002] M. Ben, R. Blouet, et F. Bimbot, 2002. A monte-carlo method for score nor-

malization in automatic speaker verification using kullback-leiber distances. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*. 28

[Bengio et Mariethoz, 2004] S. Bengio et J. Mariethoz, 2004. The Expected Performance Curve : a New Assessment Measure for Person Authentication. Dans *Odyssey'04, the Speaker Recognition Workshop, Toledo, Spain*, pp. 279–284. 13

[Bengio et al., 2001] S. Bengio, J. Mariethoz, et M. IDIAP, 2001. Learning the decision function for speaker verification. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2001)*, Volume 1. 18

[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, Special issue on biometric signal processing*. 8, 15

[Boakye et Peskin, 2004] K. Boakye et B. Peskin, 2004. Text-constrained speaker recognition on a text-independent task. Dans *Odyssey'04, the Speaker Recognition Workshop, Toledo, Spain*. 56

[Bonastre et al., 2003] J. Bonastre, F. Bimbot, L. Boë, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau, 2003. Person authentication by voice : a need for caution. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland*, pp. 33–36. 7

[Bonastre et al., 2004] J. Bonastre, N. Scheffer, C. Fredouille, et D. Matrouf, 2004. NIST04 speaker recognition evaluation campaign : new LIA speaker detection platform based on ALIZE toolkit. Dans *Proceedings of NIST speaker recognition workshop*.

[Bonastre et al., 2000] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, et C. J. Wellekens, 2000. A speaker tracking system based on speaker turn detection for NIST evaluations. Dans *International Conference on Acoustics, Speech, and Signal Processing ICASSP*. 9

[Bonastre et al., 2005] J.-F. Bonastre, F. Wils, et S. Meignier, 2005. Alize, a free toolkit for speaker recognition. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Philadelphia, USA. 68

[Brümmer et du Preez, 2006] N. Brümmer et J. du Preez, 2006. Application-independent evaluation of speaker detection. *Computer Speech and Language* *20*(2-3), 230–275. 14

[Burges, 1998] C. Burges, 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* *2*(2), 121–167. 39, 41

[Campbell et al., 2003] J. P. Campbell, D. A. Reynolds, et R. B. Dunn, 2003. Fusing high- and low-level features for speaker recognition. Dans *EUROSPEECH Conference, Geneva, Switzerland*, pp. 2665–2668. 54, 57

[Campbell et al., 1999] W. Campbell, K. Assaleh, M. SSG, et A. Scottsdale, 1999. Polynomial classifier techniques for speaker verification. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 99)*, Volume 1. 43

[Campbell et al., 2004a] W. Campbell, J. Campbell, D. Reynolds, D. Jones, et T. Leek, 2004a. Phonetic speaker recognition with support vector machines. *Advances in Neural Information Processing Systems 16*. 57

[Campbell et al., 2006] W. Campbell, J. Campbell, D. Reynolds, E. Singer, et P. Torres-Carrasquillo, 2006. Support vector machines for speaker and language recognition. *Computer Speech and Language 20*(2-3), 210–229. 43

[Campbell et al., 2006] W. Campbell, D. Sturim, et D. Reynolds, 2006. Support Vector Machines Using GMM Supervectors for Speaker Verification. *Signal Processing Letters, IEEE 13*(5), 308–311. 44

[Campbell, 2006] W. M. Campbell, 2006. Compensating for mismatch in high-level speaker recognition. Dans *Odyssey'06, the Speaker Recognition Workshop, San Juan, Puerto Rico*, San Juan, Puerto Rico. 110

[Campbell et al., 2004b] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, et T. R. Leek, 2004b. High-level speaker verification with support vector machines. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, CANADA*, pp. 73–76. 85

[Carey et Parris, 1992] M. J. Carey et E. S. Parris, 1992. Speaker verification using connected words. Dans *Proceedings of Institute of Acoustics*, Volume 146, pp. 95–100. 26

[Charlet, 1997] D. Charlet, 1997. *Authentification vocale par téléphone en mode dépendant du texte*. Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications ENST. 54

[Chollet et al., 1999] G. Chollet, J. Cernockỳ, A. Constantinescu, S. Deligne, et F. Bimbot, 1999. Computational models of speech pattern processing, chapter Towards ALISP : a proposal for Automatic Language Independent Speech Processing. *NATO ASI Series. Springer Verlag*, 375–388. 56

[Chollet et al., 1997] G. Chollet, J.-L. Cochard, A. Constantinescu, C. Jaroulet, et P. Langlais, 1997. Swiss French PolyPhone and PolyVar : Telephone speech databases to model inter- and intra-speaker variability. Dans J. Nerbonne (Ed.), *Linguistic Databases*, pp. 117–135. Stanford, California : CSLI Publications. 15

[Collet, 2006] M. Collet, 2006. *Mesures de similarité robustes dans un espace de locuteurs d'ancrage. Application pour l'indexation de documents audio*. Thèse de Doctorat, Université de Rennes 1. 29, 46

[Collobert et Bengio, 2001] R. Collobert et S. Bengio, 2001. SVMTorch : Support vector machines for large-scale regression problems. *Journal of Machine Learning Research 1*, 143–160. 86

[Cristianini et Shawe-Taylor, 2000] N. Cristianini et J. Shawe-Taylor, 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. 38

[Delacourt, 2000] P. Delacourt, 2000. *La segmentation et le regroupement par locuteurs pour l'indexation de document audio*. Thèse de Doctorat, ENST-Eurecom. 9

[Dempster et al., 1977] A. P. Dempster, N. M. Laird, et D. B. Rubin, 1977. Maximum-likelihood from incomplete data via the EM algorithm. Dans *Journal of Acoustical Society of America JASA*, Volume 39, pp. 1–38. 25

[Doddington, 2001] G. Doddington, 2001. Speaker recognition based on idiolectal differences between speakers. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark*, pp. 2521–2524. 10, 58, 59, 83

[Doddington, 1985] G. R. Doddington, 1985. Speaker recognition. identifying people by their voices. Dans *IEEE transactions*, Volume 7311, pp. 1651–1664. 8

[El Hannani et Petrovska-Delacretaz, 2005] A. El Hannani et D. Petrovska-Delacretaz, 2005. Exploiting High-Level Information Provided by ALISP in Speaker Recognition. *LECTURE NOTES IN COMPUTER SCIENCE 3817*, 66. 58

[Farinas, 2002] J. Farinas, 2002. *Une modélisation automatique du rythme pour l'identification des langues*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France. 55

[Ferrer et al., 2003] L. Ferrer, H. Bratt, V. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, et A. Venkataraman, 2003. Modeling duration patterns for speaker recognition. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland*, pp. 2017–2020. 53

[Fine et al., 2001] S. Fine, J. Navratil, et R. A. Gopinath, 2001. A hybrid gmm/svm approach to speaker identification. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2001)*, pp. 417–420. 45

[Fredouille et al., 1999] C. Fredouille, J.-F. Bonastre, et T. Merlin, 1999. Similarity normalization method based on world model and a posteriori probability for speaker verification. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99)*, Volume 2, pp. 983–986. 19

[Furui, 1981] S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. Dans *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume 292, pp. 254–272. 15, 17

[Gauvain et al., 2004] J. Gauvain, A. Messaoudi, et H. Schwenk, 2004. Language recognition using phone lattices. Dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2004)*. 54

[Gauvain et Lee, 1994] J. L. Gauvain et C. H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. Dans *IEEE Transactions on Speech and Audio Processing*, Volume 22, pp. 291–298. 28

[Grenier, 1977] Y. Grenier, 1977. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique*. Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications ENST. 16

[Hatch et al., 2005] A. Hatch, B. Peskin, et A. Stolcke, 2005. Improved Phonetic Speaker Recognition Using Lattice Decoding. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Volume 1. 54, 57

[Hatch et al., 2006] A. O. Hatch, S. Kajarekar, et A. Stolcke, 2006. Within-class covariance normalization for svm-based speaker recognition. Dans *Proceedings of Interspeech, International Conference on Spoken Language Processing (ICSLP 2006), Pittsburgh, USA*. 31

[Hermansky, 1990] H. Hermansky, 1990. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America 87*, 1738. 15, 17

[Higgins et al., 1991] A. L. Higgins, L. Bahler, et J. Porter, 1991. Speaker verification using randomized phrase prompting. Dans *Digital Signal Processing*, Volume 1, pp. 89–106. 26

[Homayounpour, 1995] M. M. Homayounpour, 1995. *Vérification vocale d'identité : dépendante et indépendante du texte*. Thèse de Doctorat, Université de Paris-Sud centre d'Orsay. 9

[Jaakkola et Haussler, 1998] T. Jaakkola et D. Haussler, 1998. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems 11*, 487–493. 45, 46, 109

[Jin et al., 2003] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, et J. Abramson, 2003. Combining cross-stream and time dimensions in phonetic speaker recognition. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China*, Volume 4. 58

[Joachims, 1999] T. Joachims, 1999. Making large-scale svm learning. Dans *Practical. Advances in Kernel Methods - Support Vector Learning,B. Schokopf and C. Burges and A. Smola, MIT Press*. 86, 101, 111

[Johnson, 1999] S. E. Johnson, 1999. Who spoke when ? - automatic segmentation and clustering for determining speaker turns. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99)*. 9

[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. 30

[Kenny et al., 2005b] P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel, 2005b. Factor Analysis Simplified. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Volume 1. 31

[Kenny et Dumouchel, 2004a] P. Kenny et P. Dumouchel, 2004a. Disentangling speaker and channel effects in speaker verification. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, CANADA*, Volume 1. 30

[Kenny et Dumouchel, 2004b] P. Kenny et P. Dumouchel, 2004b. Experiments in speaker verification using factor analysis likelihood ratios. Dans *Odyssey'04, the Speaker Recognition Workshop, Toledo, Spain*. 32

[Kenny et al., 2006] P. Kenny, V. Gupta, et G. Boulianne, 2006. Feature normalization using smoothed mixture transformations. Dans *Proceedings of Interspeech, International Conference on Spoken Language Processing (ICSLP 2006), Pittsburgh, USA*. 32

[Kuhn et al., 2000] R. Kuhn, J. Junqua, P. Nguyen, et N. Niedzielski, 2000. Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Transactions on Speech and Audio Processing* *8*(6), 695. 29, 30

[Li et Porter, 1988] K. P. Li et J. E. Porter, 1988. Normalizations and selection of speech segments for speaker recognition scoring. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 98)*, pp. 595–598. 18

[Louradour et Daoudi, 2005] J. Louradour et K. Daoudi, 2005. Svm speaker verification using a new sequence kernel. Dans *European Signal Processing Conference*. 44

[Louradour et al., 2006] J. Louradour, K. Daoudi, et F. Bach, 2006. Svm speaker verification using an incomplete cholesky decomposition sequence kernel. Dans *Odyssey'06, the Speaker Recognition Workshop, San Juan, Puerto Rico*, San Juan, Puerto Rico. 44

[Mami, 2003] Y. Mami, 2003. *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence*. Thèse de Doctorat, ENST Paris. 29

[Mami et Charlet, 2006] Y. Mami et D. Charlet, 2006. Speaker recognition by location in the space of reference speakers. *Speech communication* *48*(2), 127–141. 29

[Mariethoz, 2006] J. Mariethoz, 2006. *Algorithmes d'apprentissage discriminants en vérification du locuteur*. Thèse de Doctorat, Université Lumière - Lyon 2. 33

[Martin et Przybocki, 1997] A. F. Martin et M. A. Przybocki, 1997. The DET curve in assessment of detection task performance. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 97)*, pp. 1895–1898. 13, 14

[Martin et al., 2006] T. Martin, B. Baker, E. Wong, et S. Sridharan, 2006. A syllable-scale framework for language identification. *Computer Speech and Language* *20*(2-3), 276–302. 53, 55

[Merlin et al., 1999] T. Merlin, J. Bonastre, et C. Fredouille, 1999. Non directly acoustic process for costless speaker recognition and indexation. Dans *International Workshop on Intelligent Communication Technologies and Applications*. 29

[Moreno et al., 2004] P. Moreno, P. Ho, et N. Vasconcelos, 2004. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *Advances in Neural Information Processing Systems 16*.

[Navratil et al., 2003] J. Navratil, Q. Jin, W. Andrews, et J. Campbell, 2003. Phonetic speaker recognition using maximum-likelihood binary-decision tree models. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China*, Volume 4. 127

[NIST, 2004] NIST, 2004. The NIST year 2005 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf. 131

[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. 131

[NIST, 2006] NIST, 2006. The NIST year 2006 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf. 69, 132

[Oglesby, 1995] J. Oglesby, 1995. What's in a number ? : moving beyond the equal error rate. Dans *Speech Communication*, Volume 171-2, pp. 193–209. 13

[Openshaw et Mason, 1994] J. Openshaw et J. Mason, 1994. On the limitations of cepstral features in noise. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 94)*, Volume 2. 17

[Oppenheim et Schafer, 1989] A. Oppenheim et R. Schafer, 1989. *Discrete-time signal processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA. 16

[O'Shaughnessy, 1986] D. O'Shaughnessy, 1986. Speaker recognition. Dans *IEEE Transactions Acoustics, Speech, and Signal Processing ASSP*, pp. 4–17. 8

[Pelecanos et Sridharan, 2001] J. Pelecanos et S. Sridharan, 2001. Feature warping for robust speaker verification. Dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop, Chania, Greece*, Chania, Crete, pp. 213–218. 18

[Platt, 1999] J. Platt, 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods : support vector learning table of contents*, 185–208. 39

[Przybocki et Martin, 1998] M. A. Przybocki et A. F. Martin, 1998. NIST speaker recognition evaluation - 97. Dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications RLA2C*, pp. 120–123. 14

[Quan et Bengio, 2002] L. Quan et S. Bengio, 2002. Hybrid generative and discriminative models for speech and speaker recognition. Dans *Tech. Rep. IDIAP-RR 02-06, IDIAP*. 45

[Ramaswamy et al., 2003] G. Ramaswamy, R. Zilca, et O. Alecksandrovich, 2003. A programmable policy manager for conversational biometrics. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland*. 10

[Reynolds, 2003] D. Reynolds, 2003. Channel robust speaker verification via feature mapping. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China*, Volume 2. 31

[Reynolds et al., 2002] D. Reynolds, B. Peskin, J. Navratil, J. Campbell, W. Andrews, D. Klusacek, A. Adami, Q. Jin, J. Abramson, et R. Mihaescu, 2002. SuperSID : Exploiting high-level information for high-performance speaker recognition. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*, The Center for Language and Speech Processing, The Johns Hopkins University. http://www.clsp.jhu.edu/ws2002/groups/supersid/. 52

[Reynolds, 1994] D. A. Reynolds, 1994. Experimental evaluation of features for robust speaker identification. Dans *IEEE transactions Speech Audio Processing*, Volume 2, pp. 639–643. 15

[Reynolds, 1995] D. A. Reynolds, 1995. Speaker identification and verification using gaussian mixture speaker models. Dans *Speech Communication*, Volume 171-2, pp. 91–108. 23, 26

[Reynolds, 1997] D. A. Reynolds, 1997. Comparison of background normalization methods for

text-independent speaker verification. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 97)*. 19

[Rosenberg et al., 1992] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, et F. K. Soong, 1992. The use of cohort normalized scores for speaker verification. Dans *International Conference on Spoken Language Processing ICSLP*, pp. 599–602. 26

[Rosenberg et Soong, 1991] A. E. Rosenberg et F. K. Soong, 1991. Recent research in automatic speaker recognition. *Advances in speech signal processing*, 701–737. 8

[Rossi, 1989] M. Rossi, 1989. De la quiddité des variables. *Actes du séminaire Variabilité et spécificité du locuteur : Etudes et Applications*, 11–31. 9

[Salton et al., 1975] G. Salton, A. Wong, et C. S. Yang, 1975. A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

[Sankar et Lee, 1996] A. Sankar et C. Lee, 1996. A maximum-likelihood approach to stochastic matching for robustspeech recognition. *Speech and Audio Processing, IEEE Transactions on 4*(3), 190–202. 31

[Scheffer et Bonastre, 2005] N. Scheffer et J.-F. Bonastre, 2005. Speaker detection using acoustic event sequences. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa, Portugal*. 2, 80

[Scheffer et Bonastre, 2006a] N. Scheffer et J.-F. Bonastre, 2006a. A multi-class framework for Speaker Verification within an Acoustic Event Sequence system. Dans *Proceedings of Interspeech, International Conference on Spoken Language Processing (ICSLP 2006), Pittsburgh, USA*. 2, 97

[Scheffer et Bonastre, 2006b] N. Scheffer et J.-F. Bonastre, 2006b. {UBM}-driven discriminative approach for speaker verification. Dans *Odyssey'06, the Speaker Recognition Workshop, San Juan, Puerto Rico*, San Juan, Puerto Rico. *IBM best paper sudent award*. 3

[Schölkopf et Smola, 2002] B. Schölkopf et A. Smola, 2002. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press. 44, 45

[Shriberg et al., 2004] E. Shriberg, L. Ferrer, A. Venkataraman, et K. S., 2004. Svm modeling of snerf-grams for speaker recognition. Dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2004)*. 48, 56, 112

[Smith et al., 2001] N. Smith, M. Gales, et M. Niranjan, 2001. Data-dependent kernels in svm classification of speech patterns. Dans *Tech. Rep. CUED/F-INFENG/TR.387, Cambridge University Engineering Dept.* 45, 109

[Solomonoff et al., 2005] A. Solomonoff, W. Campbell, et I. Boardman, 2005. Advances in channel compensation for svm speaker recognition. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA*, Volume 1. 31

[Sonmez et al., 1999] K. Sonmez, L. P. Heck, et M. Weintraub, 1999. Speaker tracking and detection with multiple speakers. Dans *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99)*. 9

[Stolcke et al., 2005] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, et V. A., 2005. Mllr transforms as features in speaker recognition. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa, Portugal*. 56, 98, 112

[Sturim et al., 2001] D. Sturim, D. Reynolds, E. Singer, et J. Campbell, 2001. Speaker indexing in large audio databases using anchor models. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2001)*, Volume 1. 29

[Sturim et al., 2002] D. E. Sturim, D. A. Reynolds, R. B. Dunn, et T. F. Quatieri, 2002. Speaker Verification using Text-Constrained Gaussian Mixture Models. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*, pp. 677–680. 56

[Thyes et al., 2000] O. Thyes, R. Kuhn, P. Nguyen, et J. Junqua, 2000. Speaker identification and verification using eigenvoices. Dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2000)*.

[Torres-Carrasquillo et al., 2002] P. Torres-Carrasquillo, D. Reynolds, et J. Deller Jr, 2002. Language identification using Gaussian mixture model tokenization. Dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*, Volume 1. 58, 79

[Turk et Pentland, 1991] M. Turk et A. Pentland, 1991. *Eigenfaces for Recognition*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology. 30

[Vair et al., 2006] C. Vair, D. Colibro, et P. Laface, 2006. Channel factors compensation in model and feature domain for speaker recognition. Dans *Odyssey'06, the Speaker Recognition Workshop, San Juan, Puerto Rico*. 32

[Vapnik, 1998] V. N. Vapnik, 1998. *Statistical Learning Theory*. Wiley. 42, 108

[Vogt et al., 2005] R. Vogt, B. Baker, et S. Sridharan, 2005. Modelling Session Variability in Text-Independent Speaker Verification. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa, Portugal*. 31

[Wan, 2003] V. Wan, 2003. *Speaker Verification Using Support Vector Machines*. Thèse de Doctorat, University of Sheffield. 45, 46

[Wan et Campbell, 2000] V. Wan et W. M. Campbell, 2000. Support vector machines for speaker verification and identification. Dans *Neural Network for Signal Processing*, pp. 775–784.

[Wan et Renals, 2002] V. Wan et S. Renals, 2002. Speaker verification using sequence discriminant support vector machines. Dans *IEEE Transactions on Speech and Audio Processing*. 45, 109

[Zavaliagkos, 1995] G. Zavaliagkos, 1995. *Maximum A Posteriori Adaptation Techniques For Speech Recognition*. Thèse de Doctorat, Ph. D. Thesis, Northeastern University.

[Zissman, 1996] M. Zissman, 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing* *4*(1). 53, 57