UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
MINISTÈRE DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

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

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
en collaboration avec Swansea University
pour l'obtention du grade de Docteur

SPÉCIALITÉ : Informatique

École Doctorale 166 « Information Structures Systèmes »
Laboratoire d'Informatique (EA 4128)

Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée.

par
Anthony LARCHER

Soutenue publiquement le 24 septembre 2009 devant un jury composé de :

M^{me}	Régine ANDRÉ-OBRECHT	Professeur, IRIT, Toulouse	Rapporteurs
M.	Jan ČERNOCKÝ	Professeur, BUT, Brno (République Tchèque)	
M.	Guillaume GRAVIER	Chargé de recherches, IRISA/CNRS, Rennes	Examinateurs
M.	Sébastien MARCEL	Senior Researcher, IDIAP, Martigny (Suisse)	
M.	Patrick VERLINDE	Professeur, ERM, Brussels (Belgique)	
M.	Jean-François BONASTRE	Professeur, LIA, Avignon	Co-Directeurs
M.	John S. D. MASON	Professeur, Swansea University, Swansea (UK)	

LABORATOIRE INFORMATIQUE D'AVIGNON

Swansea University
Prifysgol Abertawe

Table des matières

Remerciements

Dank u wel, Děkuji , Diolch, Merci, Thank you, Grazie, Grandmercé, Cámo·n Dziękuję, Mauruuru...

J'ajoute quelques lignes à ce document pour remercier ceux qui ont participé de près où de loin à sa naissance.

Je pense tout particulièrement à Régine André-Obrecht et Jan « Honza » Černocký qui ont bien voulu disséquer les quelques 242 pages de cette thèse sans m'en tenir trop rigueur. Merci au Professeur Patrick Verlinde pour avoir accepté de présider mon jury de thèse, nos discussions, scientifiques ou personnelles, ont toujours été très plaisantes. Merci également à Guillaume Gravier et Sébastien Marcel qui ont consacré une partie de leur temps précieux à examiner ce rapport.
J'avoue avoir eu un grand plaisir à présenter le résultat de ces trois années devant un jury que j'estime autant pour l'excellence scientifique que pour les qualités humaines des personnalités qui le composent.

I want to deeply thank John Mason, Professor at Swansea University of Wales for his patience and kindness. Discussing with John as a co-superviser, has always been a rewarding opportunity to learn about scientific rigor. I did learnt a lot during my stay in Wales.

Je suis profondément reconnaissant envers le Professeur Bonastre pour m'avoir guidé durant ces années. Je le remercie d'avoir su rester présent malgré ses nombreuses obligations ainsi que pour l'exemple d'honnêteté scientifique qu'il m' a offert. Les opportunités offertes durant cette thèse – participation à l'organisation des JEP, nombreuses participations aux conférences – m'ont beaucoup apporté.
Merci Jef d'avoir veillé au cap tout au long de la traversé et de n'avoir pas tenu le compte exact des bugs trouvés dans mes programmes et qui dépassent de loin mes scores au bowling...

J'accorde une place à part dans ces remerciements à Corinne, qui par son dynamisme et sa bonne humeur a su égayer la gestion des « aléas »inhérents aux projets de recherche.

Du bureau à Brno, ces trois années et nombreuses pauses café auraient été nettement moins agréables sans les réunions de chantier avec Christophe. Parmi les nombreuses petites choses qui remplissent 3 ans de vie, je me souviendrai longtemps des "coups de gueule" de Nanou, des tetrinet endiablés avec Loïc, des discussions avec Laurianne, des cours d'oenologie avec Eric, des Carcassonnes avec MJ ou des "passages" de Gilles, qui font assurément partie des moments forts de ces trois ans.

Un grand merci aussi à Nico et Ben sans qui les apéros de fin de journée ne sont plus ce qu'ils étaient, à Benjamin pour ses animations en conférence, à Will pour sa bonne humeur, à Driss et Georges pour leurs conseils, à Francky sans qui tout tournerait moins rond, à Minie, Will, Nico, Nick, Mathieu, Alain, Nicole, Laure, Virginie, Vir, Tom, Phanou, Vince, Garrot, Eric, Claire et Lorène pour tous les moments partagés.

Cette thèse m'a donné la chance d'être accueilli au sein d'un laboratoire chaleureux et festif, je tiens à remercier tous ceux qui ont contribué à cette ambiance.

Merci Léa d'avoir relu une bonne partie de ce document sans « trop »râler. Merci Linda de m'avoir soutenu dans ma recherche de thèse et de m'avoir mis sur de bons rails. Merci Seb pour ton affection et ton soutien permanent.

Merci à toi aussi qui pensais que je t'avais oublié.

Et puis merci à ceux qui étaient là dès le commencement. Merci Maman et Papa pour votre soutien et votre confiance. Merci à ma famille de m'avoir accompagné tout au long de mon cursus qui s'achève ici. Je pense particulièrement à mes grands parents, à Laëtitia et Jérôme, Cédric et Sabine, Daniel et Jocelyne, et aux plus petits : Aurélien, Lorine, Baptiste et Orlane.

Enfin ces trois années n'auraient pas eu la même saveur sans l'amour et le soutien de Bérénice, merci.

Résumé

La vérification automatique du locuteur est une tâche de classification qui vise à confirmer ou infirmer l'identité d'un individu d'après une étude des caractéristiques spécifiques de sa voix. L'intégration de systèmes de vérification du locuteur sur des appareils embarqués impose de respecter deux types de contraintes, liées à cet environnement :

– les contraintes matérielles, qui limitent fortement les ressources disponibles en termes de mémoire de stockage et de puissance de calcul disponibles ;
– les contraintes ergonomiques, qui limitent la durée et le nombre des sessions d'entraînement ainsi que la durée des sessions de test.

En reconnaissance du locuteur, la structure temporelle du signal de parole n'est pas exploitée par les approches état-de-l'art. Nous proposons d'utiliser cette information, à travers l'utilisation de mots de passe personnels, afin de compenser le manque de données d'apprentissage et de test.

Une première étude nous a permis d'évaluer l'influence de la dépendance au texte sur l'approche état-de-l'art GMM/UBM (Gaussian Mixture Model/ Universal Background Model). Nous avons montré qu'une contrainte lexicale imposée à cette approche, généralement utilisée pour la reconnaissance du locuteur indépendante du texte, permet de réduire de près de 30% (en relatif) le taux d'erreurs obtenu dans le cas où les imposteurs ne connaissent pas le mot de passe des clients.

Dans ce document, nous présentons une architecture acoustique spécifique qui permet d'exploiter à moindre coût la structure temporelle des mots de passe choisis par les clients. Cette architecture hiérarchique à trois niveaux permet une spécialisation progressive des modèles acoustiques. Un modèle générique représente l'ensemble de l'espace acoustique. Chaque locuteur est représenté par une mixture de Gaussiennes qui dérive du modèle du monde générique du premier niveau. Le troisième niveau de notre architecture est formé de modèles de Markov semi-continus (SCHMM), qui permettent de modéliser la structure temporelle des mots de passe tout en intégrant l'information spécifique au locuteur, modélisée par le modèle GMM du deuxième niveau. Chaque état du modèle SCHMM d'un mot de passe est estimé, relativement au modèle indépendant du texte de ce locuteur, par adaptation des paramètres de poids des distributions Gaussiennes de ce GMM. Cette prise en compte de la structure tem-

porelle des mots de passe permet de réduire de 60% le taux d'égales erreurs obtenu lorsque les imposteurs prononcent un énoncé différent du mot de passe des clients.

Pour renforcer la modélisation de la structure temporelle des mots de passe, nous proposons d'intégrer une information issue d'un processus externe au sein de notre architecture acoustique hiérarchique. Des points de synchronisation forts, extraits du signal de parole, sont utilisés pour contraindre l'apprentissage des modèles de mots de passe durant la phase d'enrôlement. Les points de synchronisation obtenus lors de la phase de test, selon le même procédé, permettent de contraindre le décodage Viterbi utilisé, afin de faire correspondre la structure de la séquence avec celle du modèle testé. Cette approche a été évaluée sur la base de données audio-vidéo MyIdea grâce à une information issue d'un alignement phonétique. Nous avons montré que l'ajout d'une contrainte de synchronisation au sein de notre approche acoustique permet de dégrader les scores imposteurs et ainsi de diminuer le taux d'égales erreurs de 20% (en relatif) dans le cas où les imposteurs ignorent le mot de passe des clients tout en assurant des performances équivalentes à celles des approches état-de-l'art dans le cas où les imposteurs connaissent les mots de passe.

L'usage de la modalité vidéo nous apparaît difficilement conciliable avec la limitation des ressources imposée par le contexte embarqué. Nous avons proposé un traitement simple du flux vidéo, respectant ces contraintes, qui n'a cependant pas permis d'extraire une information pertinente. L'usage d'une modalité supplémentaire permettrait néanmoins d'utiliser les différentes informations structurelles pour déjouer d'éventuelles impostures par play-back. Ce travail ouvre ainsi de nombreuses perspectives, relatives à l'utilisation d'information structurelle dans le cadre de la vérification du locuteur et aux approches de reconnaissance du locuteur assistée par la modalité vidéo.

Abstract

SPEAKER verification aims to validate or invalidate identity of a person by using his/her speech characteristics. Integration of an automatic speaker verification engine on embedded devices has to respect two types of constraint, namely :

– limited material resources such as memory and computational power ;
– limited speech, both training and test sequences.

Current state-of-the-art systems do not take advantage of the temporal structure of speech. We propose to use this information through a user-customised framework, in order to compensate for the short duration speech signals that are common in the given scenario.

A preliminary study allows us to evaluate the influence of text-dependency on the state-of-the-art GMM/UBM (Gaussian Mixture Model / Universal Background Model) approach. By constraining this approach, usually dedicated to text-independent speaker recognition, we show that a lexical constraint allows a relative reduction of 30% in error rate when impostors do not know the client password.

We introduce a specific acoustic architecture which takes advantage of the temporal structure of speech through a low cost user-customised password framework. This three stage hierarchical architecture allows a layered specialization of the acoustic models. The upper layer, which is a classical UBM, aims to model the general acoustic space. The middle layer contains the text-independent specific characteristics of each speaker. These text-independent speaker models are obtained by a classical GMM/UBM adaptation. The previous text-independent speaker model is used to obtain a left-right Semi-Continuous Hidden Markov Model (SCHMM) with the goal of harnessing the Temporal Structure Information (TSI) of the utterance chosen by the given speaker. This TSI is shown to reduce the error rate by 60% when impostors do not know the client password.

In order to reinforce the temporal structure of speech, we propose a new approach for speaker verification. The speech modality is reinforced by additional temporal information. Synchronisation points extracted from an additional process are used to constrain the acoustic decoding. Such an additional modality could be used in order to add different structural information and to thwart impostor attacks such as playback.

Thanks to the specific aspects of our system, this aided-decoding shows an acceptable level of complexity. In order to reinforce the relaxed synchronisation between states and frames due to the SCHMM structure of the TSI modelling, we propose to embed an external information during the audio decoding by adding further time-constraints. This information is here labelled external to reflect that it is aimed to come from an independent process.

Experiments were performed on the BIOMET part of the MyIdea database by using an external information gathered from an automatic phonetical alignment. We show that adding a synchronisation constraint to our acoustic approach allows to reduce impostor scores and to decrease the error rate from 20% when impostor do not know the client password. In others conditions, when impostors know the passwords, the performance remains similar to the original baseline.

The extraction of the synchronisation constraint from a video stream seems difficult to accommodate with embedded limited resources. We proposed a first exploration of the use of a video stream in order to constrain the acoustic process. This simple video processing did not allow us to extract any pertinent information.

Introduction

La Déclaration Universelle des Droits de l'Homme de 1948 assure que « Toute personne (...) a droit à la propriété » (*article 17*), c'est-à-dire droit d'user, de jouir et de disposer d'une chose de manière exclusive. La loi Informatique et Libertés du 6 janvier 1978[1], relative aux données personnelles, prévoit quant à elle qu'un « traitement de données à caractère personnel doit avoir reçu le consentement de la personne concernée ». Face à la multiplication des terminaux portables, la garantie de ces droits dans le domaine des communications et des données numériques est un défi majeur.

Différentes techniques peuvent être utilisées pour la sécurisation des systèmes embarqués. La biométrie permet l'authentification d'individus à partir de leurs caractéristiques physiologiques ou comportementales. Ces caractéristiques doivent être :

- universelles : présentes chez tous les individus ;
- uniques : spécifiques à chaque individu pour permettre de le différencier par rapport aux autres ;
- permanentes : pour permettre une authentification au cours du temps ;
- mesurables : pour permettre l'enregistrement et les comparaisons futures.

L'authentification biométrique présente de nombreux avantages, puisqu'elle permet de s'affranchir des intermédiaires que constituent les clefs, cartes et autres codes personnels susceptibles d'être oubliés, perdus ou volés. Elle supprime le risque qui peut être occasionné par le prêt d'une clef ou la communication d'un mot de passe à un tiers. L'utilisation de données intrinsèques à l'utilisateur lui permet, de plus, de recourir à la biométrie en tout lieu et à tout moment.

Les principales contraintes liées à la biométrie sont dues à l'ergonomie et à l'acceptabilité de certaines modalités. Mais si la reconnaissance d'iris ou d'empreintes digitales sont généralement mal perçues par le public, il existe d'autres modalités, moins intrusives, comme la reconnaissance automatique du locuteur (RAL) et les biométries du visage. Ces modalités présentent l'avantage d'être naturelles aux êtres humains, tout en apportant un niveau de sécurité suffisant pour un grand nombre d'applications. De plus, le matériel nécessaire - microphone et caméra - est actuellement intégré à la plupart des systèmes embarqués.

Contexte

Nos travaux, réalisés dans le cadre du projet BIOBIMO [2] (cf. annexe D), ont pour objet le développement d'une application biométrique bi-modale audio-vidéo embarquée sur téléphone mobile. Outre le niveau de sécurité requis, ce cadre applicatif impose deux types de contraintes : technologiques et ergonomiques.

[1] Article 7 de la Loi n° 78-17 du 6 Janvier 1978 relative à l'informatique, aux fichiers et aux libertés (Journal Officiel de la République Française du 07-01-1978 p. 227-231)

[2] BIOBIMO : BIOmétrie BImodale sur MObile, est un projet supporté par l'ANR/RNRT 2005, `http://biobimo.eurecom.fr/`

Les technologies disponibles actuellement sur les téléphones cellulaires limitent considérablement les traitements qui peuvent être effectués en ligne ainsi que la capacité de stockage liée aux systèmes d'authentification. Les ressources disponibles sur ces appareils augmentent continuellement et de façon importante, laissant penser que les contraintes de puissance et de stockage tendent à disparaître. Cependant, le nombre et les besoins des applications embarquées croissent proportionnellement aux ressources et justifient, selon nous, une forte vigilance.
Les ressources disponibles sont difficilement quantifiables. Nous nous contenterons, dans ce document, de considérer une estimation empirique des ressources disponibles sur les appareils actuels, de manière à fixer une limite réaliste aux possibilités qui nous sont offertes.

L'utilisation quotidienne des téléphones cellulaires impose également de fortes contraintes ergonomiques. Dans le cadre des modalités audio et vidéo, la phase d'authentification doit être la plus courte possible. La limitation de la durée d'enregistrement à quelques secondes pour vérifier une identité nous semble, par exemple, être une condition à minima dans le contexte des applications embarquées.
D'autres problématiques, propres aux biométries audio et vidéo, doivent également être prises en compte. La variation des conditions d'utilisation, par exemple, dégrade fortement les performances des systèmes actuels. Des avancées importantes ont été réalisées dans ce domaine ces dernières années. Les méthodes existantes nécessitent, néanmoins, des quantités de données importantes pour accroître la robustesse des représentations des utilisateurs ainsi que des ressources calculatoires conséquentes.

Problématique

Les contraintes technologiques énoncées précédemment ne sont pas compatibles avec les systèmes d'authentification actuels, basés sur les modalités image ou vidéo. Les ressources disponibles imposent un traitement du flux vidéo plus simple que celui qui est réalisé dans la plupart des approches état-de-l'art. La reconnaissance du locuteur requiert, quant à elle, des ressources importantes mais peut plus facilement s'accommoder des contraintes liées au contexte embarqué.

La dégradation des performances des systèmes de reconnaissance automatique du locuteur, lorsque la quantité de données biométriques est restreinte, constitue un problème majeur. Qu'il s'agisse de la phase d'enrôlement ou de test, le nombre et la durée des séquences d'acquisition déterminent le niveau de performance d'un système. Cette quantité de données nécessaire influe aussi directement sur l'ergonomie du système biométrique.

La quantité de données requise doit alors être déterminée en considérant le ratio ergonomie/performances du système. Néanmoins, la faible quantité de données dispo-

nible pour le système de reconnaissance du locuteur peut être compensée par l'apport d'informations supplémentaires provenant du signal de parole. Les systèmes de reconnaissance du locuteur n'exploitent, dans leur grande majorité, que l'information acoustique à court terme du signal de parole. L'information temporelle à plus long terme, la structure du signal, peut être utilisée pour compenser la durée limitée des séquences d'enrôlement et de tests. Il est possible pour cela de s'inspirer des travaux réalisés en reconnaissance de la parole (RAP), comme le font certaines approches de reconnaissance du locuteur, dites dépendantes du texte.

Les approches dépendantes du texte exploitent en effet, pour la plupart, la structure temporelle du signal acoustique de parole. La modélisation de cette structure peut nécessiter des ressources calculatoires supplémentaires, tout comme elle requiert, généralement, une quantité de données plus importante lors de la phase d'entraînement. Le manque de données d'apprentissage, nécessaires à l'estimation de la structure temporelle du signal de parole, peut être compensé par une forte astreinte sur les énoncés comme, par exemple, l'utilisation de mots de passe personnels.
La contrainte structurelle appliquée à la modalité acoustique peut également être renforcée par l'ajout d'une information issue des flux audio ou vidéo.Les méthodes bimodales développées ces quinze dernières années ont montré que l'utilisation d'informations provenant de modalités différentes peut améliorer, d'une part, la robustesse des systèmes biométriques dans des conditions d'utilisations adverses et permettre, d'autre part, de lutter contre certains types d'impostures comme les play-backs.

Dans ce contexte, la nature bi-modale de la parole peut être exploitée pour tirer parti d'une information issue du flux vidéo. Les approches bi-modales sont nombreuses dans la littérature et doivent cependant composer avec deux difficultés majeures. Les signaux et informations audio et vidéo sont de natures différentes et leur intégration au sein d'un processus conjoint est un problème complexe. De plus, ces deux flux sont fortement corrélés et présentent une asynchronie due au processus de production de la parole, qui rend difficile un traitement simultané. Plusieurs approches sont possibles au sein desquelles la place réservée aux modalités audio et vidéo peut être très variable.
Les approches les plus répandues consistent à fusionner les informations provenant des deux modalités à différents niveaux de la chaîne de traitement. Ces méthodes ne tiennent pas compte, la plupart du temps, de la nature très différente des informations présentes dans l'un ou l'autre des flux ni, d'ailleurs, de leur forte corrélation.
D'autres approches, plus rares, exploitent au contraire la corrélation existant entre les flux de données audio et vidéo, en tenant compte de leur asynchronie.

Ces méthodes, souvent complexes, nécessitent une quantité de ressources très importante, comparativement à un système de reconnaissance du locuteur. Ce besoin est principalement dû au traitement du flux vidéo. Malgré le surcoût de la modalité vidéo, les performances obtenues par cette modalité sont nettement inférieures à celles des systèmes audio.

Contributions

Pour répondre aux contraintes ergonomiques et aux limitations de ressources, nous proposons dans cette thèse une nouvelle approche de vérification d'identité biométrique, visant à compenser le manque de données disponibles par la prise en compte de la structure temporelle du signal.

Cette approche repose sur la voix en tant que biométrie principale, pouvant être renforcée par l'apport d'autres informations provenant, par exemple, du flux vidéo.
Notre processus de reconnaissance du locuteur repose sur une architecture acoustique qui exploite la structure temporelle d'un mot de passe choisi librement par l'utilisateur. L'organisation temporelle du flux acoustique est représentée par des modèles de Markov semi-continus, nécessitant des ressources réduites, en accord avec les contraintes de l'embarqué. Les modèles de locuteur sont construits à partir d'un seul exemple du mot de passe.
Une contrainte temporelle, issue d'un processus externe, est intégrée au sein de notre architecture acoustique. Cette information a pour rôle de renforcer la modélisation de la structure temporelle issue du signal acoustique lors de la phase d'apprentissage.
La contrainte appliquée au système acoustique peut être obtenue à partir du flux audio, mais il est également possible d'extraire une information du flux vidéo. L'analyse de la cohérence des flux audio et vidéo peut, par exemple, être utilisée pour déceler des impostures de type play-back.

Structure du document

La première partie de cette thèse définit les notions d'identité et de reconnaissance biométrique. Elle introduit dans le chapitre 1 les systèmes automatiques et les enjeux qui les caractérisent avant d'en présenter une analyse plus détaillée dans le chapitre 2.

La partie II traite de la parole en tant que modalité biométrique multiple. La composante audio de la parole fait l'objet du chapitre 3 alors que le chapitre 4 traite la composante visuelle. Enfin, le chapitre 5 analyse les principales approches bi-modales existantes. La critique de ces méthodes attache une importance particulière à la place accordée à la structure temporelle du signal de parole. Celle-ci, lorsqu'elle est considérée, peut être prise en compte au sein même d'une modalité ou intégrée au processus de fusion des modalités.

La troisième partie du document est consacrée à nos contributions. Elle débute par la description des motivations qui ont guidé les travaux réalisés durant cette thèse et de l'architecture acoustique renforcée par une contrainte temporelle externe que nous avons proposée. Le chapitre 6 est une réflexion sur les particularités propres à la validation statistique des approches biométriques. Nous y commentons l'usage des corpus

d'évaluation, spécialement audio-vidéo, et justifions nos choix quant à l'évaluation de notre approche.
Les quatre chapitres suivants décrivent chacun un élément de notre architecture acoustique, renforcée par une contrainte externe.
Cette architecture repose sur le paradigme GMM/UBM décrit dans le chapitre 7. Ce paradigme ne modélise pas explicitement l'information temporelle du signal de parole mais nous proposons ici une analyse de l'influence de la dépendance au texte sur les performances des systèmes GMM/UBM, en accord avec le contexte applicatif visé, pour lequel les données d'enrôlement et de test sont limitées.
Le chapitre 8 présente notre extension du paradigme GMM/UBM permettant de modéliser, à moindre coût, la structure temporelle des mots de passe choisis par les clients. La configuration des modèles de Markov utilisés et les performances obtenues sont alors discutées.
La synchronisation du processus acoustique par une information externe est présentée dans le chapitre 9. Notre approche est validée grâce à une information provenant d'un système acoustique éprouvé avant d'être testée avec une information issue du flux vidéo selon un processus peu coûteux.

Nous conclurons finalement ce travail de thèse en présentant un résumé de nos principales contributions ainsi qu'un ensemble de perspectives.

Première partie

Introduction à la biométrie

Chapitre 1

De l'individu à la biométrie

Sommaire

Résumé

Ce chapitre propose une introduction à la biométrie. Il introduit la notion d'identité et les questions inhérentes à la reconnaissance d'un individu. Il présente ensuite les problématiques et contraintes liées à l'utilisation de systèmes automatiques. Différentes modalités peuvent être utilisées afin de reconnaître un individu et sont présentées dans ce chapitre. Finalement, la dernière partie de ce chapitre s'attache à décrire les tâches d'identification et de vérification d'identité.

1.1 Un individu - une identité

L'IDENTITÉ est une notion complexe, difficile à définir. Cette première section propose une définition des enjeux et les limitations relatifs au concept d'identité, dans le cadre des applications biométriques.

1.1.1 Une identité pour reconnaitre l'individu

L'identité renvoie à ce qu'un sujet a d'unique. D'un point de vue personnel, la caractérisation de l'identité prend en compte tout ce que l'individu considère comme faisant partie intégrante de lui et qui ne peut lui être enlevé. Cette définition inclut un certain nombre de facteurs qui peuvent évoluer dans le temps comme, d'ailleurs, la conscience de soi.
D'un point de vue externe à l'individu, son identité est la façon dont il est perçu par le monde qui l'entoure. Cette identité, en tant qu'entité, est associée à une appellation. L'individu se nomme « moi », son environnement lui associe un nom.

D'un point de vue externe, la reconnaissance d'un individu se heurte à la caractérisation de son unicité. Il n'est pas imaginable d'obtenir une description exhaustive d'un individu qui engloberait sa description physiologique complète ainsi que la description de ses connaissances, de ses possessions, de son vécu et de son expérience. Il faut alors, pour obtenir une description unique d'un individu, la restreindre aux informations nécessaires et suffisantes à sa reconnaissance au sein d'un groupe. Dans le cadre de la reconnaissance d'une identité par un système automatique, cette description ne doit intégrer que des informations susceptibles d'être vérifiées dans le contexte applicatif choisi.

Du groupe auquel appartient l'individu considéré, dépend la description minimale nécessaire à sa reconnaissance. En effet, si deux individus de ce groupe correspondent à la description courante, il faut rajouter une information permettant de les différencier. Cette information est nécessaire à la reconnaissance de chacun d'eux.
De même, le groupe considéré influe sur la quantité d'information suffisant à décrire l'individu de façon unique. Il n'est pas utile d'ajouter un élément à une description qui ne correspond qu'à une seule personne du groupe.

1.1.2 Différentes définitions de l'identité

Les informations décrivant un individu peuvent être de nature variable. Il est commun de décrire une personne par ses caractéristiques physiques, comme la couleur de ses cheveux, de ses yeux ou d'autres détails de son anatomie. Ce type de description nécessite cependant d'avoir déjà vu cet individu. Lors d'une conversation téléphonique,

il est naturel de reconnaître son interlocuteur à sa voix ou la façon dont il s'exprime. S'il existe ainsi plusieurs modes de description d'un individu, tous reposent sur la connaissance de caractéristiques qui lui appartiennent en propre. Dans son environnement social, ce lien entre identité et propriété est couramment utilisé pour identifier un individu.

Deux principaux types d'information peuvent être utiles pour décrire une personne : les informations liées à ses possessions et celles qui décrivent sa nature même. Pour les premières, il peut s'agir d'une possession matérielle comme une clef ou un passeport, mais également d'une possession intellectuelle, comme un code, un mot de passe ou, plus généralement, un souvenir. Ces informations présentent l'intérêt d'être facilement vérifiables, mais peuvent être perdues, oubliées ou usurpées.
Les informations obtenues par la mesure des caractéristiques d'une personne, ou données biométriques, font référence aux caractéristiques intrinsèques de l'individu. Leur utilisation nécessite la prise en compte de la nature changeante de l'être humain. Ces changements peuvent être dus au vieillissement, à la maladie ou à un état émotionnel différent et doivent être pris en compte dans la description biométrique d'un individu. Pour la reconnaissance des individus, les systèmes biométriques permettent d'atteindre des niveaux de performance qui sont inaccessibles aux êtres humains. La partie 1.2 décrit de façon plus détaillée les possibilités offertes par la biométrie.

1.2 Les biométries

La biométrie ou mesure (*metron*) du vivant (*bios*) est, d'après l'encyclopédie Larousse [1], « l'étude statistique des dimensions et de la croissance des êtres vivants ». L'extension de la biométrie au domaine de la reconnaissance des personnes consiste à déterminer l'identité d'un individu grâce à des mesure quantitatives. Ces mesures peuvent avoir pour objet les caractéristiques morphologiques ou les caractéristiques comportementales de cette personne.

1.2.1 La biométrie morphologique

La biométrie morphologique décrit les individus par des mesures de leurs caractéristiques biologiques ou physiologiques. Ces mesures sont moins sujettes à l'influence du stress que la biométrie comportementale. Elles sont également plus difficiles à falsifier.

Les caractéristiques mesurables qui permettent de décrire un individu sont nombreuses (Jain et al., 1999). Chaque modalité présente des avantages et inconvénients qu'il faut considérer en parallèle de ses performances et, donc, du degré de sécurité

[1] `http://www.larousse.fr/encyclopedie/`

qu'elle propose. Les biométries morphologiques les plus courantes mesurent les empreintes digitales, le réseau veineux de la rétine, l'iris, l'empreinte de la main ou certaines caractéristiques du visage. La biologie permet, quant à elle, de caractériser un individu par son ADN à travers une analyse de sa salive, de son sang ou de tout échantillon corporel.

La biométrie morphologique est, à l'heure actuelle, un des moyens les plus fiables pour reconnaître un individu, car elle mesure des caractéristiques qui sont indissociables de cet individu.
Elle présente néanmoins certains inconvénients. Elle doit, par exemple, pour être utilisable, intégrer les changements temporels intrinsèques de l'individu. L'acquisition de certaines données biométriques peut également être compliquée par des difficultés physiques ou sociétales.

1.2.2 La biométrie comportementale

La biométrie comportementale mesure et caractérise des éléments qui sont propres aux comportements d'un individu. De nombreux comportements peuvent être observés et analysés afin de caractériser une personne.

La signature dynamique constitue un exemple de biométrie comportementale. Elle consiste à mesurer certaines variables qui interviennent lorsqu'un individu signe un document. Les systèmes de biométrie utilisant cette méthode enregistrent la vitesse et les accélérations du stylo ou la pression exercée. Ils permettent aussi d'analyser, de façon plus naturelle, la forme de la signature. Il est alors possible de différencier les parties qui sont identiques à chaque réalisation de la signature de celles qui varient. Cette biométrie présente l'avantage d'être historiquement une méthode d'identification très utilisée et adaptée à l'authentification de documents manuscrits.

L'utilisation de matériels informatiques a également suscité un intérêt pour la biométrie. Par exemple, les travaux de Monrose et Rubin (2000) ont montré qu'il est possible de reconnaître une personne au rythme de sa frappe sur un clavier. Cette méthode présente l'avantage de permettre une identification continue de l'utilisateur et de détecter un changement d'utilisateur en temps réel et de façon transparente.

L'analyse de la démarche (Cunado et al., 1997) ou celle du contact du pied sur le sol (Orr et Abowd, 2000), (Rodriguez et al., 2007), (Rodríguez et al., 2008) permettent également d'authentifier un individu.

Les principaux inconvénients des biométries comportementales sont liés à la grande variabilité que peuvent générer des changements émotionnels ou environnementaux

chez l'utilisateur. Le stress ou un environnement perturbé peuvent, par exemple, affecter les comportements et ainsi perturber le résultat du test de reconnaissance.

1.2.3 Les biométries mixtes

Certaines modalités se situent à la croisée des biométries morphologiques et comportementales. La voix, qui est utilisée de façon naturelle par les êtres humains pour reconnaître un individu, est une modalité comportementale qui peut subir les influences d'une pathologie, du stress ou même d'un changement émotionnel. Elle peut également être modifiée selon la volonté du locuteur. Elle garde cependant des caractéristiques constantes qui peuvent permettent d'identifier le locuteur dans le cas où il contrefait sa voix. En effet, le phénomène complexe de la production vocale fait intervenir un grand nombre de caractéristiques intrinsèques au locuteur, qui seront abordées plus précisément dans la partie 3.1. La morphologie de l'appareil respiratoire du locuteur influe, par exemple, sur les caractéristiques de sa voix. Or, cette morphologie ne peut être modifiée de façon volontaire par l'individu.

L'analyse des battements du cœur par l'intermédiaire des signaux d'un électrocardiogramme (Israel et al., 2005), ou l'analyse de l'activité électrique du cerveau mesurée par Électro-encéphalographie (Marcel et del R. Millan., 2007) sont d'autres modalités biométriques mixtes. Les signaux acquis dans ces deux modalités sont sujets aux changements émotionnels, physiologiques et environnementaux, mais contiennent cependant des informations caractérisant respectivement le muscle cardiaque et le cerveau qui sont propres à l'individu considéré.

La biométrie vidéo exploite également les informations morphologiques et comportementales des individus. Elle permet de décrire les traits du visage de ses sujets d'étude, leur apparence physique, aussi bien que leurs mouvements.

1.3 Biométrie et systèmes automatiques

Reconnaître une personne grâce à une description de sa morphologie ou de son comportement est une opération courante pour les êtres humains. Il est naturel de reconnaître le visage ou la voix d'un individu, tout comme il peut être naturel de reconnaître son écriture ou sa démarche. Les facultés humaines à reconnaître un individu sont cependant limitées par différents facteurs.
Certaines données caractéristiques d'un individu ne peuvent pas être recueillies par des êtres humains dont les capacités sensorielles sont limitées. L'analyse de l'ADN, la description de l'iris ou les empreintes digitales ne peuvent être obtenues sans utiliser d'appareillage spécifique.

La reconnaissance de personnes par des êtres humains est également limitée par leur mémoire ou leur capacité à modéliser. Les travaux de Hollien et al. (1974) montrent qu'il est généralement plus facile de reconnaître la voix d'une personne proche que celle d'un inconnu entendue en un nombre limité d'occasions. Un être humain a besoin de côtoyer une personne suffisamment longtemps pour reconnaître en elle les informations qui lui sont propres et qui permettent une caractérisation précise. L'utilisation de systèmes automatiques permet, elle, d'acquérir en peu de temps un grand nombre de données provenant d'un individu afin de construire une représentation relativement fiable de cette personne.

L'utilisation des systèmes automatiques est avant tout justifiée par le nombre croissant de communications (téléphone, internet...) et d'échanges qui doivent être sécurisés (transactions financières, documents électroniques, accès sécurisés à des services, des locaux...). Cette quantité d'information à sécuriser a fortement augmenté du fait de la généralisation des terminaux portables. Il est impossible à un être humain de mémoriser les caractéristiques de milliers de personnes et de les reconnaître avec une confiance élevée. Par opposition, un système automatique peut mémoriser un grand nombre de descriptions qui, de plus, ne seront pas altérées par le temps comme peut l'être la mémoire humaine.

L'apparition des systèmes automatiques d'authentification, et plus particulièrement des systèmes biométriques, a amélioré considérablement le niveau de sécurité des applications qu'ils protègent. Ces systèmes présentent pourtant de nombreuses failles ou inconvénients. Certains sont dus à la nature des données biométriques utilisées, d'autres sont plus spécifiquement liés à un type d'applications.

L'automatisation des systèmes de reconnaissance biométrique implique que les données utiles puissent être prélevées sur toute personne se présentant à eux. En un mot, les données biométriques doivent être universelles.

Les systèmes biométriques traitent des données vivantes. Ils mesurent des caractéristiques qui, comme le corps humain, sont en constant changement. Ces changements sont dus au vieillissement ou à divers traumatismes. Les systèmes doivent alors assimiler ces variations intra-individu afin de permettre une utilisation prolongée dans le temps.
Pour ce faire, la plupart des systèmes automatiques utilisent des méthodes statistiques qui permettent de différencier les données propres à un individu, qui seront constantes, des données qui peuvent varier avec le temps. L'utilisation de ce type de méthodes pose la question de la fiabilité des résultats qu'elles fournissent. Les réponses fournies par ces systèmes doivent donc toujours être traitées en considérant la confiance qui peut leur être accordée. Cette confiance varie selon les méthodes utilisées, les modalités, l'acquisition et le type de données biométriques ainsi que la fiabilité des modèles statistiques utilisés.

De nombreuses contraintes sont liées à la nature, au contexte ou à l'environnement des applications qui doivent être sécurisées. Ces contraintes, telles que l'ergonomie, la vitesse d'exécution ou l'acceptation de la biométrie par les utilisateurs, doivent être prises en compte pour le choix des modalités utilisées.
Certaines approches biométriques, comme par exemple les tests ADN ou la reconnaissance d'après l'étude d'électrocardiogrammes, nécessitent une infrastructure lourde et des temps de procédures qui rendent ces modalités, pourtant assez fiables, incompatibles avec les contraintes d'une utilisation régulière.
Les tests ADN sont, de plus, très contraignants puisqu'ils requièrent des prélèvements de fluides ou de cellules directement sur le corps humain. La reconnaissance de l'iris ou du réseau veineux de la rétine sont d'autres exemples de modalités particulièrement intrusives.
La reconnaissance d'empreintes digitales est aujourd'hui plutôt bien acceptée, mais nécessite la présence physique de la personne à identifier. Des modalités comme la reconnaissance vocale ou la reconnaissance de visage n'occasionnent aucune gêne pour l'utilisateur et sont, de ce fait et par leur aspect naturel, très bien tolérées par les utilisateurs. Elles offrent un certain confort d'utilisation puisqu'elles peuvent être utilisées en mode mains libres. De plus, elles peuvent opérer à travers un réseau de communication.

De la même façon que pour la reconnaissance par un être humain, le refus de collaborer de la personne à authentifier peut nuire gravement à la fiabilité du système automatique. Un utilisateur qui ne souhaite pas être reconnu pourra, dans certains cas, falsifier les données fournies au système automatique.

L'authentification biométrique ne nécessitant aucune connaissance ou possession particulière, n'importe quel individu peut tenter d'usurper l'identité d'un client en fournissant des données biométriques au système de reconnaissance. La robustesse aux impostures constitue l'une des principales problématiques de la recherche biométrique.

La biométrie vocale, la reconnaissance de visage ou la biométrie vidéo peuvent être utilisées de façon quasiment invisible pour l'utilisateur et peuvent ainsi permettre d'obtenir des données non-falsifiées. L'utilisation de la reconnaissance biométrique pose de nombreuses questions d'ordre éthique[2] que nous ne traiterons pas ici mais qui doivent cependant être l'objet de réflexions constantes. Nous noterons toutefois que l'authentification biométrique par la voix ou les données vidéo peut être effectuée sans déranger les utilisateurs et que ces deux modalités offrent un niveau de sécurité relativement élevé (Matrouf et al., 2008), (Phillips et al., 2006), (Tan et al., 2006).

[2]encadrées par les articles 25 et 26 de la Loi n° 78-17 du 6 Janvier 1978 (Journal Officiel de la République Française du 07-01-1978 p. 227-231) relative à l'informatique, aux fichiers et aux libertés, qui stipule que « Les traitements automatisés comportant des données biométriques nécessaires au contrôle de l'identité des personnes (...) » « Sont mis en œuvre après autorisation de la Commission nationale de l'informatique et des libertés ».

1.4 Applications et tâches biométriques

L'identification et la vérification d'identité sont historiquement les principales tâches des systèmes biométriques. Cette partie décrit ces deux tâches ainsi que les caractéristiques et problématiques qui leur sont propres.

1.4.1 Identification

L'identification d'un individu est le processus qui consiste à décider quelle est, parmi une population connue, l'identité de l'utilisateur présent.
Un échantillon de données biométrique S est fourni au système automatique par l'utilisateur, sans aucune autre information sur son identité (cf. figure 1.1). Les données sont alors comparées à une référence caractéristique de chaque utilisateur I_X connu par le système. Le résultat de chaque comparaison est un score, fonction de la similitude observée par le système entre les données S prélevées sur l'utilisateur et la référence considérée. Le score le plus élevé correspond à la référence la plus proche des données de test et l'identité correspondant à cette référence est retournée par le système.

Il existe deux modes d'identification automatique, en milieu ouvert ou fermé.

En *milieu fermé* (figure 1.1), le système automatique décide de l'identité la plus probable parmi les utilisateurs connus (dont il possède une référence). Ce mode de fonctionnement tend à considérer que seules des personnes référencées peuvent accéder au système. Un tel système ne doit alors être utilisé que dans un environnement au sein duquel tous les individus sont connus.

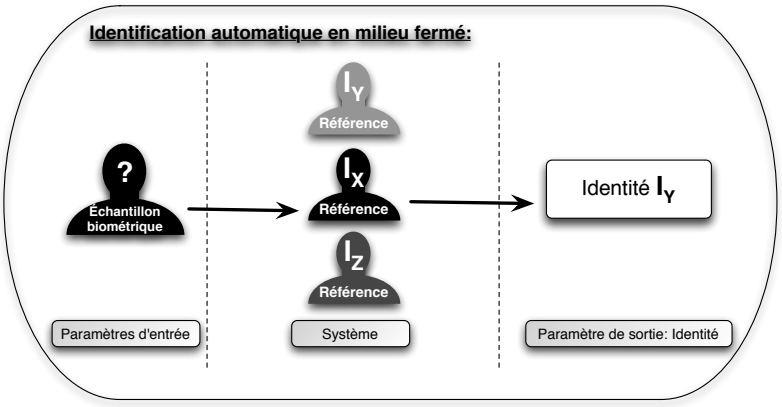


FIG. *1.1:* *Schéma de principe de la tâche d'identification automatique en milieu fermé*

L'identité I_Y retournée, correspondant à la référence $\mathcal{Y}$ est obtenue par :

$$\mathcal{Y} = \underset{\mathcal{X}}{argmax} f(\mathcal{X}|\mathcal{S}) \tag{1.1}$$

où $f(\mathcal{X}|\mathcal{S})$ est le score calculé lors de la comparaison des données S au modèle de l'individu I_X.

En *milieu ouvert* (figure 1.2), le système automatique a la possibilité de rejeter l'utilisateur dont il teste les données biométriques si elles ne correspondent à aucune des identités répertoriées. Cet utilisateur est alors considéré comme inconnu du système et est rejeté. Pour ce faire, les données biométriques S sont comparées à chaque référence $\mathcal{X}$ connue par le système. Chaque comparaison fournit un score $f(\mathcal{X}|\mathcal{S})$. Le score le plus élevé est alors comparé à un seuil Ω, fixé au préalable. Si le score est supérieur à ce seuil, le système décide qu'il s'agit de la personne correspondant à la référence sélectionnée. Si le score est inférieur à ce seuil, le système décide qu'il ne s'agit pas d'une personne « connue ».

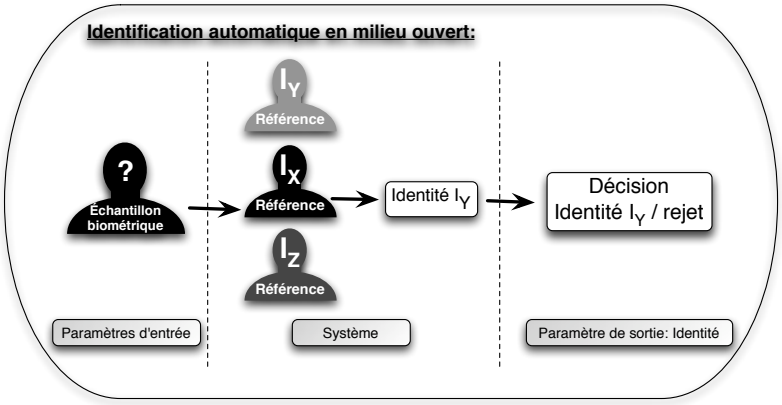


FIG. *1.2:* *Schéma de principe de la tâche d'identification automatique en milieu ouvert*

Pour résumer, en identification en milieu ouvert, le système répond à deux interrogations : « Quelle est l'identité la plus probable ? » et « Les données biométriques analysées correspondent-elles à cette identité ? », alors qu'en milieu fermé il ne répond qu'à la première.
L'identité la plus probable est obtenue, comme dans le cas de l'identification en milieu fermé, par :

$$\mathcal{Y} = \underset{\mathcal{X}}{argmax} f(\mathcal{X}|\mathcal{S}) \tag{1.2}$$

Et la réponse du système est donnée par :

$$f(\mathcal{Y}|\mathcal{S}) \lessgtr \Omega \begin{cases} \text{Identité non reconnue} \\ \text{L'identité est celle du client } I_Y \end{cases} \tag{1.3}$$

où Ω est le seuil de décision fixé au préalable.

1.4.2 Vérification d'identité

La vérification d'identité est le processus qui consiste à décider si l'identité de l'utilisateur présent correspond à celle qu'il revendique.

L'utilisateur qui se présente au système doit annoncer son identité, I_X, et fournir des données biométriques S au système. Le système compare alors la référence correspondant à l'identité clamée aux données fournies par l'utilisateur. Le système de vérification d'identité utilise deux données en entrée : l'identité et les données biométriques. L'identité revendiquée I_X doit nécessairement être connue du système.

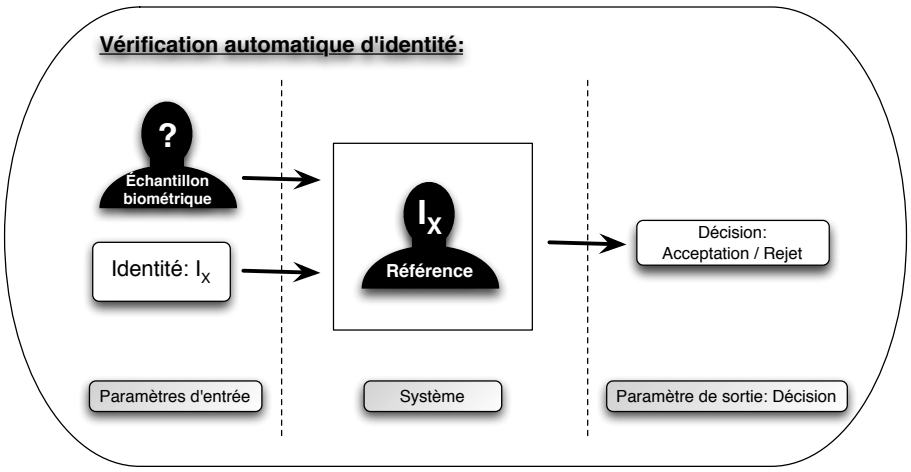


FIG. *1.3:* *Schéma de principe de la tâche de vérification automatique d'identité*

De la comparaison de ces données est issue une mesure de similarité $f(\mathcal{X}|\mathcal{S})$ qui est comparée à un seuil Ω. Si la mesure de similarité est supérieure à ce seuil, l'utilisateur est accepté, si elle est inférieure il est rejeté.
La décision du système est exprimée par :

$$f(\mathcal{X}|\mathcal{S}) \lessgtr \Omega \begin{cases} \text{Utilisateur refusé} \\ \text{Utilisateur accepté} \end{cases} \tag{1.4}$$

L'environnement des applications sécurisées est rarement un milieu fermé, restreint aux seuls utilisateurs autorisés, et les systèmes embarqués sont particulièrement exposés aux risques d'impostures. Nos travaux se situent dans le cadre du projet BIOBIMO, qui vise à sécuriser des applications embarquées sur téléphones cellulaires. Aussi, seule la tâche de vérification d'identité, qui nous paraît particulièrement adaptée à ce contexte, a fait l'objet de ces travaux.
Notons toutefois que la tâche de vérification équivaut à la deuxième étape de l'identification en milieu ouvert et que la principale différence entre ces deux tâches est le nombre de références auxquelles doivent être comparées les données biométriques de test. Lors de la vérification, les données prélevées sur l'utilisateur sont comparées au modèle correspondant à l'identité clamée, alors que pour l'identification en milieu ouvert, ce même échantillon de données est comparé au modèle de chacun des individus référencés par le système.

Chapitre 2

Description générale des systèmes de vérification biométrique d'identité

Sommaire

Résumé

Ce chapitre présente la structure commune à la plupart des systèmes de vérification d'identité biométrique. Il décrit les phases d'enrôlement et de test nécessaires au processus de vérification d'identité. Les différentes parties de ces deux phases sont présentées par ordre d'intervention dans le processus : paramétrisation, modélisation, reconnaissance et décision. La suite de ce chapitre est consacrée à la mesure de performance d'un système de vérification biométrique d'identité. Nous décrivons les différents types d'erreurs qu'un système automatique peut commettre et les outils utilisés pour représenter les performance de ces systèmes : les courbes et les fonctions de performance.

Introduction

La tâche de reconnaissance de personnes se décompose en deux phases. La première étape consiste à obtenir une représentation de l'utilisateur. Elle est appelée « enrôlement ». Cette étape joue un rôle essentiel dans le processus de reconnaissance. Lors de cette phase, le système construit sa représentation de l'individu, représentation qui sera utilisée par la suite pour autoriser ou non l'accès au service. Cette représentation de l'utilisateur doit respecter certaines contraintes pour permettre le bon fonctionnement du système : elle doit être unique et permanente.
La deuxième étape est l'étape de test. Des données provenant d'un utilisateur souhaitant être authentifié sont soumises au système. Cet utilisateur annonce, de plus, une identité connue du système. Le test consiste à mesurer la ressemblance entre les données fournies par l'utilisateur et le modèle existant correspondant à l'identité annoncée. Les étapes d'enrôlement et de test sont décrites dans la suite, sous la forme d'une succession de modules fonctionnels. Nous présentons enfin les éléments permettant d'apprécier les performances des systèmes de vérification d'identité.

2.1 Structure de la phase d'enrôlement

La phase d'enrôlement peut être décrite comme l'utilisation de deux modules : le module de paramétrisation et celui de modélisation, comme représentés par le schéma 2.1.

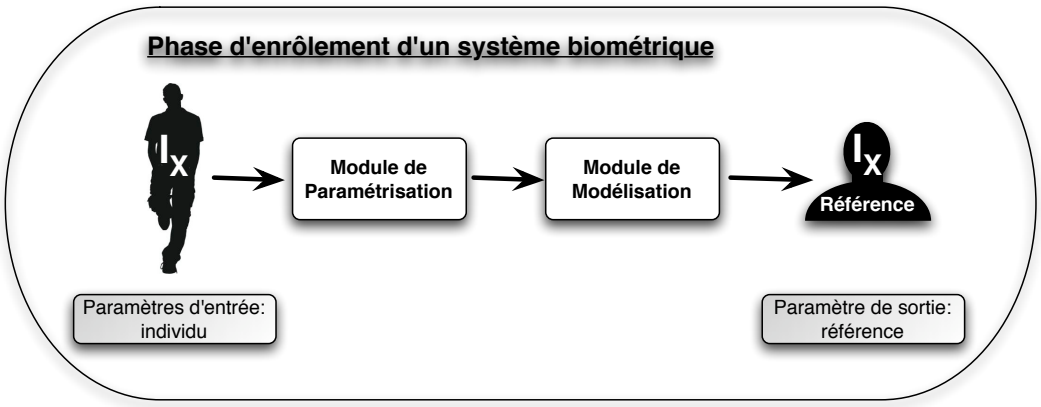


FIG. 2.1: *Schéma de principe de la phase d'enrôlement d'un système biométrique*

2.1.1 Le module de paramétrisation

Selon la modalité biométrique utilisée, la nature du signal reçu par le système biométrique varie : fluide corporel pour les analyses biologiques, signal acoustique pour la

biométrie vocale ou images pour la reconnaissance de visage, par exemple. Une caractéristique commune à toutes les modalités tient au fait qu'il est très difficile de traiter les données brutes. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées à du bruit.
Le module de paramétrisation, qui traite le signal biométrique brut, doit remplir plusieurs objectifs :

- séparer le signal du bruit ;
- extraire l'information utile à la vérification d'identité ;
- convertir les données brutes à un format directement exploitable par le système.

Chacune de ces tâches pose des problèmes complexes et influe fortement sur les résultats des systèmes automatiques de reconnaissance.

2.1.2 Le module de modélisation

Le module de modélisation exploite les données fournies par le module de paramétrisation afin de créer la représentation d'un individu qui servira, par la suite, à l'authentifier.
Selon les méthodes et la modalité biométrique utilisées, la nature du modèle varie. Pour une reconnaissance ADN, le modèle sera constitué d'une liste de certaines caractéristiques génétiques de la personne. Dans le cas de biométries utilisant la voix, les traits du visage ou des caractéristiques comportementales, le modèle utilisé est généralement une représentation statistique des données biométriques acquises lors de la phase d'enrôlement.
Quelle que soit la nature du modèle créé, il doit cependant respecter certaines contraintes :

- il doit être le plus précis possible afin de limiter l'ambiguïté inter-individus ;
- il doit prendre en compte la variabilité intra-individu afin de représenter l'individu au cours du temps.

2.2 Structure de la phase de test

Les systèmes de reconnaissance automatique d'identité présentent des caractéristiques communes. Quelle que soit la modalité biométrique ou la méthode de reconnaissance choisie, la structure de ce type de système reste la même (Jain et al., 2004). Cette structure est représentée par la figure 2.2. et comprend :

- un module de paramétrisation ;
- un module de reconnaissance ;
- un module de décision.

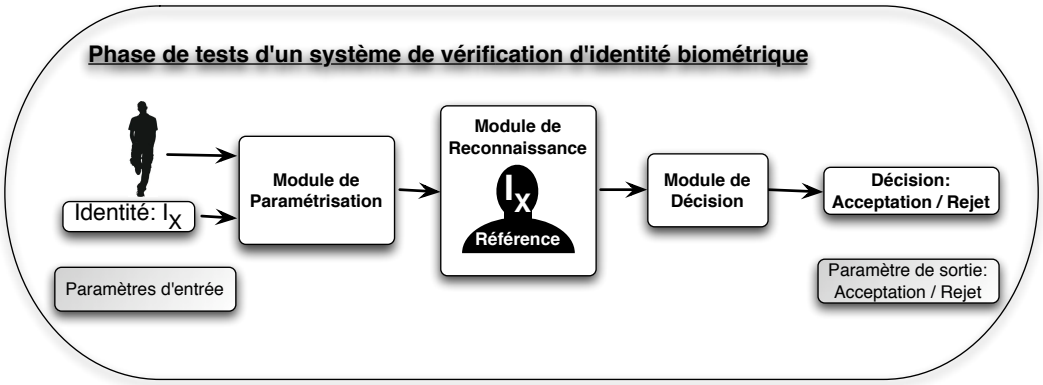


FIG. 2.2: *Schéma de principe de la phase de test d'un système de vérification d'identité biométrique*

2.2.1 Le module de paramétrisation

Le module de paramétrisation est le même que celui utilisé durant la phase d'enrôlement, décrit dans la partie 2.1.1. Il rempli, lors de la phase de test, les mêmes tâches que durant la phase de paramétrisation : séparation signal/bruit, extraction de l'information utile et, éventuellement, normalisation des données. Il est important de conserver la même paramétrisation lors des phases d'enrôlement et de test afin de fournir au système automatique des informations comparables et de même nature.

2.2.2 Le module de reconnaissance

Le module de reconnaissance a un rôle central dans le système biométrique. Il compare les paramètres extraits du signal brut à un modèle d'individu calculé lors de la phase d'enrôlement. Lors de cette comparaison, le module de reconnaissance calcule une mesure de similarité entre les données d'entrée et le modèle testé. C'est une valeur numérique, aussi appelée score, dont l'échelle et le sens varient selon les modalités, les paramètres, le type de modèle et le type de système utilisés. Cette mesure de similarité est ensuite transmise au module de décision.

2.2.3 Le module de décision

Comme le décrit la partie précédente, un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Un module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système de reconnaissance biométrique.

2.3 Quel résultat ?

2.3.1 En vérification, deux types de tests, deux types d'erreurs

Un système de vérification d'identité peut être confronté à deux types de tests :

les tests clients lors desquels l'échantillon biométrique présenté au système correspond à l'identité clamée ;

test imposteur lors desquels l'échantillon biométrique présenté au système provient d'un individu inconnu du système.

Le système automatique doit répondre à chaque tentative d'authentification auquel il fait face par une décision binaire. Il peut donc engendrer deux types d'erreurs :

Faux rejet (FR) erreur commise lorsque le système rejette, à tort, un client légitime (*i.e.* erreur commise lors d'un test client) ;

Fausse acceptation (FA) erreur commise lorsqu'un imposteur est malencontreusement accepté en tant qu'utilisateur légitime (*i.e.* erreur commise lors d'un test imposteur).

Ces deux types d'erreurs n'ont pas toujours la même incidence en terme de sécurité et de qualité de service. La fausse acceptation peut être très pénalisante dans le cas d'une application requérant un niveau de sécurité élevé. Il n'est pas tolérable par exemple que n'importe qui puisse accéder à des informations personnelles, bancaires ou même de type secret défense. Le faux rejet peut également pénaliser des applications où l'utilisateur ne peut se permettre de perdre du temps en tentant de s'authentifier à plusieurs reprises. C'est le cas, par exemple, pour des services de secours d'urgence. Un utilisateur du système doit pouvoir être reconnu par le système dans les meilleurs délais.
Nous verrons dans la suite que les taux de fausses acceptations et de faux rejets sont liés et que le réglage d'un système de vérification d'identité doit tenir compte du coût de chaque type d'erreurs dans le cadre de l'application visée.

2.3.2 Points de fonctionnement et représentations des performances d'un système de vérification d'identité

Le module de décision décrit dans la partie 2.2.3 reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et si il est inférieur au seuil, l'utilisateur est rejeté.

Le choix d'un seuil a une incidence directe sur les performances du système. Pour un système idéal, les scores obtenus par les clients seront tous plus élevés que les scores obtenus par les imposteurs. Dans ce cas, le seuil à fixer se situe entre le score imposteur le plus élevé et le score client le plus faible, assurant ainsi une authentification parfaite (cf. figure 2.3).

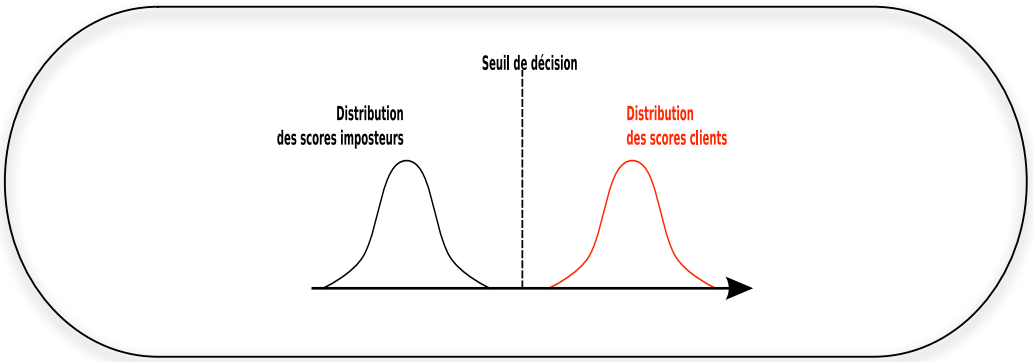


FIG. 2.3: *Répartition des scores clients et imposteurs et seuil de décision d'un système parfait*

En pratique, les distributions des scores clients et imposteurs se superposent partiellement. Ce cas ne permet pas une authentification parfaite et des erreurs de type faux rejets et fausses acceptations apparaissent. Le choix du seuil influe sur le taux de faux rejets et de fausses acceptations. Cet effet est illustré par la figure 2.4.

Pour un seuil de décision fixé les taux de faux rejet $p(FR)$ et de fausse acceptation $p(FA)$ que l'utilisation de ce seuil occasionne peuvent être calculés a posteriori.

$$p(FA) = \frac{nombre\ de\ tests\ dont\ résulte\ une\ fausse\ acceptation}{nombre\ de\ tests\ imposteurs} \tag{2.1}$$

$$p(FR) = \frac{nombre\ de\ tests\ dont\ résulte\ un\ faux\ rejet}{nombre\ de\ tests\ clients} \tag{2.2}$$

À chaque valeur de seuil est associé un couple $(p(FA), p(FR))$ et l'ensemble des couples obtenus peut être représenté sous la forme d'une courbe ROC (Receiver Operating Characteristic) (Oglesby, 1995) ou, comme sur la figure 2.5, d'une courbe DET (Detection Error Tradeoff). La courbe DET diffère principalement de la courbe ROC par l'échelle basée sur une distribution normale qui se substitue à l'échelle linéaire (Martin et al., 1997).

Selon l'application visée, le coût d'un faux rejet ou d'une fausse acceptation ne sont pas les mêmes. L'estimation des performances d'un système de vérification pour une application spécifique se doit de tenir compte des coûts différents de ces deux types

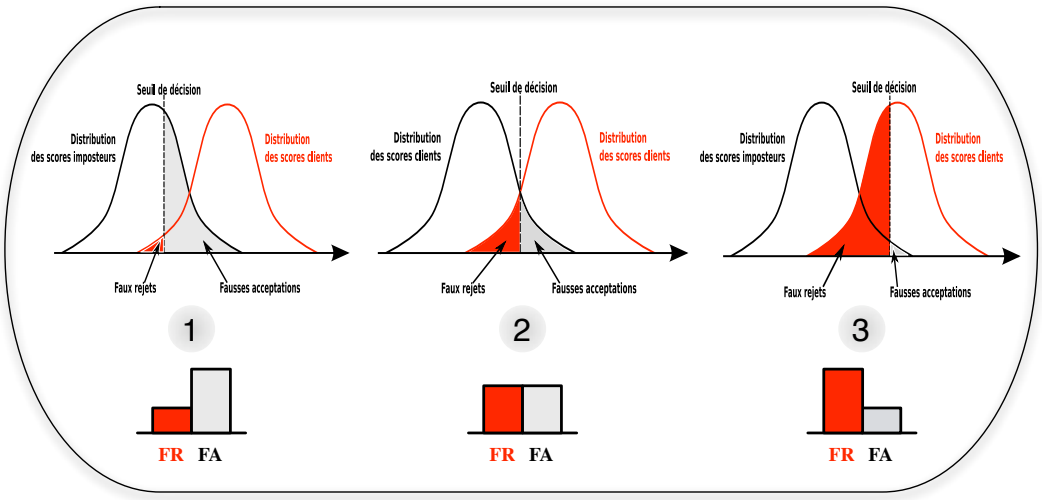


FIG. 2.4: *Influence du seuil de décision sur les erreurs d'un système de reconnaissance biométrique*
1 - Seuil de décision choisi dans le but de réduire le nombre de faux rejets
2 - Seuil de décision choisi pour obtenir autant de faux rejets que de fausses acceptations (EER)
3 - Seuil de décision choisi dans le but de réduire le nombre de fausses acceptations

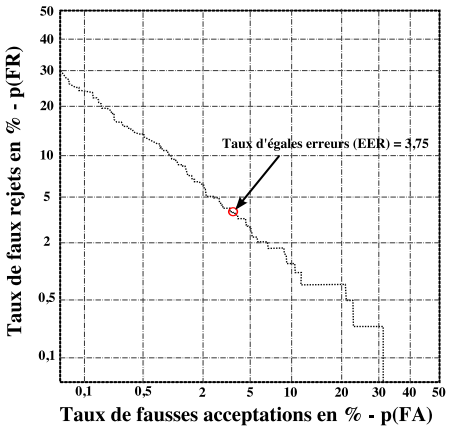


FIG. 2.5: *Exemple de représentation des performances d'un système de vérification d'identité par une courbe DET*

d'erreurs. La fonction de coût de détection (Detection Cost Function - DCT), détaillée dans l'équation 2.3.2, est souvent utilisée pour résumer les performances des systèmes de vérification biométrique d'identité (Martin et Przybocki, 2000).

$$DCF = \text{Coût}(FR)\ P(client)\ p(FR) + \text{Coût}(FA)\ P(imposteur)\ p(FA) \qquad (2.3)$$

où $\text{Coût}(FR)$ et $\text{Coût}(FA)$ sont respectivement les coûts d'un faux rejet et d'une fausse

acceptation pour l'application choisie. $P(client)$ et $P(imposteur)$ sont les probabilités qu'un client ou un imposteur utilise le système.
Les performances du système sont couramment résumées, indépendamment du contexte applicatif, par un point particulier de ces courbes pour lequel les taux de fausses acceptations et de faux rejets ont la même valeur : le taux d'égal erreur (Equal Error Rate - EER).

En pratique, le seuil de décision doit être déterminé avant utilisation du système. Il est choisi de manière empirique, après calibration sur des données de développement, pour répondre aux spécifications du système concernant le couple $(p(FA), p(FR))$. Mariethoz (2006) présente une étude très complète sur la présentation des mesures de performances par utilisation des courbes ROC et des points de fonctionnement spécifiques. Il propose une courbe spécifique, courbe de performances espérées (Expected Performance Curve - EPC). Cette représentation bien que très intéressante, n'est pas utilisée dans ce document car très peu présente dans la littérature

La plupart des systèmes de vérification du locuteur état-de-l'art intègrent une étape de normalisation avant la prise de décision. Cette étape permet de prendre en compte la variabilité des scores obtenus lors des différents tests. La variabilité provient principalement des différences de locuteurs, contenus phonétiques ou durées d'enregistrement d'un test à l'autre. La variabilité intra-locuteur doit également être prise en compte afin de fixer le seuil de décision du système automatique. La plupart des approches état-de-l'art reposent sur une normalisation des distributions de scores imposteurs. Le contexte expérimental de nos travaux (explicité dans le chapitre 6) ne nous a pas permis d'utiliser de normalisation, nous inciterons cependant le lecteur à se rapporter aux travaux de (Higgins et al., 1991), (Li et Porter, 1998), (Matsui et Furui, 1993), (Reynolds, 1996), (Auckenthaler et al., 2000) and (Fredouille et al., 1999).

Deuxième partie

La parole en biométrie

Introduction

COMME nous l'avons vu dans l'introduction à la biométrie (partie I), il existe de nombreuses façons de décrire et reconnaître un individu, différents types d'informations et différentes modalités. Aux vues des contraintes inhérentes aux différentes biométries et des performances de chacune, il apparaît que l'usage de la parole présente de nombreux avantages.

La parole est un vecteur de communication naturelle aux être humains. Son utilisation en biométrie est relativement bien acceptée du fait de son caractère peu intrusif, et les contraintes ergonomiques imposées aux utilisateurs de systèmes vocaux sont mineures. Les éléments nécessaires au déploiement d'interfaces matérielles (microphones, chaînes de traitement, etc.) sont de plus très répandus et peu coûteux. Comme nous le détaillerons dans la suite de cette partie, les performances des biométries liées à la parole fournissent un niveau de sécurité assez élevé.

L'usage de la parole en tant que modalité biométrique permet de tirer parti simultanément de trois informations.

- La parole résulte d'un processus complexe qui la rend porteuse d'une information sur les différents organes qui participent à sa production.
- Les signaux de parole contiennent aussi une information sur le vécu ou l'environnement de l'individu car la parole est un phénomène qui résulte d'un apprentissage.
- La parole est également vecteur de communication. Le message qu'elle transmet peut par exemple être utilisé comme mot de passe pour vérifier l'identité du locuteur (cf. partie 1.1.2).

Les informations liées directement aux organes de production de la parole présentent deux avantages majeurs : elles sont difficilement falsifiables et elles font partie intégrante de l'individu.

Les paragraphes suivant décrivent brièvement la production de la parole afin de mettre en lumière la nature multi-modale de la parole.

La production de la parole est un processus complexe dans lequel sont impliqués de nombreux organes dont les principaux sont représentés sur la figure 2.6. Le contenu fréquentiel du signal acoustique de parole produit par un locuteur est fortement dépendant des caractéristiques morphologiques de son appareil phonatoire. Celui-ci peut être divisé en quatre parties : le générateur, le vibrateur, le résonateur et les modulateurs (Haton et al., 2006).

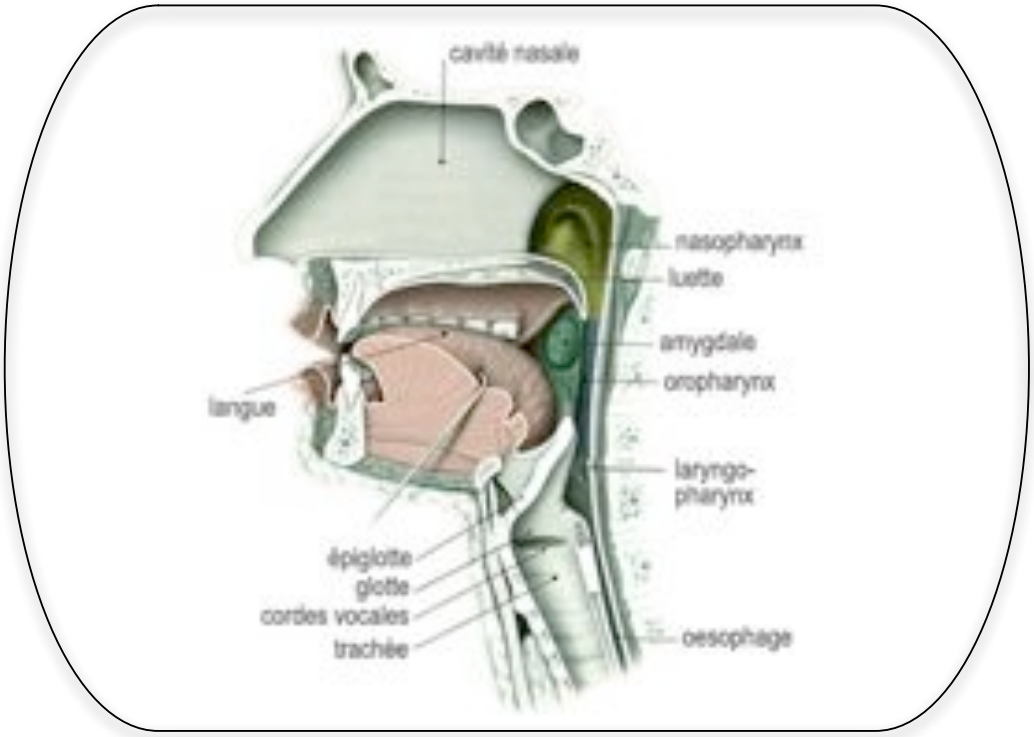


FIG. 2.6: *Schéma de l'appareil phonatoire humain. (`http://lecerveau.mcgill.ca/`)*

Le générateur : c'est l'air expulsé des poumons qui est le moteur de la parole. Cet air traverse l'appareil phonatoire comme un instrument à vent et crée la pression nécessaire à la génération d'un signal acoustique.

Le vibrateur : L'air expulsé des poumons traverse la trachée pour arriver dans le larynx où se trouvent les cordes vocales. Les cordes vocales sont une paire de muscles dont la longueur moyenne se situe entre 20 et 25 millimètres. Cette longueur varie cependant d'un individu à l'autre. L'air traversant le larynx met en vibration les cordes vocales. La fréquence de vibration des cordes vocales est modulée en fonction de leur degré de contraction. Le locuteur peut ainsi moduler la hauteur des sons qu'il émet.

Le résonateur : Ces vibrations sont modifiées par le passage de l'air dans les différentes cavités qui composent le pharynx mais aussi dans les fosses nasales, la bouche et le larynx avec qui il communique. Ces résonateurs influent sur le son en atténuant

certaines fréquences et en en amplifiant d'autres. La forme et le volume de ces cavités, spécifiques au locuteur, modifient fortement le son produit.

Les modulateurs : Enfin les organes *modulateurs* que sont la langue, les lèvres et la mâchoire sculptent le son pour produire les phonèmes qui composent la parole. La position de ces différents organes est le mécanisme final qui permet la production de parole articulée.

Certains organes impliqués dans la production de la parole se trouvent être en partie visibles pour un observateur extérieur (lèvres, langue, position des mâchoires). La configuration des organes modulateurs et leurs mouvements fournissent donc une information visuelle caractéristique de la parole produite (Luettin et Thacker, 1997).

Ainsi la parole est un phénomène temporel bi-modal structuré, dont les manifestations acoustiques et visuelles peuvent être observées au cours du temps. L'utilisation de la parole en biométrie rend possible l'acquisition simultanée du signal acoustique de parole et du signal vidéo correspondant. Cette acquisition est d'autant plus facile à réaliser que le matériel nécessaire (microphone et caméra) est bon marché et présent sur un grand nombre d'appareils commercialisés tels que les téléphones cellulaires, PDA, ordinateurs de bureau ou portables...

Cette partie présente les modalités audio et vidéo qui peuvent être utilisées lorsqu'un utilisateur s'adresse au système d'authentification. Les chapitres 3 et 4 présentent l'état-de-l'art des biométries audio et visuelle, les particularités relatives aux différents signaux utilisés, les approches existantes et leurs performances. Le chapitre 5 montre l'intérêt qu'il peut y avoir à exploiter conjointement les informations provenant des signaux audio et vidéo. Il contient également une analyse des différentes approches développées dans ce but. La structure temporelle des signaux considérés fera l'objet d'une attention particulière tout au long de cette partie.

Chapitre 3

Vérification automatique du locuteur

Sommaire

Résumé

Ce chapitre présente dans un premier temps les informations contenues dans le signal de parole et isole celles qui sont utiles à la reconnaissance du locuteur. Il décrit ensuite les approches existantes dans le domaine de la reconnaissance automatique du locuteur en distinguant, au sein de ces méthodes, celles qui tirent partie de la structure temporelle du signal.

Introduction

La voix est porteuse d'informations variées. Émission de sons structurée, la parole humaine est essentiellement un vecteur de communication. À ce titre, un signal de parole est généralement porteur d'un message à destination d'une autre personne. La parole peut cependant contenir de nombreuses informations telles que la langue parlée par le locuteur, son identité ou même des indications sur son âge ou son état émotionnel.

Les systèmes de reconnaissance automatique de la parole sont utilisés pour transcrire le message porté par le signal vocal. Il peut s'agir de reconnaître un lexique limité et déterminé à l'avance (Furui, 1986) ou de transcrire un message au vocabulaire plus large (Aubert, 2002). Ce message est prononcé dans un contexte qui, s'il est connu, peut apporter une information sur le message porté par le signal de parole.
Les systèmes de reconnaissance de la parole, cherchent à extraire du signal acoustique une information, a priori, indépendante du locuteur. Ils sont souvent perturbés par les variations inter-locuteurs.

La reconnaissance de la parole requiert la plupart du temps une étape préalable de reconnaissance de la langue parlée. Des approches automatiques comme celles décrites dans (Zissman et Singer, 1994), (Singer et al., 2003) ou (Rouas et al., 2005) permettent d'identifier la langue dans laquelle s'exprime un locuteur parmi un panel de langues connues par ces systèmes.

La reconnaissance des émotions, constitue également, depuis quelques années, un domaine de recherche en plein essor. L'analyse du signal de parole peut apporter des informations sur la volonté ou les émotions ressenties par le locuteur (Adami, 2007) . L'âge d'un locuteur peut également être estimé d'après sa voix (Minematsu et al., 2003). De la même façon, certaines pathologies influent sur les organes de production de la parole et peuvent être détectées par des systèmes automatiques (Sáenz-Lechón et al., 2006), (Pouchoulin et al., 2007).

Enfin, l'identité est l'une des informations les plus communément extraites du signal de parole. L'être humain est naturellement capable de reconnaître la voix d'une personne qui lui est proche. Il apparaît donc intuitif, dans un contexte de sécurisation des communications et des données, que des systèmes automatiques soient utilisés afin de reconnaître l'identité d'un locuteur d'après sa voix.

Le signal sonore, tel qu'il est utilisé par les systèmes automatiques, est aussi porteur d'une information relative au matériel qui compose la chaîne d'enregistrement et de transmission ou à l'environnement du locuteur. Ces informations sont généralement perçues comme nuisibles, car elles dégradent fortement les performances dans les différentes tâches des systèmes automatiques.

Dans ce chapitre, nous présentons les informations contenues dans le signal de parole et, plus particulièrement, celles qui sont utiles à la reconnaissance du locuteur (cf. section 3.1). Les paramètres acoustiques et dynamiques extraits du signal sonore constituent le matériau utilisé par les systèmes de reconnaissance automatique du locuteur (RAL). Considérant la dimension temporelle du signal de parole, nous distinguerons deux approches. Dans la section 3.2, nous supposerons, comme il est courant dans le domaine, que les paramètres acoustiques provenant du signal de parole peuvent être traités indépendamment de leur organisation temporelle. Nous verrons ensuite (cf. section 3.3) que l'ordre des vecteurs acoustiques - la structure temporelle du signal de parole - peut être prise en compte au sein des systèmes de reconnaissance du locuteur et que cette information structurale améliore sensiblement les performances de ces approches.

3.1 Extraction d'information du signal de parole

De par sa complexité, l'information portée par le signal de parole ne peut, quelle que soit la tâche considérée, être utilisée dans sa totalité. C'est pourquoi l'utilisation d'un système automatique nécessite une sélection préalable des informations à exploiter pour la tâche qui lui est confiée.

3.1.1 Utilisation des différentes informations en RAL

En reconnaissance automatique du locuteur, seules les informations présentant une forte variabilité inter-locuteurs permettent de discriminer les différents individus. À l'inverse, les informations dont la variabilité intra-locuteur est élevée rendent la tâche de RAL plus complexe.

Les informations les plus utilisées en RAL, du fait de leur fort potentiel discriminant, sont des informations acoustiques obtenues périodiquement par une analyse fréquentielle ou temporelle du signal. D'autres informations présentes dans le signal de parole et citées précédemment peuvent s'avérer discriminantes dans le cadre de la reconnaissance du locuteur. Des paramètres tels que la prosodie ou la fréquence fondamentale, par exemple, contiennent une information spécifique du locuteur (Ferrer et al., 2007), (Sönmez et al., 1997).

Les systèmes de reconnaissance du locuteur utilisent des représentations du signal de parole dans lesquelles le bruit et la redondance ont été réduits afin de ne conserver que les informations considérées comme utiles à la tâche spécifiée. Ce traitement est appelé *paramétrisation*.

3.1.2 Paramétrisation acoustique

Le signal de parole, qui résulte de la conversion d'une onde acoustique en un signal électrique par un microphone, est un signal temporel unidimensionnel. Le système de paramétrisation utilise, en entrée, le signal de parole et retourne, en sortie, des vecteurs de paramètres à intervalle de temps régulier (Reynolds, 1994). Ces vecteurs de paramètres sont calculés sur une fenêtre temporelle glissante dont la durée varie généralement entre 20 et 50 milli-secondes. Il arrive souvent que les fenêtres temporelles utilisées pour extraire deux segments de signal consécutifs se recouvrent partiellement. La fréquence couramment utilisée pour l'extraction de ces paramètres est 100Hz. Ces caractéristiques ainsi que la dimension des vecteurs de paramètres acoustiques varient selon l'application et le type d'information extraite.

Les coefficients cepstraux de prédiction linéaire

La prédiciton linéaire est une technique, issue de l'analyse de la production de la parole (Boite et al., 2000), qui repose sur l'hypothèse que le signal vocal est le résultat de l'excitation d'un filtre auto-régressif (AR) par un train pseudo périodique d'impulsions unitaires. Il en résulte qu'un échantillon de parole émis à l'instant t peut être estimé à partir des échantillons de parole précédents. Les coefficients de prédiction linéaire (Linear Prediction Coefficients - LPC) (Tremain, 1982) obtenus sous cette hypothèse peuvent être utilisés pour calculer des coefficients cepstraux LPCC (Linear Prediction Cepstral Coefficients - LPCC).

Les paramètres MFCCs (Mel Frequency Cepstral Coefficient) :

Ces coefficients font partie des paramètres les plus couramment utilisés en traitement de la parole (Furui, 1981), (Bimbot et al., 2004) et notamment en reconnaissance du locuteur. Ils sont obtenus par une analyse fréquentielle du signal et l'utilisation de bancs de filtres qui permettent de rapprocher l'information extraite de celle perçue par une oreille humaine. La chaîne d'extraction de ces paramètres est décrite par la figure 3.1. Le signal de parole est analysé localement à l'aide d'un fenêtrage temporel (souvent

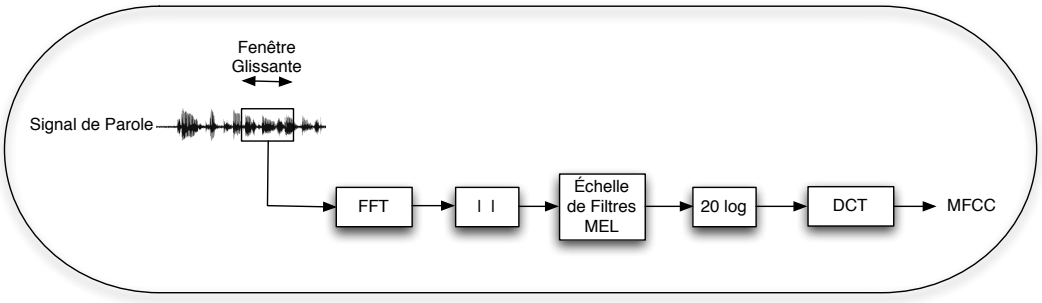


FIG. 3.1: *Représentation d'un système de paramétrisation produisant des coefficients cepstraux avec une échelle MEL (MFCC)*

Hanning ou Hamming afin de réduire les effets de bord qu'occasionne une fenêtre rectangulaire). La longueur de la fenêtre glissante (20-30 milli-secondes) est choisie pour respecter l'hypothèse de stationnarité. Le décalage des fenêtres temporelles utilisées pour extraire deux segments consécutifs de signal est choisi de manière à ce que ces fenêtres se recouvrent en partie. Au segment de signal prélevé est ensuite appliquée une transformée de Fourier rapide (Fast Fourier Transform - FFT). Le module de son spectre est filtré par un banc de filtres qui permet de réduire la dimension du vecteur spectral en calculant la moyenne du spectre sur la bande de fréquence correspondant à chacun des filtres. Les fréquences centrales de chaque filtre sont fixées par l'échelle de MEL. Le logarithme de ces valeurs est calculé et multiplié par 20 pour obtenir l'enveloppe spectrale en décibels. La dernière étape de la paramétrisation consiste à appliquer une transformée en cosinus discrète (DCT) d'où résultent les coefficients cepstraux (MFCC). La transformée en cosinus discrets est utilisée ici pour sa capacité à décorréler les données.
Différents bancs de filtres peuvent être utilisés pour analyser le signal de parole. Les échelles perceptives de MEL et Bark, inspirées de la perception humaine, sont parmi les plus utilisées.

D'autres paramètres acoustiques utilisent les caractéristiques de production ou de perception humaines. L'analyse par Prédiction Linéaire Perceptuelle (Perceptual Linear Prediction - PLP) (Hermansky, 1990), par exemple, repose sur un modèle de perception humaine de la parole. Cette analyse qui tient compte du traitement par bandes de fréquences de l'oreille humaine peut être renforcée par une analyse spectrale relative (RelAtive SpecTrAl - RASTA) qui simule l'insensibilité de l'appareil phonatoire humain aux variations temporelles lentes (Hermansky et al., 1991), (Hermansky et al., 1992).
L'approche VTLN (Vocal Tract Length Normalisation) décrite dans (Eide et Gish, 1996) modélise le conduit vocal du locuteur et peut être utilisée pour extraire des paramètres caractéristiques de ce locuteur.

3.1.3 Utilisation d'une information dynamique

L'information dynamique contenue dans le signal de parole est utile à la reconnaissance du locuteur (Furui, 1981). Il existe de nombreuses approches intégrant ces aspects dynamiques.

Ajouter une information dynamique à partir de coefficients déjà extraits :

La plus répandue consiste à utiliser les variations immédiates des paramètres acoustiques en calculant leurs dérivées temporelles premières et secondes (cf. figure 3.2). Cette méthode a été introduite par Furui (1981) et présentée de façon détaillée dans (Fredouille, 2000).

Le calcul de ces dérivées premières (Δ) et secondes ($\Delta\Delta$) est simplifié par l'utilisation

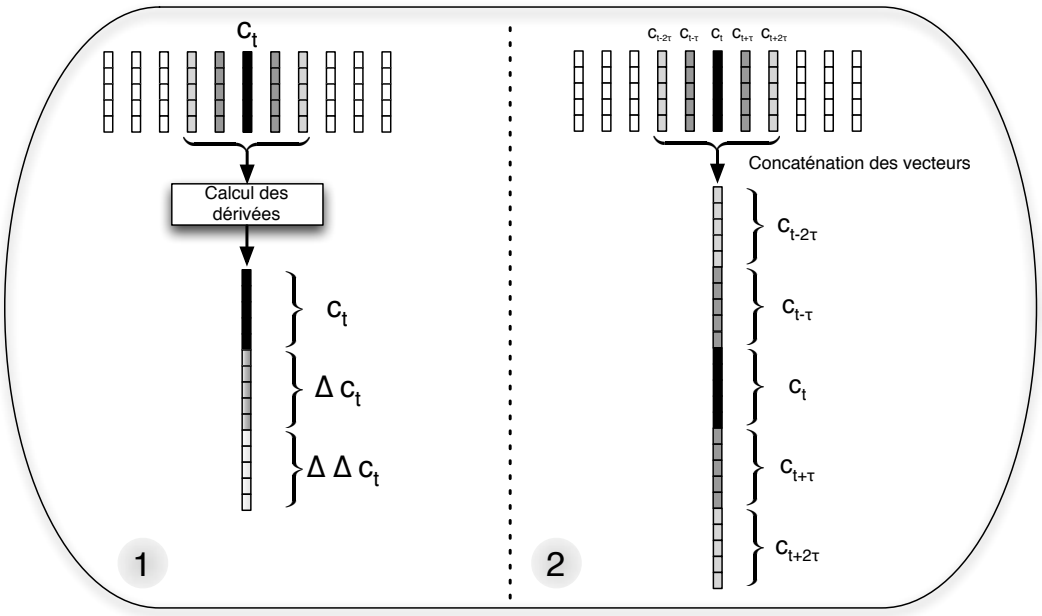


FIG. 3.2: *Différentes méthodes d'utilisation de l'information dynamique contenue dans une fenêtre temporelle glissante. Représentation de deux approches couramment utilisées en RAL.*
L'approche -1- est un calcul explicite de données dynamiques à partir des trames de la fenêtre temporelle.
L'approche -2- se contente de concaténer les trames de cette fenêtre pour augmenter la quantité d'information extraite à un moment t.

de l'approximation polynomiale suivante :

$$\frac{\delta c(t)}{\delta t} \approx \Delta c(t) = \frac{\sum_{k=-K}^{K} k.c(t+k)}{\sum_{k=-K}^{K} k^2} \tag{3.1}$$

où c est le coefficient à dériver, δc sa dérivée première à l'instant t et où les coefficients Δ sont calculés sur une fenêtre temporelle de longueur $2K+1$ trames. Le rapport entre la variable K et la longueur de la fenêtre glissante utilisée pour l'extraction des paramètres a été l'objet de nombreuses études (Furui, 1981), (Soong et Rosenberg, 1988). Le calcul des $\Delta\Delta$ utilise la même approximation à partir des coefficients Δ.

Une autre approche consiste à prendre en compte l'information extraite sur une fenêtre temporelle plus large, sans calculer explicitement de coefficients dynamiques. La concaténation des vecteurs de paramètres représentée sur la figure 3.2 illustre cette technique.

Les paramètres obtenus par ces deux approches résultent d'une analyse à court terme des signaux et ne donnent aucune indication à long terme sur le signal. Cette paramétrisation constitue pourtant souvent la seule information utilisée par les systèmes état-de-l'art en RAL (Bimbot et al., 2004).

Paramètres prenant directement en compte une information temporelle :

D'autres approches permettent d'extraire directement une information sur la structure temporelle du signal de parole. Les paramètres TRAPS (Temporal Patterns) présentés dans (Hermansky et Sharma, 1999) et (Schwarz et al., 2006) sont obtenus par utilisation de réseaux de neurones qui extraient la structure locale du signal temporel.

3.1.4 Sélection des informations (speech activity detection)

Une fois le signal de parole paramétrisé, les trames utiles au processus de reconnaissance du locuteur doivent être soigneusement sélectionnées. En effet, les performances des systèmes de RAL sont très dépendantes de cette sélection comme l'ont montré notamment Besacier et al. (2000) ou Scheffer (2006). Les trames sont généralement étiquetées *parole* ou *non-parole* par un module de détection d'activité (Voice Activity Detection - VAD), même si la classification tend à distinguer les trames utiles ou non à la reconnaissance du locuteur. L'étiquette *non-parole* correspond plus exactement au silence ou au bruit. Seules les trames identifiées *parole* sont utiles à la tâche de RAL.

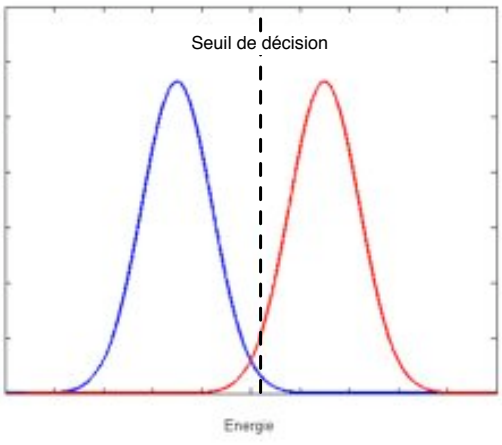


FIG. 3.3: *Modélisation des distributions énergétiques des trames parole (Gaussienne d'énergie la plus élevée) et non-parole (Gaussienne d'énergie la plus faible). Le seuil utilisé pour la classification de trames parole et non-parole est fixé en fonction des paramètres de ces Gaussiennes.*

Une façon de sélectionner ces trames de *parole* consiste à utiliser l'énergie, en faisant l'hypothèse que les trames les plus énergétiques, correspondant principalement aux zones stables des voyelles et aux zones pour lesquelles le rapport signal à bruit est élevé, sont les plus intéressantes.
Une façon d'obtenir la classification des trames *parole non-parole*, consiste à utiliser un modèle de mélange de gaussiennes. Dans l'exemple de la figure 3.3, la distribution énergétique des trames est réalisée par un mélange de deux gaussiennes (cf. chapitre 7). La Gaussienne de plus faible énergie modélise les trames de *non-parole* et la Gaussienne de plus haute énergie les trames de *parole*. Une fois ces Gaussiennes apprises, un seuil est calculé pour attribuer les trames à l'une ou l'autre des classes. Cette méthode est simple à mettre en oeuvre et obtient de bons résultats sur des séquences courtes (quelques secondes). Elle ne prend cependant pas en compte les variations locales qui sont moyennées et ne peut pas être utilisée de façon simple pour un traitement en flux.

D'autres classifieurs comme les machines à vecteur support (Enqing et al., 2002), présentées dans la partie 3.2.2, ou les réseaux de neurones (Ikedo, 1998) peuvent être utilisés pour la détection des trames de *parole*.
Une variante de cette méthode, basée sur le standard Aurora (Technical Specification Digital, European Telecommunications Standards Institute (ETSI), 2005) et utilisable en ligne, est présentée dans (Preti, 2008). Cette approche permet de sélectionner les trames utiles en parallèle de l'extraction de paramètres.

3.2 Vérification du locuteur non-structurale

La vérification automatique du locuteur, comme toutes les biométries, utilise généralement l'organisation décrite dans le chapitre 2. Une modélisation de chaque client du système est obtenue à partir d'enregistrements réalisés lors de la phase d'entraînement. Une mesure de similarité est ensuite calculée entre un modèle de client et une séquence de test avant d'être traitée par un module de décision.
Parmi les approches non-structurales de RAL, nous distinguerons deux grandes catégories : les approches génératives et les approches vectorielles. La suite de cette partie donnera un bref aperçu des principales approches existant en RAL.

3.2.1 Approche générative

L'approche générative tend à modéliser la distribution statistique qui a pu produire les vecteurs de paramètres de la séquence d'apprentissage. L'ordre des statistiques peut être variable mais les systèmes état-de-l'art actuels utilisent des statistiques d'ordre 2 qui permettent de représenter les variations des paramètres acoustiques.

Modélisation Gaussienne

Les approches génératives en reconnaissance du locuteur reposent sur l'hypothèse qu'il existe une bijection entre l'ensemble des locuteurs et l'espace des fonctions de densité de probabilité. Cela signifie que les vecteurs de paramètres provenant d'un locuteur suivent une loi de probabilité propre à ce locuteur.
Dans (Bimbot et al., 1995), les auteurs utilisent une distribution Gaussienne multi-dimensionnelle pour représenter chaque locuteur. Considérons que chaque vecteur de paramètres extrait d'un signal de parole est une réalisation d'une variable aléatoire multi-dimensionnelle.
Dans cette approche, les locuteurs sont représentés par une loi Gaussienne, un triplet (μ, Σ, X) où μ est le vecteur moyen de la Gaussienne et Σ la matrice de covariance estimée à partir de la séquence acoustique d'apprentissage X. La densité d'une distribution normale pour une variable X à d dimensions est exprimée par :

$$\mathcal{N}(X, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \, exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right] \tag{3.2}$$

Les Méthodes Statistiques du Second Ordre (MSSO) ont l'avantage d'offrir une modélisation simple et donc peu de paramètres à estimer. Elles fournissent de bons résultats sur des courtes durées (3s) mais ne permettent pas de modéliser, dans l'espace acoustique, les variations locales.

Modèles de mélanges de Gaussiennes (GMM)

Une meilleure approximation de la distribution correspondant au locuteur est obtenue en combinant plusieurs modèles Gaussiens dans une somme pondérée. Ces mélanges de Gaussiennes multi-variées (Gaussian Mixture Model - GMM), introduits en reconnaissance du locuteur par Reynolds et Rose (1995), constituent à l'heure actuelle l'essentiel des méthodes état-de-l'art et sont à la base des systèmes offrant les meilleurs performances lors des évaluations internationales NIST-SRE (Doddington et al., 2000). Leur utilisation, et principalement l'approche GMM/UBM, seront décrits de façon détaillée dans le chapitre 7. Cependant, comme les MSSO, les modèles GMMs n'exploitent pas la structure temporelle du signal de parole. Seule l'information dynamique à court terme, calculée lors de la paramétrisation, est prise en compte. Les méthodes utilisant des mélanges de Gaussiennes obtiennent de bonnes performances en reconnaissance du locuteur indépendante du texte pour des durées d'enregistrement moyennes à longues lors des phases d'enrôlement et de test (Doddington et al., 2000).

Modèles de mélanges de segments

Starpert et Mason (2001) associent l'approche générative GMM avec l'approche DTW (Dynamic Time Warping), prenant en compte la structure temporelle du signal de parole (cf. section 3.3.3), en créeant des *Modèles de Mélanges de Segments* (SMM). Ces mo-

dèles SMMs, utilisent la structure des modèles GMMs standards au sein de laquelle chaque composante du modèle GMM est remplacée, par une courte séquence temporelle λ_i appelée segment (cf. figure 3.4).

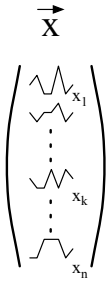


FIG. 3.4: *Représentation d'un segment utilisé pour l'apprentissage des modèles SMMs.*

La mesure de similarité calculée entre un segment $\overrightarrow{X}$ et la composante λ_i du modèle SMM est donnée par :

$$b_i(\overrightarrow{X}) = \ln\left[\prod_k^K |\Sigma_i|^{-\frac{1}{2}}\, e^{(-\frac{1}{2}\, d_i)}\right] \tag{3.3}$$

où $\prod_k^K |\Sigma_i|^{-\frac{1}{2}}$ est le produit de la matrice diagonale de covariances le long du chemin calculé par le DTW. K est la longueur des segments considérés, donnée en nombre de vecteurs, d_i est la mesure DTW calculée entre le segment d'entrée $\overrightarrow{X}$ et la composante λ_i du modèle SMM. Le terme d_i est calculé par :

$$d_i = W_k^K((\overrightarrow{X_k} - \overrightarrow{\lambda_{i,k}})^T \Sigma_i^{-1} (\overrightarrow{X_k} - \overrightarrow{\lambda_{i,k}})) \tag{3.4}$$

où W est un terme de normalisation considérant la déformation du chemin DTW.
Le résultat de cette comparaison, $b_i(X)$, est équivalent, dans le cas d'un modèle GMM standard (cf. chapitre 7.1.1), à la vraisemblance calculée entre un vecteur acoustique et une distribution Gaussienne. La similarité entre le segment X et le modèle de mélange de segments λ est donnée par :

$$s(X|\Lambda) = \sum_{i=0}^{n} w_i\, b_i(X) \tag{3.5}$$

3.2.2 Approche vectorielle

L'approche vectorielle représente les locuteurs ou les séquences de parole par un ou plusieurs points dans un espace de grande dimension. Les méthodes vectorielles présentent généralement l'avantage d'être symétriques : les représentations des locuteurs et des séquences de test sont de même nature. La diversité des espaces de représentation et des mesures de distances qui leur sont associées font des approches vectorielles

un champs ouvert de la recherche en RAL. Parmi les approches les plus courantes, nous détaillons ici la quantification vectorielle, les modèles d'ancrage et les machines à vecteurs supports (SVM).

Quantification vectorielle

La quantification vectorielle (Vector Quantization -VQ) (Soong et al., 1985), (Mason et al., 1989) est une méthode provenant du domaine de la compression d'images. En effet, la quantité de données acquises pour un locuteur donné est trop importante pour être utilisée en l'état pour une tâche de vérification.

La quantification vectorielle consiste à représenter au mieux un séquence de parole $(a_0 \ldots a_n)$ à partir d'un ensemble $\{d_k\}_{k \in K}$ de vecteurs de paramètres appelé dictionnaire de quantification. La séquence de parole est ainsi approximée par une séquence $(d_{k_0} \ldots d_{k_n})$ où $k_j \in K$.
Cette méthode de compression avec perte est adaptée à la reconnaissance du locuteur puisqu'elle permet de réduire le nombre de vecteurs d_k à stocker pour représenter n'importe quelle séquence de parole. L'erreur de quantification est donnée par :

$$D = \sum_i d(a_i, d_{k_i}) \tag{3.6}$$

où $d(a_i, b_{k_i})$ est la distance entre les vecteurs a_i et d_{k_i}. Cette erreur peut cependant être minimisée par le choix d'un dictionnaire adapté. Une technique utilisée en compression pour minimiser cette erreur consiste à choisir les éléments du dictionnaire de quantification parmi les données à compresser.

En reconnaissance du locuteur, les données d'enrôlement de chaque locuteur sont utilisées pour obtenir un dictionnaire de quantification. Un tel dictionnaire est représentatif des données provenant de ce locuteur et ainsi particulièrement adapté à leur compression.
Lors de la phase de test, chaque trame a_i de la séquence $(a_0 \ldots a_n)$ est remplacée par l'élément d_{k_i} du dictionnaire de quantification, dont elle est le plus proche.
L'erreur de quantification, dont l'expression est donnée par l'équation 3.6, est utilisée, en reconnaissance du locuteur, pour mesurer la similarité entre la référence d'un locuteur et les données de test. Plus le dictionnaire est adapté à la compression, plus les données décompressées sont proches des données originales et plus l'erreur de quantification est faible.

Comme en compression, la rapidité et les performances de cette méthode dépendent fortement de la taille des dictionnaires choisis. Un dictionnaire de taille importante ralenti le calcul de distance mais améliore la description du locuteur.

Modèles d'ancrage

Les modèles d'ancrage ont été introduits par Merlin et al. (1999) afin d'accélérer le processus de RAL utilisant une approche générative (notamment en identification). Le principe en est simple, le modèle de locuteur n'est plus appris directement mais il devient une représentation relative du locuteur dans l'espace des modèles (Mami et Charlet, 2004). Cette représentation est relative à des modèles existants et calculés précédemment en utilisant une quantité importante de données : les modèles d'ancrage. Cette méthode est à rapprocher des EigenVoices présentés dans la section 7.2.3. Dans le cas des modèles d'ancrage, la représentation d'un locuteur est située dans l'espace des scores et correspond à un vecteur dont la dimension est le nombre de modèles d'ancrage. Chaque composante de ce vecteur est le score obtenu par comparaison d'un modèle d'ancrage avec les données d'apprentissage du locuteur. Ainsi, la représentation d'un locuteur par la méthode des modèles d'ancrage n'est pas une modélisation acoustique de ce locuteur mais une fonction des modèles d'ancrages utilisés.

Durant la phase de test, la séquence de test peut être projetée dans la base des modèles d'ancrage. Le score de cette séquence de test pour un locuteur connu est ensuite calculé par une mesure de distance entre la projection M de la séquence dans la base des modèles d'ancrage et le vecteur L du locuteur, calculé durant la phase d'enrôlement. Il est également possible de projeter chaque vecteur de paramètres de la séquence de test, indépendamment, dans la base des modèles d'ancrage pour calculer ensuite la similarité avec le modèle du locuteur testé.

Quelle que soit la mesure de distance choisie, la comparaison d'une séquence de test avec l'ensemble des modèles de locuteurs connus reste relativement rapide. La difficulté liée à cette approche provient du choix des modèles d'ancrages. Un nombre restreint de ces modèles doit être choisi de façon à maximiser la variance autour des axes qu'ils définissent.

Machines à Vecteurs Support (SVM)

Les SVMs sont des classifieurs binaires développés pour permettre la séparation de données complexes dans des espaces de grandes dimensions (Wan et Campbell, 2000), (Fine et al., 2001). En tant que classifieurs binaires, ils sont très bien adaptés à la tâche de vérification du locuteur.

Classification binaire :
L'approche par SVMs consiste, étant donnée un jeu de paramètres d'apprentissage $(X_t, Y_t)_{t=1..T} \in \mathbb{R}^d \times \{-1, 1\}$, à trouver une fonction $f : \mathbb{R}^d \rightarrow \{-1, 1\}$ telle que pour

tout élément (X, Y) de l'ensemble d'apprentissage :

$$\begin{cases} f(X) = \mathbf{W}^T X + \mathbf{b} \\ Y = sgn(f(X)) \end{cases} \tag{3.7}$$

où $\mathbf{W} \in \mathbb{R}^d$ et $\mathbf{b} \in \mathbb{R}$. Cette fonction est un séparateur linéaire. Si chaque échantillon X d'apprentissage est considéré dans un espace de dimension $\mathbb{R}^d$ divisé en deux classes C et I par un hyperplan, la fonction f prend la valeur 1 lorsque $x \in C$ et -1 lorsque $X \in I$. Bien entendu, il existe une bijection entre l'hyperplan et la fonction f qui lui correspond.
Les SVMs présentent l'énorme avantage de rechercher l'hyperplan qui maximise la marge, distance entre l'hyperplan optimal et les vecteurs supports (cf. figure 3.5).

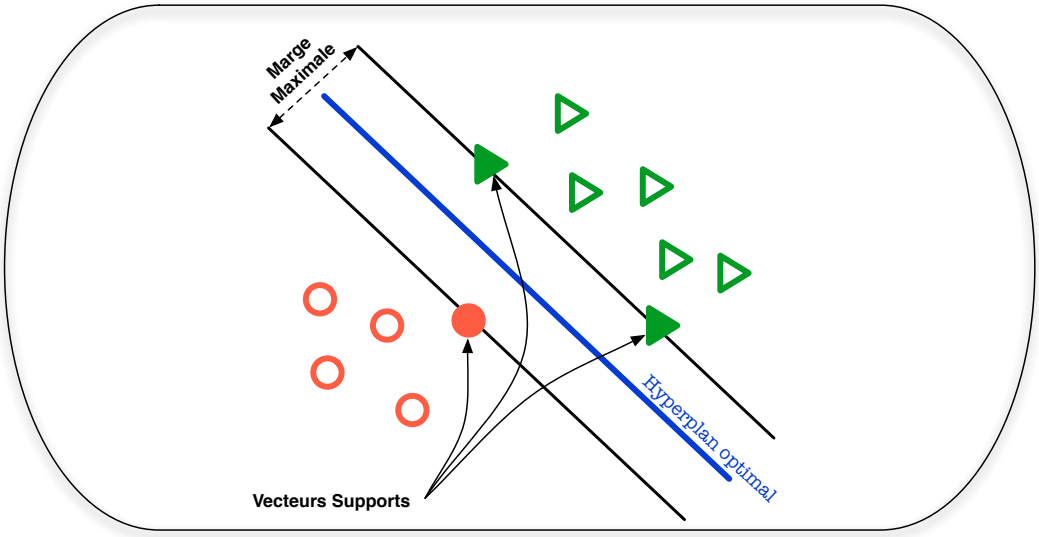


FIG. 3.5: *Séparation de classes par utilisation d'une machine à vecteurs supports. Les SVMs maximisent la distance entre l'hyperplan optimal, qui sépare les sous-espaces des exemples positifs et négatifs, et les vecteurs supports.*

Les données concrètes sur lesquelles sont utilisés les SVMs ne sont généralement pas linéairement séparables. Les SVMs peuvent cependant être utilisés pour des problèmes non-linéairement séparables, en conservant le formalisme décrit ci-dessus. Ces problèmes sont, la plupart du temps, traités en projetant d'un point de vue théorique ces données non-linéairement séparables dans un espace de plus grande dimension où cette séparation est possible. Dans la pratique, cette projection n'est pas nécessaire, les opérations effectuées dans l'espace de projection peuvent être réalisées sur les données non-projetées grâce à des fonctions bien choisies : les *noyaux* (Vapnik, 1998).

Appliquée à la RAL, le SVM est entraîné en utilisant des données provenant du client (pour lesquelles $f(X) = 1$), mais aussi des données appartenant à d'autres lo-

cuteurs ($f(X) = -1$). Les données fournies aux SVMs peuvent être de deux types : paramètres (Campbell et al., 2006b) ou scores (Jaakkola et Haussler, 1999).

Noyaux de séquences et GLDS Une manière intuitive d'utiliser les SVMs consisterait à traiter chaque segment de parole comme une suite d'exemples d'échantillons clients ou imposteurs. Ainsi pour chaque vecteur de paramètres X issu des données d'apprentissage, $f(X) = 1$ et pour les vecteurs acoustiques provenant d'exemples imposteurs, $f(X) = -1$. Cependant, la quantité importante de données en reconnaissance du locuteur ne permet pas d'employer les SVMs de cette façon, cette méthode s'avèrant très coûteuse et ne donnant pas de bons résultats.

Différentes méthodes ont été développées pour traiter les séquences de vecteurs dans leur ensemble grâce à des *noyaux de séquence* qui permettent l'utilisation de SVMs à un coût relativement faible. L'appellation *noyaux de séquence* est employée, même si ces noyaux ne tiennent pas compte de l'ordre des vecteurs de paramètres dans la séquence.
Parmi les différents noyaux, citons les pseudo-noyaux GLDS (Generalized Linear Discriminant Sequence Kernel) développés par Campbell et al. (2006a) et généralisés aux expansions autres que polynomiales par Louradour et Daoudi (2005). Ces noyaux utilisent une projection dans un espace de dimension fixe en utilisant une moyenne des expansions des vecteurs cepstraux suivit d'un produit scalaire linéaire.

Noyaux de produit de probabilités Les modèles génératifs s'avèrent très intéressant dans le cadre des SVMs, car ils permettent de « projeter » des séquences de paramètres de longueurs variables dans un espace de dimension fixe. Les modèles génératifs sont utilisés pour calculer les *super-vecteurs* qui sont fournis aux SVMs. Dans les approches GMM/SVM, les super-vecteurs sont constitués d'une partie des paramètres du modèle GMM du locuteur considéré (généralement les paramètres de moyennes). Parmi les noyaux utilisés, citons les noyaux de *Fisher* (Jaakkola et Haussler, 1999). Les approches SVM/super-vecteurs incluant la technique NAP (Nuisance Attribute Projection) ont montré des résultats équivalents à ceux obtenus par l'approche GMM/UBM utilisant le *Joint Factor Analysis* durant les dernières évaluations NIST-SRE (Fauve et al., 2007). Pour plus de détails sur ces méthodes, nous conseillons la lecture de (Louradour, 2007).

Les performances des machines à vecteurs supports sont aujourd'hui considérées comme équivalentes ou supérieures aux approches GMM/UBM décrites dans le chapitre 7.1.1 (Matrouf et al., 2008).

3.2.3 Performance des systèmes

Les bonnes performances obtenues par les approches état-de-l'art lors des campagnes d'évaluation de vérification du locuteur font de la vérification du locuteur une

des modalités biométrique les plus sûres après les analyses ADN, les empreintes digitales et la reconnaissance d'Iris. Ces performances, entre 1 et 5% de taux d'égales erreurs à l'évaluation NIST-SRE 2008 pour le système LIA (Matrouf et al., 2008) selon les conditions considérées, conjuguées à de faibles contraintes, rendent la biométrie vocale particulièrement intéressante.

L'analyse détaillée des résultats de ces mêmes évaluations souligne cependant certains points faibles des systèmes état-de-l'art. Les résultats obtenus à cette même évaluation NIST-SRE 08 [1] par notre système *LIA* (Matrouf et al., 2008), montrent que les performances sont fortement liées à la quantité de données utilisées lors de la phase d'apprentissage.

Considérons deux conditions évaluées lors de cette campagne NIST-SRE08 organisée par l'institut Américain NIST.

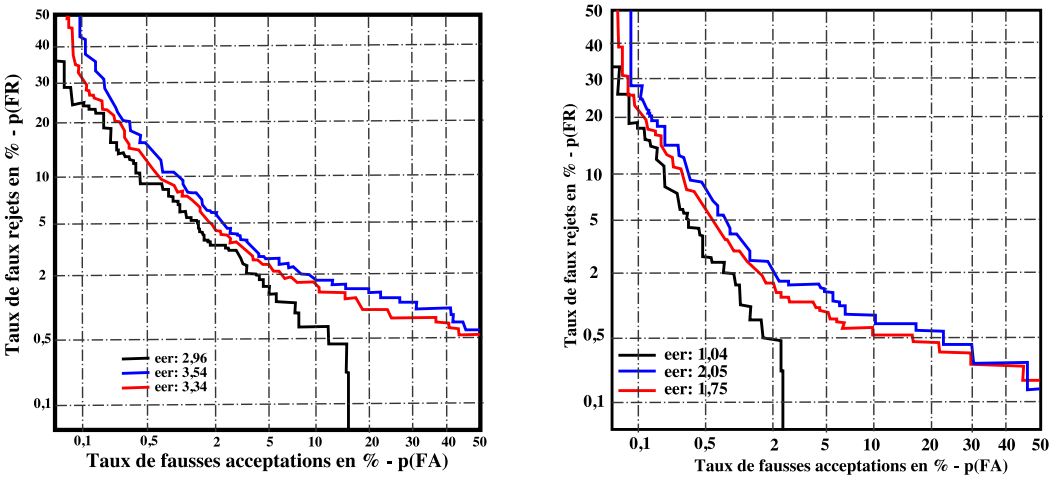
- pour la première, **short2-short3**, les données d'apprentissage de chaque client sont composées d'un fichier de conversation téléphonique ou d'un fichier d'interview d'environ trois minutes. Durant ces trois minutes, le client et son interlocuteur discutent, enregistrés chacun sur un canal différent. Les données utilisées pour la phase de test sont de même nature (téléphone et interview mélangés).
- pour la deuxième condition, **3conv-short3**, le système dispose de trois fichiers d'apprentissage par locuteur. Les enregistrements sont de même nature que lors de la première tâche, téléphonique ou interview. Les données de test sont les mêmes que pour la première condition.

Pour les deux conditions, les tests portent sur un groupe d'hommes et de femmes. Les courbes 3.6(a) et 3.6(b) présentent les résultats obtenus par le système LIA dans ces deux tâches pour les hommes et les femmes séparés ou ensembles (Matrouf et al., 2008).

Lorsque la quantité de données d'entraînement diminue d'un facteur trois, les taux d'égales erreur sont multipliés par presque trois (2,84%) pour les tests hommes et presque doublés pour les tests femmes (1,72%) et les tests mixtes (1,90%).

Cette perte de performance est due au manque de données d'apprentissage. La faible quantité de données disponibles pour la tâche short2-short3 ne permet pas au modèle de bien prendre en compte la variabilité inter-sessions. L'amélioration observée dans la condition 3conv-short3 est due à l'augmentation de la quantité de données d'entraînement mais aussi à l'utilisation de trois sessions différentes pour l'apprentissage des modèles, ce qui permet de mieux modéliser la variabilité intra-locuteur.
La perte de performance, due à la variabilité intra-locuteur ou inter-sessions, peut être réduite par l'augmentation de la quantité de données d'enrôlement mais également par la réduction de la variabilité existant entre les sessions d'enrôlement et les sessions de test. Il est possible d'agir sur de nombreux facteurs qui sont la cause de la variabilité inter-sessions comme le bruit, le canal de transmission ou le contenu linguistique.

[1] `http://www.itl.nist.gov/iad/mig/tests/sre/`



(a) Résultats NIST-SRE 2008, tâche s2-s3

(b) Résultats NIST-SRE 2008, tâche 3conv-s3

FIG. 3.6: *Résultats obtenus par le système LIA dans les tâches short2-short3 et 3conv-short3 de l'évaluation NIST-SRE 08 pour les hommes (en noir) et les femmes (en bleu) séparés ou ensembles (en rouge).*

Le système du LIA intègre une modélisation de la variabilité inter-sessions (cf. section 7.2.5) utilisant le *Factor Analysis* (Matrouf et al., 2007).

3.3 Vérification du locuteur structurale

Cette partie présente les méthodes de reconnaissance automatique du locuteur structurale, qui exploitent une information dépendante du texte. Cette information permet de réduire la variabilité linguistique en ajoutant une contrainte faible sur l'ergonomie des méthodes proposées.

3.3.1 Structure du signal de parole et dépendance au texte

Nous avons introduit dans la partie 3.1.2 les paramètres acoustiques utilisés dans le cadre de la reconnaissance du locuteur, tout en observant que ceux-ci reposent sur une analyse à court terme du signal de parole. De ce fait, ils ne contiennent aucune information sur la structure à « long terme » dudit signal. Cette structure temporelle du signal de parole, essentiellement dépendante du texte énoncé, peut être utile à la tâche de reconnaissance du locuteur en réduisant la variabilité inter-sessions.

La reconnaissance automatique du locuteur dépendante du texte consiste à imposer au client d'utiliser un texte fixe lorsqu'il souhaite être authentifié. Ce mode de fonctionnement implique la coopération du locuteur.

Si la contrainte du texte permet dans ce cas de réduire la variabilité inter-sessions pour le client, elle diminue de même la variabilité pour un éventuel imposteur. Cependant, l'information temporelle introduite par le texte fixé peut être utilisée pour caractériser le client. Ce mode de fonctionnement exploite de façon implicite l'hypothèse suivante : de la même façon qu'il existe une variabilité acoustique inter-locuteurs, la structure temporelle du signal de parole doit être caractéristique de chaque locuteur.
La mise en œuvre d'un tel système peut revêtir différentes formes :

- le texte prononcé est unique, commun à tous les locuteurs et à toutes les sessions ;
- le texte des sessions d'enrôlement, comme celui des sessions de tests, peuvent être affiché par un prompteur ;
- le texte peut être dépendant du locuteur, attribué par le système lors de la phase d'enrôlement. ;
- le texte peut également être dépendant du locuteur mais laissé au libre choix du client lors de la phase d'enrôlement.

Ces différentes configurations permettent toutes de tirer partie de la variabilité inter-locuteurs liée à l'énoncé.

L'approche par texte prompté peut être utilisée pour contrer les impostures rejouant un enregistrement du client. Ainsi l'utilisation de prompteur permet de s'assurer de la présence effective du locuteur (Matsui et Furui, 1993), (Delacretaz et Hennebert, 1998).

Une autre approche combine les avantages de la reconnaissance biométrique avec la vérification d'un élément appartenant au locuteur (Sharma et Mammone, 1996). Le client choisit un texte lors de son enrôlement (BenZeghiba et Bourlard, 2006) (Gagnon et al., 2001). Ce mot de passe peut être choisi parmi un certain nombre de propositions du système ou être laissé au libre choix du client. Il est possible d'imposer également certaines contraintes comme la longueur minimale, maximale ou des contraintes linguistiques. Le modèle créé pour ce client est une représentation des caractéristiques acoustiques du locuteur et de la structure propre à son mot de passe.
Lors de la phase de test, le client prononce son mot de passe pour réduire la variabilité inter-sessions alors qu'un imposteur ne connaissant pas ce mot de passe prononcera un autre texte et augmentera, lui, cette variabilité.
Quelle que soit la méthode choisie pour fixer le texte des phases d'enrôlement et de test, il est possible de classer les approches dépendantes du texte en deux catégories comparables à celles décrites pour la RAL indépendante du texte : l'approche générative, reposant essentiellement sur les modèles de Markov cachés, et l'approche vectorielle, que nous qualifierons plutôt de séquentielle .

3.3.2 Approche générative structurale

Les approches génératives en reconnaissance du locuteur dépendante du texte sont proches du domaine de la reconnaissance de la parole. Elles décomposent le signal de parole en sous-unités lexicales (phonèmes, mots...) qui sont représentées par des modèles de Markov cachés (Hidden Markov Models - HMM) (Rabiner, 1989).

Les HMMs sont des modèles à état fini pour lesquels chaque état est un modèle génératif, souvent approximé par un mélange de Gaussiennes. Le formalisme des modèles de Markov cachés, décrit ci-après, est l'objet d'une étude plus détaillée dans la partie 8.1.1.

Les modèles de Markov cachés :

Un modèle de Markov caché ou HMM (Hidden Markov Models) est un modèle statistique dans lequel le système modélisé est supposé être un processus Markovien de paramètres inconnus. Il s'agit d'un automate à états finis caractérisé par un quadruplet $\Lambda = \{\mathcal{S}, \Pi, \tau, f\}$ où $\mathcal{S}$ est un ensemble d'états S_i, $i \in [1, E]$, π_i est la probabilité que l'état i soit l'état initial, $\tau_{i,j}$ la probabilité de transition de l'état i à l'état j et $f(X|i)$ la probabilité d'émettre le symbole X en étant dans l'état $\mathcal{S}_i$.

La somme des probabilités des états initiaux est égale à 1 :

$$\sum_{i=1}^{E} \pi_i = 1 \tag{3.8}$$

La somme des probabilités de quitter un état i est égale à 1 :

$$\forall i \in [1, E]\,,\ \sum_{j=1}^{E} \tau_{i,j} = 1 \tag{3.9}$$

Un tel automate est représenté par la figure 3.7

Leur utilisation en reconnaissance du locuteur

En RAL, les modèles HMMs utilisés sont dit d'ordre 1. Dans un HMM d'ordre 1, la possibilité d'être dans un état E_j au temps $t+1$ ne dépend que de l'état E_i dans lequel se trouvait le système au temps t. Nous verrons dans le chapitre 5 que des HMMs plus complexes peuvent être utilisés en biométrie bi-modale.

Ces automates sont utilisés pour modéliser des sous-unités lexicales (Rabiner, 1989). La caractérisation de chaque état i par une fonction de densité de probabilité $f(\cdot|i)$ permet de modéliser la variabilité inter-sessions, ce qui ne sera pas le cas des approches séquentielles décrites par la suite. Cette fonction de densité de probabilité est généralement approximée par un mélange de Gaussiennes (GMM).

Certaines approches utilisent un modèle HMM pour modéliser la totalité du texte que le locuteur doit prononcer pour être authentifié. C'est le cas de Rosenberg et al.

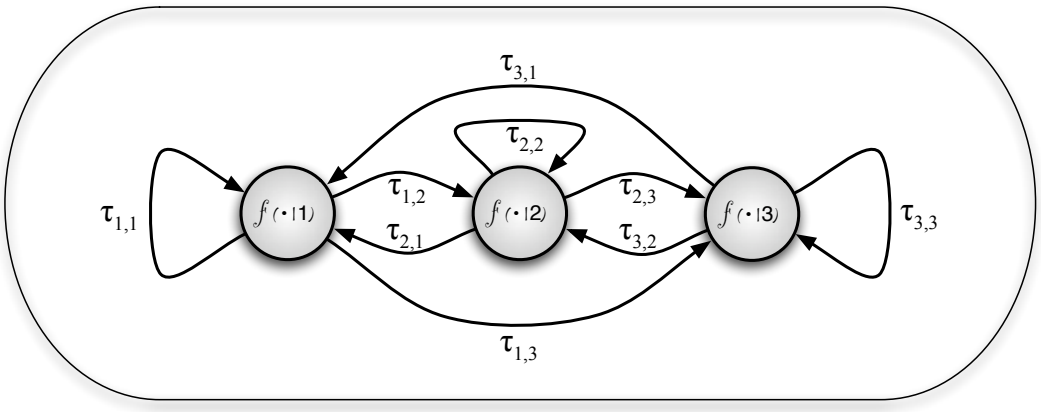


FIG. 3.7: *Représentation d'un modèle de Markov caché à trois états caractérisé par le quadruplet de paramètres $\Lambda = \{\mathcal{S}, \Pi, \tau, f\}$ où $\mathcal{S}$ est un ensemble d'états S_i, $i \in [1, E]$, Π est l'ensemble des probabilités π_i que l'état i soit l'état initial, τ l'ensemble des $\tau_{i,j}$, probabilité de transition de l'état i à l'état j et f l'ensemble des fonctions de densité de probabilité associée aux états $\mathcal{S}$.*

(1991), qui considère chaque mot dans son ensemble et produit un HMM pour chacun à partir des enregistrements du client prononçant ce mot pendant la phase d'enrôlement ou dans (BenZeghiba et Bourlard, 2006), où les auteurs utilisent des réseaux de neurones pour optimiser l'entraînement de la structure de ce HMM.

Les performances de ces approches sont intéressantes mais l'entraînement des modèles pose des problèmes car il nécessite une quantité importante de données d'apprentissage. C'est pourquoi, à ces approches « unicistes », sont souvent préférées des méthodes qui utilisent une réelle décomposition du signal modélisée par un ensemble de sous-HMMs.

Modèles HMM de sous-unités lexicales

Les méthodes qui décomposent la représentation du signal en une suite de modèles HMMs représentant des sous-unités lexicales permettent de faciliter l'apprentissage des modèles. Ces approches tirent parti de la mutualisation des données d'apprentissages. Différents niveaux de segmentation peuvent être utilisés. Ce type d'approche est présenté dans (Nakagawa et al., 2004), (Charlet, 1997) ou encore (Matsui et Furui, 1993). Chacune de ces approches utilise un modèle phonémique indépendant du locuteur comme ceux utilisés en reconnaissance de la parole. Ces modèles de phonèmes sont ensuite adaptés à chaque locuteur et utilisés pour la phase de test.

L'adaptation d'un ensemble des modèles de phonèmes permet d'utiliser une infinité de textes différents. Cette approche implique le stockage du modèle de phonème ainsi que l'utilisation d'un lexique.

L'adaptation de modèles de phonèmes à un locuteur donné nécessite cependant une quantité de données importante, même si elle demeure relativement moins conséquente que dans le cas de modèles « unicistes ».

3.3.3 Approche séquentielle

Les méthodes introduites dans cette section, sont comparables aux approches vectorielles développées précédemment (cf. section 3.2.2). Nous les qualifierons cependant d'approches séquentielles puisque, contrairement aux approches présentées dans la partie 3.2.2, l'organisation des vecteurs utilisés revêt ici une importance réelle.

Alignement dynamique

L'algorithme DTW (Dynamic Time Warping) appliqué à la reconnaissance vocale (Myers et al., 1980), (Rabiner et al., 1978) permet de mesurer la similarité entre deux séquences de paramètres.
Il procède à l'alignement temporel d'une séquence de paramètres par rapport à une séquence de référence, afin de comparer la structure temporelle de ces deux séquences.

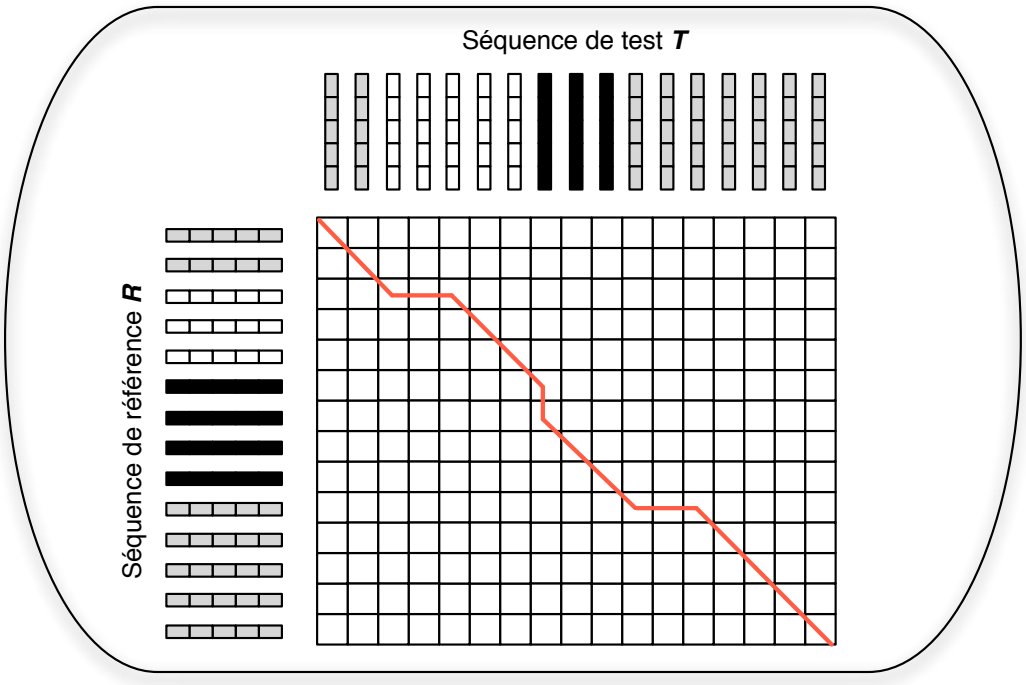


FIG. 3.8: *Représentation du meilleur alignement obtenu par l'algorithme DTW. La séquence de trames T est alignée sur la référence R afin de minimiser la distance cumulée trame à trame.*

Étant donnée une séquence de référence $R(n)$ composée de N trames ($1 \leq n \leq N$), une séquence de test T composée de M trames ($1 \leq m \leq M$), considérons une distance D définie pour tout couple (n, m).
L'approche DTW permet de trouver une fonction d'alignement w qui fasse correspondre les indices de temps m de la séquence de test aux indices n de la référence ($w(n) = m$) en minimisant la distance D_t cumulée le long de w, ainsi :

$$D_T = \min_{\{w(n)\}} \sum_{n=0}^{N} D(R(n), T(w(n))) \tag{3.10}$$

Les extrémités de la séquence de test doivent être les images respectives des extrémités de la séquence de référence. Cette contrainte est traduite dans les équations suivantes :

$$w(0) = 0$$
$$w(N) = M$$

La figure 3.8 illustre l'alignement de la séquence de test par la fonction w.

L'approche DTW permet de mesurer la similarité entre les structures de deux signaux. La distance mesurée lors de l'alignement des deux séquences augmente lorsque la structure de la séquence testée s'éloigne de la séquence de référence.
L'approche DTW n'intègre pas la variabilité acoustique intra-locuteur comme peuvent le faire les approches génératives. C'est pour cette raison que DTW est plus utilisé en reconnaissance de mots isolés qu'en reconnaissance du locuteur.

SVM à noyaux de séquences ordonnées

Certaines approches utilisant des machines à vecteurs supports (cf. section 3.2.2) utilisent des noyaux qui prennent en compte la structure des séquences de paramètres où plus précisément l'ordre des vecteurs de paramètres. Ces méthodes n'ont, jusqu'alors, pas été utilisées en reconnaissance du locuteur, mais comme l'indique Louradour (2007), pourraient être utilisées dans ce cadre.
Parmi ces méthodes, Louradour (2007) distingue tout d'abord celles dont les noyaux exploitent des densités de probabilité et plus particulièrement des modèles de Markov cachés comme dans (Lyngso et al., 1999) et (Lyngso et Pedersen, 2001). L'adaptation de modèles HMMs continus par la méthode MLLR (*Maximum Likelihood Linear Regression*), décrite dans (Leggetter et Woodland, 1995), a également été utilisée par Stolcke et al. (2005) pour développer un noyau SVM prenant en compte les paramètres d'adaptation de chaque état du modèle de Markov. Les méthodes proposées par (Shimodaira et al., 2001) ou (Wan et Carmichael, 2005) qui utilisent les résultats d'alignements dynamiques (DTW, cf. section 3.3.3) pourraient, de même, être exploitées dans le cadre de cette problématique.

Approche mixte : GDW

L'approche GDW (Gaussian Dynamic Warping) de Bonastre et al. (2003) est un hybride qui exploite la généralisation propre aux approches génératives et la modélisation structurelle du DTW. Une structure à trois couches permet d'optimiser l'apprentissage des modèles génératifs et de pallier le manque de données d'apprentissage qui constitue souvent le point faible de ces modèles. Les nœuds de cette structure sont des modèles GMMs. Le premier niveau de cette architecture est un modèle acoustique sensé modéliser la totalité de l'espace acoustique. Les modèles du deuxième niveau sont spécifiques à un locuteur, mais indépendants du texte. Enfin, les modèles de la troisième couche combinent l'information spécifique au locuteur, mais également au texte qu'il prononce lors de sa phase d'enrôlement. Les deux premières couches utilisent le paradigme GMM/UBM qui est détaillé dans le chapitre 7.1.
Le modèle de la troisième couche est entraîné en utilisant l'ensemble des enregistrements du locuteur prononçant le texte d'enrôlement pour calculer un modèle moyen : la référence de ce locuteur.
L'approche développée au cours de cette thèse et présentée dans le chapitre III est proche du GDW.

Conclusion

Nous avons présenté dans ce chapitre les principales approches développées en reconnaissance automatique du locuteur. Les approches non-structurales permettent d'obtenir un haut niveau de performance avec peu de contraintes. Les méthodes structurales permettent d'améliorer encore ces résultats mais impliquent des contraintes d'utilisation supplémentaires notamment lors de la procédure d'enrôlement, ainsi qu'un surcoût calculatoire du à la complexification des modèles utilisés. Ces méthodes présentent cependant l'avantage de réellement prendre en considération la dimension dynamique de la parole. Nous verrons dans la partie 7.3.2, que l'information structurelle apporte un gain substantiel.

Chapitre 4

Reconnaissance visuelle de personnes

Sommaire

Résumé

Nous décrivons dans ce chapitre les approches biométriques existantes dans le domaine vidéo. Les approches issues de la reconnaissance faciale à partir d'images statiques étendues à la vidéo sont présentées dans un premier temps. Elles appartiennent pour la plupart à la biométrie morphologique. Les approches présentées dans un second temps ont été développées spécifiquement pour la vidéo et exploitent les mouvements caractéristiques des individus. Elles sont de ce fait à classer, en majorité, dans les biométries comportementales. Nous conclurons ce chapitre en discutant les performances de ces deux types d'approches dans le champ plus vaste de la biométrie.

Introduction

La biométrie faciale constitue depuis de nombreuses années un domaine de recherche très actif (Zhao et al., 2003), (Chellappa et al., 1995). Dans ce domaine, la plupart des méthodes existantes utilisent des images, que ce soit pour la phase d'enrôlement ou pour procéder aux tests d'authentification. Ces approches sont exposées aux nombreuses variations qui apparaissent sur les images, qu'il s'agisse de la luminosité, des différentes poses d'un sujet ou des différentes expressions que peut afficher un individu. La biométrie faciale a longtemps ignoré les données vidéo du fait de la difficulté de l'époque à traiter l'importante quantité de données liées à cette modalité.

Cette quantité de données constitue cependant le principal attrait de la vidéo en biométrie. Un flux vidéo est une séquence temporelle d'images. À ce titre, il inclue l'information contenue dans les images fixes tout en offrant une variabilité proportionnelle au nombre d'images de la séquence. L'usage de la vidéo en reconnaissance biométrique offre, en terme de robustesse, un potentiel supérieur aux méthodes qui se contentent d'images fixes. La vidéo permet, de plus, d'exploiter une information absente des images fixes : la dynamique.

Alors que la majorité des approches biométriques vidéo demeurent des extensions de méthodes développées pour traiter des images fixes, plusieurs méthodes de reconnaissance comportementales, basées sur les caractéristiques dynamiques des individus (mouvement de tête, des lèvres, ou du corps dans son ensemble) ont été proposées dans les dernières décennies.

Les problématiques de détection (Hjelmas et Low, 2001), (Yang et al., 2002) et de suivi des visages (Steffens et al., 1998), (Schwerdt et Crowley, 2000), (Goecke et Asthana, 2008) doivent être prises en compte dans l'élaboration d'un système de reconnaissance faciale car ces méthodes permettent des gains importants en terme de performance. Elles permettent, en effet, de supprimer une partie du bruit en sélectionnant les données utiles au processus de reconnaissance des individus. Ces problématiques constituent des domaines de recherche à part entière. Les tâches de détection et de suivi de visage ne sont cependant pas abordées dans ce document car elles nécessitent une quantité importante de ressources, due à la complexité des algorithmes utilisés et la dimension des données, qui les rend difficilement compatible avec notre contexte applicatif.

Dans ce chapitre, nous présentons un état-de-l'art de la reconnaissance de personnes à partir d'un flux vidéo. L'utilisation de la vidéo est privilégiée, du fait de la quantité de données présente dans les séquences vidéo et des possibilités offertes par la présence d'information dynamique.

4.1 La vidéo, un signal à 2+1 dimensions

Cette section décrit des approches provenant du traitement des images statiques et qui ont fait l'objet d'extensions vidéo.

4.1.1 Mise en correspondance de grilles élastiques

Principe

Les travaux de Lades et al. (1993) sur les réseaux de neurones appliqués à la reconnaissance de forme ont abouti à la méthode de mise en correspondance de grilles élastiques (Elastic Graph Matching - EGM). Une grille rectangulaire est appliquée sur le visage de référence. En chacun des nœuds de cette grille est calculé un jeu de paramètres fréquentiels locaux. Durant le test, la grille est placée sur l'image à tester et déformée globalement, puis localement, afin de faire correspondre au mieux les coefficients de la grille de référence et les particularités fréquentielles de l'image. La déformation de la grille est mesurée. Si la déformation n'est pas trop importante et que les jeux de paramètres sont assez proches, le test est positif. L'un des points essentiels des grilles élastiques réside dans la position des nœuds sur l'image de référence.

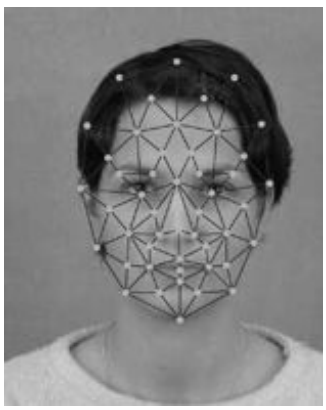


FIG. 4.1: *Résultat de l'alignement automatique d'une grille élastique sur une image (Boehringer et al., 2006)*

Une évolution des grilles élastiques a été développée dans (Wiskott et al., 1997), où les auteurs associent les nœuds de la grille aux points importants du visage, comme les pupilles ou les coins des lèvres. Cette méthode EBGM (Elastic Bunch Graph Matching) s'avère plus performante que les grilles élastiques originales mais nécessite la détection de ces points particuliers de façon automatique.

Extension vidéo

Steffens et al. (1998) ont proposé d'étendre l'EBGM à une tâche de détection et reconnaissance de visage dans le cadre de la modalité vidéo. Les grilles élastiques simples sont appliquées sur les images de la séquence vidéo afin de déterminer les images les plus adaptées à la reconnaissance de visage. La sélection est effectuée par passes successives en augmentant progressivement la complexité des grilles. Seul le score obtenu avec la grille la plus complexe est retenu pour la prise de décision.
Cette méthode utilise la grande quantité de données disponibles dans la vidéo sans prendre en compte ni la corrélation existante entre les images successives, ni la structure temporelle qui en résulte. Elle nécessite de plus un processus très coûteux qui consiste à parcourir partiellement la vidéo plusieurs fois pour affiner la recherche des meilleurs images.

4.1.2 Eigenfaces

Principe

Les Eigenfaces (Turk et Pentland, 1991b), (Turk et Pentland, 1991a) sont une des approches les plus répandues en reconnaissance faciale. Les variantes de cette approche sont nombreuses. Notre but ici n'est pas de les détailler toutes mais d'exposer le principe sous-jacent à ces méthodes, l'Analyse en Composante Principales (ACP).

Une image Γ de dimensions $n \times n$ peut être vue comme un vecteur de dimension n^2. L'idée de base des Eigenfaces consiste à représenter Γ, ou plutôt représenter Φ, tel que $\Phi = \Gamma - \textit{Visage moyen}$, dans un espace de dimension réduite qui permette une classification simple des visages.

$$\Phi = (\Gamma - \textit{visage moyen}) = w_1.U_1 + w_2.U_2 + w_3.U_3 + ... + w_K.U_K \quad \text{avec } K << n^2 \quad (4.1)$$

Le calcul du *visage moyen* : Ψ ainsi que celui de la base de vecteurs U_i, $i = 1, ..., K$ nécessite un grand nombre d'images d'apprentissage I_i représentées par un vecteur Γ_i. Ces images de visage doivent de plus être de même dimension ($n \times n$). Le *visage moyen* devient :

$$\Psi = \frac{1}{M}.\sum_{i=1}^{M} \Gamma_i \quad \text{où M est le nombre d'images utilisées pour l'apprentissage} \quad (4.2)$$

Le visage moyen est soustrait de toutes les images d'apprentissage afin de procéder au calcul de la base de vecteurs u_i. Les vecteurs u_i sont en fait les vecteurs propres de la matrice de covariance C :

$$C = \frac{1}{m}.\sum_{i=1}^{m} \Phi_i.\Phi_i^t = A.A^t \quad \text{matrice de dimension } n^2 \times n^2 \quad (4.3)$$

où

$$A = [\Phi_1, \Phi_2, \Phi_3, ..., \Phi_M], \quad \text{matrice de dimension } n^2 \times m \quad (4.4)$$

Le calcul des vecteurs propres de C ne s'effectue pas de manière directe car les dimensions de C sont trop grandes pour permettre de réaliser ce calcul précisément. La matrice $A^t.A$ de dimension $m \times m$, qui possède les mêmes valeurs propres que $C = A.A^t$, est utilisée (sauf si $m >> n$). Plus exactement, les m valeurs propres de la matrice $A^t.A$ correspondent aux m plus grandes valeurs propres de C. Les vecteurs propres U_i de C et v_i de $A^t.A$ correspondant à ces valeurs sont liés par la relation :

$$U_i = A.v_i \tag{4.5}$$

Une fois les vecteurs $U_{1 \leq i \leq m}$ obtenus, seuls les K premiers, qui forment la base de l'espace où seront projetées les images de visages à classifier, sont conservés.

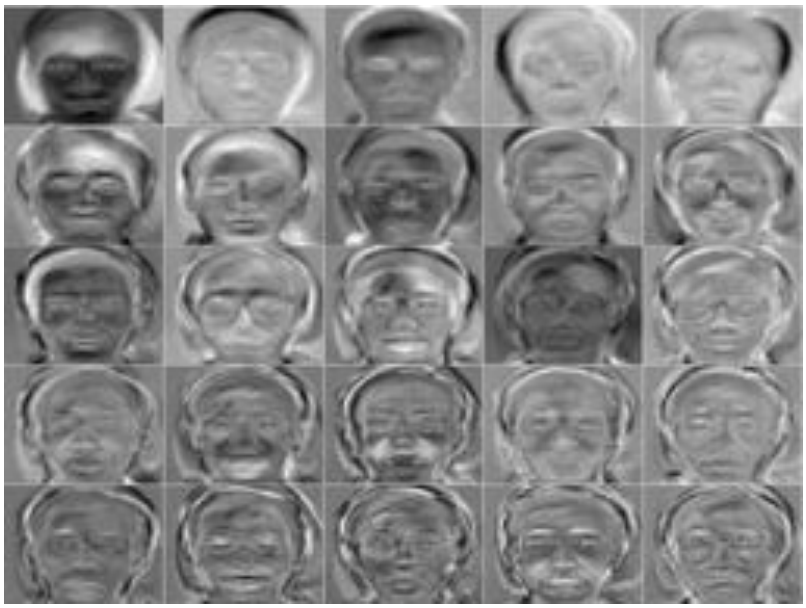


FIG. 4.2: *Base des EigenFaces obtenue pour le corpus Yale (Georghiades et al., 2001)*

Lors d'un test, l'image I du client est redimensionnée et normalisée géométriquement. L'image moyenne Ψ est retiré à son vecteur Γ, qui est projeté dans l'espace des EigenFaces.

$$\Phi = \Gamma - \Psi \tag{4.6}$$

$$\hat{\Phi} = \sum_{i=1}^{K} w_i.U_i \quad \text{avec} \quad w_i = U_i^t.\Phi \tag{4.7}$$

La représentation de l'image I dans l'espace des EigenFaces est alors :

$$\Omega = \begin{pmatrix} w_1 \\ w_2 \\ . \\ . \\ w_K \end{pmatrix} \tag{4.8}$$

La comparaison de cette image à la référence R de l'identité revendiquée par l'utilisateur dans l'espace des EigenFaces est calculée par un simple calcul de distance dans cet espace. Les distances les plus utilisées sont la distance Euclidienne et la distance de Mahalanobis (Mahalanobis, 1936).

Extensions

Le principe des EigenFaces a été utilisé par Bregler et Konig (1994) en reconnaissance de la parole pour représenter la position des lèvres dans un espace de dimension réduite (EigenLips).

Dans (Moghaddam et Pentland, 1997), les auteurs utilisent la représentation discriminante obtenue dans l'espaces des vecteurs propres en tant que paramètres. Les vecteurs obtenus dans l'espace réduit servent de vecteurs de paramètres pour une approche statistique à base de GMMs (cf. section 7.1).

Les Eigenfaces ont étés adaptés à la reconnaissance du locuteur par Kuhn et al. (1998) dans un premier temps et repris ensuite par Kenny et al. (2005) qui en ont donné une lecture différente. Ces approches rappellent également les modèles d'ancrage présentés dans la section 3.2.2.

Extensions vidéo

Les extensions des EigenFaces au domaine de la biométrie vidéo sont nombreuses du fait des bonnes performances fournies dans le domaine de l'imagerie statique. La plupart, comme Satoh (2000), n'utilisent pas vraiment la dimension temporelle de la vidéo. Dans cette approche, les auteurs projettent chaque image d'une séquence vidéo dans l'espace des EigenFaces et définissent la distance entre cette vidéo et la représentation d'une autre séquence comme le minimum des distances entre l'ensemble des images de la vidéo et cette référence. Cette méthode consiste donc à utiliser la redondance présente dans la vidéo pour calculer le meilleur score possible pour cette vidéo. Huang et Trivedi (2002) utilisent le même procédé mais calculent une décision par image de la séquence. Ils procèdent ensuite à un vote majoritaire pour déterminer la décision finale.

Dans (Torres et Vilà, 2002) les auteurs utilisent la variabilité présente dans les vidéos d'enrôlement des clients en calculant pour chaque utilisateur enrôlé un espace propre. Les images d'apprentissage sont issues de la séquence d'apprentissage. De la projection d'une image dans un de ces espaces résulte une approximation de l'image de départ qui est une somme pondérée des images propres. Cette approximation peut être perçue comme la reconstruction de l'image de départ par sommation d'images de l'utilisateur correspondant à l'espace considéré.

La mesure de similarité utilisée dans cette approche est une mesure de distance entre l'image originale et sa reconstruction dans l'espace des EigenFaces choisi. L'hypothèse qui motive cette approche considère que si l'image de test provient bien de l'utilisateur correspondant à l'espace de projection, elle peut être reconstruite exactement car les EigenFaces utilisés proviennent d'images de cet utilisateur. Cette méthode est similaire aux travaux de (Soong et al., 1985) en reconnaissance du locuteur par quantification vectorielle.

Les extensions des EigenFaces présentées ici utilisent la variabilité présente dans la vidéo mais en aucun cas la structure dynamique de cette modalité.

4.1.3 Fisherfaces

Principe

Les Fisherfaces (Belhumeur et al., 1997), (Martinez et Kak, 2001) sont un autre outil utilisé en reconnaissance faciale. Ils permettent de projeter les images dans un espace de dimension réduite de la même façon que les Eigenfaces. La base de projection n'est cependant pas calculée de la même façon puisque l'ACP est remplacée par une Analyse Linéaire Discriminante (Linear Discriminant Analysis - LDA). Cette méthode d'analyse statistique permet de trouver un espace de projection dans lequel des classes définies à l'apprentissage seront plus facilement séparables que dans l'espace de départ. Cette méthode, utile en identification, nécessite pour être optimale, de recalculer la base de projection à chaque ajout d'une référence de personne dans le système. Elle est de ce fait peu adaptée à la tâche de vérification d'identité.

Extension vidéo

Dans (Satoh, 2000), l'auteur propose la même extension que pour les Eigenfaces (voir section 4.1.2). La similarité entre la séquence de test et la référence connue est calculée en prenant le minimum de la distance entre chaque image de la séquence et la référence dans l'espace de projection.

4.1.4 Approche générative et paramètres locaux

Principe

L'approche présentée dans (Sanderson et al., 2005), utilise le paradigme GMM/UBM (cf. chapitre 7). La spécificité de cette méthode vient de l'extraction des paramètres. Contrairement à la plupart des méthodes de reconnaissance faciale, les paramètres sont calculés sur des blocs de l'image par une transformée en cosinus discrète (DCT) ou par

une ACP. Ces approches donnent des résultats intéressants, d'autant plus que Sanderson présente également des méthodes de normalisation locale qui permettent de réduire l'effet des variations d'illumination.
McCool et Marcel (2009) extraient les paramètres DCT locaux afin de décomposer l'image en une somme de sous images correspondant à chaque bande de fréquences. Des modèles GMMs appris sur chaque bande de fréquences sont ensuite combinés dans une somme pondérée, afin de constituer un modèle GMM de l'individu. Cette approche par bandes de fréquences améliore les performances du système de vérification d'identité utilisant les coefficients DCT locaux.

Extension vidéo

L'utilisation de paramètres locaux pourrait être étendu à la vidéo de la même façon que les modèles génératifs le sont en reconnaissance du locuteur (cf. section 3.3.2). Pour la vidéo, ces approches sont décrites dans la partie 4.2.2.

4.2 La vidéo, un signal temporel

De nombreux travaux exploitent la dimension temporelle du flux vidéo, particulièrement dans le domaine de la parole. Considérant le caractère discret du flux vidéo, nous distinguerons dans cette partie deux types d'approches selon leur granularité. Dans un premier temps, nous verrons qu'il est possible d'extraire des paramètres dynamiques à la fréquence d'échantillonnage du flux original. Ces paramètres, caractérisant le mouvement au sein d'un groupe d'images consécutives, peuvent alors être exploités dans différentes architectures qui seront abordées.
Dans un deuxième temps, nous considérerons le flux vidéo comme une suite structurée d'événements finis dont le nombre est inférieur au nombre d'images de la vidéo.

4.2.1 Exploitation de la dynamique

Flot optique

Méthode spécifiquement développée pour la vidéo, la détection du flot optique (Quenot, 1992) estime localement le mouvement survenu entre deux ou plusieurs images consécutives d'une séquence vidéo. Cette estimation repose sur deux hypothèses :

- la couleur des pixels reste inchangée au cours du déplacement,
- les déplacements sont petits par rapport aux dimensions de l'image (de 5 à 10%).

L'estimation de mouvement est issue des méthodes de programmation dynamique déjà évoquées dans la section 3.3.3 puisqu'il s'agit encore une fois de minimiser les

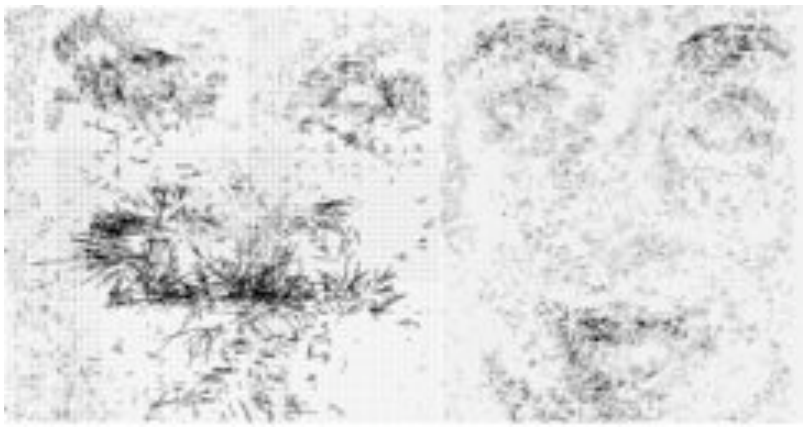


FIG. 4.3: *Représentation de deux champs de vecteurs de mouvement obtenus par calcul du flux optique (Matta, 2008)*

changements à appliquer à un objet pour le faire correspondre à une référence. Ce procédé tend à optimiser la fonction d'énergie suivante :

$$E = \int\int (I_x.u + I_y.v + I_t)^2 + \alpha.(|\nabla u|^2 + |\nabla v|^2)dxdy \tag{4.9}$$

où $I(x, y, t)$ est la luminosité du pixel (x, y) de l'image au temps t. $[u, v] = [u(x, y), v(x, y)]$ est le vecteur de mouvement. ∇ est l'opérateur gradient, I_x et I_y sont les dérivées partielles de $I(x, y, t)$ par rapport à x et y respectivement. Le premier terme de l'expression de E est la contrainte de l'image, le second terme est la contrainte de régularité et de continuité et le coefficient α est le poids qui permet de renforcer l'une ou l'autre des contraintes. Cette équation provient d'un développement de Taylor au premier ordre de l'hypothèse d'invariance de la luminosité décrite par l'équation 4.10.

$$I(x + u\delta t, y + v\delta t, t + \delta t) = I(x, y, t) \tag{4.10}$$

En biométrie vidéo, la détection de flot optique est utilisée pour caractériser un individu par la déformation de son visage au cours du temps (cf. figure 4.3). Cette déformation est représentée par le champ de vecteurs de mouvements calculé sur l'intervalle de temps considéré. La détection du flot optique se révèle plus robuste aux changements de luminosité que de nombreux paramètres fréquentiels.

Dans (Chen et al., 2001), les vecteurs de mouvements calculés entre deux images sont concaténés. Une ACP et une LDA réduisent la dimension de l'espace de représentation, et un classifieur permet durant la phase de test de déterminer la similarité entre les données de références et la représentation d'une séquence de test qui a subi le même traitement.

Reconnaissance de personnes par l'étude des mouvements de tête

Les paramètres dynamiques décrits par Matta et Dugelay (2006) reposent sur la détection semi-automatique de certains éléments caractéristiques d'un visage comme, par exemple, le contour des lèvres, la position du nez ou des pupilles. Le suivi de ces points d'intérêt sur la séquence vidéo permet de calculer un vecteur de déplacement par couple d'images pour chacun de ces points. Contrairement au flot optique, les paramètres estimés ici reposent sur le mouvement global de la tête et non sur des déformations locales. Le vecteur résultant de la concaténation des vecteurs de mouvement est finalement dérivé deux fois selon la méthode décrite dans la section 3.1.3.
Des paramètres relatifs à la géométrie de la bouche et au contour des lèvres sont ajoutés aux précédents dans (Saeed et al., 2006). Les vecteurs de paramètres proposés dans ces deux articles sont utilisés dans des modélisations GMMs (cf. section 7.1).

Les résultats présentés par les auteurs sont encourageants mais nécessiteraient une évaluation sur une base de données de taille plus importante. De plus, la détection des points particuliers du visage est réalisée selon un processus semi-automatique qui nécessite une détection manuelle de ces points sur la première image de chaque séquence.

Paramètres dynamiques en cascade

L'extraction des paramètres dynamiques décrite dans (Potamianos et al., 2001) s'inspire des méthodes utilisées en reconnaissance vocale. Les auteurs proposent deux approches utilisant des paramètres obtenus par une analyse fréquentielle d'une région d'intérêt de l'image (Region Of Interest - ROI).
Dans la première, ces paramètres sont dérivés deux fois (voir section 3.1.3), ce qui apporte une première information dynamique.
Dans la deuxième méthode, les auteurs proposent de concaténer ces vecteurs de paramètres (comme décrit dans la section 3.1.3) et d'utiliser une analyse linéaire discriminante pour réduire la dimension des vecteurs résultant de ce processus. Les vecteurs obtenus subissent également une double dérivation.

Ces paramètres, développés pour la reconnaissance, de parole ont été utilisés pour la reconnaissance biométrique par Dean et al. (2008b), qui ont démontré leur efficacité dans ce domaine. Ils ont été employés, comme c'est le cas en reconnaissance vocale, dans des approches génératives (GMM). La méthode de Dean et al. (2008b) utilise pour seule information dynamique celle qui est contenue dans ces paramètres.

4.2.2 Structure du flux vidéo en parole

Utilisation des HMMs en vidéo

Les modèles de Markov cachés ont déjà été cités dans la partie 3.3.2 pour leur capacité à modéliser la structure temporelle des séquences de paramètres et à intégrer la variabilité intra-individu. Ils sont proposés par Liu et Chen (2003) dans le cadre de la biométrie vidéo. Les paramètres d'entrée peuvent être le résultat d'une analyse fréquentielle (DCT) ou simplement les valeurs des pixels de l'image. Dans leur approche, Liu et Chen utilisent des paramètres résultant d'une ACP pour leur compacité. Chaque image de la vidéo fournit un vecteur de paramètres.
L'apprentissage des modèles HMMs nécessite une quantité importante de données spécifiques à chaque individu. Afin de mutualiser les données, cet apprentissage est remplacé par un modèle générique indépendant de l'individu qui est adapté, ensuite, à chaque client. Dans le cas où le résultat du test confirme l'identité revendiquée, les données de tests sont utilisées pour renforcer l'adaptation du modèle (apprentissage non-supervisé).

Comme de nombreuses méthodes proposées en reconnaissance faciale, cette approche a été évaluée sur une base de données pour laquelle la détection de visage est effectuée au préalable et est considérée comme parfaite. Il est donc difficile d'estimer les performances d'un système complet, qui détecte et reconnaît les visages.

Segmentation en Visèmes

De la même façon qu'il est possible de segmenter la parole en phonèmes, le flux vidéo associé peut être découpé en une suite d'états finie. Ces états appelés visèmes (Fisher, 1968) représentent les plus petites entités permettant de décrire la structure temporelle du flux vidéo de parole.
Malgré cette analogie, il n'existe pas de correspondance simple entre phonèmes et visèmes. En effet, les variations du flux visuel, plus rapides que leurs homologues acoustiques, rendent difficile la correspondance. Si, par exemple, les phonèmes /m/, /b/ et /p/ peuvent être différenciés acoustiquement, les mouvements de la bouche lors de leur production ne permettent pas de les discriminer de manière fiable. Il en résulte que ces trois phonèmes sont regroupés en un visème commun. Un groupement des phonèmes est proposé dans (Binnie et al., 1974) et d'autres peuvent être utilisés selon le contexte applicatif (Morishima et al., 2002) (Foo et al., 2003).

La problématique des visèmes apparaît en reconnaissance de la parole (Foo et al., 2003), (Dong et al., 2005). Sumby et Pollack (1954) ont d'ailleurs montré que, pour cette tâche, l'utilisation de l'information visuelle est équivalente à un gain de 12dB pour le rapport signal sur bruit.
Cependant, les études relatives aux visèmes sont plus nombreuses dans la littérature traitant des têtes parlantes ou de la synthèse vocale (Goff et Benoit, 1996), (Dongmei

et al., 2002), (Benoît et al., 1991). La plupart des approches proposées utilisent des modèles HMMs. Les informations utilisées peuvent provenir de différents paramètres. Les plus courants sont de deux types : des paramètres morphologiques extraits de manière semi-automatique, ou des paramètres issus d'une analyse en composante principale sur les images entières ou, plus précisément, sur la zone de la bouche.
La difficulté liée à ces approches réside dans la vitesse de variation du flux visuel, qui est, de plus, généralement échantillonné à des fréquences assez basses (environ 25 images par seconde). Il est alors difficile d'obtenir une modélisation pertinente des visèmes grâce à des modèles de Markov appris de manière classique (cf. section 8.1.1). C'est pourquoi différentes techniques d'apprentissage ont été développées pour ce type de modèles (Foo et al., 2003), (Verma et al., 2003). Ces contraintes d'apprentissage rendent difficile un apprentissage de modèles de Markov dépendant du locuteur mais des approches d'adaptation des modèles HMMs pourraient être envisagées dans ce cadre.

Conclusion

Nous avons décrit dans ce chapitre les principales approches existantes en biométrie vidéo. Nous avons fait le choix de ne pas expliciter les performances des différents systèmes car les résultats présentés dans la littérature sont difficilement comparables. L'utilisation de bases de données et de protocoles divers pour évaluer les différentes approches ne permettent pas un classement objectif des méthodes les plus performantes.

Cependant, les performances des systèmes de reconnaissance vidéo sont généralement perçues, dans la communauté biométrique, comme inférieures aux performances de la reconnaissance automatique du locuteur. Une comparaison des performances de systèmes audio et vidéo est présentée dans (Fauve et al., 2008). Il y apparaît que, face à divers types d'impostures, les performances des systèmes audio surpassent leurs équivalents vidéo. Cette « supériorité » de la biométrie vocale sur la vidéo est principalement due au manque de robustesse des systèmes vidéo face à des variations importantes (illumination, environnement, changement d'apparences des personnes à reconnaître...).
Il demeure néanmoins que la reconnaissance vidéo révèle de bonnes performances dans des environnements contrôlés. Les résultats de la campagne d'évaluation organisée par le NIST (Phillips et al., 2005) et présentées dans (Phillips et al., 2006) donnent un aperçu des performances actuelles en reconnaissance de visage.

Chapitre 5

Authentification bi-modale audio-visuelle

Sommaire

Résumé

La multi-modalité offre plusieurs avantages. Associer plusieurs modalités peut permettre de renforcer la sécurité en termes de performances, mais peut également améliorer la robustesse des systèmes biométriques par le biais de scénarii où une modalité peut pallier le manque de la modalité principale. L'utilisation conjointe des modalités audio et vidéo nécessite tout d'abord une réflexion sur la nature des informations à exploiter. Des possibilités offertes par ces modalités découlent différentes approches, comme la fusion des informations ou les systèmes que nous appelons multi-flux. Celles-ci sont décrites en détails dans ce chapitre.

Introduction

UNE réflexion sur l'état-de-l'art dans les domaines des biométries vocales (chapitre 3) et vidéo (chapitre 4) a montré que chacune de ces modalités possède ses propres limitations. Les performances des approches biométriques souffrent notamment des variations dues aux environnements au sein desquels s'effectuent les phases d'enrôlement et de vérification. En effet, si la phase d'enrôlement peut aisément être réalisée dans un environnement contrôlé, l'environnement de test dépend, lui, de l'application visée.
D'une modalité à l'autre, les contraintes sont différentes. Alors qu'un système biométrique audio peut être perturbé par un environnement bruyant, un système utilisant un flux vidéo sera perturbé par des variations lumineuses. Une façon de remédier à ces problèmes est d'utiliser plusieurs biométries conjointement. L'utilisation conjointe de deux modalités biométriques se justifie également par la volonté de déjouer des impostures en augmentant la quantité et la diversité des informations vérifiées.

Comme nous l'avons évoqué dans l'introduction à la partie II, des données audio et vidéo peuvent être acquises simultanément lors de la production de parole. Mais si les systèmes d'acquisition sont séparés, par nature, les signaux enregistrés sont - dans ce cas - intimement liés.
Néanmoins, si les mouvements du visage sont liés à la production du signal sonore et si, de fait, ces deux phénomènes sont fortement corrélés, ils présentent une asynchronie qui peut gêner la réalisation d'un traitement simultané et qui doit être prise en compte pour un usage conjoint des modalités audio et vidéo. La différence d'échantillonnage est également un facteur à prendre en compte. Lors de l'acquisition des signaux audio et vidéo, le taux d'échantillonnage est de l'ordre de 25 images par secondes pour la vidéo tandis que la fréquence d'échantillonnage audio généralement utilisée en biométrie varie entre 8 et 16 kHz, ramenée à 100 vecteurs de paramètres par seconde après paramétrisation.

Les possibilités d'intégration des deux approches, audio et vidéo, au sein d'un même système de vérification d'identité sont nombreuses, du fait de la grande diversité des approches uni-modales existantes.
De plus, la structure temporelle des flux audio et vidéo peut être utilisée au sein de chacune des modalités biométriques, pour la vérification d'identité.
Nous continuerons dans cette partie à nous interroger sur les possibilités de prise en compte de la structure temporelle des différents flux au sein du processus bi-modal de vérification d'identité. Nous verrons que la possibilité de prise en compte de cette structure est fortement conditionnée par les différentes méthodes d'intégration de chaque modalités dans ce processus.

Dans ce chapitre nous présentons les caractéristiques communes ou spécifiques aux signaux audio et vidéo acquis lors de la production de parole. Les différentes façons de tirer parti de ces deux modalités dans une approche commune sont ensuite décrites à la lumière de cette analyse, ces multiples approches conjointes influant bien évidemment sur la structure générale des systèmes biométriques. Enfin après avoir présenté les principales méthodes existant dans la littérature, nous livrons nos conclusions quant

à l'utilisation combinée des modalités audio et vidéo.

5.1 Audio et Vidéo, un lien étroit

L'introduction de la partie II suggère que les informations portées par les modalités audio et vidéo du signal de parole se recouvrent partiellement. Cette section décrit les caractéristiques communes et les différences des modalités audio et vidéo.

5.1.1 Des signaux redondants

Certains mouvements du visage sont intimement liés aux sons émis lors de la production de parole. En effet, ils ont une source commune : les organes modulateurs. L'étude menée dans (Yehia et al., 1997) montre l'importante corrélation existant entre les mouvements du visage, des articulateurs (langue, lèvres) et le signal acoustique. D'après leurs résultats, 91% de la variance des mouvements du visage observés peut être prédite à partir du mouvement du conduit vocal. La prédiction réciproque est vérifiée à 80%. De même, ils montrent qu'entre 72% et 85 % de l'enveloppe spectrale du signal acoustique produit peuvent être prédits à partir des mouvements du visage du locuteur.
Cependant, l'acquisition des mouvements dans cette étude utilise des capteurs particulièrement intrusifs et des mesures moins intrusives ne permettent pas d'obtenir les mêmes résultats. Lors d'une autre étude, pour laquelle l'information visuelle est limitée à la forme des lèvres (Barker et Berthommier, 1999), la prédiction de l'enveloppe spectrale chute à $\sim$ 60%.

Ces deux études prouvent néanmoins que les signaux visuels et acoustiques sont fortement corrélés. Des résultats similaires sont présentés dans (Goecke et Millar, 2003) et cette corrélation est utilisée dans (Siracusa et Fisher, 2007) pour détecter le locuteur parmi plusieurs personnes présentes sur la vidéo.

5.1.2 Des signaux complémentaires

Les signaux de parole visuels et acoustiques ne sont pas équivalents. Malgré leur forte corrélation, ils contiennent chacun une information qui leur est propre. De nombreuses études, en reconnaissance de la parole (Zhang et al., 2002), (Dean et al., 2008a) ou en reconnaissance du locuteur (Jourlin et al., 1997), (Bengio, 2003a), montrent que l'association de ces deux modalités permet d'améliorer les résultats, comparativement à des systèmes utilisant une seule des deux modalités.

Les résultats de Yehia et al. (1997) et de Barker et Berthommier (1999), comparés précédemment, montrent que l'information provenant d'une deuxième modalité peut être

présente dans la première modalité et mal caractérisée par la paramétrisation choisie. La deuxième modalité peut cependant apporter une information complètement absente de la première.

5.1.3 Asynchronie des modalités audio et vidéo

En parole, les signaux audio et vidéo ne sont pas exactement synchrones, en effet les mouvements des articulateurs peuvent commencer avant ou après la production sonore (Rogozan et Deléglise, 1998). Ces phénomènes d'anticipation et de rétention (Abry et Lallouache, 1991) ont été illustrés par Eveno et Besacier (2005). Des mesures de CO-Inertia Analysis (Doledec et Chessel, 1994), (Goecke et Millar, 2003) et CANonical CORrelation (Hotelling, 1936) ont mis en évidence le décalage temporel existant entre les paramètres extraits des flux audio et vidéo. La figure 5.1 montre les scores obtenus pour ces deux mesures lorsque les signaux audio et vidéo sont décalés dans le temps, dans le cas où il s'agit d'un enregistrement simultané des deux modalités, ou d'une imposture mêlant deux signaux de sessions différentes. Ces courbes mettent en évidence deux phénomènes : d'une part les scores obtenus par le couple audio-vidéo « légitime » sont plus élevés que ceux obtenus par des impostures, d'autre part, les maxima de ces scores sont obtenus lorsqu'un léger décalage temporel est autorisé entre les signaux, faisant ainsi apparaître l'asynchronie des deux modalités.

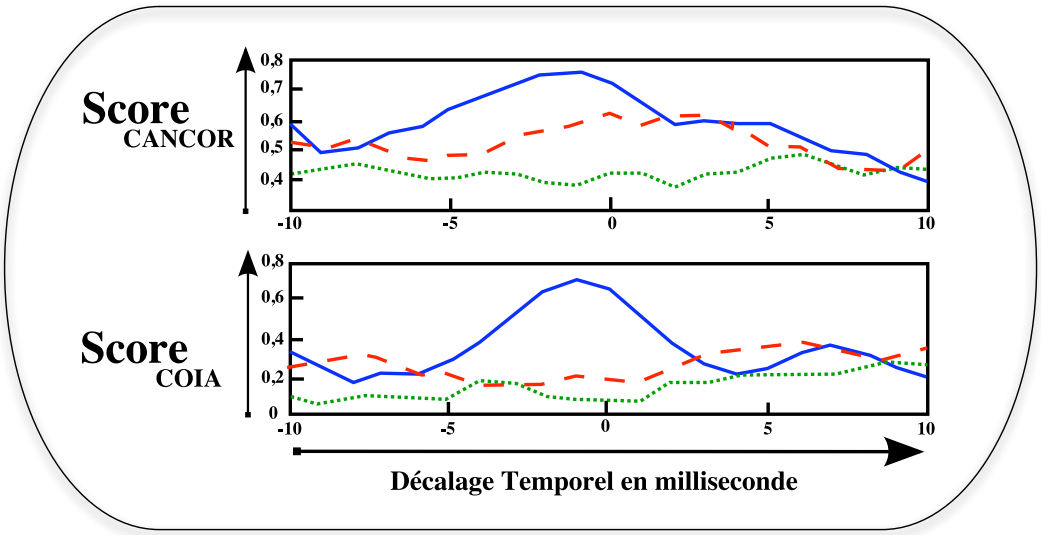


FIG. 5.1: *Effet du décalage temporel sur les mesures de CANonical CORrelation (CANCOR) et CO-Inertia Analysis (COIA) pour un enregistrement en direct (trait plein) ou une imposture rejouée avec le même contenu syntaxique (tirets) ou un contenu différent (pointillés) (courbes provenant de (Eveno et Besacier, 2005))*

La problématique de l'asynchronie entre audio et vidéo a essentiellement été traitée dans le cadre de la reconnaissance de la parole, (Andre-Obrecht et Jacob, 1997), (Deleglise et al., 1996), (Bengio, 2004) du fait du caractère temporel inhérent à cette théma-

tique. Il existe cependant quelques travaux réalisés dans le cadre de l'authentification biométrique bi-modale comme (Bengio, 2003b) ou (Jourlin et al., 1997).

Plusieurs études ont tenté de définir à quel moment l'intégration des deux modalités a lieu pour de la perception humaine et à quel moment elle doit être réalisée lors d'un traitement automatique. D'un point de vue sensoriel, aucune réponse catégorique n'a été apportée à cette question. Certaines études penchent en faveur d'une intégration précoce, d'autres pour une intégration tardive ou intermédiaire (Robert-Ribes, 1995). Les parties 5.2 et 5.3 montrent que cette question reste également ouverte dans le domaine de approches automatiques.

5.2 Bi-Modalité et fusion

La fusion d'informations est un domaine de recherche spécifique qui étudie les différentes façons de fusionner des informations de nature différentes en exploitant au mieux les caractéristiques de chaque source.
Dans la littérature, les architectures de systèmes biométriques multi-modaux sont nombreuses à utiliser une fusion d'information. Ces architectures peuvent être divisées en trois catégories selon le niveau de la fusion dans le système biométrique. Les trois types de fusions, fusion de données, de scores et de décisions, sont représentés par la figure 5.2 dans le cas d'un système bi-modale audio-vidéo.
Dans la suite, nous présentons sommairement ces trois catégories de fusions en nous intéressant aux divers exemples existant dans le domaine de la biométrie bi-modale audio-vidéo. Notons toutefois que ces trois catégories de fusions peuvent être combinées (Verlinde et Cholet, 1999).

5.2.1 Fusion de données

Fusion de données brutes

Fusionner les données brutes provenant de différentes sources permet de simplifier l'architecture des systèmes biométriques. Les sorties des capteurs sont regroupées pour ne former qu'un seul signal qui est alors utilisé comme entrée d'un système de reconnaissance automatique. Les signaux provenant des différentes modalités doivent partager certaines caractéristiques, comme la fréquence d'échantillonnage ou la dimension pour permettre ce type de fusion (Iyengar et al., 1995). Il est alors possible de sommer les signaux ou de concaténer les données pour en faire un signal unique qui est ensuite paramétrisé (Sanderson et Paliwal, 2004), (Matthews et al., 1998), (Adjoudani et Benoît, 1995).

La fusion de données brutes simplifie l'architecture des systèmes biométriques car seuls les capteurs sont dédoublés. Les modules de paramétrisation, reconnaissance et

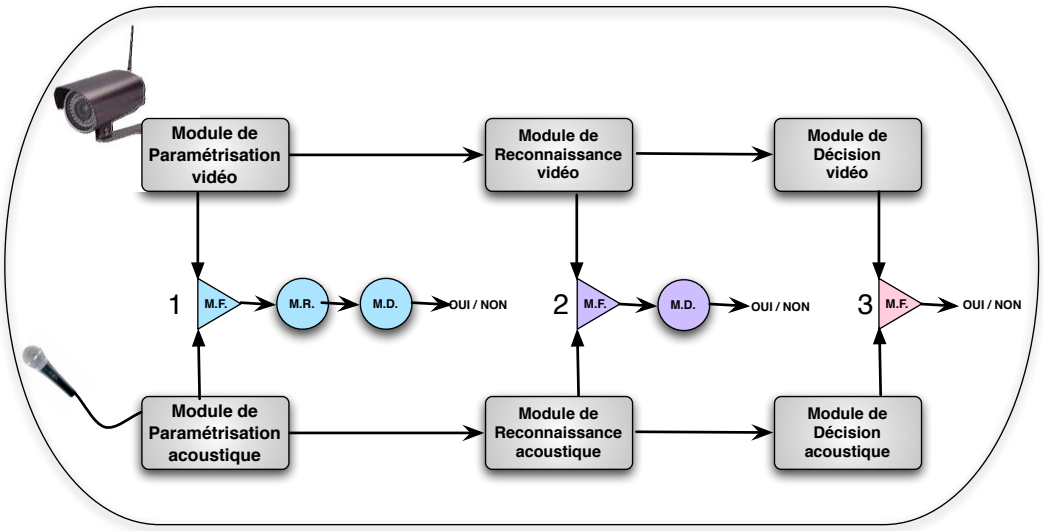


FIG. 5.2: *Différents niveaux de fusion d'informations possibles en biométrie bi-modale audio-vidéo.*
Niveau 1 : Fusion de Données ;
Niveau 2 : Fusion de Scores ;
Niveau 3 : Fusion de Décisions.
M.F. → Module de Fusion ; M.R. → Module de Reconnaissance ; M.D. → Module de Décision

décision sont uniques et n'utilisent qu'un flux de données. De plus, cette fusion autorise l'usage de la quasi totalité des méthodes développées pour l'une ou l'autre des modalité considérées. Seule la taille des données est modifiée. Le traitement reste le même.

De ce fait, les approches utilisant une fusion de données brutes peuvent améliorer les performances des systèmes, mais demeurent limitées car elles ne considèrent pas les natures différentes des informations d'entrée. En effet, des données présentant des valeurs et des dynamiques très différentes sont traitées simultanément.

Fusion de paramètres

La fusion de paramètres consiste à fusionner les données après une première étape de paramétrisation, spécifique à chaque modalité. Cette approche permet d'appliquer aux données un traitement en lien avec leur nature et prend donc en considération leur caractéristiques propres (dynamique, dimension etc.). Ce traitement spécifique dépend, la plupart du temps, de la seule nature des données, mais peut également être dépendant des autre modalités avec lesquelles elles vont être fusionnées.

Dans (Fox et Reilly, 2003), les auteurs présentent une fusion de paramètres audio et vidéo. Ils extraient des paramètres MFCC (cf. section 3.1.2) à une fréquence de 100Hz.

De chaque image de la séquence vidéo sont extraits des paramètres fréquentiels résultant d'une transformée en cosinus discrète. Fox et Reilly (2003) augmentent la fréquence des vecteurs de paramètres vidéo par une interpolation qui permet d'égaler la fréquence des paramètres audio et de compenser l'asynchronie des deux modalités. Les vecteurs de paramètres audio et vidéo synchrones sont finalement concaténés.
Le module de reconnaissance utilisé dans cette approche est un HMM identique à ceux décrits dans les sections 3.3.2 ou 4.2.2. Les résultas décrits par Fox et Reilly montrent que les performances de cette approche audio-vidéo sont très supérieures à celles d'une approche vidéo identique, mais ne parviennent pas au niveau de l'approche audio correspondante, en reconnaissance de parole comme en reconnaissance du locuteur.
Ce phénomène peut avoir plusieurs explications. Il peut être du à une mauvaise synchronisation des signaux audio et vidéo. Il peut également provenir d'un manque de données d'apprentissage pour le modèle HMM. En effet l'augmentation de la dimension vectorielle nécessite une augmentation de la quantité de données d'apprentissage. À défaut, cette augmentation peut entraîner une dégradation des performances du système.
Enfin, il est regrettable que la base de données utilisée ne contienne que des données enregistrées dans un environnement contrôlé, alors que l'un des intérêts des systèmes multi-modaux réside dans leur robustesse supposée aux environnements bruités.

Une fusion de paramètres MFCC avec des paramètres représentant le mouvement des lèvres permet, en revanche, à Faraj et Bigun (2007) d'améliorer les performances de leur système de reconnaissance du locuteur. Le gain observé reste cependant modéré, comparativement au surcoût calculatoire du à la complexité d'extraction des paramètres vidéo.

Dans (Patterson et Gowdy, 2003), les paramètres audio et vidéo utilisés sont les mêmes que dans (Fox et Reilly, 2003). Le système de reconnaissance ne repose plus, cette fois, sur des modèles HMMs, mais sur des modèles GMMs qui, contrairement aux HMMs, améliorent les performances. De cette observation, il faut conclure que quelle que soit la fusion de paramètres opérée, chaque module qui compose un système de reconnaissance biométrique doit être adapté aux données considérées. L'adaptation d'un seul GMM par locuteur et non d'un HMM complet, peut expliquer que, contrairement aux résultats de Fox et Reilly (2003), l'approche multi-modale améliore ici les performances.

Comme nous l'avons mentionné plus haut, la grande dimension des vecteurs de paramètres, obtenue par la concaténation des différentes modalités, peut nuire aux performances des approches. Notons que cette dimension entraîne, quoi qu'il en soit, un surcoût calculatoire important.
Des solutions réduisant la dimension de ces vecteurs ont été présentées dans les sections 4.1.2 et 4.1.3. Cependant, ces méthodes ignorent les natures différentes des informations représentées dans chaque dimension des vecteurs de paramètres.
L'analyse de la co-inertie (COIA) proposée par Doledec et Chessel (1994) a été appli-

quée avec succès en reconnaissance bi-modale audio-vidéo par Goecke (2005) et dans (Eveno et al., 2004) où elle est jugée plus stable que l'analyse canonique (CANCOR) (Gittins, 1985).

5.2.2 Fusion de scores

La fusion de scores est certainement l'approche la plus utilisée. Chaque modalité dispose de ses propres modules de paramétrisation et de reconnaissance. Lors d'une tentative d'authentification, un score est calculé pour chaque modalité.
Comme pour la fusion de décisions, l'idée consiste à retarder le processus de fusion afin de conserver un maximum d'information le plus longtemps possible. Cependant, il est possible que de la fusion d'un système efficace avec un système moins efficace résulte un système moyen. Il faut alors trouver une stratégie de fusion qui tire avantage des différents scores et qui incorpore également des informations a priori sur ces scores pour permettre une fusion efficace (Verlinde et al., 2000).

Comme nous l'avons déjà expliqué, les scores proviennent de méthodes et de traitements différents. Il en résulte que les natures de ces scores ne sont pas directement comparables et que ces différences doivent être prises en considération au sein de la stratégie de fusion.

Dans la littérature, certains modules de fusion calculent un score unique à partir des différentes sorties des modules de reconnaissance, d'autres utilisent l'ensemble des scores de chaque modalité pour prendre une décision.

Fusion en un score unique

La fusion de plusieurs scores en un seul qui soit le plus représentatif de l'ensemble des scores de départ fait logiquement intervenir la notion de moyenne. Il apparaît cependant plus prudent d'accorder au score de chacune de ces modalités un poids en rapport avec la confiance que l'on porte à celle-ci (cf. équation 5.1).
Il est possible de fusionner les scores au moyen d'une somme ou d'un produit pondéré, cependant, Kittler et al. (1998) ont montré que la fusion obtenue par somme pondérée est plus robuste qu'une fusion par produit pondéré. De nombreux travaux ont été réalisés afin de calculer le poids optimal de chaque modalité dans une fusion.

$$S_{fusion} = \frac{\alpha_{audio}.S_{audio} + \alpha_{video}.S_{video}}{\alpha_{audio} + \alpha_{video}} \tag{5.1}$$

Dans (Fox et al., 2005), les poids alloués aux deux modalités sont déterminés par la dynamique des scores obtenus par le système lors d'une phase de calibration. Cette

approche tend à accorder à une modalité une confiance proportionnelle à son pouvoir discriminant.

De façon à peu près similaire, Maison et al. (1999) utilisent les valeurs de $cos(\alpha)$ et $sin(\alpha)$ pour pondérer les modalités audio et vidéo. α est déterminé de façon empirique à partir des résultats obtenus sur les données de développement.

L'approche développée par Wark et Sridharan (2001) permet de fixer le poids de chaque modalité en fonction de probabilités à posteriori et du signal lui même. Le but visé est de diminuer le poids de la composante audio du système lorsque le signal acoustique est bruité.

Prise de décision à partir de plusieurs scores

Les différents scores calculés par les modules de reconnaissance de chaque modalité peuvent être les entrées d'un module de décision qui fourni une décision unique. C'est le cas dans (Teoh et al., 2004), où les scores des différentes modalités sont perçus comme les différentes composantes d'un score multidimensionnel. Chaque test d'authentification est représenté par un point unique dans l'espace des scores. Les distances entre ce point de l'espace et des scores clients et imposteurs résultants de la phase de développement du système sont calculés. La méthode des k-plus proches voisins détermine s'il s'agit d'un test imposteur ou client. Les scores des différentes modalités sont normalisés, ce qui est une façon d'attribuer plus de poids à une modalité ou à une autre.

Cette approche peut être étendue à d'autres classifieurs binaires qui seraient utilisés pour séparer les scores imposteurs et clients dans l'espace des scores. Un certain nombre de ces classifieurs sont comparés dans (Ben-Yacoub et al., 1999). Il apparaît que les SVMs et les classifieurs Bayésiens obtiennent de bons résultats.

5.2.3 Fusion de Décisions

Pour la fusion de décisions (Achermann et Bunke, 1996), (Ho et al., 1992), (Verlinde et Cholet, 1999), chaque système de vérification uni-modal fournit une décision binaire. Ces décisions peuvent être combinées avec des opérateurs OU et ET, ou par un vote majoritaire.

D'après Daugman (2000), l'utilisation des opérateurs OU et ET est conditionnée par les performances des systèmes et le type d'erreurs qu'ils produisent. Les paragraphes ci-dessous expliquent ces contraintes dans le cas d'un système bimodal pour lequel des tests sont effectués pour deux modalités, 1 et 2. Pour chacun de ces tests, il est possible

d'obtenir une fausse acceptation avec une probabilité $P(FA)$ et un faux rejet avec une probabilité $P(FR)$ (cf. section 2.3.1).

Fusion par un opérateur OU

Dans ce cas, l'utilisateur est accepté si au moins un des deux tests est positif. Dans cette configuration, un faux rejet ne peut exister que si les deux tests produisent un faux rejet. La probabilité finale de faux rejet $P(FR)$ est le produit des deux probabilités de faux rejet

$$P(FR) = P_1(FR).P_2(FR) \tag{5.2}$$

La probabilité de fausse acceptation finale est décrite par :

$$P(FA) = 1 - [1 - P_1(FA)].[1 - P_2(FA)] \tag{5.3}$$
$$= P_1(FA) + P_2(FA) - P_1(FA).P_2(FA) \tag{5.4}$$

Cette expression provient de ce que l'événement "fausse acceptation finale" est complémentaire de l'événement « le test 1 ne produit pas de fausse acceptation ET le test 2 ne produit pas de fausse acceptation ».
La probabilité finale de faux rejet est donc clairement plus faible que pour les modalités 1 et 2 considérées isolément, alors que la probabilité de fausse acceptation est plus élevée que dans le cas uni-modal.

Fusion par un opérateur ET

Avec un opérateur ET, une fausse acceptation ne survient que si le résultat de chaque test est une fausse acceptation. La probabilité de fausse acceptation est donc le produit des probabilités obtenues pour chacun des tests :

$$P(FA) = P_1(FA).P_2(FA) \tag{5.5}$$

Mais de façon symétrique, la probabilité de faux rejets devient :

$$P(FA) = 1 - [1 - P_1(FR)].[1 - P_2(FR)] \tag{5.6}$$
$$= P_1(FR) + P_2(FR) - P_1(FR).P_2(FR) \tag{5.7}$$

Les conclusions dans ce cas sont symétriques à celles obtenues pour l'usage de l'opérateur OU.

Exemple

Supposons maintenant que, dans l'approche bi-modale considérée, la première modalité produise des taux de faux rejets et d'acceptations de 1% alors que la deuxième modalité, plus sûre, obtienne des taux égaux à 1‰. Si 100 000 tests imposteurs et autant de tests clients sont réalisés, la modalité 1 conduira à 2000 erreurs alors que la deuxième

modalité produira seulement 200 erreurs.

Dans le cas où les deux systèmes fonctionnent à un point de fonctionnement particulier, qui est le point correspondant au taux d'égales erreurs (voir la section 2.3.2), la combinaison des deux modalités par un opérateur *OU* produira 1099 fausses acceptations et 1 faux rejet. La fusion par un opérateur *ET* produira, elle, 1 fausse acceptation et 1099 faux rejets. Les deux opérateurs mènent donc au même nombre d'erreurs et celui-ci est plus de 5,5 fois supérieur au nombre d'erreurs réalisées par la meilleure des deux modalités.
Cet exemple montre qu'il est plus intéressant d'utiliser une modalité sûre seule plutôt que de la coupler avec une autre modalité moins sûre. Ceci n'est cependant vrai qu'au point de fonctionnement particulier correspondant au taux d'égales erreurs.

D'après les équations 5.2 à 5.7, l'utilisation d'une seconde modalité moins sûre apporte une amélioration pour une fusion par opérateur *OU* si le point de fonctionnement de ce second système permet d'obtenir un taux de fausse acceptation plus petit que la moitié du taux d'égales erreurs du premier système.
De la même façon, pour un opérateur *ET*, un gain est obtenu si le taux de faux rejets du système le moins sûr est inférieur à la moitié du taux d'égales erreurs du premier système.

5.3 Traitement conjoint des modalités audio et vidéo

Certaines approches développées en biométrie bi-modale audio-vidéo ont privilégié une utilisation conjointe des deux modalités plutôt qu'une fusion de celles-ci. Les approches, présentées ici, exploitent les informations spécifiques à chaque modalité en considérant leurs différentes natures mais également leur corrélation temporelle.

5.3.1 Systèmes biométriques multi-flux

L'architecture correspondant à un système que nous appellerons « multi-flux » est présentée par la figure 5.3.

Nous décrivons dans cette partie différentes approches biométriques multi-flux. Les algorithmes développés pour ce type d'approches sont généralement assez complexes du fait du traitement conjoint de deux flux de natures différentes.
Les paramètres sont extraits séparément. Un module de reconnaissance unique admet ces deux flux en entrée et retourne un score unique et un module de décision fournit la décision finale.

La structure des HMMs se prête facilement à des aménagements en vue d'une utilisation conjointe de deux modalités. Les variations de ce thème sont nombreuses dans

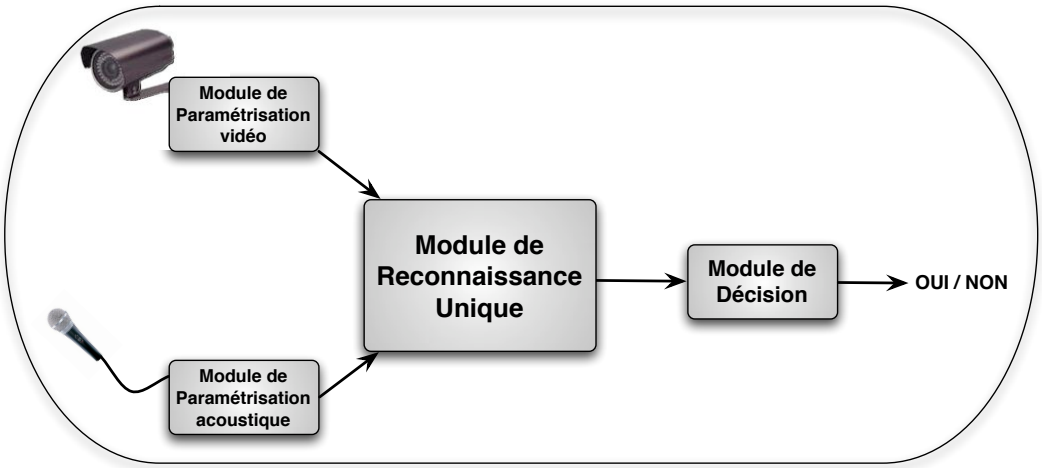


FIG. 5.3: *Architecture des systèmes de vérification biométrique bi-modaux multi-flux appliquée à l'audio-vidéo.*

la littérature et particulièrement dans le domaine de la reconnaissance de parole, où la prise en considération de la corrélation temporelle de ces deux modalités est maintenant bien ancrée.

Une première possibilité, décrite dans (Dean et al., 2006), intègre au sein d'une architecture HMM classique les données audio et vidéo. Pour ce faire, chaque état du FHMM (Fused HMM) est un doublet (G_a, G_v) où G_a et G_v sont des GMMs représentant respectivement l'état audio et vidéo. Soit deux observations audio et vidéo simultanées : O^A et O^V. La probabilité de ces observations est donnée par :

$$p_A(O^A, O^V) = p(O^A).p(O^V|\hat{S}^A) \tag{5.8}$$

où p_A indique que la modalité audio est considérée comme la modalité dominante et $p(O^V|\hat{S}^A)$ est la probabilité d'observer O^V étant donnée l'estimation $\hat{S}^A$ de la séquence des états audio qui a produit l'observation O^A. Dans ce cas, $p(O^A)$ peut être obtenue par un HMM de même structure, mais pour lequel les états seraient réduits au GMM acoustique.

Cette modélisation permet d'intégrer la dépendance statistique des données audio et vidéo au sein des probabilités acoustiques. Ce terme est décrit par les auteurs comme une pondération corrélationnelle qui est d'autant plus élevée que la probabilité d'observer O^A et O^V à un instant donné est forte.

Le modèle de HMMs couplés présenté par Nefian et al. (2003) comprend, lui, un HMM pour chaque modalité. La probabilité de transiter dans un état donné d'une modalité au temps t est conditionnée par l'état du système au temps $t-1$ pour chaque modalité, cette architecture est représentée par la figure 5.4.

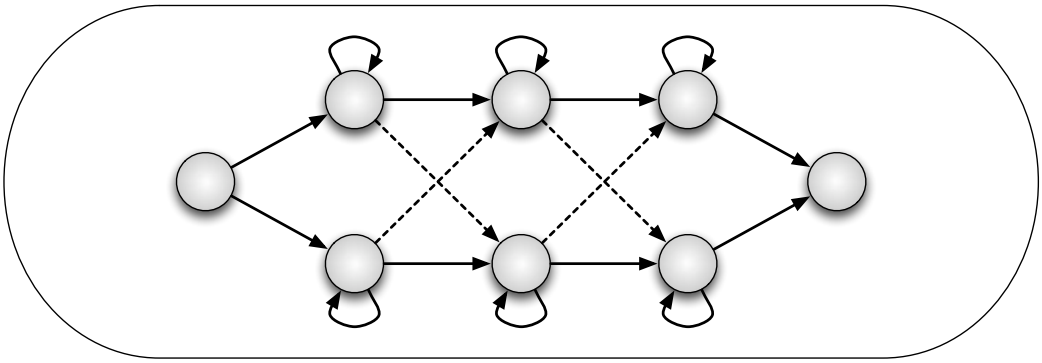


FIG. 5.4: *Architecture d'un modèle à modèles de Markov cachés couplés (CHMM) pour une application bi-modale.*

Nous ne présentons pas ici l'ensemble des architectures développées dans les dernières décennies à partir des modèles de Markov cachés, cependant il faut indiquer que certains travaux, comme ceux de Bengio (2003b) ou Wang et al. (2008), se sont attachés à considérer l'asynchronie des modalités audio et vidéo en parole. Les travaux de Jourlin (1998) présentent également un intérêt pour le formalisme des produits de HMMs qui y est introduit.

Dans (Hazen, 2006) et (Hazen et al., 2004), la modalité audio est utilisée pour segmenter le flux vidéo afin de contraindre le modèle HMM pour la reconnaissance de la parole. La modalité dominante est dans ce cas la vidéo, puisque le flux audio n'est utilisé que comme référence temporelle. Ces travaux ont démontré que l'utilisation d'une contrainte temporelle provenant d'une seconde modalité améliore la reconnaissance vidéo de la parole.

Il nous paraît intéressant de considérer, pour la suite de ce travail, la modalité audio comme modalité dominante pour tirer parti de ses performances supérieures à celles des approches vidéo. Cependant il nous paraît plus approprié d'utiliser une référence temporelle audio dans un système vidéo plutôt que le contraire. En effet, nous considérons qu'il existe une fonction injective de l'ensemble des phonèmes dans l'ensemble des visèmes et nous avons vu dans la section 4.2.2 que la réciproque n'est pas vraie.

5.3.2 Détection d'impostures : "liveness test"

Les tests de « liveness »vérifient la cohérence entre les flux audio et vidéo présentés à un système de reconnaissance biométrique bi-modale. Même s'ils ne constituent pas un test de vérification, ils peuvent être utiles à la détection des impostures. Les méthodes d'analyse de données CANCOR et COIA décrites dans la section 5.2.1 sont utilisées dans (Eveno et al., 2004) et (Chetty et Wagner, 2005) pour mesurer la corrélation existant entre les deux flux.

Chetty et Wagner (2004b) montrent qu'un système de reconnaissance utilisant le paradigme GMM/UBM est capable de détecter une imposture utilisant une image fixe et un signal audio grâce à une fusion de paramètres MFCC avec des paramètres décrivant la géométrie des lèvres.

Conclusion

Nous avons présenté dans ce chapitre les deux classes d'approches existantes en biométrie bi-modale audio-vidéo : les approches par fusion d'informations et celles qui reposent sur un module de reconnaissance multi-flux.

Améliorer les performances en authentification par une fusion, quel que soit le niveau auquel elle est réalisée, relève actuellement d'un processus supervisé qui reste perfectible. De plus, ces approches ne permettent généralement pas de mettre à profit la corrélation temporelle des données audio et vidéo due au processus de production de la parole et nécessitent l'emploi de deux systèmes de reconnaissance plus ou moins disjoints. Cet état de faits implique, lors de l'ajout d'une nouvelle modalité, une augmentation proportionnelle des ressources nécessaires.
L'emploi de la fusion souffre également de l'asynchronie des deux modalités et de la difficulté à traiter conjointement des informations de natures différentes.

Les approches multi-flux, qui prennent en compte l'information temporelle des deux modalités, sont en général assez complexes et les processus d'enrôlement et de test sont très gourmands en termes de ressources et de temps. Elles sont donc difficilement exploitables pour des applications réelles.

L'utilisation conjointe des modalités audio et vidéo permet toutefois d'améliorer sensiblement les performances des systèmes de reconnaissance. L'apport de la bi-modalité est cependant difficile à évaluer car, comme nous le verrons dans le chapitre 6, il n'existe pas de base de données contenant les mêmes enregistrements à divers niveaux de bruit et suffisamment de sessions pour obtenir des résultats statistiquement fiables. L'ajout de bruit sur des sessions enregistrées dans un environnement contrôlé permet bien sûr de se faire une idée des performances mais ne remplace en aucun cas des données enregistrées en conditions réelles. En effet, le bruit présent dans l'environnement, où sont effectuées les acquisitions biométriques, peut perturber les systèmes automatiques, mais il peut également être la cause de changements marqués chez l'individu à reconnaitre. L'effet Lombard, qui fait que les locuteurs parlent plus fort, mais aussi plus haut et qu'ils augmentent naturellement leur fréquence fondamentale quand ils parlent dans le bruit, est une bonne illustration de ce phénomène (Doddington et al., 2000), (Hansen et Bria, 1990).

Troisième partie

Vérification du locuteur et synchronisation contrainte

Introduction

Motivations et contraintes

LES travaux que nous présentons sont motivés par la volonté d'améliorer les performances et la robustesse des systèmes de reconnaissance du locuteur en environnement embarqué. Il s'agit de développer une approche adaptée à la forte variabilité de l'environnement mais également aux contraintes ergonomiques limitant le nombre et la durée des sessions d'acquisition. Pour ce faire, nous choisissons d'exploiter la structure temporelle du signal de parole et de la renforcer par une contrainte supplémentaire qui peut avoir pour origine le flux vidéo du signal de parole.

L'état-de-l'art des approches audio et vidéo des chapitres 3 et 4 a montré que ces modalités permettent d'obtenir des performances d'authentification satisfaisantes.
Malgré la difficulté inhérente à la comparaison de ces deux modalités, nous avons conclu que la modalité audio surpasse la modalité vidéo en terme de fiabilité et de stabilité des résultats. Cette conclusion peut, bien évidemment, être nuancée selon les applications et les conditions d'utilisation considérées.
L'influence de l'environnement peut justement être atténuée par l'utilisation conjointe d'une deuxième modalité biométrique comme nous l'avons vu dans le chapitre 5.
La combinaison de deux modalités au sein d'un système biométrique peut présenter de nombreux aspects, selon la forme que prend la « fusion » des informations, le niveau auquel est effectuée leur intégration et le type d'approche choisi. Enfin, il est possible d'exploiter la complémentarité ou la redondance de ces modalités.

Nous avons choisi de développer une approche de vérification du locuteur synchronisée par un processus externe. Le cœur de cette architecture est constitué d'un système de reconnaissance du locuteur dépendant du texte. Cette architecture est renforcée par une information temporelle issue d'un processus additionnel, indépendant du coeur acoustique de notre approche et que nous appellerons processus externe. La redondance des informations temporelles issues des flux audio et du processus externe est utilisée pour structurer les modèles acoustiques, d'après l'exemple du DTW ou, plus précisément, de l'approche GDW (cf. partie 3.3.3).

La dépendance au texte est introduite dans le but d'exploiter la structure temporelle

commune aux deux processus. Elle permet également de réduire la variabilité inter-sessions. Le texte énoncé pour l'authentification est laissé au libre choix de chaque utilisateur. L'utilisation de mots de passe personnels augmente potentiellement la variabilité inter-locuteurs en introduisant une information connue du seul client.

Nous ajoutons enfin une contrainte supplémentaire : l'utilisation du contenu linguistique et du processus externe ne doivent pas engendrer un surcoût trop important en terme d'utilisation mémoire et de temps de calcul, comme c'est souvent le cas dans les approches présentées dans le chapitre 5.

Description de l'architecture

L'approche que nous avons développée est fondée sur une architecture hiérarchique à trois niveaux, représentée par la figure 5.5. Cette architecture s'inspire des travaux de Bonastre et al. (2003) et Lévy et al. (2006), respectivement dans le contexte de la reconnaissance du locuteur (utilisant un algorithme DTW) et de la reconnaissance de mots isolés. Les trois niveaux, décrits ci-après, permettent une spécialisation hiérarchique.

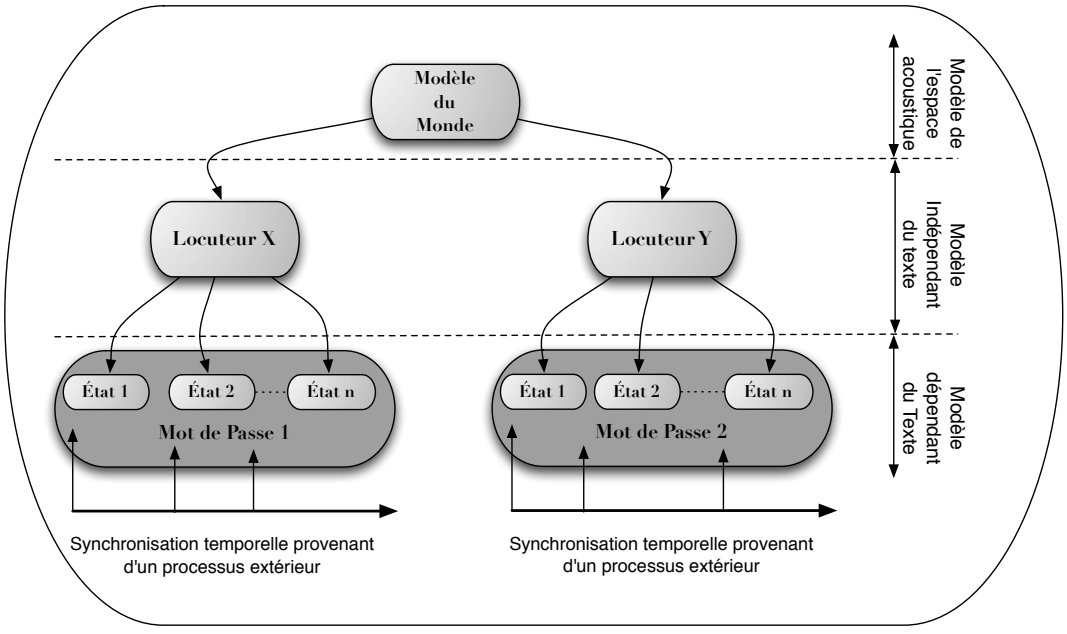


FIG. 5.5: *Architecture acoustique hiérarchique intégrant une synchronisation externe fournie par un processus extérieur.*

Premier niveau : modèle de l'espace acoustique

Le premier niveau est le moins spécialisé. Il s'agit d'un modèle représentant l'ensemble de l'espace acoustique. La justification de ce modèle ainsi que son utilisation sont détaillées dans la section 7.1.3.

Deuxième niveau : modèle de locuteur indépendant du texte

Le niveau intermédiaire contient les modèles de locuteurs indépendants du texte. Ces modèles sont dérivés du modèle de l'espace acoustique de la couche supérieure. Ils sont obtenus grâce à l'ensemble des enregistrements audio du locuteur. L'architecture des deux premiers niveaux est décrite de façon détaillée dans le chapitre 7.

Troisième niveau : modèle de locuteur dépendant du texte

Le dernier niveau regroupe les modèles de locuteurs dépendants du texte. Chaque modèle de mot de passe est obtenu à partir d'un modèle de locuteur indépendant du texte. Il représente les informations spécifiques à ce locuteur mais contient également une information sur la structure temporelle du mot de passe (cf. chapitre 8).

Synchronisation externe

La synchronisation temporelle issue d'un processus extérieur est utilisée pour déterminer et renforcer la structure temporelle du modèle acoustique de mot de passe du troisième niveau. Ce procédé est décrit dans le chapitre 9.

Réalisation

L'ensemble des développements réalisés pour nos travaux repose sur la plate-forme logicielle libre MISTRAL[1] (Meigner et al., 2008) et plus particulièrement sur la librairie ALIZE (Fauve et al., 2007), (Bonastre et al., 2008), développée au Laboratoire Informatique d'Avignon (LIA).

[1] `http://mistral.univ-avignon.fr/`

Chapitre 6

Corpus et protocole expérimental

Sommaire

Résumé

Ce chapitre justifie le choix de la base de données MyIdea, utilisée pour la validation de notre approche et présente le protocole expérimental que nous avons défini. Comme ces deux sujets sont des points critiques de la recherche en biométrie, nous discutons leurs avantages et inconvénients.

Introduction

LES bases de données constituent un point critique pour les recherches en biométrie bi-modale audio-vidéo. Pour pallier le manque de données, de nombreux corpus ont été enregistrés durant les deux dernières décennies. La quantité et la représentativité des données restent cependant des points bloquants. En effet, la validation expérimentale des méthodes de vérification d'identité bi-modale requiert plusieurs sessions d'acquisition et un nombre important de locuteurs différents, si possible représentatifs de la population visée. Le protocole d'enregistrement des bases de données doit également être défini avec soin, en fonction des tâches visées, tant en termes de contenu qu'en termes de conditions d'enregistrements.

Dans ce chapitre, nous introduisons les critères qui nous ont guidé pour le choix d'un corpus. Nous confrontons ensuite les principales bases de données existantes à ces contraintes avant de présenter la base de données MyIdea que nous avons utilisée. Le protocole de test mis en place est enfin présenté et critiqué.

6.1 Contraintes fixées

La base de données audio-vidéo nécessaire à la validation de notre approche doit répondre aux critères suivants :

- les locuteurs enregistrés doivent être aussi nombreux que possible ;
- la base doit intégrer des séquences au cours desquels les locuteurs prononcent un texte fixe ;
- le nombre et la variabilité de ces énoncés fixes doivent être aussi importants que possible afin d'augmenter la variabilité lexicale ;
- le nombre de sessions par locuteur doit être le plus important possible afin d'augmenter la variabilité inter-sessions et de permettre la réalisation d'un grand nombre de tests clients et imposteurs.

Dans le cadre de la reconnaissance du locuteur, et en considérant que la détection audio de genre est une technique arrivée à maturité, les tests réalisés sont dépendants du genre.

6.2 Bases de données existantes

Le tableau 6.1 propose un aperçu des principales bases de données audio-vidéo existantes. Une liste plus complète des corpus disponibles dans le domaine image et vidéo est présentée sur le site `http://www.face-rec.org/databases/`. La description donnée dans le tableau suivant n'est pas exhaustive, mais est centrée sur les principaux critères qui ont retenu notre attention.

Corpus	#Locuteurs	#Hommes	#Sessions	Conditions d'enregistrement	Texte prononcé
AVOZES	20	10	2	Studio	3 phrases
AV-TIMIT	43	-	3	Studio	10 phrases dépendantes de la session
BANCA	208	104	12	Studio, réalistes et adverses	Informations personnelles + 1 série de digits
BIOMET	91	45	3	Studio	Nombreuses phrases dont 10 phrases courtes + 2 phrases longues
Extended M2VTS	37	-	5	Studio	1 phrase + 1 série de digits
MVGL-AVD	50	-	-	Studio	nom prénom + données de 5 autres clients
MyIdea	30	30	3	Studio	Nombreuses phrases dont 10 phrases courtes + 2 phrases longues
ValidDB	106	76	5	1 session studio + 4 sessions en environnements de travail	1 phrase + 1 série de digits
Mobio[a]	~ 160	~ 54	12	environnements de travail	Réponses libres à une série de questions

TAB. 6.1: *Présentation des principales bases de données audio-vidéo existantes*
a - La base de données MOBIO sera disponible courant 2010

Parmi les corpus présentés ici, deux seulement sont susceptibles de répondre à nos besoins : BIOMET (Garcia-Salicetti et al., 2003) et MyIdea.
En effet, le corpus AV-TIMIT (Sanderson, 2008) ne contient aucune répétition des énoncés fixés, les 10 phrases imposées par le protocole d'enregistrement étant dépendantes de la session.
Pour les bases AVOZES (Goecke et al., 2000), BANCA (Bailly-Bailliere et al., 2003), XM2VTS (Messer et al., 1999), MVGL-AVD (Erzin et al., 2005) et ValidDB (Fox et al., 2005), les protocoles ne prévoient pas assez de contenus lexicaux différents ou de sessions pour permettre la validation d'une approche reposant sur l'utilisation de mots de passe personnels.

La base de données MyIdea qui reprend le protocole d'enregistrement de BIOMET a été choisie pour une raison de disponibilité. La section suivante décrit brièvement le corpus vidéo et le protocole mis en place pour valider nos travaux. Une description complète est disponible dans l'annexe A.

6.3 La base de données MyIdea

Tous les résultats présentés dans ce document ont été obtenus à partir de la base de données MyIdea qui présente cependant de nombreux inconvénients. Ce corpus est le plus à même, à notre connaissance, de répondre à nos attentes, notamment à l'utilisation de mots de passe personnalisés. Dans cette section, nous proposons une brève description de cette base de données, accompagnée d'une discussion critique.

6.3.1 Description du corpus

MyIdea se présente comme une extension de différentes bases de données multimodales existantes : BANCA , BIOMET, XM2VTS, Mcyt (Ortega-Garcia et al., 2003) et IAM (Marti et Bunke, 2002). La partie audio-vidéo de cette base contient des enregistrements d'une trentaine de locuteurs masculins réalisés lors de trois sessions dans un environnement contrôlé. Les enregistrements sonores sont exempts de bruit et l'éclairage est contrôlé.
Chaque locuteur prononce 28 énoncés différents comprenant 12 phrases indépendantes du locuteur et de la session, qui tiennent un rôle particulier pour notre étude. Ces 12 phrases comprennent 2 phrases longues et 10 phrases courtes utilisées comme mots de passe :

phrases longues ($\sim$ 5 secondes)

1. Alors que Monsieur Gorbatchev regagnait Moscou au terme d'un difficile voyage en Lituanie, une partie du Caucase s'est embrasée.
2. Chaque jour ils reçoivent, dans la bonne humeur, la visite du commissaire des renseignements généraux, qui suit de loin l'opération.

phrases courtes ($\sim$ 2 secondes)

1. Il se garantira du froid avec un bon capuchon.
2. Annie s'ennuie loin de mes parents.
3. Les deux camions se sont heurtés de face.
4. Un loup s'est jeté immédiatement sur la petite chèvre.
5. Dés que le tambour bat, les gens accourent.
6. Mon père m'a donné l'autorisation.
7. Vous poussez des cris de colère.
8. Ce petit canard apprend à nager.
9. La voiture s'est arrêtée au feu rouge.
10. La vaisselle propre est mise sur l'évier.

Nous présentons maintenant les principaux points forts et points faibles de la base de données MyIdea, choisie pour valider nos travaux.

Points forts

- MyIdea contient 10 phrases fixes indépendantes du locuteur et de la session qui peuvent être considérées comme des mots de passe.
- Le contenu lexical des 10 phrases mots de passe est varié.

Points faibles

- Le nombre de locuteurs enregistrés (30) reste limité.
- Les enregistrements sont tous réalisés dans un environnement sans bruit et dont la luminosité est contrôlée. Ces conditions ne permettent pas d'éprouver la robustesse de notre approche aux différents environnements auxquels peut être confronté un système embarqué.

6.3.2 Discussion

Les caractéristiques de la base de données MyIdea imposent certaines limites expérimentales. Nous proposons ici une discussion de ces principales caractéristiques.

Le nombre d'individus

L'estimation des performances d'un système biométrique passe par une connaissance des distributions des scores clients et imposteurs obtenus par ce système. D'après l'interprétation généralement consentie du Théorème Central Limite, la moyenne d'une variable aléatoire peut être estimée à partir de 30 de ses réalisations. Dans le contexte biométrique, il ne s'agit pas seulement d'obtenir trente scores clients et trente scores imposteurs pour connaître les performances du système. En effet, cette approximation peut être appliquée aux scores calculés pour un modèle en particulier. Connaître la « réponse » de ce modèle nécessiterait alors la réalisation de trente tests imposteurs et trente tests clients pour ce seul modèle. Il est possible, en raisonnant sur les séquences de tests, de justifier la comparaison de chaque séquence de tests avec trente modèles de locuteurs différents. Un tel raisonnement poussé à l'extrême rendrait certainement la réalisation d'un test *grandeur nature* plus aisée que l'obtention de données statistiquement *fiables*.

Le choix d'une base de données intégrant seulement trente locuteurs n'est évidemment pas fait pour répondre aux conditions minimales de l'approximation normale. Ce nombre réduit limite fortement la variabilité inter-locuteurs.
Le paradigme GMM/UBM suggère qu'il existe, au sein de l'espace acoustique, un sous-espace acoustique (de dimension relativement faible) dans lequel il est possible de représenter l'ensemble des locuteurs (cf. section 7.2.3). Cette interprétation soulève immédiatement plusieurs questions :

– *Est-il possible d'estimer ce sous-espace avec le nombre limité de locuteurs disponibles ?*
– *Les clients/imposteurs disponibles suffisent-ils à « paver » cet espace ?*
– *Ces locuteurs sont ils répartis de façon représentative dans le sous-espace défini ?*

Le faible nombre de locuteur a une autre conséquence immédiate puisqu'il rend impossible la distinction entre données de développement et données de validation. Cet amalgame introduit certainement un biais dans les résultats obtenus.

Nous n'apportons pas de réponse à toutes ces questions dans ce document mais il est important de les garder à l'esprit, lors de la lecture de résultats comme dans un cadre plus général.

Le contenu lexical

De même que les variabilités inter et intra-locuteur peuvent être sujettes à questionnement, la variabilité du contenu lexical des phrases utilisées dans le corpus MyIdea doit être prise en compte. La durée des 10 phrases courtes, toujours comprise entre 2 et 3 secondes, et le contenu lexical de ces phrases ne sont peut être pas représentatifs de mots de passe laissés au libre choix des utilisateurs.

Le protocole d'enregistrement

Les conditions d'enregistrement, dans un environnement isolé, sans bruit et à la luminosité contrôlée, ne permettent pas une généralisation des résultats obtenus à d'autres conditions d'utilisation plus « difficiles ».

6.4 Protocole expérimental

Nous présentons à présent le protocole expérimental mis en place. Cette description met en lumière les motivations qui ont guidé nos choix et les points critiques qui doivent être pris en compte à la lecture des résultats présentés dans ce document.

6.4.1 Description

Les trente hommes sont séparés en deux groupes, A et B de chacun quinze locuteurs. Chaque groupe est successivement considéré comme groupe de clients/imposteurs tandis que l'ensemble des enregistrements de l'autre groupe est utilisé pour entraîner le modèle du monde (Universal Background Model - UBM) supposé représenté l'ensemble de tous les locuteurs (cf. section 7.1).

Considérons, pour la suite, que le groupe A est le groupe client et que le groupe B est utilisé pour entraîner le modèle du monde (UBM). En raison du faible nombre de

locuteurs, nous utilisons un protocole *leave one out*. Chaque locuteur du groupe A est successivement considéré comme le client alors que les 14 autres locuteurs représentent les imposteurs qui tentent d'usurper son identité. Un protocole symétrique est utilisé lorsque le groupe B est le groupe clients/imposteurs et le groupe A le groupe UBM.

Le protocole mis en place intègre trois configurations d'entraînement et trois conditions de test qui peuvent être associées selon les tâches considérées.

Configurations d'entraînement

Rappelons tout d'abord que les modèles de locuteur appris durant cette phase sont de deux types :

- un modèle de locuteur, indépendant du texte ;
- un modèle de mot de passe, dépendant du locuteur et du texte prononcé.

Seules les phrases courtes et les phrases longues présentées précédemment sont utilisées lors de la phase d'apprentissage. Les phrases courtes jouent le rôle des mots de passe que chaque locuteur doit prononcer afin de s'authentifier. Les phrases longues sont utilisées pour augmenter la quantité de données d'apprentissage des modèles de locuteur indépendant du texte. L'utilisation de ces phrases longues, communes à tous les locuteurs, permet de ne pas introduire de variabilité lexicale entre les modèles indépendants du texte. La seule variabilité due au contenu lexical provient des phrases courtes utilisées, donc des mots de passe laissés au libre choix des utilisateurs. Le nombre de modèles de locuteurs simulés (900) est identique pour les trois configurations suivantes.

Configuration 1-occ Chaque locuteur dispose des 2 phrases longues et d'une occurrence d'une phrase courte.

Le modèle indépendant du texte est appris avec les 2 phrases longues et la phrase courte disponibles, ce qui constitue environ 12 secondes de parole.

Le modèle de mot de passe est appris avec une seule occurrence du mot de passe choisi (la phrase courte disponible), soit environ 2 secondes de parole.

Cette configuration constitue notre référence en terme de données d'apprentissage.

Configuration 2-occ Chaque locuteur dispose des 2 phrases longues et de 2 occurrences d'une même phrase courte.

Le modèle indépendant du texte est appris avec les 2 phrases longues et les 2 occurrences disponibles de la phrase courte, ce qui constitue environ 14 secondes de parole.

Le modèle de mot de passe est appris avec les 2 occurrences du mot de passe, soit environ 4 secondes de parole.

Cette configuration a pour but d'évaluer l'effet de l'augmentation de la quantité de données d'apprentissage (par rapport à la configuration **1-occ**) pour les deux modèles de locuteur puisque le contenu linguistique ajouté est une seconde occurrence du mot de passe.

Configuration 1-occ + aléatoire Chaque locuteur dispose des 2 phrases longues, d'une occurrence d'une phrase courte qui constitue son mot de passe et d'une occurrence d'une phrase courte supplémentaire, différente de son mot de passe.

Le modèle indépendant du texte est appris avec les 2 phrases longues et les 2 phrases courtes disponibles, ce qui constitue environ 14 secondes de parole.

Le modèle de mot de passe est appris avec une seule occurrence du mot de passe choisi, soit environ 2 secondes de parole.

Cette configuration a pour but d'évaluer l'effet de l'augmentation de la quantité de données d'apprentissage (par rapport à la configuration **1-occ**) pour le seul modèle indépendant du texte, puisque le contenu linguistique ajouté est une phrase courte différente du mot de passe choisi.

Conditions de test

Pour chacune des trois conditions de test définies dans ce protocole, les clients, pour être authentifiés, prononcent leur mot de passe. Le nombre de tests clients varie selon la configuration choisie pour l'apprentissage. Dans les conditions **1-occ** et **1-occ + aléatoire**, le nombre de tests clients total est 1800, (15 clients $\times$ 14 imposteurs $\times$ 3 sessions $\times$ 2 sessions de test $\times$ 2 groupes de locuteurs) alors que pour la condition **2-occ** ce nombre est réduit à 900 (1 seule session de tests disponible pour chaque modèle).

Condition MDP Cette condition correspond à un imposture réalisée par des locuteurs qui connaissent le mot de passe des clients. Chaque modèle est comparé à trois occurrences du mot de passe prononcées par chacun des 14 locuteurs. 37 800 tests sont ainsi réalisés (900 modèles de mots de passe $\times$ 14 imposteurs $\times$ 3 sessions). Cette condition est supposée être la plus difficile pour les système de vérification d'identité structuraux du fait de la faible variabilité lexicale.

Condition FAUX Les imposteurs ne connaissent pas le mot de passe des clients. Il prononcent donc l'une des neuf autres phrases courtes du corpus. Afin d'obtenir le même nombre d'accès imposteur que dans la condition MDP, chaque modèle est comparé à trois séquences prononcées par chacun des 14 locuteurs (15 locuteurs du groupe moins le client). 37 800 tests sont ainsi réalisés (900 modèles de mots de passe $\times$ 3 sessions $\times$ 14 locuteurs).

Condition TOUS Les deux conditions de tests précédentes sont associées pour obtenir la condition **TOUS**. Le nombre de tests imposteurs cumulés est : 75 600.

Récapitulatif

Configurations		Données utilisées
Enrôlement	**1-occ**	**indépendant du texte :** 2 phrases longues + 1 mot de passe (∼12s) **dépendant du texte :** 1 mot de passe (∼2s)
	2-occ	**indépendant du texte :** 2 phrases longues + 2 mots de passe (∼14s) **dépendant du texte :** 2 mots de passe (∼4s)
	1-occ + aléatoire	**indépendant du texte :** 2 phrases longues + 1 phrase courte + 1 mot de passe (∼14s) **dépendant du texte :** 1 mot de passe (∼2s)
Test	**MDP**	phrases courtes identiques au mot de passe du client
	FAUX	phrases courtes différentes du mot de passe du client
	TOUS	phrases courtes identiques et phrases courtes différentes du mot de passe du client

TAB. 6.2: *Récapitulatif des configurations d'apprentissage et de test du protocole expérimental*

6.4.2 Discussion

Devant les dangers liés aux évaluations - généralisation abusive ou conclusions liées à la base de données - la critique du protocole expérimental est essentielle. Le protocole développé doit être critiqué. L'utilisation du *leave one out* induit très probablement un biais. Les tests croisés entre locuteurs ne sont pas symétriques, mais il est peu probable qu'ils soient totalement décorrélés. L'effet de cette corrélation n'a pas fait, à notre connaissance, l'objet d'une étude précise et il est difficile d'apprécier le biais introduit par ce procédé.

L'apprentissage de plusieurs modèles par locuteur, par utilisation de séquences d'apprentissage différentes, permet d'effectuer un grand nombre de tests. Il est cependant illusoire de considérer que tous les modèles de locuteur appris soient réellement

indépendants et puissent pallier le manque de diversité en termes de locuteurs. Cette simulation introduit certainement un biais qu'il nous est impossible de quantifier.

La quantité limitée de répétitions des mots de passe, par un même locuteur, ne permet pas une étude approfondie de l'impact de la quantité de données sur le taux d'authentification.

Pour finir, la condition **TOUS** est supposée être la plus proche des conditions réelles d'utilisation d'un tel système, du fait de la présence conjuguée d'imposteurs connaissant le mot de passe et d'imposteurs l'ignorant. Pourtant, il est difficile de déterminer le ratio d'imposteurs ayant connaissance du mot de passe d'un client, de la même façon qu'il est généralement difficile d'estimer les ratios de tests clients et imposteurs. Néanmoins, ces informations ont une importance considérable sur les performances affichées des systèmes biométriques.

Conclusion

Nous avons présenté dans ce chapitre le corpus MyIdea sur lequel sont réalisées toutes les expériences présentées dans ce document.

Malgré les motivations qui nous ont fait choisir ce corpus, nous sommes conscients de ses limites. Cependant, il existe peu de bases de données audio-vidéo permettant d'évaluer une approche d'authentification bi-modale dépendante du texte.

Le protocole expérimental développé à cet effet exploite au maximum les données disponibles, mais introduit un biais principalement dû aux manques de locuteurs et de données et à leurs conséquences.

Chapitre 7

Représentation des locuteurs

Sommaire

Résumé

Ce chapitre présente le paradigme GMM/UBM utilisé en reconnaissance du locuteur. Nous décrivons les mélanges de Gaussiennes et leur capacité à structurer l'espace acoustique. Une première étude est réalisée afin d'évaluer la capacité de cette approche à tirer parti du contenu lexical prononcé par les locuteurs.

7.1 Le paradigme GMM/UBM

LES approches génératives utilisées en reconnaissance du locuteur reposent essentiellement sur le paradigme GMM/UBM qui a été introduit dans la section 3.2.1. Cette partie présente ce paradigme de façon détaillée tout en se focalisant sur son utilisation dans le cadre de la reconnaissance du locuteur dépendante du texte.

7.1.1 Les Mixtures de Gaussiennes en RAL

La reconnaissance du locuteur s'appuie sur une représentation discrète du signal de parole. Celui-ci est transformé en une séquence de vecteurs de paramètres, dont la fréquence d'échantillonnage est généralement 100Hz.
Considérons que chaque vecteur de paramètres extrait d'un signal de parole est une réalisation d'une variable aléatoire multi-dimensionnelle. Les approches génératives en reconnaissance du locuteur reposent sur l'hypothèse qu'il existe une fonction injective de l'ensemble des locuteurs dans l'espace des fonctions de densité de probabilité. Cette hypothèse suppose, plus précisément, que les vecteurs de paramètres provenant d'un locuteur suivent une loi de probabilité propre à ce locuteur.

La complexité de ces fonctions de densité nous conduit à rechercher une approximation suffisant à la résolution du problème de reconnaissance du locuteur. Nous avons présenté dans la partie 3.2.1 les Méthodes Statistiques du Second Ordre (MSSO) Bimbot et al. (1995). Dans cette approche, les locuteurs sont représentés par une loi Gaussienne, c'est à dire un doublet (μ, Σ) où μ est le vecteur moyen de la Gaussienne et Σ la matrice de covariance, estimée à partir de la séquence acoustique d'apprentissage X. Nous avons souligné la simplicité de la modélisation des locuteurs par MSSO et le fait qu'elle limite la granularité de modélisation des variations acoustiques.

L'utilisation de mélanges de Gaussiennes (GMMs) permet d'obtenir une approximation plus précise de la fonction de densité de probabilité caractéristique des locuteurs, tout en restant relativement simple à estimer (Reynolds et Rose, 1995), (Reynolds et al., 2000), (Bimbot et al., 2004). La densité de probabilité d'un mélange de N distributions Gaussienne est :

$$p(\mathcal{X}|\Theta) = \sum_{i=1}^{N} \gamma_i \mathcal{N}(\mathcal{X}, \mu_i, \Sigma_i) \tag{7.1}$$

telle que : $\sum_i \gamma_i = 1$ et $\forall i,\ \gamma_i \geq 0$. γ_i, μ_i et Σ_i sont respectivement le poids, le vecteur moyen et la matrice de covariance de la distribution i dans la mixture.
$\Theta = [\mu, \Sigma, \gamma]^T$ est le vecteur de paramètres global de la mixture de Gaussiennes. La densité de probabilité Gaussienne $\mathcal{N}(\mathcal{X}, \mu, \Sigma)$ a été définie dans l'équation 3.2. En reconnaissance du locuteur, la matrice de covariance est généralement supposée diagonale.

La vraisemblance pour qu'un vecteur de paramètres X ait été produit par le GMM de vecteur de paramètres Θ est :

$$f(X|\Theta) = \sum_{i=1}^{N} \gamma_i \mathcal{N}(X, \mu_i, \Sigma_i) \tag{7.2}$$

La valeur moyenne de la log-vraisemblance pour une séquence X de paramètres X_t, $t \in [1; T]$ et un GMM Θ, que nous notons $\ell(X|\Theta)$, est[1] :

$$\ell(X|\Theta) = \log\left[f(X|\Theta)\right] = \frac{1}{T}\sum_{t} \log f(X_t|\Theta) \tag{7.3}$$

7.1.2 Le rapport d'hypothèses Bayésien

En reconnaissance du locuteur, le processus de décision est basé sur un test d'hypothèses. Étant donné un signal de parole S et une identité I_X revendiquée par l'utilisateur, le système doit décider laquelle des deux hypothèses suivantes est la plus vraisemblable :

- H_0 : le signal S a été produit par I_X
- H_1 : le signal S n'a pas été produit par I_X

Le rapport de vraisemblance (Likelihood Ratio - LR) entre les deux hypothèses H_0 et H_1 pour l'identité I_X est noté $\mathcal{LR}(X, H_0, H_1)$. Le test bayésien est la comparaison du rapport de vraisemblance avec un seuil de décision Ω.

$$\mathcal{LR}(\mathcal{S}, H_0, H_1) = \frac{p(H_0|\mathcal{S})}{p(H_1|S)} \tag{7.4}$$

En pratique, il est plus facile d'estimer $p(\mathcal{S}|H_0)$ que $p(H_0|\mathcal{S})$ et le théorème de Bayes permet d'écrire :

$$p(H_0|\mathcal{S}) = \frac{p(\mathcal{S}|H_0)\, p(H_0)}{P(X)} \tag{7.5}$$

ainsi l'équation 7.4 devient :

$$\mathcal{LR}(\mathcal{S}, H_0, H_1) = \frac{p(\mathcal{S}|H_0)\, p(H_0)}{p(\mathcal{S}|H_1)\, p(H_1)} \tag{7.6}$$

les probabilités a priori $p(H_0)$ et $p(H_1)$ sont incorporées au seuil de décision Ω, et finalement, le rapport entre les deux hypothèses H_0 et H_1 s'écrit :

$$\mathcal{LR}(S, H_0, H_1) = \frac{p(\mathcal{S}|H_0)}{p(S|H_1)} \lessgtr \Omega \begin{cases} H_1 \text{ est acceptée} \\ H_0 \text{ est acceptée} \end{cases} \tag{7.7}$$

[1] Cette expression est obtenue sous l'hypothèse d'indépendance des observations X_t, qui permet d'écrire la log-vraisemblance de la séquence X comme la somme des log-vraisemblances de chaque vecteur x_t.

La valeur $p(H_0|S)$ est généralement la vraisemblance moyenne du signal S avec le modèle GMM de I_X. La valeur choisie au dénominateur du deuxième terme de l'équation 7.7 représente la vraisemblance du modèle de non-locuteur (tous les locuteurs hormis I_X). Cette valeur n'est pas observable, elle est seulement la probabilité complémentaire de $p(H_0|S)$. Nous verrons par la suite que cette valeur, probabilité d'un événement non observable, est difficile à estimer et ne peut qu'être approximée.

7.1.3 Modélisation de l'hypothèse de non-locuteur

La probabilité de l'hypothèse de non-locuteur dans le test Bayésien est souvent approximée grâce à une cohorte d'imposteurs ou un modèle du monde (Universal Background Model - UBM).

La cohorte d'imposteurs

Higgins et al. (1991) et plus tard Rosenberg et al. (1992) suggèrent d'approximer la probabilité $p(H_1|S)$ en utilisant une cohorte d'imposteurs. A chaque locuteur I_X est associé un groupe d'imposteurs, dont les modèles sont « proches du sien ». Ces modèles sont ensuite utilisés pour exprimer l'hypothèse de « non-locuteur I_X ». Plusieurs facteurs sont à considérer pour cette cohorte :

- la cohorte peut être dépendante ou indépendante du locuteur ;
- le nombre d'imposteurs dans la cohorte doit être choisi ;
- il faut utiliser soit un modèle unique, appris avec l'ensemble des données des imposteurs, soit plusieurs modèles ;
- il faut estimer la « proximité » entre les modèles d'imposteurs et celui du locuteur.

Le nombre de facteurs à considérer et principalement la difficulté du choix des imposteurs, a conduit la plupart des systèmes état-de-l'art à utiliser un modèle unique appelé modèle du monde et décrit dans la partie suivante.

Le modèle du monde (Universal Background Model - UBM)

Alors que Higgins et al. (1991) préconisent l'utilisation d'imposteurs dont les modèles sont proches de celui du client considéré, afin de rendre la reconnaissance robuste à des imposteurs dont la voix est « proche » de celle du client, Reynolds (1995) choisit des imposteurs plus ou moins « proches » du client. D'après Reynolds, le choix d'imposteurs « proches » du client permet d'améliorer la robustesse face à des voix « proches » mais n'assure aucune garantie face à des voix très « éloignées » puisqu'elle le seront autant du client que des imposteurs.

Le paradigme de reconnaissance du locuteur, utilisant des modèles GMMs et un modèle du monde (UBM), est introduit par Carey et Parris (1992) et Reynolds (1995). Il consiste à modéliser l'hypothèse de non-locuteur par un modèle universel $\mathcal{W}$ représentant l'ensemble des locuteurs, excepté le locuteur considéré et ce, pour tous les

locuteurs. Le rapport de vraisemblance pour un signal S et un modèle du locuteur I_X devient :

$$\mathcal{LR}(S, H_0, H_1) = \frac{p(S|\mathcal{X})}{p(S|\bar{\mathcal{X}})} \approx \mathcal{LR}(S, \mathcal{X}, \mathcal{W}) = \frac{p(S|\mathcal{X})}{p(S|\mathcal{W})} \tag{7.8}$$

Il faut noter que, de la même façon que pour les cohortes d'imposteurs, le choix des locuteurs dont les données sont utilisées pour l'entraînement de l'UBM est primordial. Il est généralement considéré que plus la variabilité et le nombre de locuteurs sont élevés, meilleur sera le modèle UBM. La quantité de données disponible et le temps d'apprentissage d'un tel modèle fait cependant resurgir la question de la sélection des données utilisées pour apprendre le modèle du monde, que ce soit la sélection des locuteurs ou des vecteurs de paramètres.

7.1.4 Calcul des scores

En considérant le paradigme GMM/UBM, le score obtenu par une séquence de test $\mathcal{O} = \{o_t\}, t \in [1, T]$ pour une identité clamée I_X est donnée par :

$$LLR(S) = \frac{1}{T}.\sum_t \log\left(\frac{p(o_t|\mathcal{X})}{p(o_t|\mathcal{W})}\right) \tag{7.9}$$

$p(o_t|\mathcal{X})$ et $p(o_t|\mathcal{W})$ sont respectivement les vraisemblances du vecteur de paramètres o_t pour les modèles du locuteur $\mathcal{X}$ et du monde $\mathcal{W}$. Le score calculé est l'espérance mathématique logarithmique d'un rapport de vraisemblance (Log-Likelihood Ratio - LLR) sur le segment de test $\mathcal{O}$.
Considérant qu'une première normalisation résulte de l'utilisation du modèle du monde à travers le rapport de vraisemblance, la normalisation sur le segment $\mathcal{O}$ n'est pas indispensable mais permet de comparer des segments de longueurs différentes.

Le calcul de ce score n'est pas symétrique. Soient deux segments de parole donnés : X et Y, les deux modèles GMMs $\mathcal{X}$ et $\mathcal{Y}$ appris sur ces segments et un modèle du monde $\mathcal{W}$, l'asymétrie du calcul des scores est visible dans l'inéquation 7.10 :

$$\frac{1}{T_y}.\sum_{t_y} \log\left(\frac{p(Y_t|\mathcal{X})}{p(Y_t|\mathcal{W})}\right) \neq \frac{1}{T_x}.\sum_{t_x} \log\left(\frac{p(X_t|\mathcal{Y})}{p(X_t|\mathcal{W})}\right) \tag{7.10}$$

7.2 Place des modèles de locuteurs dans l'espace acoustique

Les modèles GMMs permettent de représenter les locuteurs dans un espace acoustique de grande dimension. Les performances des systèmes de reconnaissance du locuteur dépendent de la pertinence de la représentation du monde et des locuteurs dans cet espace. Cette section traite de la structure de l'espace acoustique et des méthodes état-de-l'art permettant d'obtenir une modélisation pertinente des locuteurs.

7.2.1 Apprentissage du modèle du monde

Le modèle du monde est appris à partir de plusieurs centaines d'heures de parole provenant d'un maximum de locuteurs différents. Son apprentissage repose sur l'algorithme d'Espérance Maximisation (Expectation-Maximisation algorithm - EM) avec comme critère d'optimisation, le critère de maximum de vraisemblance (Maximum Likelihood - ML). Celui-ci permet de maximiser la vraisemblance des données d'apprentissage par rapport au modèle GMM du monde. Cet algorithme itératif converge vers un maximum local et présente l'avantage d'augmenter la vraisemblance des données avec le modèle estimé à chaque itération. Une description approfondie de l'algorithme EM est donnée dans l'annexe B.

7.2.2 Adaptation de modèle, critère du Maximum a Posteriori

L'algorithme $EM - ML$, utilisant un critère de maximum de vraisemblance, n'est pas adapté à l'apprentissage des modèles de locuteurs. La quantité de données disponible pour un locuteur n'est pas suffisante. Une méthode courante consiste à apprendre le modèle GMM d'un locuteur en adaptant le modèle du monde avec les données de ce locuteur. Différents critères d'adaptation existent dans la littérature. La méthode la plus utilisée en reconnaissance du locuteur est celle du *Maximum a Posteriori* (MAP) (Gauvain et Lee, 1994), (Reynolds et al., 2000).

L'adaptation selon le critère MAP utilise l'algorithme EM, mais considère un modèle a priori. Si Θ est le vecteur de paramètres qui doit être estimé d'après les données $\mathcal{X}$ avec la fonction de densité de probabilité $f(.|\Theta)$ et si g est la fonction de densité de probabilité a priori de Θ, alors le vecteur de paramètres, Θ_{MAP}, estimé est :

$$\hat{\Theta}_{MAP} = \underset{\Theta}{argmax}\ p(\Theta|\mathcal{X}) = \underset{\Theta}{argmax}\ p(\Theta|\mathcal{X})p(\Theta) \tag{7.11}$$

L'adaptation MAP peut être interprétée comme une modification des paramètres du modèle GMM du monde en vue de le « rapprocher » d'un modèle appris sur l'ensemble des données d'apprentissage.

Chaque paramètre de chaque distribution est adapté par une transformation spécifique, indépendante des autres paramètres. Cette adaptation indépendante nécessite une quantité de données importante pour estimer chaque paramètre avec une confiance suffisante.

En pratique, pour un mélange de Gaussiennes, les nouveaux paramètres sont estimés par :

$$\hat{\mu_i} = \alpha_i^{\mu} E_i(X) + (1 - \alpha_i^{\mu})\mu_i \tag{7.12}$$

$$\hat{\sigma_i^2} = \alpha_i^{\sigma} E_i(X^2) + (1 - \alpha_i^{\sigma})(\sigma_i^2 + \mu_i^2) - \hat{\mu_i^2} \tag{7.13}$$

$$\hat{w}_i = \left[\alpha_i^w \frac{n_i}{T} + (1 - \alpha_i^w) w_i\right] \gamma \tag{7.14}$$

où γ permet d'obtenir une somme de poids égale à 1, n_i représente l'occupation de la Gaussienne i, T est le nombre total de trames d'apprentissage, $E_i(X)$ et $E_i(X^2)$ sont respectivement les moments d'ordre 1 et 2 des données d'adaptation obtenues par maximum de vraisemblance.

Le coefficient $\alpha_i^{\circledast}$ ($\circledast$ = μ, σ ou w) est égal à :

$$\alpha_i^{\circledast} = \frac{n_i}{n_i + r^{\circledast}} \tag{7.15}$$

où $r^{\circledast}$ est un facteur de régulation. Ce facteur indique en pratique la confiance accordée aux statistiques provenant des données d'apprentissage par rapport à un a priori que sont les statistiques issues du modèle du monde. Le facteur $r^{\circledast}$ correspond au nombre de trames d'apprentissage nécessaires à l'adaptation d'un paramètre pour accorder le même poids au paramètre appris d'après les données d'apprentissage qu'au paramètre a priori (si l'occupation de la Gaussienne considérée est égale à ce paramètre, alors le paramètre résultant du MAP est une moyenne du paramètre a priori et du paramètre appris par maximum de vraisemblance). En pratique, seules les moyennes des distributions Gaussiennes sont adaptées pour la reconnaissance du locuteur.

7.2.3 Autre formulation de l'estimation MAP et EigenVoices

Dans la suite, nous appelons super-vecteur le vecteur contenant tous les paramètres de moyennes d'un modèle GMM. Soit un mélange de Gaussiennes S à N composantes de dimension d. Le vecteur de moyennes de la distribution i du mélange est de la forme :

$$\mu^i = [\mu_1^i ... \mu_d^i] \tag{7.16}$$

Alors, le super-vecteur associé au modèle S et noté $\mu(S)$ est de la forme :

$$\mu(S) = [\mu^1\ \mu^2 ...\ \mu^N] = [\mu_1^1\ \mu_2^1 ...\ \mu_d^1\ \mu_1^2 ...\ \mu_d^N] \tag{7.17}$$

L'apprentissage d'un modèle de locuteur avec le critère MAP, à partir du modèle du monde W, peut alors être mis sous la forme :

$$\mu(S) = \mu(\mathcal{W}) + D\,Z \tag{7.18}$$

où la matrice D, diagonale dans le cas de l'adaptation MAP, vérifie $I = r^{\mu}\, D^t\, \Sigma^{-1}\, D$ où I est l'identité, Σ la matrice de covariance du mélange de Gaussiennes et r^{μ} est le facteur de régulation du MAP.

Remarque Le terme $\mu(\mathcal{W})$ est indépendant du choix du locuteur S et des données d'apprentissage utilisées, donc de la session d'apprentissage.

L'absence de prise en compte de l'inter-corrélation entre les Gaussiennes peut être légitimée par le rôle structurant du modèle du monde dans les approches génératives, considérant qu'un modèle du monde bien appris « segmente » l'espace acoustique en régions cohérentes.

Dans certaines approches, comme par exemple (Zavaliagkos, 1995), la matrice D est considérée sous sa forme pleine. L'approche par *EigenVoices* ou *voix propres* développée par Kuhn et al. (1998) s'inspirent des travaux de Turk et Pentland (1991a) (cf. section 4.1.2). Elle permet de simplifier l'estimation de ces paramètres en formulant l'hypothèse selon laquelle la variabilité locuteur est restreinte à un sous-espace acoustique de dimension réduite. La formulation de cette hypothèse prend alors la forme :

$$\mu(S) = \mu(\mathcal{W}) + V\,Y \tag{7.19}$$

où la matrice V est une matrice rectangulaire pleine de rang faible. Le rang faible correspond à la dimension du sous espace au sein duquel se trouvent les modèles des locuteurs. Chaque locuteur est caractérisé par un vecteur Y de dimension réduite. Les colonnes de la matrice V constituent les *voix propres* et les composantes de Y sont appelées *speaker factors*.

Il existe une similitude entre les modèles d'ancrage (cf. partie 3.2.1) et les *voix propres*. La différence entre ces deux approches réside dans l'espace d'application de la décomposition. Les *voix propres* sont utilisées pour obtenir la modélisation d'un locuteur, quand les modèles d'ancrage permettent une représentation du locuteur dans l'espace des scores.

7.2.4 Mise en évidence de la composante canal

L'équation 7.19 présente le modèle de locuteur comme la somme d'une composante invariante (le modèle du monde) et d'une composante spécifique au locuteur considéré. Les variations acoustiques dues au canal, ou à la spécificité de la session d'apprentissage, n'apparaissent pas explicitement dans ce formalisme.

Différentes approches, développées ces dernières années, utilisent le postulat selon lequel tout super-vecteur provenant de l'enregistrement d'un locuteur S dans des conditions E peut être mis sous la forme :

$$\mu(S) = s + c \tag{7.20}$$

où s est une composante dépendante du locuteur S, et c appelée composante canal, est un vecteur qui ne dépend que des conditions d'enregistrement E. Ce formalisme est illustré par la figure 7.1.

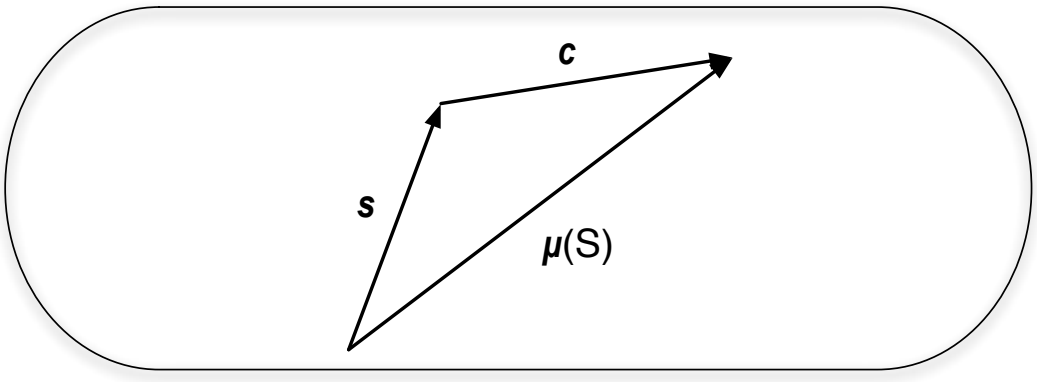


FIG. 7.1: *Plusieurs approches telles que la synthèse de modèles de locuteurs ou les EigenChannels représentent un modèle de locuteur $\mu(S)$ comme la somme de deux composantes s et c, la première ne dépendant que du locuteur et la deuxième que des conditions d'enregistrement.*

Modélisation discrète de la composante canal

Dans (Teunen et al., 2000) les auteurs considèrent que le vecteur c est la composante canal spécifique à chaque couple microphone/canal. Dans ce contexte, la suppression de la composante canal est rendue difficile par le caractère discret de la composante c. Sous cette hypothèse, la suppression de la composante c nécessite la connaissance explicite de tous les canaux qui peuvent intervenir.

Modélisation continue de la composante canal

La théorie des EigenChannels (Kenny et al., 2003), (Burget et al., 2007) reprend l'expression précédente du super-vecteur de S (cf. equation 7.20). Toutefois, elle considère que la composante canal c n'est plus discrète, mais qu'elle appartient à un sous-espace acoustique de faible dimension et qu'elle peut être exprimée de la façon suivante :

$$c = U\,X \tag{7.21}$$

où U est une matrice rectangulaire de rang faible. Les vecteurs propres non-nuls de la matrice UU^T sont appelés EigenChannels. L'interprétation physique de cette approche équivaut à considérer tout canal comme une combinaison linéaire des colonnes de la matrice U.

7.2.5 Joint Factor Analysis

Dans le formalisme du *Joint Factor Analysis*, introduit par Kenny et Dumouchel (2004), l'expression du super-vecteur d'un locuteur, pour une session donnée, devient

la somme de quatre termes :

$$\mu(S) = \mu(\mathcal{W}) + U\,X + V\,Y + D\,Z \tag{7.22}$$

Dans cette expression, le terme Y représente le locuteur dans le sous espace acoustique de variation locuteur, X est la représentation de la session, ou composante canal, dans le sous espace de variation canal, le terme $\mu(W)$ est indépendant du locuteur et de la session et le terme $D\,Z$ est appris de la même façon que le terme $D\,Z$ du MAP (cf. équation 7.18). La puissance de cette approche réside dans les simplifications théoriques réalisées par Kenny et al. (2005) pour l'estimation des paramètres des matrices V, et principalement U, qui permettent une modélisation continue des variabilités locuteur et session.

7.3 Performances des systèmes GMM/UBM

Le paradigme GMM/UBM constitue, depuis une quinzaine d'années, la clef de voûte de la plupart des systèmes état-de-l'art de reconnaissance du locuteur. À l'heure actuelle, les systèmes état-de-l'art sont à même de répondre aux besoins d'applications biométriques réelles, sous réserve d'être développés spécifiquement pour l'application visée. Ceci sous-entend que les conditions d'utilisation, telles que l'environnement, le degré de sécurité requis, le nombre et le type d'individus concernés, ainsi que les contraintes ergonomiques d'utilisation, soient connus.
Nous allons, dans cette partie, étudier les performances d'un système GMM/UBM soumis aux contraintes exposées dans le chapitre 6, à savoir une quantité de données d'apprentissage réduite et une durée de test de l'ordre de 2 à 3 secondes.

L'objectif de cette étude est double. Dans un premier temps, nous souhaitons déterminer, pour ce système GMM/UBM, la configuration optimale dans le contexte applicatif choisi. Ce système sera utilisé par la suite comme référence pour la tâche de reconnaissance du locuteur. Dans un deuxième temps, nous souhaitons estimer les performances de ce même système GMM/UBM pour la reconnaissance du locuteur dépendante du texte.

Remarque

Nous n'utilisons pas, dans notre étude, de méthode s'apparentant au *Joint Factor Analysis* décrit dans 7.2.5. En effet, nous estimons que le protocole utilisé dans cette étude n'est pas approprié à ce type de méthode.
En comparant notre architecture à un système de référence, n'utilisant tout deux aucune normalisation de type *Factor Analysis*, nous conservons des approches comparables. Nous supposons, de fait, que le *Factor Analysis* peut être utilisé en complément de notre approche comme il peut l'être actuellement avec des systèmes GMM/UBM.

7.3.1 Incidence du nombre de distributions

Le nombre de distributions constituant les modèles GMMs du modèle du monde et des locuteurs est un paramètre important à plusieurs égards. Il détermine les ressources nécessaires à l'utilisation de ce système, puisque la mémoire et la puissance de calcul sont directement proportionnelles au nombre de distributions et à leur dimension.
Le nombre de distributions conditionne également les performances du système. Nous avons déjà observé, dans la partie 3.2.3, l'impact de la quantité de données d'apprentissage et de test sur les performances des systèmes de reconnaissance du locuteur. Mason et al. (2005) montrent que le ratio *quantité de données / taille des modèles* influe fortement sur les performances des systèmes.

Dans une première expérience, nous faisons varier ce ratio afin de déterminer la configuration la plus adaptée à nos contraintes. En utilisant le protocole **1-occ** décrit dans l'annexe A, nous faisons varier le nombre de distributions des modèles GMMs. Rappelons que ce protocole prévoit d'utiliser environ 12 secondes de parole pour l'apprentissage des modèles GMMs des clients. La quantité de données d'apprentissage et de test est constante tout au long de l'expérience. Le tableau 9.7 présente le taux d'égales erreurs en fonction du nombre de distributions des modèles GMMs. Les scores présentés dans ce tableau, ainsi que dans le suite de ce document sont calculés en utilisant l'ensemble des distributions Gaussiennes des modèles.

	Nombre de distributions par modèle GMM							
	16	32	64	128	256	512	1024	2048
Taux d'égales erreurs (EER)	7,01	5,28	4,33	3,67	3,22	3,06	3,16	3,17

TAB. 7.1: *Évolution du taux d'égales erreurs d'un système GMM/UBM état-de-l'art pour différentes tailles de modèles (base de données MyIdea). Le nombre de distributions Gaussiennes dans les mélanges varie de 16 à 2048. Les imposteurs prononcent indifféremment le mot de passe du client ou une autre phrase (configuration 1-occ TOUS, annexe A).*

Parmi les configurations testées, les meilleurs résultats sont obtenus pour des modèles GMMs à 512 distributions. L'EER augmente rapidement lorsque cette valeur diminue. Un explication bien admise dans le domaine de la reconnaissance du locuteur est qu'en dessous d'une certaine valeur, le nombre de distributions Gaussiennes est insuffisant pour modéliser précisément les densités de probabilité de chaque locuteur.
Lorsque le nombre de distributions dépasse 512, les performances stagnent ou chutent légèrement. Cette légère baisse de performance peut s'expliquer par un manque de données d'apprentissage, nécessaires à l'estimation de tous les paramètres des modèles GMMs. Ceux-ci sont alors sous-optimaux.
Les résultats obtenus dans le tableau 9.7 pour des modèles GMMs à 1024 ou 2048 distributions montrent cependant que la dégradation des performances ne suit pas l'augmentation du nombre de distributions. Il semble que l'information modélisée par les

modèles GMMs reste la même si la taille des modèles augmente encore. Ce phénomène peut être dû au fait que les données d'apprentissage ne correspondent qu'à un certain nombre de distributions Gaussiennes, laissant ainsi les distributions supplémentaires inchangées lors de l'adaptation des modèles. Cette hypothèse pourrait être vérifiée en comparant les paramètres des modèles de différentes tailles. Cette expérience n'a pas été réalisée ici car n'ayant pas pas directement trait à notre problématique.
Dans la suite de cette partie, nous ne considérons que des modèles GMMs à 512 distributions.

7.3.2 Dépendance au texte des systèmes GMM/UBM

Le paradigme GMM/UBM est utilisé la plupart du temps pour la reconnaissance du locuteur indépendante du texte. Nous désirons cependant évaluer les performances de notre système de référence dans le cadre de la reconnaissance du locuteur dépendante du texte. Comme précédemment, notre motivation est double du fait de l'utilisation de ce système en tant que référence et de l'intégration du paradigme GMM/UBM dans l'architecture proposée dans ce document.
Dans cette section, nous tentons de répondre à un certain nombre de questions, afin de déterminer l'influence du contenu linguistique dans le cadre de la vérification d'identité par un système GMM/UBM.

Les performances du système GMM/UBM sont elles dépendantes du texte ?

D'un point de vue structurel, les modèles GMMs tels qu'ils sont généralement utilisés en reconnaissance du locuteur ne prennent pas en compte la structure temporelle du signal de parole. Dans notre cas, l'utilisation de paramètres calculés sur une fenêtre temporelle de quelques dizaines de milli-secondes et de leur dérivée du premier ordre renforce encore l'idée selon laquelle notre système GMM/UBM de référence n'exploite pas l'information structurelle présente pour la tâche de reconnaissance du locuteur dépendante du texte mais seulement une information temporelle à court terme.
Cependant, la faible quantité de données d'apprentissage disponible et la faible variabilité lexicale entre données d'enrôlement et données de test des clients devraient, en réduisant la variabilité inter-sessions, améliorer les performances du système GMM/UBM. Les performances de notre système de référence devraient donc être dépendantes du contenu lexical d'apprentissage ou de test.

Afin de vérifier cette assertion, nous réalisons 3 séries de tests. Dans ces trois expériences, les clients prononcent la même phrase lors de l'entraînement des modèles et de la phase de test. Dans la première condition (**MDP**), les imposteurs prononcent la phrase qui a été utilisée par le client lors de la phase d'apprentissage. Dans la condition **FAUX**, les imposteurs prononcent une phrase différente de celle d'entraînement. La troisième condition (**TOUS**) regroupe l'ensemble des tests des deux premières conditions. Le nombre de tests imposteurs est le même pour les conditions **MPD** et **FAUX**. Il est doublé pour la condition **TOUS**. Les protocoles pour les conditions **MDP**, **FAUX**

et **TOUS** sont détaillés dans l'annexe A. Les résultats de ces expériences sont présentés dans le tableau 9.8.

Conditions de test	Taux d'égales erreurs
MDP	3,68
FAUX	2,11
TOUS	3,06

TAB. 7.2: *Performances d'un système de reconnaissance du locuteur pour différentes conditions de dépendance au texte.*

Conformément à l'hypothèse que nous avions évoqué ci-dessus, les résultats obtenus par le système GMM/UBM, dans les trois conditions testées, sont dépendants du texte. Notre système GMM/UBM de référence obtient de meilleures performances lorsque le contenu lexical des tests imposteurs est différent de celui utilisé par les clients pour s'enrôler. Le Système GMM/UBM exploite donc la variabilité lexicale en plus de la variabilité inter-locuteurs pour reconnaître les locuteurs.

Ces résultats s'expliquent par le protocole utilisé. Dans cette expérience, les modèles GMMs des clients sont appris avec environ 12 secondes de parole. Cette durée correspond à deux phrases, d'environ 5 secondes chacune, en plus d'un phrase courte. C'est le contenu lexical de cette phrase courte qui est utilisé par les clients lors des tests et par les imposteurs dans la condition **MDP**.
Ainsi le modèle GMM d'un client ne modélise pas l'ensemble des phonèmes que peut produire ce client, mais principalement ceux présents dans les 3 phrases qu'il prononce lors de l'enrôlement.
Il est cependant probable que l'écart observé entre les résultats des conditions **MDP** et **FAUX** tende à se réduire si la quantité de données d'apprentissage des modèles GMMs et la variabilité lexicale des données utilisées augmentent.

Quelle est l'influence de la quantité de données sur la dépendance au texte du système GMM/UBM ?

D'après les résultats précédents, nous pouvons penser qu'augmenter la quantité de données d'apprentissage améliore les performances du système GMM/UBM dans toutes les conditions. Cependant, l'effet de cette augmentation pourrait différer selon le contenu lexical ajouté.

Nous testons 3 configurations d'apprentissage différentes. La première est identique à celle de l'expérience 1, à savoir que chaque modèle de client est appris avec deux phrases longues ($\sim$5 secondes) et une phrase courte ($\sim$ 2 secondes). Dans la deuxième configuration, nous ajoutons aux données d'apprentissage une phrase courte ($\sim$ 2 secondes). Cette phrase est choisie aléatoirement parmi neuf phrases différentes de la première phrase courte utilisée. La dernière configuration d'apprentissage (**2-occ**) est

décrite dans l'annexe A. Chaque modèle y est appris avec les deux phrases longues et 2 occurrences de la même phrase courte. Le protocole de test est identique à celui de l'expérience 1. Le tableau 9.9 présente les résultats obtenus pour chacune de ces trois configurations dans les conditions **TOUS**, **MDP** et **FAUX**.

		Condition de tests		
		1-occ	1-occ + aléatoire	2-occ
Taux d'égales erreurs	TOUS	3,06	2,33	1,31
	MDP	3,68	2,89	2,00
	FAUX	2,11	1,78	0,56

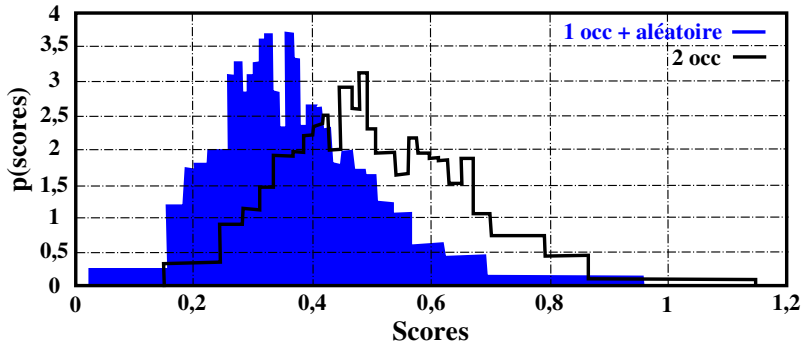
TAB. 7.3: *Influence du contenu lexical des enregistrements utilisés pour entraîner les modèles GMMs des locuteurs dans un contexte dépendant du texte.*

Ces résultats montrent que l'augmentation de la quantité de données d'apprentissage améliore considérablement les performances du système considéré quelle que soit la nature lexicale du matériel utilisé.

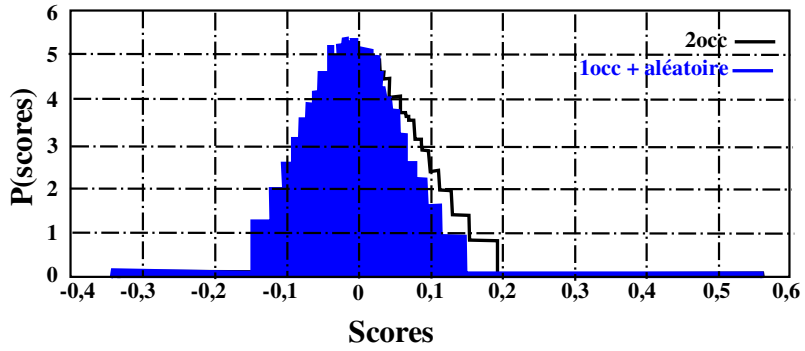
Les résultats obtenus sont cependant meilleurs lorsque le contenu lexical ajouté est identique à celui de la première phrase courte utilisée. Si ce résultat s'explique facilement dans la condition **FAUX**, où les imposteurs prononcent un texte différent, le cas de la condition **MDP** est plus délicat à interpréter. En effet, le tableau 9.9 ne permet pas de préciser si l'ajout d'une seconde phrase avec le même contenu lexical permet d'augmenter les scores des tests clients ou au contraire de diminuer les scores imposteurs. Si ces deux effets entraînent la même conséquence, à savoir la meilleure séparation des clients et des imposteurs, les conclusions qui en découlent peuvent différer. Les figures 9.10(b) et 9.10(a) présentent les distributions des scores clients et imposteurs obtenues dans les conditions **1-occ + aléatoire** et **2-occ** pour la condition **TOUS**.

Nous observons sur la figure 9.10(b) que les distributions des scores imposteurs varient peu en fonction du contenu lexical de la deuxième phrase courte utilisée pour l'apprentissage des modèles clients. L'analyse de la très légère tendance visible sur cette courbe tend à montrer que plus la variabilité des séquences d'apprentissage est importante et plus les tests des imposteurs obtiennent des scores faibles. Cette assertion ne repose cependant que sur la lecture graphique d'une très faible tendance et ne peut pas être considérée comme démontrée. Il nous a pourtant semblé opportun de décrire cette observation puisqu'elle va dans le sens de la littérature existante en reconnaissance du locuteur.

La figure 9.10(a) montre une dépendance plus importante des scores clients à la variabilité des données d'apprentissage. En effet, lorsque la phrase ajoutée pour l'apprentissage des modèles est celle utilisée lors des tests clients, les scores sont plus élevés que s'il s'agit d'une phrase différente (i.e. avec un contenu phonétique différent). Cette observation tend à montrer que le système GMM/UBM tire partie de l'informa-



(a) Distributions des scores clients



(b) Distributions des scores imposteurs

FIG. 7.2: *Distributions des scores imposteurs et clients en fonction des données d'apprentissage pour les conditions* ***1-occ + aléatoire*** *et* ***2-occ***

tion phonétique des données d'apprentissage plus que de la simple augmentation de la quantité de données.
La connaissance du texte prononcé par les clients d'un système de reconnaissance du locuteur peut alors permettre d'améliorer les performances de ce système.

Malgré les limites de cette expérience, les conclusions qu'elle nous permet d'élaborer sont cependant cohérentes avec notre connaissance de l'approche GMM/UBM et les comportements observés nous paraissent logiques, compte tenu du protocole expérimental.

Quelle est la capacité du système GMM/UBM à reconnaître un énoncé ?

Afin de pousser plus loin notre expérience, nous souhaitons tester la capacité du système GMM/UBM à reconnaître le contenu linguistique prononcé. Il s'agit donc de tester ici notre système de RAL de référence pour une tâche de vérification de mot de passe dépendante du locuteur (tous les tests réalisés sont des tests intra-locuteurs).
Dans l'expérience suivante, les modèles de client sont appris en utilisant le même protocole : **1-occ**, décrit dans le chapitre 6. Pour chaque test, le client correspondant au

modèle GMM testé prononce la phrase de l'apprentissage ou une phrase différente.
Le nombre de tests réalisés pour cette évaluation est fortement limité par l'utilisation de la base de données MyIdea. Le nombre de tests pour lesquels le client prononce la phrase d'apprentissage est égal au nombre de tests pour lesquels il prononce une phrase différente : 1800.

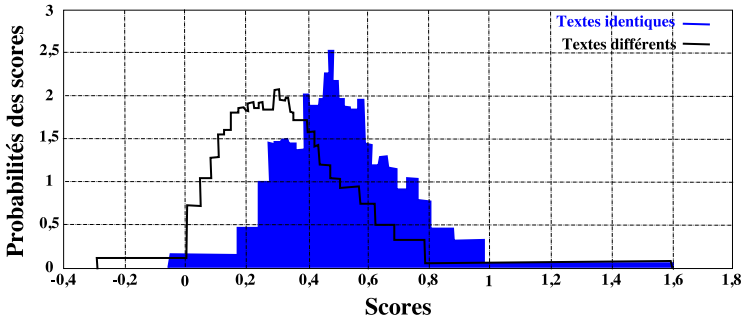


FIG. 7.3: *Distributions des scores de reconnaissance de texte obtenues par un système GMM/UBM*

Le taux d'égales erreurs pour cette expérience est de 31,4% et l'histogramme 7.3 présente les distributions des scores obtenus par le système GMM/UBM lorsque le client prononce la même phrase (en bleu) ou lorsqu'il prononce une phrase différente (en noir).
Le système GMM/UBM ne permet pas de reconnaître un énoncé spécifique dans ces conditions, mais l'observation des distributions des scores nous permet d'affirmer que le système GMM/UBM affiche, dans ces conditions, un comportement différent, selon le contenu lexical prononcé. En effet, les distributions des scores apparaissent comme décalées, les scores sont plus élevés lorsque le client prononce un contenu lexical identique à celui de la phase d'apprentissage.

Conclusion

Les méthodes de reconnaissance du locuteur présentées dans ce chapitre et reposant sur le paradigme GMM/UBM sont à la base de la plupart des approches état-de-l'art actuelles. La popularité de ces approches est due, entre autres, à la capacité des modèles GMMs à structurer l'espace acoustique. De nombreux travaux réalisés ces dernières années ont d'ailleurs exploité cette capacité pour améliorer la robustesse des systèmes. Un grand nombre de questions restent ouvertes, traitant notamment du choix et de la quantité des données d'apprentissage nécessaires. Nous avons montré (cf. section 7.3.1) l'importance du ratio *quantité des données d'apprentissage/taille des modèles GMMs*.

Nous nous sommes intéressé particulièrement à l'incidence des contraintes lexicales sur les performances des systèmes GMM/UBM. Les modèles GMMs n'exploitent pas la structure temporelle des séquences de parole utilisées lors des phases d'apprentissage

et de tests. Cependant, les conclusions des expériences présentées dans la section 7.3.2 montrent que le système GMM/UBM de reconnaissance du locuteur utilisé n'est pas, comme nous le supposions, totalement indépendant du texte. Il apparaît que ce système tire avantage du contenu lexical des séquences vocales utilisées. Cette conclusion est particulièrement vérifiée dans le cas où le système dispose de peu de données pour entraîner les modèles. Elle est probablement due à l'importante variabilité existante entre sessions d'apprentissage et de tests imposteurs lorsque le texte prononcé diffère.

Toutefois, la faible quantité de données d'apprentissage et la courte durée des tests font partie des contraintes que nous avons fixées durant notre étude. Nous devons donc prendre en compte cet aspect.

Chapitre 8

Structuration temporelle de la séquence acoustique

Sommaire

Résumé

Afin d'exploiter la corrélation temporelle des signaux audio et vidéo, nous souhaitons introduire une dimension temporelle au sein du modèle acoustique. Parmi les méthodes possibles, nous choisissons d'utiliser les modèles de Markov cachés et plus particulièrement leur version semi-continue qui est présentée dans ce chapitre. Les problématiques propres à ce type d'approches : estimation des paramètres des modèles ou de la probabilité d'une séquence d'observation sont détaillées. La deuxième partie de ce chapitre présente la méthode itérative que nous avons développée afin d'estimer les paramètres des modèles semi-continus.

Introduction

DANS ce chapitre, nous désirons étendre les modèles de locuteurs indépendants du texte issus du paradigme GMM/UBM afin d'obtenir une représentation du locuteur prononçant un mot de passe personnel. L'architecture que nous proposons est une extension du paradigme GMM/UBM, au sein de laquelle un troisième niveau de spécialisation permet de modéliser la structure des mots de passe énoncés, tout en conservant l'information acoustique spécifique du locuteur.
L'encyclopédie Larousse[1] fournit plusieurs définitions du concept de structure :

1. Manière dont les **parties d'un tout** sont arrangées entre elles.
2. **Organisation** des parties d'un système, qui lui donne sa cohérence et en est la **caractéristique permanente**.
3. Organisation, système complexe considéré dans ses **éléments fondamentaux**.

Définir et modéliser la structure d'un mot de passe nécessite, d'après la première définition, la segmentation du signal en sous-parties que nous appellerons « cellules acoustiques » et qui composent le mot de passe.
Une description de l'organisation du mot de passe, robuste à la variabilité inter-sessions, permet d'en extraire les caractéristiques permanentes.
Pour que la description de l'organisation soit robuste, celle-ci doit intégrer un niveau de détail suffisant, tout en restant assez générale pour être indépendante de la session. Les sous-parties du mot de passe correspondant à ce niveau de description sont les éléments fondamentaux qu'il nous faut déterminer.

Pour représenter les mots de passe, nous utilisons par la suite des modèles de Markov cachés. Ce sont des modèles à états finis au sein desquels chaque état modélise une entité acoustique fondamentale. L'organisation des états est définie de façon explicite et permanente.

Ce chapitre est organisé en quatre parties. Nous présentons, tout d'abord, les modèles de Markov cachés et leur version semi-continue.
Nous présentons ensuite le processus itératif d'apprentissage des modèles de mots de passe que nous avons développé dans le cadre de nos travaux, tout en mettant en évidence les différents paramètres qui définissent la structure de ces modèles en fonction des besoins et des contraintes propres à notre tâche.
Dans la section suivante, nous cherchons à optimiser les paramètres de notre système, pour la tâche que nous nous sommes fixée.
Enfin, nous développons, dans une quatrième section, les possibilités offertes par notre architecture à trois niveaux et comparons les performances obtenues à celles du système GMM/UBM que nous avons choisi pour référence.

[1] `http://www.larousse.fr/encyclopedie/`

8.1 Modélisation des mots de passe

Rappelons tout d'abord que notre approche repose sur une architecture acoustique à trois niveaux, permettant une spécialisation hiérarchique et un partage des données. Cette architecture est rappelée par la figure 8.1.

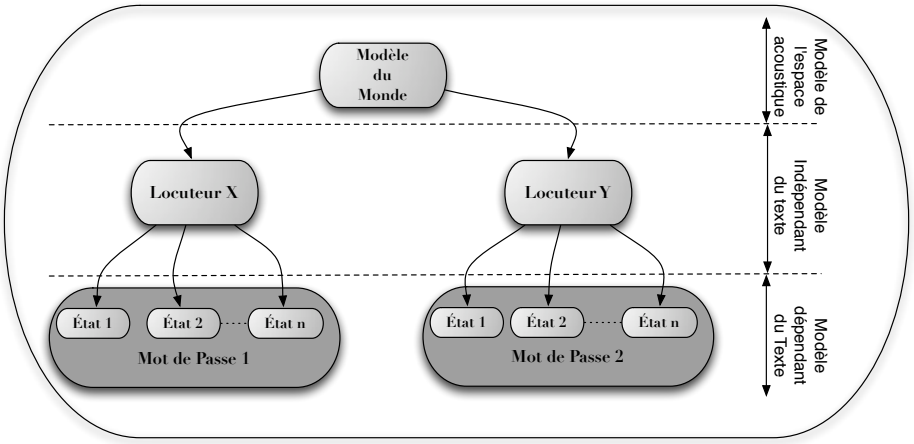


FIG. *8.1:* *Architecture acoustique hiérarchique à trois niveaux*

La différence entre les modèles GMMs et les modèles à états finis, des deuxième et troisième couches de cette architecture, réside dans la structuration temporelle des informations acoustiques portées par les probabilités d'émission des modèles à états finis.

Comme symbolisé par la figure 8.2, le modèle GMM du second niveau de l'architecture constitue une représentation du locuteur indépendante du texte prononcé. Il est obtenu par adaptation du modèle du monde, du premier niveau de l'architecture. Le vecteur moyen de chaque distribution composant le modèle du monde est déplacé pour modéliser au mieux les spécificité du locuteur considéré. Le modèle GMM indépendant du texte du locuteur modélise l'ensemble des sons que celui-ci peut prononcer. À l'inverse, les fonctions de probabilité d'émission du modèle à états finis du troisième niveau modélisent plus précisément la suite de sons prononcés par ce locuteur, qui composent le mot de passe choisi. Chaque état du modèle SCHMM du dernier niveau modélise un des sons réalisables par le locuteur considéré.

8.1.1 Modélisation par des modèles de Markov

La troisième couche de l'architecture que nous avons développée est constituée d'un modèle de Markov caché (HMM). Ce choix est expliqué par plusieurs motivations.
Nous avons vu d'une part dans la partie 3.3.2 que les approches génératives, contrairement aux approches vectorielles, permettent naturellement de modéliser la variabilité inter-sessions.

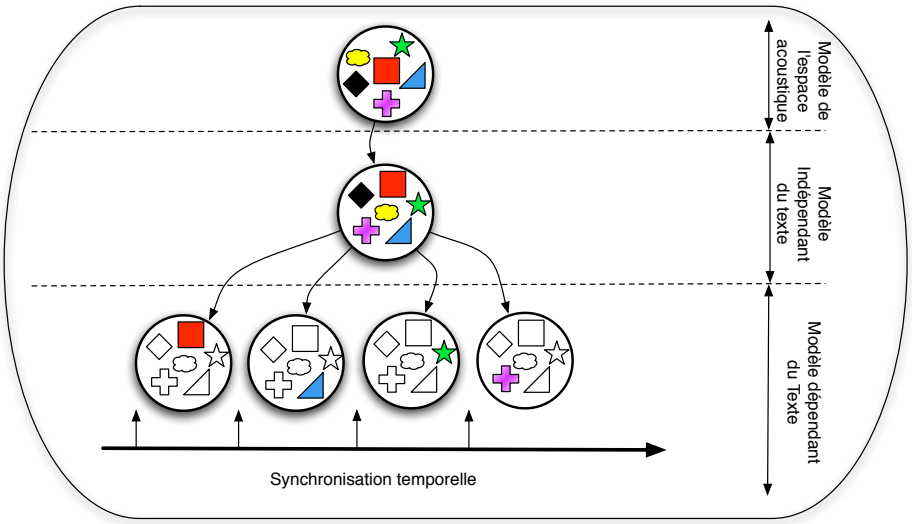


FIG. 8.2: *Spécialisation des modèles de locuteur dépendant du texte du troisième niveau de l'architecture hiérarchique. Alors que le modèle de locuteur indépendant du texte modélise l'ensemble des sons que peut émettre celui-ci, chaque état du modèle dépendant du texte est spécialisé pour modéliser la suite de son qui compose le mot de passe.*

De plus les modèles HMMs constituent la référence dans le domaine de la reconnaissance automatique de la parole (RAP). Leur structure d'automate à états finis, qui permet de représenter le signal de parole comme une succession de sons distincts, a montré son efficacité. Les HMMs autorisent une certaine flexibilité temporelle, puisqu'ils font intervenir un processus stochastique lors de l'alignement d'une séquence audio avec un modèle. Cette caractéristique est détaillée par la suite.

Un modèle de Markov caché est un modèle statistique dans lequel le phénomène modélisé est supposé être un processus Markovien de paramètres inconnus. C'est un automate à états finis caractérisé par un quadruplet $\Lambda = \{\mathcal{S}, \Pi, \tau, f\}$ où $\mathcal{S}$ est un ensemble d'états $\{S_i\}$, $i \in [1, E]$, π_i est la probabilité que l'état i soit l'état initial, $\tau_{i,j}$ la probabilité de transiter de l'état i à l'état j et $f(X|i)$ la probabilité d'émettre le symbole X en étant dans l'état $\mathcal{S}_i$.

La résolution des trois problèmes suivants constitue un préalable à l'utilisation des modèles HMMs :

- estimation de la probabilité d'émission d'une séquence d'observations $\mathcal{O}$ par un HMM de paramètres Λ, réalisée à l'aide des algorithmes *Forward* et *Backward* (Baum et al., 1970) ;
- alignement d'une séquence d'observations sur un modèle HMM pour faire correspondre observations et états émetteurs, réalisé à l'aide de l'algorithme de Viterbi (Viterbi, 1967) ;
- estimation des paramètres du HMM à partir de données observées.

Dans cette partie nous appelons implicitement HMM les modèles de Markov cachés continus.

Estimation de la probabilité d'une séquence

Soit $\mathcal{O} = \{O_0, O_1, ..., O_T\}$ une séquence d'observations alignée sur une suite d'états $\mathcal{A} = \{a_0, a_1, ..., a_T\}$. La probabilité de cette séquence est :

$$\begin{aligned} p(\mathcal{O}|\mathcal{A}, \Lambda) &= \prod_{t=0}^{T} p(O_t|a_t, \Lambda) \\ &= f(O_0|a_0)\, f(O_1|a_1) \ldots f(O_T|a_T) \end{aligned} \tag{8.1}$$

La probabilité du chemin $\mathcal{A}$ est donnée par :

$$\begin{aligned} p(\mathcal{A}|\Lambda) &= \pi_{a_0} \prod_{t=1}^{T} \tau_{a_{t-1}, a_t} \\ &= \pi_{a_0}\, \tau_{a_0, a_1}\, \tau_{a_1, a_2} \ldots \tau_{a_{t-1}, a_t} \end{aligned} \tag{8.2}$$

La probabilité conjointe du chemin $\mathcal{A}$ et de l'observation $\mathcal{O}$ est :

$$p(\mathcal{A}, \mathcal{O}|\Lambda) = p(\mathcal{O}|\mathcal{A}, \Lambda)\, p(\mathcal{A}|\Lambda) \tag{8.3}$$

et pour tous les chemins :

$$p(\mathcal{O}|\Lambda) = \sum_{\mathcal{A}} p(\mathcal{O}|\mathcal{A}, \Lambda)\, p(\mathcal{A}|\Lambda) \tag{8.4}$$

Le coût de calcul important de cette estimation peut être réduit grâce aux algorithmes *Forward* et *Backward* qui sont développés dans l'annexe C.

Alignement d'une séquence d'observations par l'algorithme de Viterbi

Étant donnée une séquence $\mathcal{O} = \{O_0, O_1, ..., O_T\}$ et un modèle HMM de paramètres Λ connus, aligner la séquence sur le HMM consiste à chercher la suite d'état $\mathcal{A} = \{a_0, a_1, ..., a_T\}$ qui maximise la probabilité conjointe :

$$\hat{\mathcal{A}} = \underset{\mathcal{A}}{argmax}\; p(\mathcal{O}, \mathcal{A}|\Lambda) \tag{8.5}$$

Comme l'estimation de la probabilité, la complexité de ce problème peut être réduite par l'utilisation d'un algorithme de programmation dynamique : l'algorithme de Viterbi, (Viterbi, 1967), (Forney, 1973).

L'algorithme de Viterbi a pour but de trouver la séquence d'états $\mathcal{A} = \{a_0, a_1, ..., a_T\}$ la plus probable ayant émis la séquence observée $\mathcal{O} = \{O_0, O_1, ..., O_T\}$. Cette séquence est calculée par récurrence pour chacun des états. Pour l'état e' à l'instant t :

$$\epsilon_t(e') = max\ p(s_0, s_1, ..., s_{t-1}, s_t = e'; O_0, O_1, ..., O_t) \tag{8.6}$$

Ce maximum est calculé sur l'ensemble des séquences $[s_0, s_1, ..., s_t]$ possibles.

Initialisation : à l'instant $t = 0$:

$$\epsilon_0(e') = \pi'_e\, f(O_0|e') \tag{8.7}$$

où $f(O_0|e')$ est la probabilité d'émettre le symbole O_0 en étant dans l'état e'.

Récurrence : La valeur $\epsilon_{t-1}(\cdot)$ est supposée calculée pour chacun des E états du modèle HMM. Pour chaque état e' de ce modèle :

$$\epsilon_t(e') = \max_e \left\{\epsilon_{t-1}(e)\ \tau_{e,e'}\ f(O_t|e')\right\} \tag{8.8}$$

où $\tau_{e,e'}$ est la probabilité de transition de l'état e à l'état e'. Cette étape est illustrée par la figure 8.3.

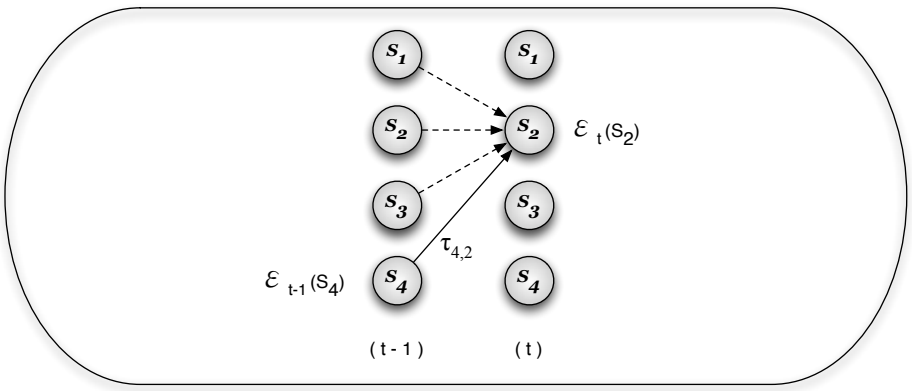


FIG. 8.3: *Sélection d'un chemin dans le treillis entre les instants $t-1$ et t par l'algorithme de Viterbi.*

Cette étape permet de déterminer l'état e le plus probable au temps $t-1$ à partir duquel l'automate a évolué vers l'état e' au temps t. Pour chaque état du HMM, le prédécesseur $q_t(e')$ qui permet de maximiser $\epsilon_t(e') = max\ p(s_0, s_1, ..., s_{t-1}, s_t = e'; O_0, O_1, ..., O_t)$ est déterminé.

Ce prédécesseur peut servir à trouver la séquence d'états $\{a_0, a_1, ..., a_{T-1}\}$ qui a engendré la séquence $\{O_0, O_1, ..., O_{T-1}\}$. Le processus récursif nous permet de retrouver a_0 :

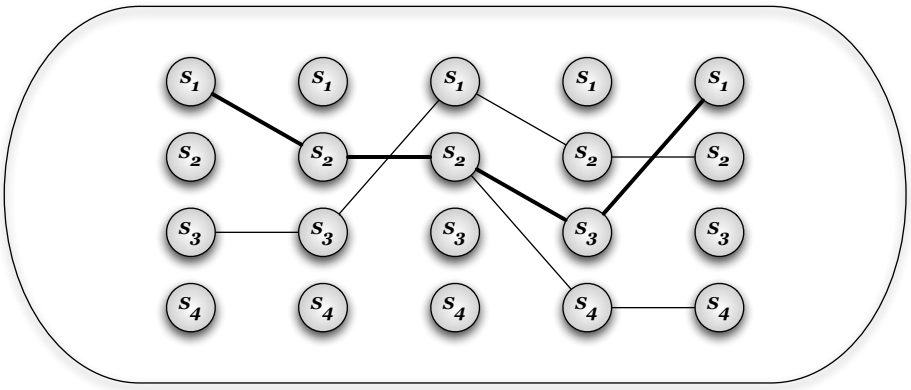


FIG. 8.4: *Reconstitution du chemin correspondant à la séquence optimale $\{a_0, a_1, ..., a_T\}$ si la séquence $\{O_0, O_1, ..., O_T\}$ a été engendrée par le HMM de paramètres Λ.*

$$
\begin{aligned}
a_{T-1} &= q_t e' \\
a_{T-2} &= q_{t-1}(a_{T-1}) \\
&\vdots \\
a_1 &= q_1(a_2) \\
a_0 &= q_0(a_1)
\end{aligned}
$$

L'algorithme de Viterbi est optimal : pour une séquence d'observations $\{O_0, O_1, ..., O_T\}$ et un modèle HMM de paramètres Λ, il fournit l'alignement $\mathcal{A}$ qui maximise : $p(\mathcal{O}, \mathcal{A}|\Lambda)$. Nous verrons dans la partie 9 que cet algorithme, parce qu'il maximise ce score, peut présenter un inconvénient majeur dans le contexte de la reconnaissance automatique du locuteur. Aussi, il peut être intéressant, d'imposer une contrainte qui limite la maximisation de la vraisemblance comme peut le faire le critère du maximum a posteriori lors de l'adaptation des modèles de locuteurs.

Estimation des paramètres des HMMs

L'apprentissage des paramètres du modèle HMM est certainement le plus complexe des trois problèmes posés. Considérons une architecture donnée. Elle est définie par le nombre d'états du modèle et la possibilité ou non de passer d'un état à un autre. Les éléments du modèle HMM, tels que le nombre d'états et l'organisation de ceux-ci, sont supposés fixés a priori. Il reste à déterminer les probabilités de transition entre les états autorisés $\{\tau_{i,j}\}$, et les probabilités d'émission des états du modèle $\{f(\cdot|s_i)\}$.

L'approche utilisée généralement pour estimer les paramètres Λ du modèle consiste à maximiser la vraisemblance du modèle avec un ensemble de données d'apprentis-

sage. Le critère de Maximum de Vraissemblance (Maximum Likelihood Estimation - MLE) ré-estime Λ selon la formule suivante :

$$\hat{\Lambda} = \underset{\Lambda}{argmax}\ p(\mathcal{O}|\Lambda) \tag{8.9}$$

D'autres critères peuvent être considérés pour ré-estimer Λ, comme le critère du Maximum A Posteriori ou le critère MMIE (Maximum Mutual Information Etimation) (Bahl et al., 1986). Cette estimation par le critère du maximum de vraisemblance est obtenue par l'algorithme de Baum-Welch (Baum et al., 1970) connu aussi sous le nom de *Forward-Backward* décrit dans l'annexe C. Comme l'algorithme *EM*, celui de Baum-Welch est un processus itératif qui assure la convergence des paramètres estimés vers un maximum local. Il présente de même l'avantage de garantir l'augmentation de la vraisemblance à chaque itération :

$$p(\mathcal{O}|\Lambda^{(n-1)}) \leqslant p(\mathcal{O}|\Lambda^{(n)}) \tag{8.10}$$

où $\Lambda^{(n)}$ sont les paramètres après n itérations de l'algorithme de Baum-Welch.

L'estimation des paramètres d'un modèle HMM par l'algorithme de Baum-Welch nécessite une quantité importante de données et représente un processus long et complexe. Cette approche n'est pas adaptée à nos contraintes, notamment à la faible quantité de données disponible pour estimer correctement les probabilités d'émission des états ainsi que les probabilités de transition.

8.1.2 Modélisation par des modèles de Markov semi-continus

Cette partie présente une variante des modèles de Markov cachés qui est plus adaptée aux contraintes que nous nous sommes fixées.

Présentation des modèles semi-continus

Les modèles de Markov cachés semi-continus (Semi-Continuous Hidden Markov Model - SCHMM) introduits par Young (1992) présentent une alternative à l'utilisation des HMMs continus. Nous avons évoqué précédemment la quantité importante de données que nécessite l'apprentissage des paramètres des HMMs continus. Les SCHMMs diminuent le nombre de paramètres à estimer et donc la quantité de données nécessaire, tout en conservant une modélisation détaillée de la structure temporelle du signal de parole.

Un HMM continu est défini comme nous l'avons vu précédemment (cf. section 3.3.2) par un quadruplet $\Lambda = \{\mathcal{S}, \Pi, \tau, f\}$ où $\mathcal{S}$ est un ensemble d'états $\{S_i\}$, $i \in [1, E]$, π_i est la probabilité que l'état i soit l'état initial, $\tau_{i,j}$ la probabilité de transition de l'état i à l'état j et $f(x|i)$ la probabilité d'émettre le symbole x en étant dans l'état $\mathcal{S}_i$.

Les probabilités d'émission des états sont généralement approximées par des mélanges de Gaussiennes. Chaque état du modèle markovien est modélisé par un mélange de Gaussiennes qui lui est propre. Les paramètres à estimer pour chaque état du HMM lors de l'apprentissage sont dans ce cas les vecteurs de moyennes et variances de chaque distribution ainsi que le vecteur des poids de chacune d'elles dans le mélange.

Les modèles SCHMMs sont fondés sur le partage d'une partie des paramètres des probabilités d'émission entre les états. Chaque état ne se différencie des autres que par le poids attribué à chaque Gaussienne au sein du GMM modélisant la probabilité d'émission des états. Les paramètres de moyenne et variance des états sont partagés. Le gain en nombre de paramètres à estimer est conséquent puisque dans le cas d'un HMM à E états, où chaque état est un GMM à N distributions de dimension n le nombre de paramètres qui caractérisent les probabilités d'émission de ce HMM est :

$$\#\text{Paramètres}_{HMM} = E \times N \times \underbrace{(2\,n+1)}_{\text{Nombre de paramètres par distribution Gaussienne}} \tag{8.11}$$

alors que le nombre de paramètres à estimer pour un HMM semi-continu possédant les mêmes caractéristiques est :

$$\#\text{Paramètres}_{SCHMM} = E \times N \tag{8.12}$$

Mutualisation de l'information

Les modèles de mots de passe que nous adjoignons au paradigme GMM/UBM sont à la fois des représentations d'un locuteur et du contenu linguistique correspondant au mot de passe qu'il choisi. Chacun des états du SCHMM qui composent la troisième couche de notre architecture (figure 5.5) modélise la prononciation d'une cellule acoustique par le locuteur considéré. Comme présentée sur la figure 8.2, l'information modélisée par chaque état des mots de passe du troisième niveau de l'architecture est présente dans la modélisation du locuteur indépendante du texte du niveau supérieur.

C'est pour cette raison que nous avons fait le choix, dans notre architecture, d'utiliser le modèle GMM indépendant du locuteur, obtenu par adaptation d'un modèle du monde (cf. chapitre 7) comme base pour les états des modèles SCHMMs de ce locuteur. Cette architecture, inspirée de celles développées dans (Lévy et al., 2006) ou (Bonastre et al., 2003), est illustrée par la figure 5.5. Les probabilités d'émission des états du SCHMM sont approximées par des modèles GMMs obtenus à partir des distributions Gaussiennes du modèle du locuteur indépendant du texte. Le vecteur des poids des distributions Gaussiennes, au sein du mélange constituant la probabilité d'émission d'un état, permet de spécialiser l'état. Cette étape, symbolisée par la figure 8.2 est illustrée dans le cas des modèles GMMs par la figure 8.5.

Concernant les paramètres de moyenne et variance, cette approche réduit le nombre de paramètres à estimer pour un modèle de mot de passe, puisqu'elle permet de mutualiser l'information entre tous les états et le modèle GMM indépendant du locuteur

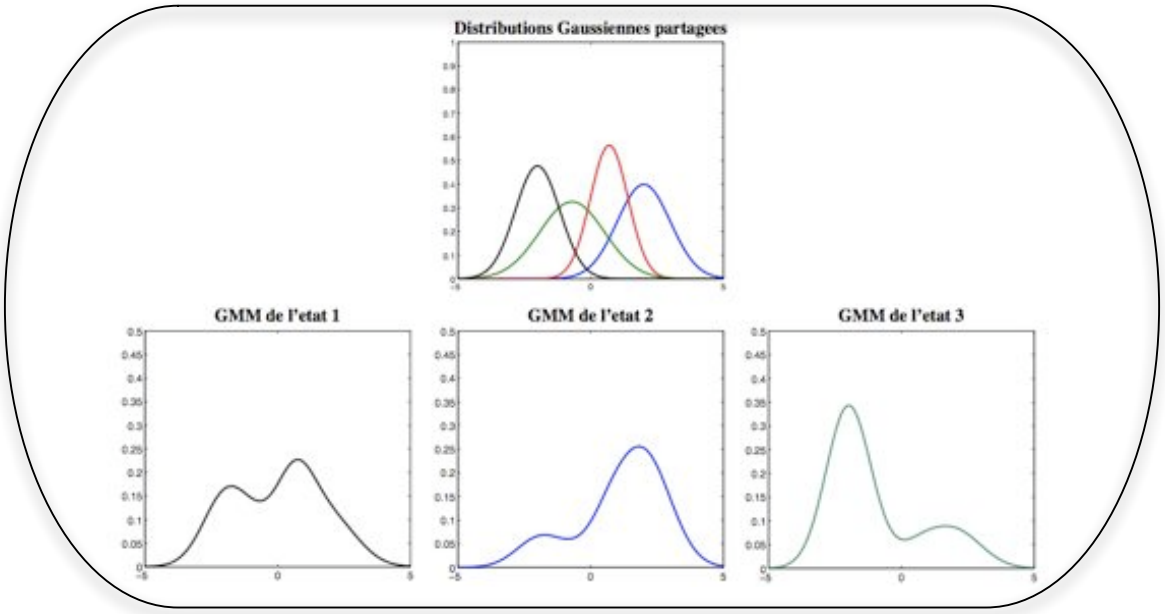


FIG. 8.5: *Exemple de distributions de probabilités modélisées par différents vecteurs de poids à partir des mêmes distributions Gaussiennes.*

qui constitue le niveau intermédiaire de notre architecture. De ce fait, les données disponibles pour l'apprentissage du modèle du locuteur sont mutualisées, permettant un meilleur apprentissage.

Enfin le partage des distributions Gaussiennes entre les modèles de locuteur indépendant du texte et les états des modèles SCHMMs constitue un avantage non négligeable en terme de temps de calcul et autorise à moindre coût le calcul de deux scores lors d'un test d'authentification, l'un dépendant du texte et l'autre qui ne dépend que du locuteur. Ce principe est détaillé dans la suite (cf. section 8.4.1).

8.2 Apprentissage itératif des modèles de mot de passe

Nous avons proposé précédemment (cf. section 8.1) d'utiliser les probabilités d'émission des états du modèle SCHMM comme une version spécialisée du modèle du locuteur indépendant du texte du deuxième niveau de notre architecture.
L'apprentissage des modèles de Markov cachés est un problème complexe qui nécessite des ressources importantes, comme nous l'avons vu dans la partie 8.1.1. Nous proposons dans cette partie un processus d'apprentissage itératif des modèles de mots de passe et nous mettons en exergue les choix auxquels nous avons été confrontés.

Parmi les questions auxquelles nous devons répondre pour optimiser les performances de notre approche de reconnaissance du locuteur structurale, il nous faut déterminer le degré de spécialisation des modèles SCHMMs. La modélisation doit être la plus précise possible (nombre d'états, dimension, paramètres des probabilités d'émission, etc.), tout en tenant compte de la quantité limitée des données d'apprentissage et de test.
La place des trames *non-parole* (introduites dans la section 3.1.4) au sein de l'organisation temporelle de la séquence acoustique constitue une autre problématique qui doit être abordée dans le cadre de nos travaux. En effet, nous avons expliqué dans la section 3.1.4 qu'utiliser les trames étiquetées *non-parole*, par les système de détection d'activité vocale, dégrade généralement les performances des systèmes de reconnaissance du locuteur. Ces trames participent pourtant à l'organisation temporelle de la séquence acoustique. La section 8.3.1 traite de l'utilisation des trames *parole/non-parole* lors des phases d'enrôlement et de test.

8.2.1 Initialisation

Les processus d'apprentissage du modèle du monde et des modèles de locuteur indépendants du texte, qui composent les premier et deuxième niveau de notre architecture, ont été décrit dans le chapitre 7. L'apprentissage du modèle du monde ou des modèles de locuteurs indépendants du texte ne fait pas directement partie des problématiques de cette thèse. Nous avons donc utilisé les méthodes présentées précédemment (cf. chapitre 7), à savoir : apprentissage du modèle UBM par maximisation de la vraisemblance et adaptation des modèles de locuteur par une adaptation MAP.

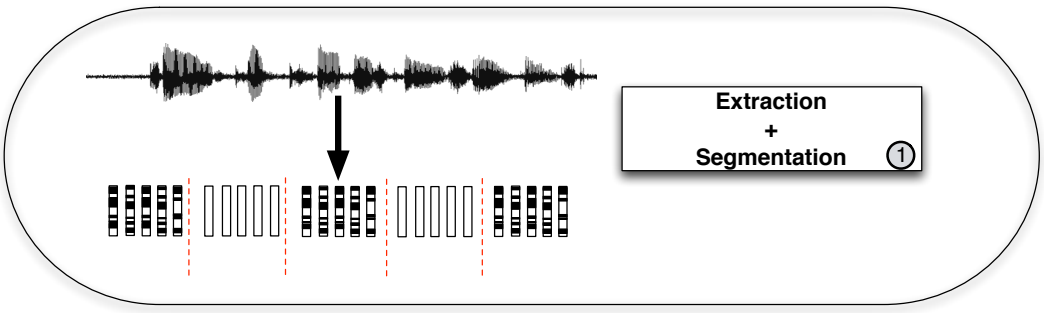


FIG. 8.6: *Première étape du processus itératif d'apprentissage des mots de passe. Les cellules acoustiques composant la séquence d'apprentissage sont isolés.*

Lors de l'apprentissage d'un modèle de mot de passe, la première étape consiste à séparer les éléments fondamentaux du signal de parole, qui constituent la structure du mot de passe considéré (cf. figure 8.6). Cette décomposition nécessite la connaissance du nombre et de la durée de ces cellules acoustiques.
Une première segmentation de la séquence acoustique d'apprentissage en cellules acoustiques doit être déterminée. Le choix de cette segmentation peut reposer sur une connais-

sance a priori de la structure du mot de passe ou être calculée selon d'autres critères qu'il nous faut déterminer. Cette question est traitée par la suite.

Chaque cellule acoustique est utilisée pour apprendre un état du modèle SCHMM (cf. figure 8.7), *i.e.* un modèle GMM, par adaptation des paramètres de poids du modèle de locuteur indépendant du texte. L'apprentissage de ces modèles introduit plusieurs problématiques :

- une fois déterminées les limites temporelles des cellules acoustiques, chaque état peut être adapté en utilisant l'ensemble des paramètres extraits sur le segment qui lui est alloué. Il est également possible de n'utiliser, pour l'adaptation, que certaines trames sélectionnées au préalable ;
- les critères d'adaptation des densités de probabilités d'émissions doivent être fixés en fonction de la faible quantité de données d'apprentissage disponible.

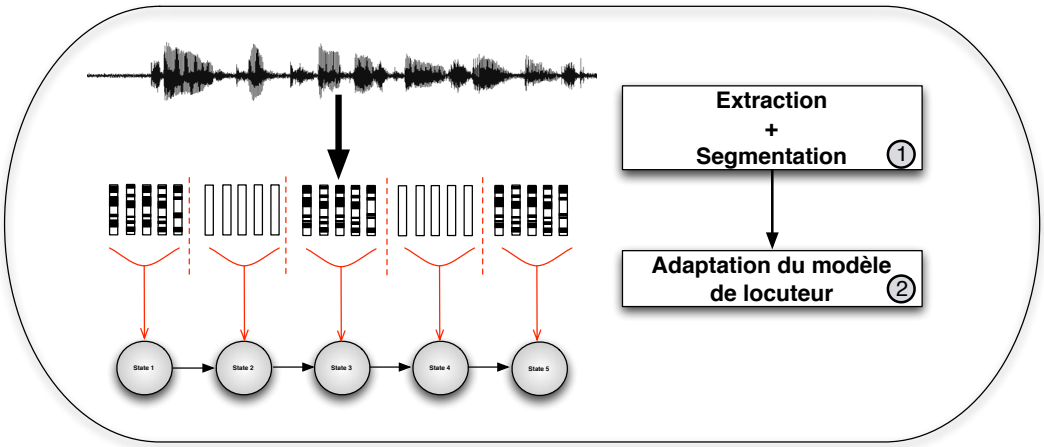


FIG. 8.7: *Deuxième étape du processus itératif d'apprentissage des mots de passe, un état du modèle SCHMM est appris à partir de chaque cellule acoustique.*

Une fois les états du SCHMM appris, nous devons déterminer l'organisation de ces états au sein du modèle de Markov semi-continu. Le choix de la structure des modèles mais également le calcul des probabilités de transitions entre les états doivent prendre en compte la nature et les contraintes de la tâche que nous nous sommes fixée.

À ce stade de l'apprentissage, nous disposons d'une première modélisation du mot de passe. Cependant il est peu probable que le nombre et la répartition des cellules acoustiques composant le mot de passe soient connus exactement. Aussi, considérant que le modèle SCHMM disponible peut être amélioré, nous choisissons d'utiliser un processus itératif au cours duquel nous espérons converger vers la segmentation optimale en cellules acoustiques.

8.2.2 Itérations

Chaque itération débute par un décodage Viterbi de la séquence d'apprentissage avec le modèle obtenu à l'itération précédente (cf. figure 8.8). Ce processus fournit une nouvelle segmentation. Selon la segmentation obtenue, des états peuvent être ajoutés ou supprimés du modèle SCHMM, de manière automatique.

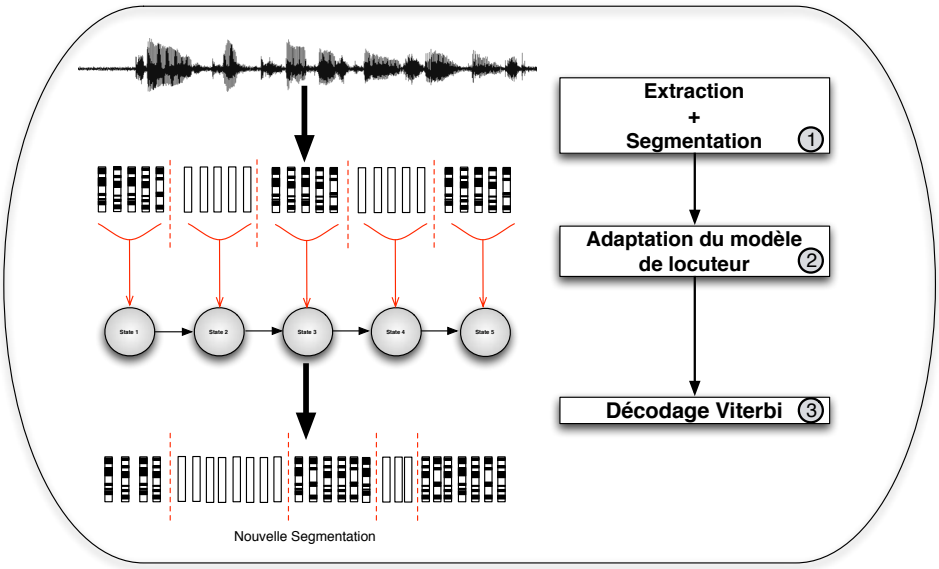


FIG. 8.8: *Troisième étape du processus itératif d'apprentissage des mots de passe, après avoir déterminé l'architecture du modèle SCHMM, un décodage Viterbi est effectué afin d'obtenir un nouveau découpage en cellules acoustiques.*

Chacun des segments qui compose la nouvelle segmentation est utilisé pour adapter un état du modèle SCHMM. Il est possible d'adapter le nouveau modèle d'état à partir du modèle de l'étape précédente ou, comme lors de l'initialisation, à partir du modèle de locuteur indépendant du texte (cf. figure 8.9).

Une étude expérimentale, visant à déterminer le nombre optimal d'itération à effectuer en fonction des résultats et des contraintes ergonomiques, est présentée dans la section 8.3.3.

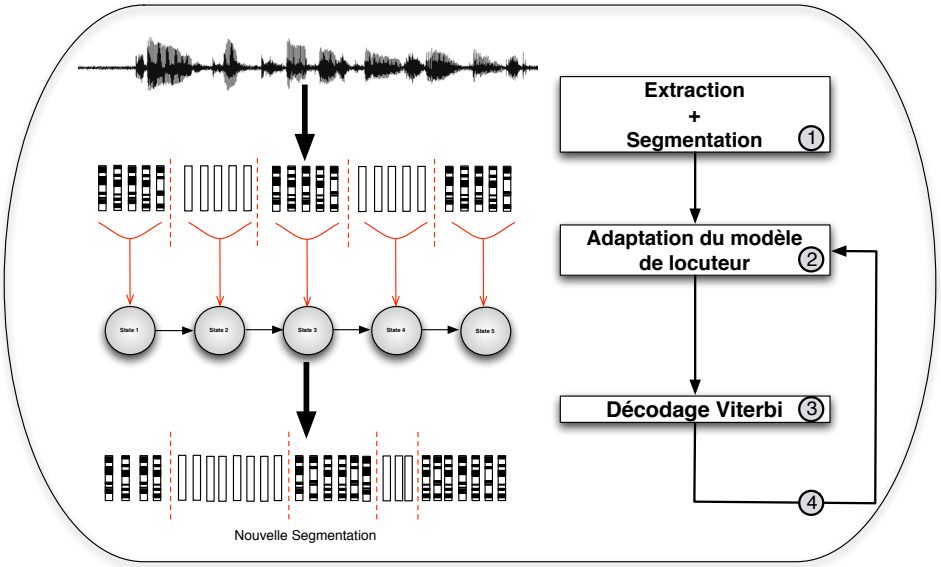


FIG. 8.9: *Quatrième étape du processus itératif d'apprentissage des mots de passe, les états du modèles SCHMM sont ré-estimés à partir de la nouvelle segmentation.*

8.3 Améliorations dues à la structuration du modèle acoustique

Du fait du nombre important de facteurs à déterminer pour parvenir à la configuration optimale de notre architecture, nous ne pouvons réaliser une étude exhaustive. Dans la suite de cette partie, nous présentons les choix et principales expériences réalisées pour déterminer la meilleure configuration possible. Cette configuration n'est donc pas optimale au sens théorique mais tend à s'en approcher de manière expérimentale. Le calcul des scores résultant de la comparaison d'une séquence avec un modèle de mot de passe est détaillé dans la section 8.4.

Nous rappelons ici que tous les résultats produits dans cette partie doivent être considérés avec vigilance (cf. chapitre 6). En effet, la taille réduite des corpus disponibles ne nous a pas permis de disposer de données de développement et de données de validation distinctes et a pu entraîner une sur-estimation des performances.

8.3.1 États du modèle de mot de passe

L'apprentissage de modèles de locuteur à états finis dépendants du texte nécessite la définition des cellules acoustiques, éléments de base de la structure du signal de parole. Comme nous associons toujours, dans ce document, une cellule acoustique à un état du modèle de Markov semi-continu, nous ne considérons que le nombre d'états de ce modèle, celui-ci se confondant avec le nombre de cellules acoustiques.

Quelle est la place des pauses dans la structure des mots de passe ?

Les approches état-de-l'art en reconnaissance du locuteur utilisent un détecteur d'activité vocale (Voice Activity Detection - VAD), présenté dans la section 3.1.4, afin de ne sélectionner que les vecteurs acoustiques utiles au processus de reconnaissance. La sélection des vecteurs de données permet un gain significatif. Il faut cependant s'interroger sur la pertinence de cette sélection dans le cadre de notre approche structurale. Intuitivement, l'alternance des segments de *parole* et de *non-parole* au sein d'une séquence acoustique nous paraît contenir une information relative à la structure du mot de passe prononcé. Il est également possible que cette alternance renferme une information dépendante du locuteur.

La figure 8.10 illustre l'étape de paramétrisation et de détection d'activité vocale. Cette figure permet de distinguer deux flux de données : le *flux brut*, qui résulte de la paramétrisation du signal (cf. section 3.1.2) et le *flux de parole* qui ne contient que les vecteurs de paramètres étiquetés *parole* par le VAD. Les deux flux obtenus peuvent être utilisés en entrée de notre système de vérification du locuteur.

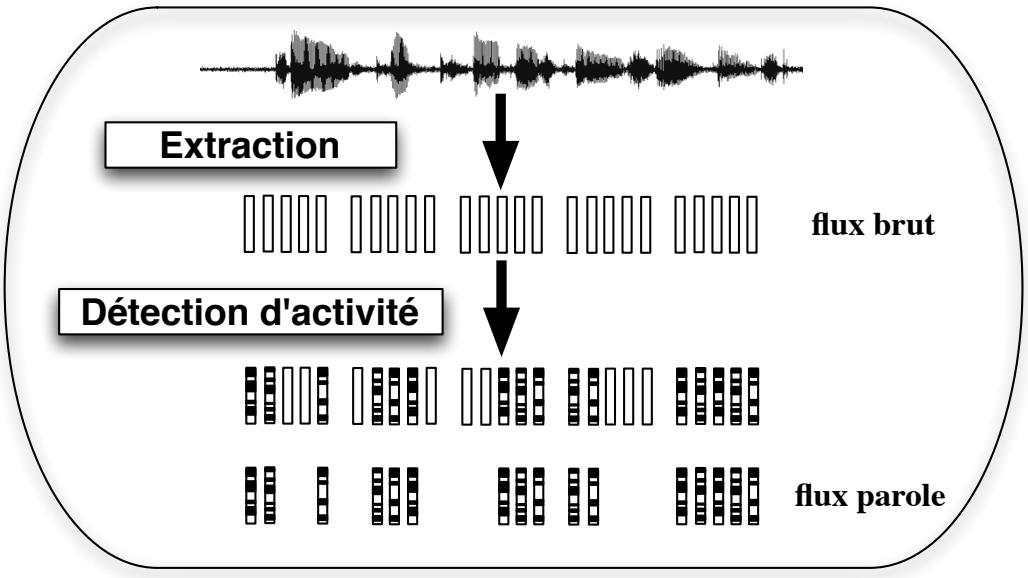


FIG. *8.10:* *Sélection des trames d'entrée par un détecteur d'activité*

Afin d'évaluer l'effet de la sélection des vecteurs de paramètres, nous avons réalisé deux expériences identiques avec le *flux brut* et le *flux de parole* pour la configuration suivante : modèles GMMs à 256 distributions, modèles SCHMMs initialisés à 20 états par modèle, une structure *Gauche-Droite* et des probabilités de transitions équiprobables. Seul le choix des vecteurs de paramètres utilisés pour l'apprentissage des mots de passe diffère.
Tous les alignements phonétiques sont effectués sur le *flux brut*. Les scores calculés lors

de la phase de test ne tiennent compte que des vraisemblances cumulées sur le *flux parole*, puisque ce flux ne contient que les informations spécifiques du locuteur.

En utilisant le *flux brut* pour adapter les modèles, nous considérons que les segments de trames étiquetées silences et pauses par le détecteur d'activité sont des cellules acoustiques à part entière et que leur présence est caractéristique de la structure du mot de passe choisi. La figure 9.12 présente la segmentation *parole/non-parole* calculée par un système VAD et la segmentation résultant du processus itératif d'apprentissage d'un mot de passe utilisant le *flux brut* pour adapter les modèles. Cette seconde segmentation est donc obtenue sans aucune intervention du module de VAD.

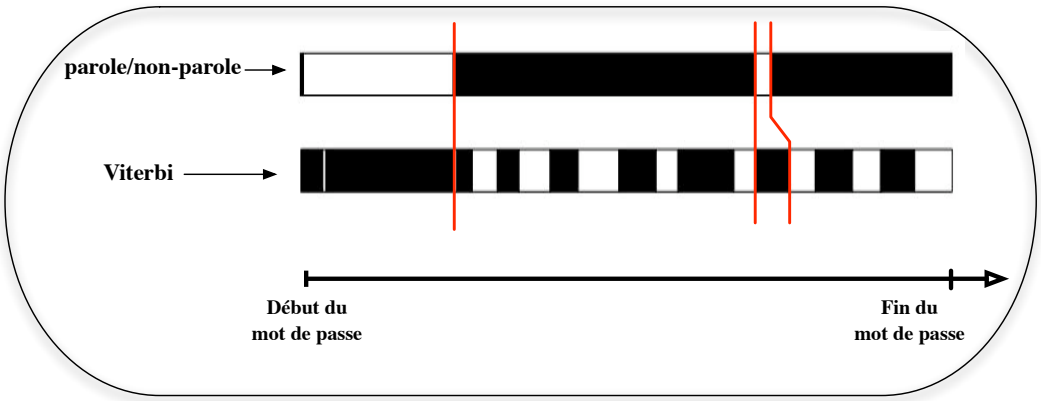


FIG. 8.11: *Comparaison de la segmentation parole/non-parole avec l'alignement résultant du processus d'apprentissage des modèles d'états.*

Le détecteur d'activité délimite cinq segments (3 segments de *parole* et 2 segments de *non-parole*). Notre système, dont le nombre initial d'états par modèle SCHMM est fixé à 20, fournit une segmentation permettant l'apprentissage de ses 20 états. Les deux segmentations présentent certaines similitudes. Certains segments déterminés par le décodage Viterbi semblent correspondre aux limites *parole/non-parole* fixées par le VAD. Cet exemple nous laisse penser que lors de l'apprentissage du modèle de mot de passe, les segments *parole* et *non-parole* sont modélisés par différents états du SCHMM. Les segments *non-parole* paraissent, sur cet exemple, être considérés comme des cellules acoustiques spécifiques.
Cependant, la suppression du détecteur d'activité entraîne une dégradation des performances dans toutes les conditions de tests. Ainsi, le taux d'égales erreurs pour la condition **TOUS**, qui était de $3,17\%$ en utilisant un module VAD, passe à $3,89\%$ lorsque le *flux brut* est utilisé en entrée du système.

Cet effet prévisible est certainement du au fait que pour certains fichiers, la segmentation obtenue ne correspond pas aux cellules acoustiques. Aussi, lors de la création des modèles SCHMMs, les états sont appris avec un ensemble de vecteurs de paramètres issus de *parole* mais aussi de *non-parole*. Par la suite, nous utiliserons toujours,

sauf mention explicite, un module de détection d'activité afin de sélectionner les vecteurs de paramètres utiles à l'apprentissage des modèles.

D'autres approches sont envisageables. Dans le cas où les états du modèle SCHMM sont appris sur les seules trames étiquetées *parole*, aucun état ne modélise les segments de *non-parole*. Il est possible que ce manque nuise au décodage Viterbi et que celui-ci fournisse une segmentation inadaptée. Une solution pourrait être d'ajouter, au sein du modèle SCHMM, des états modélisant explicitement les segments *non-parole* appris sur une quantité importante de données et permettant de prendre en compte la variabilité des environnements à modéliser. Cette méthode est utilisée dans le domaine de la reconnaissance de parole.
L'utilisation de deux modèles SCHMMs par mot de passe est une autre possibilité. Le premier, adapté à partir du *flux brut*, peut être utilisé pour obtenir un alignement lors du décodage Viterbi et l'autre, appris sur le *flux de parole*, serait utilisé pour le calcul du score.

Quelle granularité et quelle dimension pour les modèles ?

Le nombre d'états utilisés pour modéliser un mot de passe détermine la quantité de données correspondant à chaque état ; plus le nombre de cellules acoustiques est important et plus leur durée est courte. L'importance du ratio *quantité de données d'apprentissage/nombre de distributions des modèles* a été mise en évidence dans le chapitre 7. Aussi, en influant sur la quantité de données disponible pour chaque état, le choix du nombre d'états guide celui de la dimension des modèles GMMs à utiliser. Nous préférons donc considérer ce ratio plutôt que chacun des paramètres séparément.

Afin de fixer la granularité - le nombre de cellules acoustiques - nécessaire à la modélisation de chacun des énoncés, nous disposons de connaissances provenant de différents domaines, notamment du domaine de la reconnaissance de la parole. En effet, la plupart des systèmes état-de-l'art de ce domaine utilisent des modèles phonémiques qui sont composés, chacun en moyenne, de trois états émetteurs. Les mots de passe issus de la base de données MyIdea comptent entre 17 et 36 phonèmes par phrase, soit un total allant de 51 à 108 états émetteurs pour des durées à peu près identiques.

Cependant, les modèles phonémiques utilisés en reconnaissance de la parole sont appris en utilisant une grande quantité de données (de nombreuses répétitions d'un même phonème par différents locuteurs). Il est probable que la quantité de données disponible dans notre cas ne permette pas un apprentissage des modèles aussi performant. L'apprentissage, plus robuste, d'un plus faible nombre d'états par modèle permettrait sans doute d'améliorer les performances de notre approche.

Le critère de choix du nombre d'états par modèle ne peut cependant pas tenir compte du texte prononcé puisque celui-ci, laissé au libre choix du client, n'est pas connu. La méthode que nous avons privilégiée ici consiste à déterminer un nombre initial d'états

fixe par modèle, en considérant que tous les mots de passe ont une durée approximativement équivalente. Comme détaillé dans la section 8.2, le nombre des états du mot de passe peut évoluer durant le processus d'apprentissage. Le nombre final d'états est alors variable pour chaque modèle SCHMM. La figure 9.13 illustre les performances de notre approche pour différentes valeurs du ratio *nombre d'états des modèles / dimension des GMMs*.

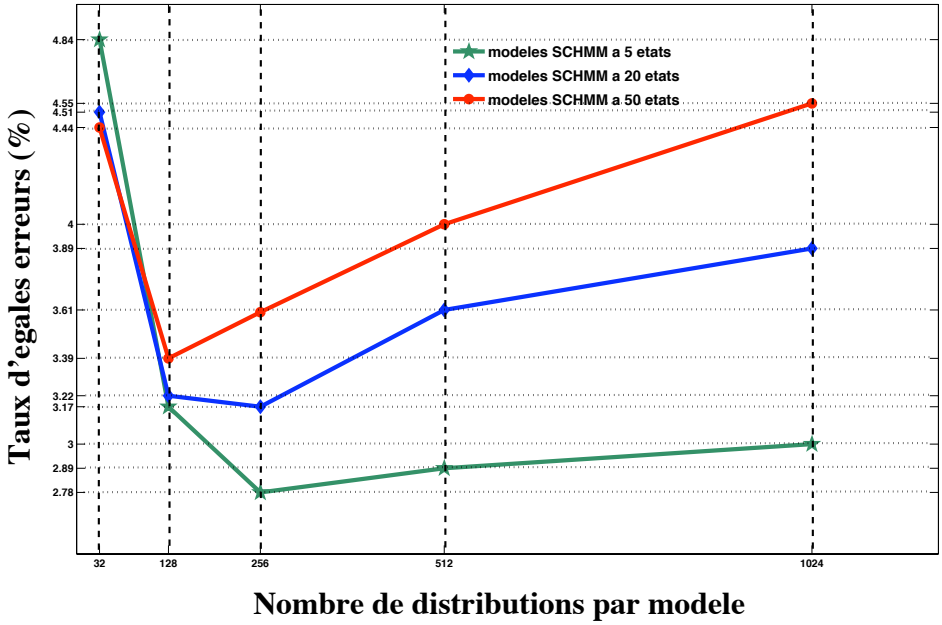


FIG. 8.12: *Évolution du taux d'égales erreurs (%) en fonction de la dimension des modèles GMMs pour différentes valeurs du nombre d'états initial des modèles SCHMMs dans la condition de test TOUS.*

Nous observons que pour une initialisation à nombre d'états constant, l'EER en fonction du nombre de distributions des modèles GMMs est une fonction convexe. La valeur pour laquelle cette fonction atteint son minimum dépend du nombre d'états choisi pour l'initialisation. Plus le nombre d'états à l'initialisation est élevé et plus la dimension des modèles GMMs pour laquelle le taux d'erreur est minimum est petit. Cette observation est en accord avec nos conclusions quant au ratio *nombre d'états des modèles / dimension des GMMs*. Plus la quantité de données d'apprentissage est importante et plus la dimension des modèles GMMs peut être élevée afin de tirer parti de ces données.

Dans le cadre de nos travaux, nous aurions souhaité utiliser un nombre élevé d'états par modèle, de manière à modéliser la structure temporelle des mots de passe le plus précisément possible. Cependant, notre approche impose de conserver une dimension

des modèles GMMs constante entre les trois niveaux de l'architecture. Or les résultats présentés dans le tableau 9.7 montrent que des modèles GMMs de dimension trop faible ne permettent pas une bonne modélisation des locuteurs.

À la lumière des résultats obtenus, et considérant que notre approche mêle le paradigme GMM/UBM à une structuration temporelle des mots de passe utilisant des modèles Markoviens, nous choisissons de privilégier dans la suite des modèles à 256 distributions Gaussiennes. Nous continuerons cependant à présenter des résultats obtenus pour différents nombres d'états par modèle SCHMM.

Condition de test	Nombre initial d'états par modèle SCHMM	
	5	20
FAUX	1,56	0,95
TOUS	2,78	3,17
MDP	3,72	4,56

TAB. *8.1:* *Influence du nombre d'états initial par modèle de mot de passe pour des modèles à 256 distributions par GMM. Les résultats sont exprimés en taux d'égales erreurs.*

Les résultats présentés par la figure 9.13 sont obtenus dans la condition de tests TOUS. Dans cette condition, le nombre optimal d'états pour l'initialisation des modèles de mots de passe est 5. Cependant, la même expérience réalisée dans les conditions de tests FAUX et MDP (cf. tableau 8.1) montrent que l'initialisation des modèles de mots de passe à 20 états permet mieux exploiter la structure temporelle des énoncés. dans la suite de notre étude, nous choisissons de privilégier cette information temporelle et d'utiliser des modèles de mot de passe initialisés à 20 états.

D'autres approches peuvent être envisagées. Une approche scalable pourrait permettre d'utiliser différentes dimensions de modèles GMMs, selon le niveau de l'architecture considéré.

Pourquoi faire varier le nombre d'états au cours de l'apprentissage ?

Nous avons indiqué précédemment que le nombre d'états de chaque modèle SCHMM varie au cours du processus d'apprentissage. Ce phénomène résulte d'un choix qui, comme d'autres, présentés dans cette section, peut être discuté. La diversité des approches envisageables ne nous a cependant pas permis d'évaluer la pertinence de chaque décision.

Comme nous l'avons indiqué précédemment, le décodage Viterbi est effectué sur le *flux brut*, qui comprend les vecteurs étiquetés *parole* et *non-parole* alors que l'appren-

tissage des modèles GMMs de probabilité d'émission n'utilise que les trames *parole*. Il est donc possible dans ces conditions que le nombre de vecteurs de paramètres, étiquetés *parole*, attribués à un état soit très limité ou même nul. Or nous avons montré que la quantité de données d'adaptation conditionne la qualité des modèles GMMs appris (cf. section 7.3.2 et section 8.3.1). C'est pourquoi nous avons fait le choix, pour nos travaux, de supprimer automatiquement les états des modèles SCHMMs ne disposant pas d'un nombre minimum de vecteurs de paramètres pour leur apprentissage.

La figure 8.13 présente le nombre final d'états des modèles SCHMMs après apprentissage.

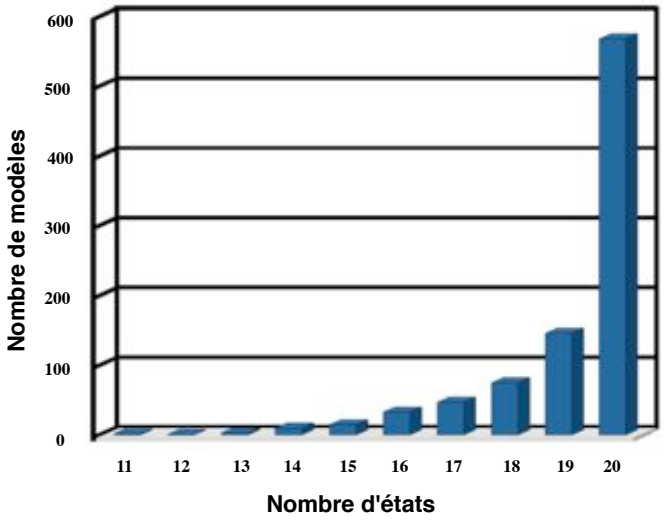


FIG. 8.13: *Nombre d'états final par modèle de mot de passe. Le nombre de ces états est initialisé à 20 au début de la phase d'apprentissage. Les modèles GMMs sont composés de 256 distributions.*

La majorité des modèles appris ont conservé leurs 20 états initiaux, mais environ 36% des SCHMMs ont perdu des états. Nous désirons à présent savoir si le nombre final d'états par modèle est lié au contenu lexical ou à la durée des mots de passes. La figure 8.14 présente le nombre final moyen d'états par modèle pour chacun des dix mots de passes différents utilisés (cf. section 6.2).

La lecture du graphique 8.14 ne permet pas d'établir un lien évident entre le texte prononcé et le nombre final d'états des modèles. Ceci est probablement du, selon nous, à différents paramètres tels que la diction des locuteurs. Le nombre insuffisant de sessions par locuteur ne nous permet cependant pas d'évaluer la corrélation entre locuteur et nombre d'états.

Dans le cadre des expériences présentées dans ce document, la quantité minimale de données a été fixée de manière empirique, garantissant que chaque état est appris

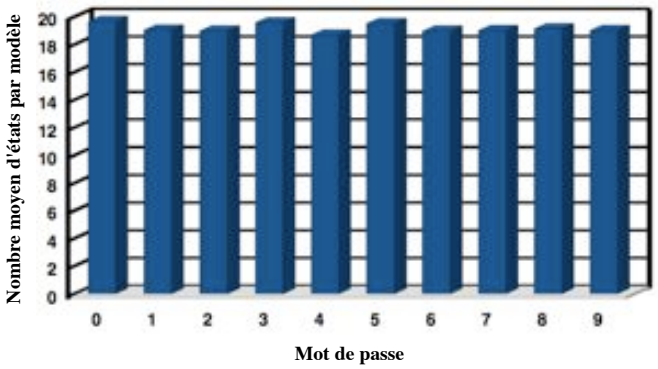


FIG. *8.14:* *Nombre moyen d'états final par modèle de mot de passe pour chaque énoncé. Le nombre de ces états est initialisé à 20 au début de la phase d'apprentissage.*

avec au mois 2% des vecteurs du *flux de parole*. Sa valeur ainsi que le critère de modification du nombre d'états par modèle peuvent être optimisés. L'enjeu est de déterminer de façon précise, et sans aucune connaissance a priori, le nombre et la répartition des cellules acoustiques qui composent une séquence audio.

Les modèles d'états doivent-ils être adaptés à partir du modèle de locuteur indépendant du texte ou à partir des états de l'itération précédente ?

Lors de l'initialisation des modèles SCHMMs, le GMM de chaque état est créé à partir du modèle de locuteur indépendant du texte qui compose le deuxième niveau de notre architecture. À chaque itération du procédé d'apprentissage, les nouveaux modèles GMMs des états peuvent être adaptés, comme lors de l'initialisation, à partir du modèle de locuteur indépendant du texte. Il est également possible d'utiliser une adaptation MAP pour laquelle l'a priori provient des statistiques issues du modèle appris à l'itération précédente.
Il ne s'agit plus alors exactement d'une adaptation MAP puisque ce processus converge vers le maximum de vraisemblance et risque de dégrader les performances du système.

Le tableau 8.2 présente les performances obtenues pour différentes configurations d'adaptation des modèles de mot de passe. Dans le premier cas, les états des modèles SCHMMs sont appris en adaptant, à chaque itération, les modèles de locuteur indépendant du texte. Une seule itération est effectuée. Dans les trois autres configurations, les modèles sont adaptés à chaque itération à partir des modèles calculés à l'itération précédente, pour une, trois ou cinq itérations successives.

Comme nous l'avions supposé, l'apprentissage des modèles à parti des états calculés à l'étape précédente dégrade les résultats de la vérification d'identité. Par la suite,

	Configuration d'apprentissage des modèles d'états			
Modèle utilisé pour l'apprentissage	locuteur indépendant du texte	états de l'itération précédente		
Nombre d'itérations effectuées	1	1	3	5
EER obtenu en condition 1-occ **TOUS**	3,17	3,39	3,34	3,34

TAB. 8.2: *Influence du choix du modèle adapté pour l'apprentissage des états des mots de passe. Les résultats sont exprimés en taux d'égales erreurs.*

les modèles d'états seront adaptés à partir des modèles de locuteur indépendants du texte.

Comment adapter les paramètres de poids des états ?

Nous avons fait le choix d'utiliser des modèles de Markov semi-continus en arguant que l'apprentissage de leurs homologues continus nécessitent une importante quantité de données. Chaque état des modèles SCHMMs est obtenu par adaptation du modèle de locuteur indépendant du texte de la deuxième couche de notre architecture. Il est important de trouver un critère d'adaptation adapté à la faible quantité de données disponible pour chaque état.
Nous choisissons le critère d'adaptation utilisé généralement en reconnaissance du locuteur, le critère du Maximum A Posteriori, présenté dans la partie 7.2.2. Une étude plus approfondie pourrait être menée afin de trouver un critère d'adaptation plus pertinent dans ce cadre applicatif. En effet, l'utilisation du critère MAP pour l'adaptation des poids fait apparaître un phénomène de récursivité (cf. équation 7.14). De plus, ce critère ne tire pas partie du contexte de l'état adapté, au sein du modèle SCHMM, comme c'est le cas lors de l'apprentissage d'un modèle HMM avec le critère d'estimation de l'information mutuelle maximale (Maximum Mutual Information Estimation - MMIE) (Bahl et al., 1986).

Afin de déterminer la meilleure adaptation possible avec le critère MAP, nous avons testé différentes configurations. Notre démarche a consisté à accorder plus ou moins d'importance aux données d'adaptation. Le graphique 8.15 présente les taux d'égales erreurs obtenus pour les trois configurations suivantes :

Critère MAP : L'adaptation de modèles GMMs selon le critère MAP a été présentée dans la partie 7.2.2. Une formulation simplifiée est donnée ci-dessous.

$$W_{adapt} = \alpha\, W_c + (1 - \alpha)\, W_{init} \tag{8.13}$$

W_{adapt} est le paramètre de poids résultant de l'adaptation, il est le résultat d'une somme pondérée de W_c, poids estimé sur les données d'apprentissage en utilisant l'algorithme

EM (cf. annexe B) avec W_{init}, paramètre provenant du modèle a priori (ici le modèle du locuteur indépendant du texte).

Critère du Maximum de vraisemblance (MLE) : Le critère du maximum de vraisemblance consiste à négliger le terme a priori dans l'équation 8.13. Ainsi :

$$W_{adapt} = 1 \times W_c + 0 \times W_{init} = W_c \tag{8.14}$$

Critère MAP contraint (CMAP) : Afin de modifier la confiance sur la distribution a priori des paramètres de poids, nous avons ajouté un paramètre β. Ce paramètre a le même rôle que le *relevance factor* du MAP mais permet une meilleure lisibilité. La valeur du paramètre β du critère CMAP est fixée à $0,3$ dans la suite.

$$W_{adapt} = \alpha\,\beta\,W_c + (1 - \alpha\,\beta)\,W_{init} \tag{8.15}$$

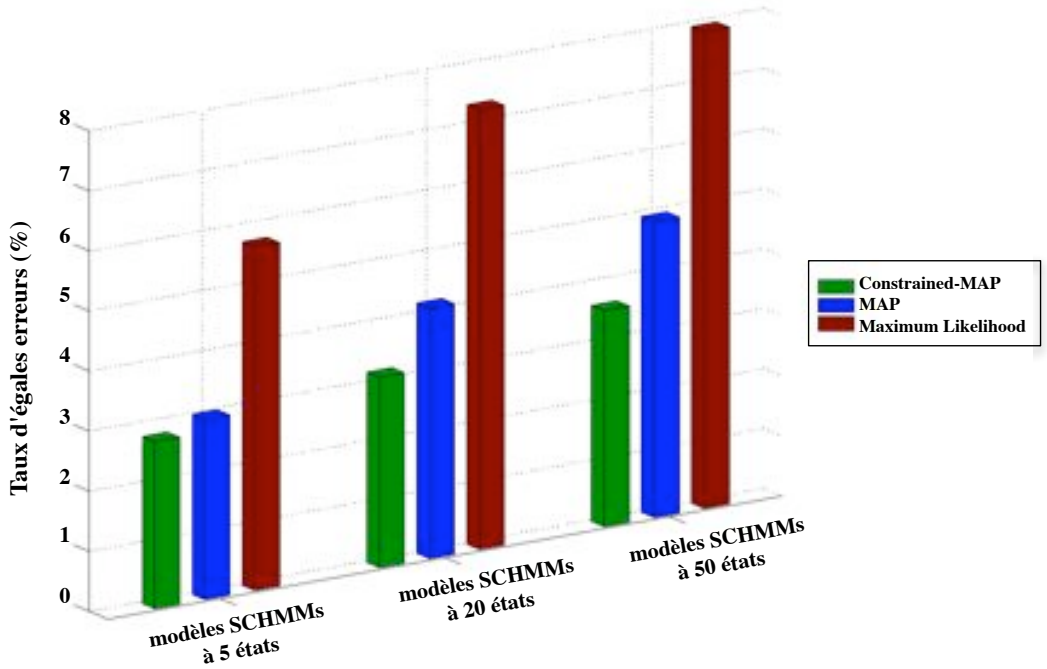


FIG. 8.15: *Évolution du taux d'erreur du système de vérification du locuteur en fonction du critère d'adaptation utilisé pour l'apprentissage des états du modèle SCHMM*

L'adaptation par un critère MAP contraint fournit les meilleurs résultats. Ainsi, pour les trois critères retenus, plus la confiance dans les données d'apprentissage est élevée et plus les résultats se dégradent. La quantité de données est trop faible pour leur pouvoir accorder une confiance élevée. Pour le suite de notre étude, le critère d'adaptation des états sera le critère MAP contraint avec $\beta = 0,3$.

Faut-il adapter tous les paramètres de poids des modèles d'états ?

Chaque état des modèles SCHMMs est une mixture de Gaussiennes caractérisée par un vecteur de poids. Nous avons vu précédemment que réduire la dimension des modèles GMMs augmente le taux d'erreurs de vérification d'identité. La réduction du nombre de paramètres caractérisant chaque état des SCHMMs pourrait permettre de réduire considérablement les ressources nécessaires au stockage des modèles tout en permettant une bonne représentation du locuteur. Mais la réduction du nombre de paramètres risque de provoquer une perte d'information.
Afin d'estimer les dégradations engendrées par cette réduction, nous avons mené les expériences suivantes. Pour une configuration donnée, nous avons réalisé 5 expériences. Pour chacune d'elle, nous faisons varier le nombre de paramètres de poids adaptés pour chaque état des modèles SCHMMs. Afin de respecter la condition selon laquelle la somme des poids des distributions d'un GMM est égale à 1, les paramètres non-adaptés sont obtenus par une normalisation des paramètres de poids du modèle de locuteur indépendant du texte de la deuxième couche de notre architecture. Les poids adaptés sont les N plus importants. La figure 9.14 présente les performances obtenues pour des modèles SCHMMs à 20 ou 50 états.

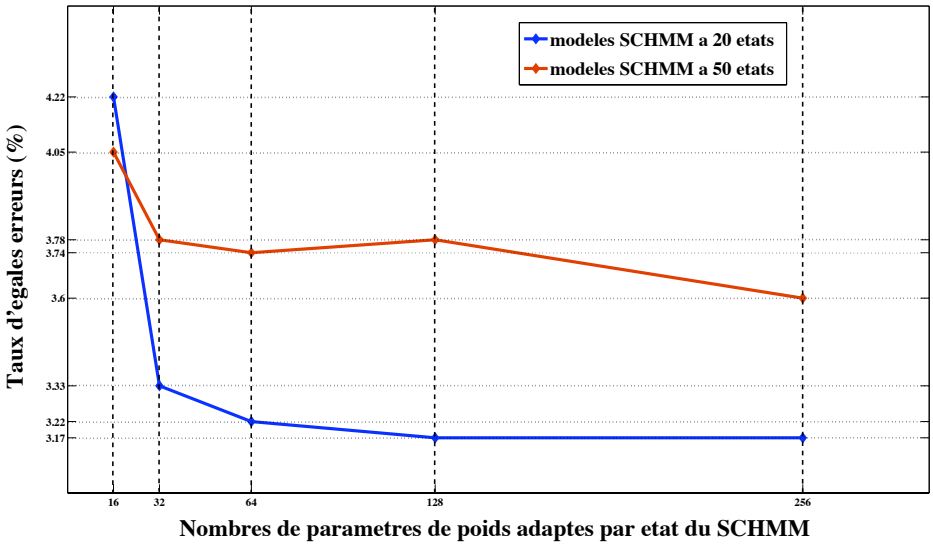


FIG. 8.16: *Analyse de la dégradation des performances en fonction du nombre de poids adaptés pour des modèles de mot de passe à 256 distributions.*

D'après les expériences réalisées, il semble que l'adaptation de seulement 128 paramètres de poids, sur les 256 que comptent chaque état des modèles SCHMMs, permette de conserver des performances acceptables tout en diminuant d'un facteur 2 le nombre de paramètres à stocker.
Cette économie de ressources paraît cependant négligeable au regard de la dégradation observée, surtout en considérant l'évolution des capacités de stockage des équipements

sur lesquelles pourrait être embarqué notre système.
Nous continuons donc, dans la suite de notre étude, à adapter l'ensemble des paramètres de poids des modèles SCHMMs.

8.3.2 Structure du modèle SCHMM

Les modèles de Markov peuvent revêtir différentes structures selon les transitions qui sont autorisées entre les états. Les trois principales architectures existantes sont représentées sur la figure 8.17

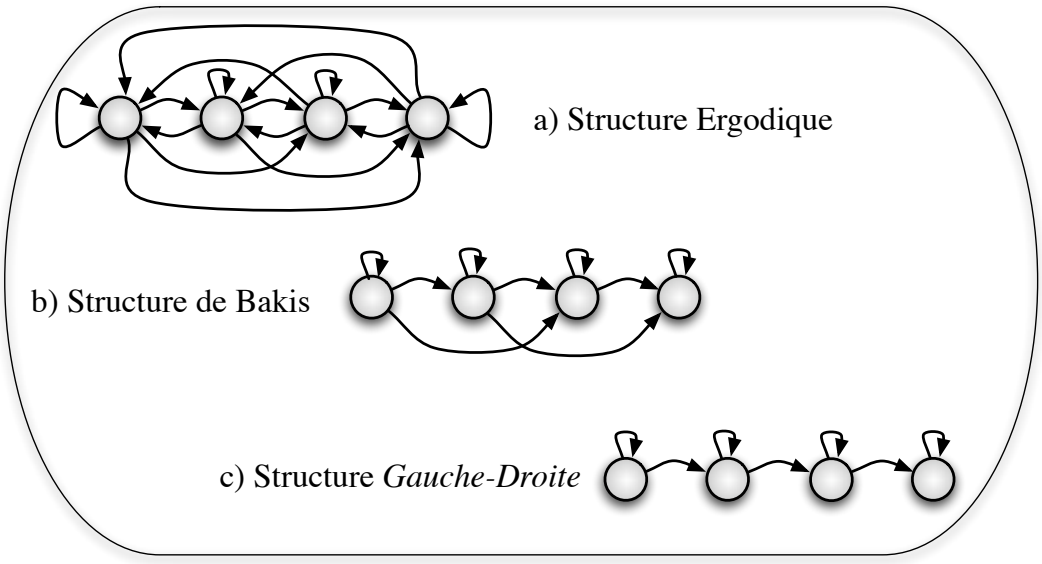


FIG. *8.17:* *Structures possibles des modèles de Markov*

Nous pouvons d'ores et déjà distinguer la structure ergodique, qui autorise des retours en arrière, des structures de Bakis et *Gauche-Droite* qui n'autorisent qu'une progression chronologique. Ces deux dernières nous paraissent plus adaptées à la tâche fixée, puisqu'elles respectent la chronologie des événements acoustiques. La structure *Gauche-Droite* impose une contrainte chronologique plus forte que la structure de Bakis, puisque cette dernière permet de « sauter » certains états et offre une plus grande liberté lors de l'alignement d'une séquence.

La section 8.3.1 a permis de déterminer un certain nombre de paramètres permettant un apprentissage des états du modèle SCHMM adapté à nos contraintes. Comme nous l'avons précisé dans la partie 8.1.1, l'estimation des paramètres des modèles de Markov inclue également l'estimation des probabilités de transitions entre états.
Comme pour l'apprentissage des états, nous ne disposons pas de suffisamment de données d'apprentissage pour permettre une estimation par les techniques décrites précédemment (cf. annexe C). Deux options s'offrent à nous :

– Nous pouvons considérer que toutes les transitions sont équiprobables. La seule

contrainte imposée provient dans ce cas de l'architecture même du modèle de Markov.

– Il est également possible de calculer pour chaque état i la probabilité de transiter vers un état j en tenant compte d'un premier alignement $\mathcal{A} = \{a_0, a_1, ..., a_T\}$ obtenu avec des transitions équiprobables. Dans ce cas, la probabilité dépendante de la durée, $\tau_{i,j}$, de transiter de l'état i vers l'état j est donnée par :

$$\tau_{i,j} = \frac{p_{i \to j}}{\sum_{\mathcal{S}_i} p_{i \to s}} \tag{8.16}$$

où $\mathcal{S}_i$ est l'ensemble des états vers lesquels il est possible de transiter à partir de l'état i et $p_{i \to s}$ est le nombre de transitions de l'état i vers l'état s pour le chemin $\mathcal{A}$.

Nous testons ces deux configurations avec les trois types de structures décrits précédemment, les résultats obtenus sont présentés dans le tableau 8.3.

Type de transitions	Référence GMM/UBM	Architecture des modèles SCHMM		
		Gauche-Droite	Structure de Bakis	Ergodique
Équiprobables	3.22	3,17	3,18	4,01
Dépendantes de la durée		3,68	3,78	4,60

TAB. 8.3: *Influence de la structure du modèle SCHMM pour différentes estimations des probabilités de transitions entre états sur le taux d'égales erreurs (%) dans la condition de tests TOUS. Les modèles SCHMMs sont initialisés avec 20 états et chaque mixture contient 256 distributions Gaussiennes.*

Comme nous l'avons supposé plus tôt, l'architecture Ergodique n'est pas adaptée à la tâche qui est la nôtre. L'organisation chronologique de la structure de Bakis et de la structure *Gauche-Droite*, offrent une meilleure modélisation de l'organisation temporelle des mots de passe. Le calcul des probabilités de transition dépendantes de la durée, tel que nous l'avons effectué, ne semble pas adapté. Peut être est-ce du au fait que ces transitions ne sont calculées qu'à partir d'une occurrence du mot de passe et n'intègrent donc pas la variabilité inter-sessions nécessaire.
Pour la suite de notre étude nous nous contentons d'utiliser une structure *Gauche-Droite*, moins coûteuse qu'une structure de Bakis, lors de l'alignement des séquences, pour laquelle les transitions seront équiprobables.

8.3.3 Apprentissage itératif

Nous avons décrit dans la partie 8.2 un processus itératif supposé permettre l'obtention d'une segmentation en cellules acoustiques pertinente. Jusqu'alors, la première

segmentation permettant l'initialisation du processus était réalisée sans aucune information a priori. La séquence était découpée en segments de longueurs égales. Des expériences réalisées en faisant varier le nombre d'itérations du processus ont fait apparaître deux comportements. Dans certains cas, la segmentation converge après deux ou trois itérations alors que pour d'autres séquences, la convergence n'est jamais atteinte et le système oscille entre deux alignements. Les performances du système ne variant pas de façon significative avec l'augmentation du nombre d'itérations, nous fixons le nombre d'itérations à 1.
La recherche d'une segmentation initiale ainsi que le développement d'un algorithme permettant l'optimisation de cette segmentation nous apparaît néanmoins comme un champ d'investigation important qui permettrait peut être d'améliorer la modélisation des mots de passe de manière significative.

8.4 Exploiter pleinement l'architecture à trois niveaux

Le partage des distributions Gaussiennes entre le modèle du locuteur indépendant du texte et les états du SCHMMs qui modélisent le mot de passe de ce locuteur permet de réduire significativement le coût de calcul du score lors de la comparaison d'une séquence de test avec ce modèle. Cette architecture permet également de calculer non pas un mais deux scores.
Les deux premières couches de l'architecture sont issues du paradigme GMM/UBM et permettent le calcul d'un premier score, n'exploitant pas la structure de la séquence acoustique.

8.4.1 Calcul d'un score double

Soit un locuteur I_X dont le modèle GMM est de la forme $\sum_{i=1}^{N} \gamma_i^{loc} \mathcal{N}(O_t, \mu_i^{loc}, \Sigma_i^{loc})$, où γ_i^{loc}, μ_i^{loc} et Σ_i^{loc} sont respectivement le poids, le vecteur de moyennes et la matrice de co-variance de la distribution i. Le résultat de la comparaison d'une séquence de test $\mathcal{O} = \{O_t\}, t \in [1, T]$ avec le modèle de I_X est un score dont l'expression est donnée par :

$$
\begin{aligned}
LLR(\mathcal{S}) &= \frac{1}{T}.\sum_t \log\left(\frac{p(O_t|X)}{p(O_t|\mathcal{W})}\right) \\
&= \frac{1}{T}.\sum_t \log\left(\frac{\sum_{i=1}^{N} \gamma_i^{loc} \mathcal{N}(s_t, \mu_i^{loc}, \Sigma_i^{loc})}{\sum_{i=1}^{N} \gamma_i^{UBM} \mathcal{N}(O_t, \mu_i^{UBM}, \Sigma_i^{UBM})}\right)
\end{aligned}
$$

On pose :

$$L_i(O_t) = \frac{\mathcal{N}(O_t, \mu_i^{loc}, \Sigma_i^{loc})}{\sum_{i=1}^{N} \gamma_i^{UBM} \mathcal{N}(O_t, \mu_i^{UBM}, \Sigma_i^{UBM})} \tag{8.17}$$

d'où :

$$LLR(\mathcal{O}) = \frac{1}{T}.\sum_t \log\left(\sum_{i=1}^{N} \gamma_i^{loc}\, L_i(O_t)\right) \tag{8.18}$$

Lors de la comparaison de la même séquence avec le modèle SCHMM de mot de passe du locuteur I_X, un décodage de Viterbi fournit un alignement $\mathcal{A} = \{a_0, a_1, ..., a_T\}$. Le score de ce test est un rapport entre la vraisemblance de la séquence $\mathcal{A}$ avec les états alignés et la vraisemblance obtenue pour le modèle du monde du premier niveau de l'architecture. L'expression de ce score est :

$$\begin{aligned} LLR(\mathcal{S}) &= \frac{1}{T}.\sum_t \log\left(\frac{p(O_t|a_t)}{p(O_t|\mathcal{W})}\right) \\ &= \frac{1}{T}.\sum_t \log\left(\frac{\sum_{i=1}^{N} \gamma_i^{a_t} \mathcal{N}(O_t, \mu_i^{a_t}, \Sigma_i^{a_t})}{\sum_{i=1}^{N} \gamma_i^{UBM} \mathcal{N}(O_t, \mu_i^{UBM}, \Sigma_i^{UBM})}\right) \end{aligned}$$

Or, les distributions Gaussiennes des états du SCHMM et du modèle du locuteur indépendant du texte sont les mêmes, donc :

$$\mu_i^{a_t} = \mu_i^{loc} \;\text{ et }\; \Sigma_i^{a_t} = \Sigma_i^{loc} \tag{8.19}$$

d'où :

$$\begin{aligned} LLR(\mathcal{S}) &= \frac{1}{T}.\sum_t \log\left(\frac{\sum_{i=1}^{N} \gamma_i^{a_t} \mathcal{N}(O_t, \mu_i^{loc}, \Sigma_i^{loc})}{\sum_{i=1}^{N} \gamma_i^{UBM} \mathcal{N}(O_t, \mu_i^{UBM}, \Sigma_{UBM_i})}\right) \\ &= \frac{1}{T}.\sum_t \log\left(\sum_{i=1}^{N} \gamma_i^{a_t}\, L_i(O_t)\right) \end{aligned}$$

On observe que les expressions des deux scores sont semblables à l'exception des paramètres de poids. Aussi le surcoût du au calcul du score dépendant de la structure du mot de passe se limite à une somme pondérée par vecteur de données et par état.

Pour une séquence acoustique d'entrée, chacun des scores, dépendant ou indépendant de la structure temporelle, décrits précédemment peut être calculé à partir du *flux brut* ou du *flux de parole* (cf. section 8.3.1). Le traitement d'une même séquence peut donc générer quatre scores :

- un score Sc_{GMM}^{parole} indépendant de la structure temporelle et ne prenant en compte que les trames étiquetées parole par le VAD ;
- un score Sc_{GMM}^{brut} indépendant de la structure temporelle et prenant en compte toutes les trames ;
- un score Sc_{SCHMM}^{parole} dépendant de la structure temporelle et ne prenant en compte que les trames étiquetées parole par le VAD ;
- un score Sc_{SCHMM}^{parole} dépendant de la structure temporelle et prenant en compte toutes les trames.

Considérant que la conservation des trames *non-parole* est motivée par une volonté de prendre en compte les pauses et silences au sein du modèle structural, il ne sera plus question dans ce document du score Sc_{GMM}^{brut}. De nombreux travaux (Besacier et al., 2000) ont déjà traité de la sélection de trames pour les systèmes GMM/UBM et la suppression du VAD pour ce type de système dégrade fortement leurs performances. Par la suite, toutes les références au score indépendant de la structure temporelle renverront au score Sc_{GMM}^{parole}. Toutes les expériences, sauf mention explicite, seront réalisées à partir du *flux de parole*
Nous reviendrons dans le chapitre 9 sur les effets de la suppression du VAD et comparerons à cette occasion les scores obtenus pour l'approche structurale avec les *flux brut* et *flux de parole*.

8.4.2 Comparaison des scores dépendants et indépendants de la structure temporelle

Les deux modèles de locuteur, dépendant et indépendant du texte (deuxième et troisième couche de l'architecture hiérarchique), modélisent des informations communes. Les deux scores, dépendant et indépendant de la structure temporelle contiennent pourtant deux informations différentes. Nous avons donc testé une fusion de scores (cf. section 5.2).

Le tableau 9.10 présente les résultats obtenus par l'approche non-structurale, correspondant au système GMM/UBM de référence, mais également aux deux premiers niveaux de notre architecture, les performances de l'approche structurale (SCHMMs du troisième niveau), ainsi qu'une fusion de ces approches. Il s'agit simplement d'une fusion de scores par une somme pondérée. Les poids attribués empiriquement aux deux approches sont les suivants :

- approche structurale : $0,3$
- approche non structurale : $0,7$

Nous avons analysé les performances du système GMM/UBM dans ces trois conditions dans la section 7.3.2. L'introduction de l'information structurelle permet un gain important dans la condition **FAUX** puisque le taux d'égales erreurs diminue de 60% relatifs (de 2,46% à 0,94%). L'utilisation des modèles SCHMMs permet une réelle discrimination des structures temporelles. Cette discrimination est certainement la cause

Conditions de test	Taux d'égales erreurs		
	Approche non structurale	Fusion	Approche structurale
MDP	4,00	4,06	4,62
FAUX	2,46	1,11	0,94
TOUS	3,22	2,83	3,17

TAB. ***8.4:*** *Performances obtenues par les approches structurales, non-structurales et par une fusion des scores de celles-ci.*

des mauvaises performances de l'approche structurale dans la condition **MDP** où les imposteurs prononcent le mot de passe des clients. Il semble que dans cette condition, le modèle « reconnaît » la structure au détriment du locuteur. La réunion de ces deux expériences au sein de la condition **TOUS** montre que la différence entre les deux approches n'est pas significative.

Les résultats obtenus par fusion des deux approches montrent en revanche un gain relatif de 10% dans la condition **TOUS**. En effet, les performances de la fusion de systèmes pour la condition **MDP** sont équivalentes à celles de l'approche GMM/UBM alors qu'elles restent, dans la condition **FAUX**, très proches de ceux de l'approche structurale.
À la lecture du tableau 9.10, nous voyons que la fusion des informations de chaque niveau de notre architecture permet de tirer le meilleur parti des deux approches.

Comme nous l'avons indiqué, la fusion des scores est une simple somme pondérée, pour laquelle les poids ont été déterminés empiriquement. Il est certainement possible d'améliorer les performances en choisissant une méthode de fusion plus adaptée.

Conclusion

Nous avons introduit dans ce chapitre une approche permettant d'intégrer une information structurelle au sein d'un processus de reconnaissance du locuteur. Cette méthode repose sur une segmentation en cellules acoustiques des séquences sonores. Nous avons proposé un algorithme itératif permettant de déterminer les cellules acoustiques composant les mots de passe des clients. L'architecture proposée a été évaluée sur la base de données MyIdea et a montré des capacités intéressantes en terme de vérification d'identité, notamment dans le cas où les imposteurs ne connaissent pas le mot de passe des clients. L'utilisation des trois niveaux de notre système permet d'égaler ou de surpasser les performances des systèmes GMM/UBM état-de-l'art dans toutes les conditions testées.

Un certain nombre de points pourraient être améliorés comme, par exemple, l'apprentissage itératif des modèles SCHMMs modélisant la structure temporelle des mots

de passe. Le processus de segmentation en cellules acoustiques nécessiterait une meilleure prise en compte de la structure globale des séquences sonores incluant un processus discriminant du type MMIE (Bahl et al., 1986). Le procédé de fusion des scores pourrait aussi être perfectionné et il serait certainement intéressant de fusionner les informations issues des approches structurales et non-structurales au cours du processus de scoring.

Chapitre 9

Renforcement de la structure temporelle par une contrainte de synchronisation

Sommaire

Résumé

Ce chapitre présente une nouvelle méthode de vérification du locuteur dépendante du texte basée sur la contrainte du processus acoustique par une synchronisation extérieure. Cette approche est validée dans un premier temps par l'utilisation d'une information provenant d'un alignement phonétique. La suite de ce chapitre est consacrée à la synchronisation du décodage acoustique par une information provenant du flux vidéo. Cette étude est réalisée dans le cadre de nos contraintes applicatives. Les expériences présentées corroborent l'analyse déjà exposée dans le chapitre 4 : l'extraction d'une information structurelle à partir d'un flux vidéo nécessite un traitement coûteux.

Introduction

Ce chapitre est consacré au renforcement de la structure temporelle des modèles acoustiques de mots de passe par l'ajout d'une information synchrone. L'ajout de cette contrainte a pour but de pallier le manque de données d'entraînement et la courte durée des séquences de test.
L'approche acoustique structurale développée dans le chapitre 8 a montré des performances intéressantes dans ce cadre applicatif. Nous avons cependant souligné le fait que la segmentation temporelle en cellules acoustiques, nécessaire à l'initialisation de cette méthode, doit être améliorée. Nous proposons d'exploiter une connaissance a priori sur la segmentation.

Les travaux de Yehia et al. (1997) et de Eveno et Besacier (2005) ont démontré la forte corrélation existant entre les flux audio et vidéo lors de la production de parole. Eveno et Besacier (2005) ont également exploité cette corrélation pour détecter des impostures par play-back.

Comme nous l'avons vu précédemment, la plupart des systèmes embarqués disposent d'une caméra et permettent des traitements simples du flux vidéo. Les travaux de Barker et Berthommier (1999), Goecke et Millar (2003) ou encore Siracusa et Fisher (2007), déjà évoqués, ont montré que les mouvements des articulateurs sont très fortement corrélés avec le signal de parole.
Un signal caractérisant les mouvements des articulateurs, ou la quantité de mouvement visible sur la vidéo, possède un fort potentiel pour apporter une information permettant la synchronisation du décodage acoustique. Cette information pourrait également être exploitée afin de détecter certains types d'impostures (Eveno et Besacier, 2005).

Afin de respecter le cadre applicatif que nous nous sommes fixé, l'extraction de cette information vidéo ne doit pas nécessiter de traitement du flux vidéo trop onéreux. Cette contrainte a des implications fortes, du fait de la nature même du signal. Il s'agit en effet d'un signal de grande dimension (nombre de pixels par image $\times$ temps) qui requiert, de plus, une prise en compte de la nature bi-dimensionnelle des images, au sens où l'information portée par un pixel au cours du temps est fortement corrélée avec celle des pixels voisins.

Nous présentons dans la section 9.1.1 le processus d'intégration de l'information temporelle dans le cadre de l'architecture acoustique. La section 9.1.2 décrit plus particulièrement l'effet de cette information sur le décodage Viterbi lors des phases d'apprentissage et de test. Cette approche est validée grâce à une information issue d'un alignement phonétique, les conditions et conclusions des expériences sont décrites dans la partie 9.2.

Nous revenons dans la section 9.3 sur les techniques vidéo état-de-l'art, qui permettent d'extraire un signal caractéristique du mouvement des articulateurs, tout en veillant à limiter la quantité de calcul requise. Nous présentons ensuite une proposition pour caractériser simplement la quantité de mouvement liée à la production de parole, ainsi que les résultats obtenus par cette méthode.

9.1 Intégration d'une information temporelle externe au sein de l'architecture acoustique

L'intégration d'une contrainte temporelle provenant d'un flux externe, au sein du système acoustique présenté dans les chapitres 7 et 8 est motivée, d'une part, par la volonté d'améliorer la modélisation de la structure des mots de passe et, d'autre part, par la volonté de pénaliser les séquences prononcées par des imposteurs lors de la phase de test. Ces deux aspects sont détaillés dans les paragraphes suivants.

9.1.1 Améliorer l'apprentissage des modèles en améliorant la segmentation en cellules acoustiques

La structure temporelle des mots de passe est modélisée par des modèles à états finis (cf. chapitre 8). Chaque état de ces modèles est supposé modéliser une cellule acoustique, élément de base de la structure du signal de parole. Nous avons conclu, dans le chapitre 8, que la segmentation en cellules acoustiques élémentaires gagnerait à profiter d'une information a priori.

Idéalement, l'information apportée au système acoustique permettrait de fixer les limites temporelles de chacune des cellules acoustiques composant un mot de passe. En pratique, l'extraction des points de synchronisation ne permet pas de garantir ce résultat. Dans le cas où le flux de synchronisation est extrait du flux vidéo, par exemple, le taux d'échantillonnage du flux vidéo est inférieur à celui de la modalité audio. De plus, nous avons vu dans la section 4.2.2, que la correspondance entre entités acoustiques et entités visuelles est rarement vérifiée.
Nous émettons l'hypothèse que les points de synchronisation correspondent, non pas aux débuts et fins de chaque cellule acoustique, mais à des points caractéristiques de la structure temporelle du mot de passe.
Le processus d'apprentissage des modèles SCHMMs est adapté par la méthode itérative décrite dans la section 8.2. Comme expliqué dans la partie 8.2, les modèles SCHMMs utilisés ont une structure *Gauche-Droite*.

Initialisation

La première version du modèle SCHMM d'un mot de passe est obtenue par une adaptation des états qui utilise une segmentation issue de la contrainte temporelle. Cette procédure d'initialisation est présentée en trois étapes sur la figure 9.16 :

Étape 1 la contrainte temporelle permet une première segmentation de la séquence acoustique d'un mot de passe ;

Étape 2 les états du modèle SCHMM de ce mot de passe sont répartis entre chacun des segments, selon la durée de ces derniers. Chaque segment initial est alors lui même découpé en sous-segments de même longueur, permettant l'adaptation des états du modèle SCHMM ;

Étape 3 une fois apprise les probabilités d'émission de chaque état, nous fixons les valeurs des transitions entre états. À partir de la solution précédemment utilisée (transitions équiprobables), nous distinguons deux types de transitions, représentés à l'étape 3 de la figure 9.16.

Les transitions de type *A*, dont la probabilité, fixée ci-dessus ne subit aucune modification liée à l'usage du flux vidéo ;

Les transitions de type *V*, dont la probabilité est déterminée directement par le flux vidéo.

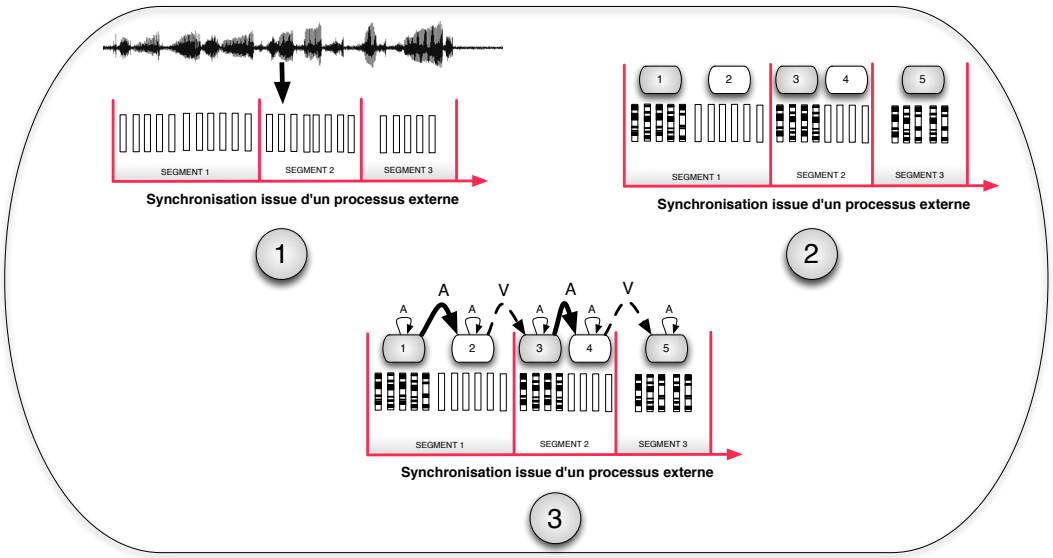


FIG. *9.1:* *Représentation des trois étapes principales du processus d'initialisation contraint d'un modèle SCHMM.*

Une première version du modèle SCHMM est ainsi obtenue.

Itérations

Chaque itération débute par un décodage Viterbi contraint par la synchronisation vidéo de la séquence d'apprentissage avec le modèle obtenu à l'itération précédente. Il s'agit d'un décodage Viterbi classique pour lequel la valeur des transitions de type *V* du modèle SCHMM varie au cours du temps. Ainsi ces transitions, correspondant aux points de synchronisation vidéo, sont presque toujours interdites (elles ont pour

valeur 0). Elles ne sont autorisées qu'à l'instant qui correspond à leur point de synchronisation. Dans la pratique, la transition correspondant au point de synchronisation au temps τ reste activée durant l'intervalle de temps $[\tau - \Delta; \tau + \Delta]$, où Δ permet de pallier l'asynchronie éventuelle des deux sources.
Cet alignement fournit une nouvelle segmentation. Chacun des segments de cette nouvelle segmentation est utilisé pour ré-apprendre un état du modèle SCHMM.

9.1.2 Effet d'une contrainte temporelle en phase de test

En plus du gain escompté, grâce au processus d'apprentissage contraint par le signal de synchronisation, nous proposons d'utiliser la contrainte de synchronisation durant la phase de test. La synchronisation imposée est cette fois utilisée pour dégrader les scores obtenus lors de la comparaison d'un modèle de mot de passe avec une séquence de structure temporelle différente.

Dans la section 8.1.1, nous avons présenté l'algorithme de Viterbi. Nous avons souligné son caractère optimal : l'alignement obtenu par cette méthode maximise la vraisemblance cumulée sur la séquence de paramètres. Dans notre cas, en phase de test, l'algorithme de Viterbi maximise le score de chaque test effectué en calculant l'alignement optimal entre les données d'entrée et le modèle de Markov. Or, si la maximisation du score d'un test client est bénéfique, maximiser le score des tests imposteurs peut engendrer une augmentation du nombre d'erreurs commises pour la tâche de vérification.
Nous proposons d'utiliser la contrainte temporelle comme un a priori sur le type de test effectué. Cette contrainte a pour but d'obtenir un score moins élevé dans le cas où la structure des séquences de test diffère de celle du modèle testé. La suite de cette partie présente les effets auxquels nous souhaitons parvenir grâce à l'usage de la synchronisation contrainte.

La figure 9.2 illustre les alignements obtenus sans contrainte, sur un même modèle de mot de passe, pour trois séquences :

- la séquence qui a permis l'apprentissage de ce modèle ;
- une séquence de test du client prononçant son mot de passe, donc proche de la séquence d'apprentissage ;
- la séquence d'un imposteur prononçant un contenu lexical différent de celui du mot de passe auquel il est comparé.

L'alignement des séquences consiste à déterminer le meilleur chemin à travers le graphe représenté sur cette figure. Sans contrainte extérieure, l'algorithme de Viterbi garanti la maximisation du score de chacune des séquences testées.

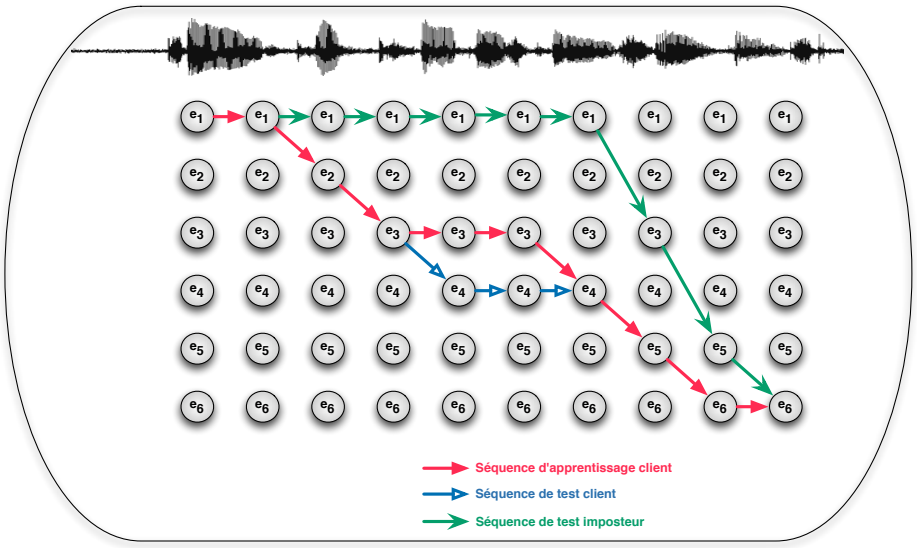


FIG. 9.2: *Illustration de l'alignement de différentes séquences pour un même modèle de mot de passe dans le cadre d'un décodage de Viterbi (optimal).*

La contrainte extérieure est illustrée par la figure 9.17. Cette contrainte interdit, au chemin calculé selon l'algorithme de Viterbi, tout passage dans les zones où les états sont noircis. Les zones autorisées correspondent à l'intersection de la synchronisations imposée lors de l'apprentissage du modèle avec la synchronisation imposée par la séquence de test.

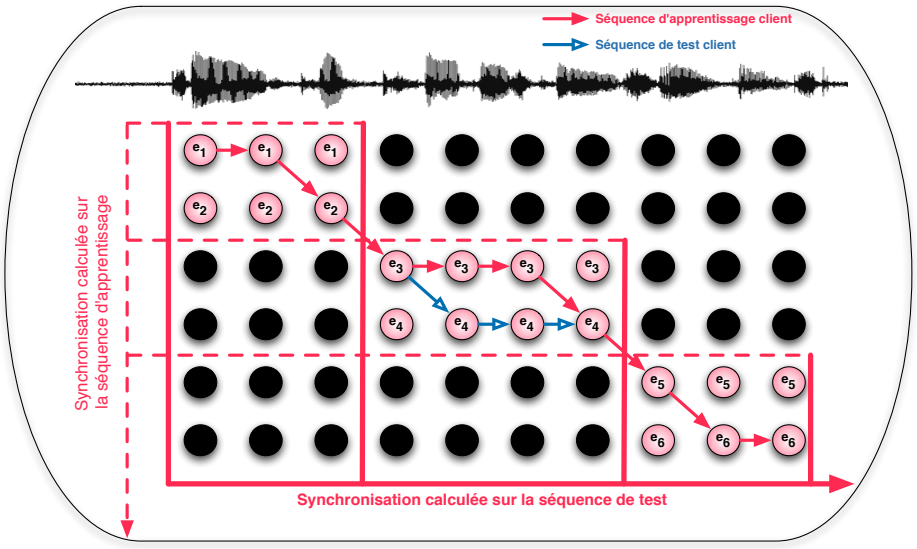


FIG. 9.3: *Illustration de l'alignement de la séquences d'apprentissage du mot de passe et d'une séquence de test client pour le modèle de mot de passe correspondant, dans le cadre d'un décodage de Viterbi contraint par une synchronisation externe*

Nous observons sur la figure 9.17 que l'alignement obtenu pour la séquence de test client n'est pas perturbé par la synchronisation contrainte. Le score calculé est donc le même que celui calculé sans cette contrainte, c'est le score maximisé par le décodage de Viterbi.

La figure 9.18 présente les alignements obtenus pour un test imposteur, avec et sans la synchronisation contrainte. L'alignement obtenu sans synchronisation contrainte (identique à celui de la figure 9.2) traverse des zones interdites par la synchronisation extérieure. Cette contrainte impose alors de passer par un autre chemin qui, cette fois, est confiné aux parties du graphe autorisées. Le chemin obtenu sous la contrainte n'est plus celui du Viterbi original. Le score du test imposteur n'est plus maximal.

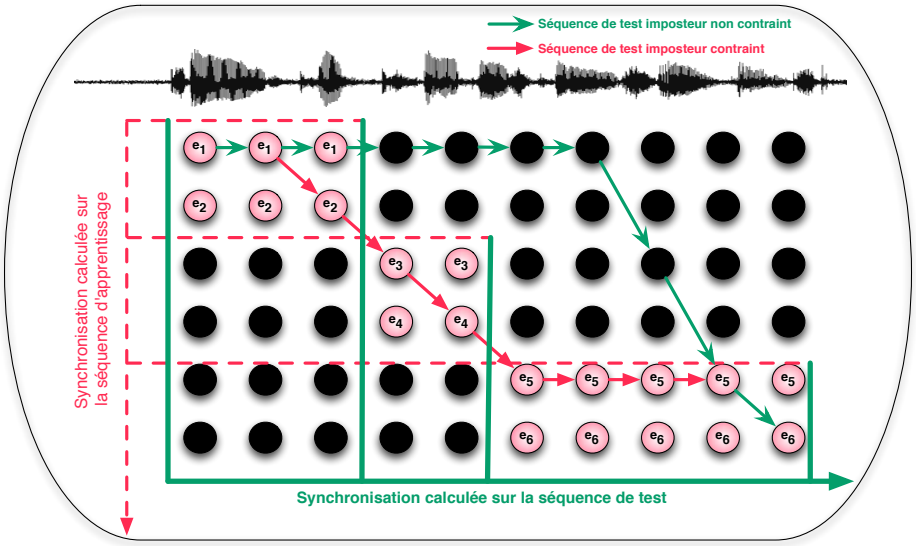


FIG. 9.4: *Illustration de l'alignement d'une séquence de test imposteur pour un modèle de mot de passe, dans le cadre d'un décodage de Viterbi contraint ou non par une synchronisation externe,*

9.2 Validation expérimentale avec un alignement phonétique

Afin de valider les hypothèses formulées dans la section précédente (cf. section 9.1), nous utilisons, dans un premier temps, une synchronisation issue d'un alignement phonétique automatique (Bürki et al., 2008) dont nous extrayons des points de synchronisation forts, correspondant à des frontières inter-mots. De cette façon, nous sommes assurés de la cohérence des deux flux de données.

9.2.1 Configuration de test et choix de l'alignement phonétique

L'alignement phonétique permet, connaissant le texte prononcé par les locuteurs, d'aligner temporellement la séquence phonétique correspondante avec le signal audio. Dans notre cas, ce procédé nous permet de déterminer 5 segments pour chaque mot de passe. Le nombre de segments phonétiques utilisés a été choisi afin de se rapprocher des conditions qui nous paraissent envisageables lors du remplacement de l'alignement phonétique par une source externe, comme la vidéo.
L'alignement phonétique pourrait permettre une segmentation en phonèmes, qui paraissent les cellules acoustiques les plus pertinentes, au vue de la littérature du domaine de la RAP. Mais les travaux présentés dans la partie 4.2.2 montrent que la vidéo ne permet pas une telle segmentation. Il nous paraît donc pertinent de considérer que l'information externe ne délimite qu'un nombre restreint de segments pour une séquence de parole de longueur comprise entre 2 et 3 secondes.
Nous avons choisi un nombre de segments phonétiques indépendant du contenu linguistique des mots de passe afin de restreindre la variabilité inter mots de passe et de ne pas sur-estimer les possibilités de notre approche.
La segmentation phonétique utilisée est décrite dans l'annexe A.

Les expériences sont réalisées avec des modèles GMMs à 256 distributions, les 256 vecteurs moyens des modèle de locuteurs indépendants du texte sont adaptés selon le critère MAP. Les modèles SCHMMs sont initialisés avec 20 états, dont les 256 paramètres de poids sont adaptés selon le critère MAP contraint. Le paramètre Δ (cf. section 9.1.1), qui détermine la *liberté* du système, a pour valeur $10ms$. Cette valeur a été fixée d'après les travaux d'Eveno et Besacier (2005).

9.2.2 Influence d'une contrainte synchrone sur les performances de notre approche

Une première expérience est réalisée afin de déterminer l'effet de la synchronisation acoustique sur les performances de notre système de vérification du locuteur. Nous comparons les performances obtenues en utilisant notre architecture avec ou sans contrainte externe. Les expériences sont réalisées dans les trois conditions **TOUS**, **MDP** et **FAUX**.

Afin de vérifier l'effet de la contrainte structurelle, une expérience supplémentaire est réalisée. Pour cette expérience, la phase d'apprentissage est inchangée. Le protocole de test des trois conditions : **MDP**, **FAUX** et **TOUS**, est modifié. Pour chaque test - client ou imposteur - la séquence audio testée est associée aléatoirement à une synchronisation provenant de l'alignement phonétique d'un imposteur prononçant une phrase différente. Dans ces conditions, la contrainte temporelle devrait dégrader les résultats.

Le tableau 9.11 présente les résultats comparés à ceux du système GMM/UBM de

	Configurations			
	GMM-UBM	Aucune segmentation	Segmentation phonétique	Segmentation aléatoire
TOUS	3,22	3,17	3,17	3,78
MDP	4,00	4,62	4,56	5,56
FAUX	2,46	0,94	0,62	1,38

TAB. *9.1:* *Incidence d'une contrainte externe provenant d'un alignement phonétique sur les performances en vérification du locuteur de l'approche structurale. Résultats exprimés en terme de taux d'égales erreurs (EER).*

référence.
Lorsque les imposteurs prononcent le mot de passe des clients (condition **MDP**) l'ajout de la contrainte temporelle externe n'influe pas sur le taux d'égales erreurs. En revanche, dans le cas où les imposteurs ne connaissent pas le mot de passe d'un client, l'ajout d'une contrainte structurelle permet de réduire l'EER de 34% relatif (de $0,94\%$ à $0,62\%$). Notons également que le taux d'erreurs obtenu est relativement faible dans cette condition (-85% relatif par rapport à la condition **MDP** et -80% par rapport à la condition **TOUS**).

La baisse de performances (cf. tableau 9.11) entraînée par l'utilisation d'une synchronisation aléatoire lors de la phase de test montre que le gain apporté par la synchronisation externe est lié à la corrélation des différentes informations.

Fusion des scores dépendants et indépendants de la structure

Nous avons montré (cf. section 8.4) que la fusion des scores des approches structurale et non-structurale permet, lors des expériences réalisées, de toujours égaler ou surpasser les performances du système GMM/UBM de référence. L'ajout d'une information temporelle, qui contraint les modèles SCHMMs dépendants du texte, devrait apporter une information absente du processus de vérification du locuteur indépendant du texte. Une complémentarité des informations fournies par les deux scores, devrait ainsi bénéficier à la fusion.
Le tableau 9.12 montre les taux d'égales erreurs obtenus en opérant une fusion (décrite dans la section 8.4) entre les scores présentés dans le tableau 9.11 et les scores indépendant de la structure temporelle.

La fusion, par somme pondérée, des scores dépendants et indépendants du texte, n'apporte qu'un gain marginal. Plusieurs raisons peuvent expliquer cela. Tout d'abord le mode de fusion. La somme pondérée ne prend pas en compte la présence supposée d'une information différente qui serait introduite par la contrainte externe. Mais ces résultats peuvent être dus au fait que les informations sont redondantes. En effet, la synchronisation provient d'un alignement phonétique. Il est possible que l'information

	Configurations		
	GMM-UBM	Aucune segmentation	Segmentation phonétique
TOUS	3,22	2,83	2,83
MDP	4,00	4,06	4,07
FAUX	2,46	1,11	0,89

TAB. 9.2: *Incidence d'une contrainte externe provenant d'un alignement phonétique sur les performances en vérification du locuteur d'un système résultant de la fusion des approches structurale et non-structurales. Résultats exprimés en terme de taux d'égales erreurs (EER).*

introduite soit trop corrélée à celle qui est portée par la séquence acoustique.

9.2.3 Vérification des hypothèses

La modification des processus d'entraînement des modèles et de tests, par l'ajout d'une contrainte structurelle au sein du décodage de Viterbi, a répondu aux attentes qui étaient les nôtres. Elle permet une relative amélioration des performances en vérification d'identité, dans le cas où les imposteurs ne connaissent pas le mot de passe des clients.
Les résultats présentés jusque là n'ont cependant pas permis de valider ou d'invalider les deux hypothèses qui ont motivé notre démarche (cf. sections 9.1.1 et 9.1.2). Nous proposons maintenant une analyse détaillée de ces deux hypothèses, à la lumière des résultats obtenus.

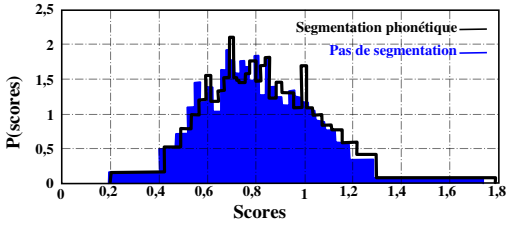
La contrainte externe améliore l'apprentissage des modèles

L'amélioration de l'apprentissage des modèles de mots de passe devrait rapprocher les modèles acoustiques appris des données à modéliser et donc augmenter la valeur des scores clients.

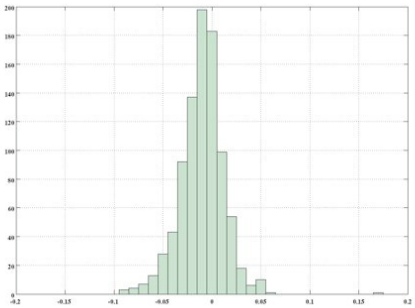
La figure 9.19(a), qui illustre l'évolution des scores clients, lorsque la contrainte temporelle est ajoutée, ne laisse pas apparaître de différence entre les distributions des scores avec et sans synchronisation externe. Cette impression est renforcée par la lecture de la figure 9.19(b) qui présente la distribution des différences entre les scores calculés avec et sans contrainte temporelle. L'évolution des scores des tests client ne permet aucune conclusion quant à la pertinence de la segmentation obtenue pour l'apprentissage des états du modèle SCHMM.

La contrainte temporelle permet de réduire les scores imposteurs

La contrainte temporelle imposée au cours du décodage Viterbi est supposée, lors de la phase de test, dégrader les scores des tests imposteurs (particulièrement lorsque



(a) Évolution des scores clients

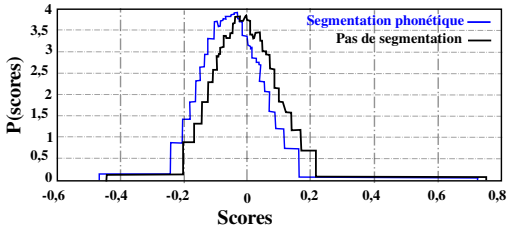


(b) Différence des scores clients

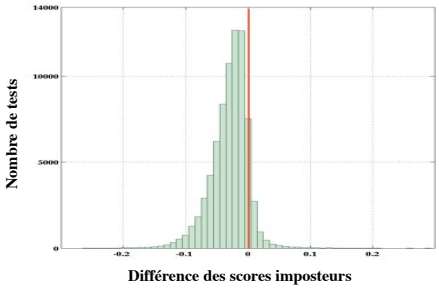
FIG. 9.5: *Évolution des scores clients, dépendants de la structure sous l'effet d'une contrainte temporelle imposée lors du décodage de Viterbi.*
La figure (a) montre les distributions des scores obtenus avec ou sans la synchronisation externe.
La figure (b) présente la distribution des différences entre les scores obtenus avec une synchronisation externe et ceux obtenus sans contrainte.

le mot de passe prononcé n'est pas celui du client).

Condition *FAUX* les figures 9.20(a) et 9.20(b) représentent les distributions des scores imposteurs dépendants de la structure temporelle, obtenus dans la condition **FAUX**, avec et sans contrainte externe. Sur la figure 9.20(a), nous observons que la distribution



(a) Évolution des scores imposteurs



(b) Différence des scores imposteurs

FIG. 9.6: *Évolution des scores imposteurs dépendants de la structure, dans la condition **FAUX**, sous l'effet d'une contrainte temporelle imposée lors du décodage de Viterbi.*
La figure (a) montre les distributions des scores obtenus avec ou sans la synchronisation externe.
La figure (b) présente la distribution des différences entre les scores obtenus avec une synchronisation externe et ceux obtenus sans contrainte.

des scores imposteurs se décale vers la gauche sous l'effet de la synchronisation externe.

Ceci laisse supposer que la synchronisation externe permet de dégrader les scores imposteurs en rendant l'algorithme de Viterbi sous-optimal dans ce cas. Ce constat est confirmé par la figure 9.20(b), puisque cette figure montre que les différences calculées entre les scores résultant d'un algorithme de Viterbi contraint et ceux résultant d'un algorithme de Viterbi non contraint sont majoritairement négatives. La contrainte imposée lors du décodage de Viterbi dégrade donc le score imposteur dans une grande majorité des cas.

Condition *MDP* : la même analyse dans la condition **MDP**, pour laquelle les imposteurs prononcent le mot de passe du client, montre que la synchronisation par l'alignement phonétique affecte beaucoup moins la distribution des scores imposteurs (cf. figure 9.7).

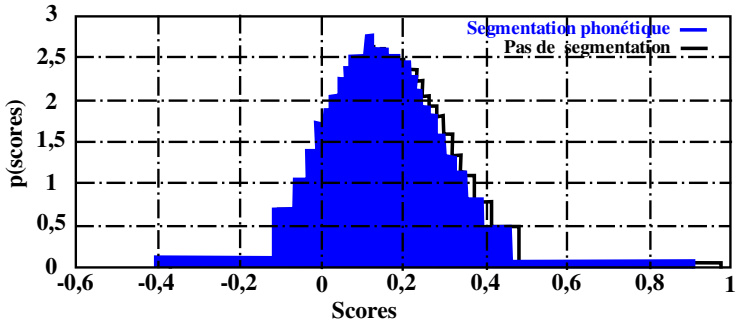


FIG. *9.7:* *Évolution des scores imposteurs dépendants de la structure, dans la condition **MDP**, sous l'effet d'une contrainte temporelle imposée lors du décodage de Viterbi.*

Nous observons toutefois que la distribution des scores imposteurs calculés sous la contrainte phonétique est décalée très légèrement vers la gauche. Nous tirons trois conclusions de cette observation :

- la dégradation des scores moins importante pour les tests imposteurs où ceux-ci prononcent le bon mot de passe montre que leur structure temporelle est plus proche de celle des données d'apprentissage ;
- nous pensons qu'il existe, pour un même énoncé, une variabilité inter-locuteurs liée à la façon de le prononcer. Le léger décalage observé entre les distributions pourrait illustrer cette variabilité. Néanmoins, ce décalage n'est pas significatif compte tenu du protocole expérimental et de la base de données utilisée ;
- la contrainte de synchronisation, contient une information dépendante de l'énoncé, mais ne semble pas assez précise pour apporter une information dépendante du locuteur.

9.2.4 Retour sur la place des « silences »

La question de l'importance des silences, pauses ou plus généralement des trames acoustiques étiquetées *non-parole* (cf. section 3.1.4), a déjà été posée dans le chapitre précédent (cf. section 8.3.1). Nous avons montré que supprimer la détection d'activité dégrade fortement les performances de notre système structural.
Les trames étiquetées *non-parole* comprennent, bien évidemment, les trames correspondant aux interruptions du signal de parole, mais peuvent également être des trames correspondant à certaines consonnes. En effet, la partie plosive des consonnes occlusives (/p/, /t/ ou /k/) possède un niveau d'énergie assez bas et peut être étiquetée *non-parole* par des systèmes n'utilisant que l'énergie en guise de VAD, comme celui que nous utilisons.
Il n'est donc pas exclu que certaines trames *non-parole* contiennent de l'information et nous proposons ici de supprimer le module de détection *parole/non-parole* de notre système, afin de tester la capacité de notre approche à exploiter cette information.

Les expériences réalisées sont identiques à celles de la section précédente à l'exception de l'étape de détection *parole/non-parole* qui est supprimée (les trames *non-parole sont utilisées pour estimer les probabilités d'émission des états*). Les résultats obtenus par notre approche structurale, avec et sans synchronisation externe, sont présentés dans le tableau 9.3

	Configurations		
	GMM-UBM	Aucune segmentation	Segmentation phonétique
TOUS	3,22	3,89	3,23
MDP	4,00	5,67	4,89
FAUX	2,46	1,40	0,83

TAB. 9.3: *Incidence d'une contrainte externe provenant d'un alignement phonétique sur les performances en vérification du locuteur de l'approche structurale sans détection d'activité. Résultats exprimés en terme de taux d'égales erreurs (EER).*

L'apport de la synchronisation externe s'avère assez surprenant. La colonne centrale du tableau reprend les résultats de la partie 8.3.1. La suppression de la détection *parole/non-parole* au sein de l'approche structurale augmente considérablement les taux d'erreurs dans toutes les conditions de tests. L'introduction de la synchronisation issue de l'alignement phonétique améliore globalement les performances de ce système, jusqu'à obtenir des résultats comparables au système GMM/UBM dans la condition **TOUS** et même des résultats comparables à ceux de nos meilleures approches dans la condition **FAUX**. Le taux d'erreur obtenu pour cette condition est comparable à celui obtenu pour une fusion de l'approche structurale incluant une synchronisation externe et une détection *parole/non-parole* avec une approche non-structurale (cf. section 9.2.2). Cependant, si les imposteurs connaissent le mot de passe des client (condition **MDP**), le taux d'égales erreurs reste très élevé. Le tableau 9.4 présente les résultats obtenus

pour les mêmes expériences après fusion des deux scores (dépendant et indépendant du texte) calculés par notre système.

	Configurations		
	GMM-UBM	Aucune segmentation	Segmentation phonétique
TOUS	3,22	3,33	2,99
MDP	4,00	4,50	4,22
FAUX	2,46	1,44	1,06

TAB. 9.4: *Incidence d'une contrainte externe provenant d'un alignement phonétique sur les performances en vérification du locuteur d'un système résultant de la fusion des approches structurale et non-structurale sans détection d'activité. Résultats exprimés en terme de taux d'égales erreurs (EER).*

Comme précédemment, cette configuration de notre système obtient de bons résultats dans les conditions **TOUS** et **FAUX**. Les résultats présentés semblent indiquer que le système ne disposant pas du module de détection *parole/non-parole* est plus à même de modéliser la structure temporelle des séquences acoustiques. Ceci explique les faibles taux d'erreurs observés dans la condition **FAUX**, pour laquelle la structure temporelle des séquences de tests est différente de celle du mot de passe client. Cette conclusion permet aussi d'expliquer les mauvaises performances obtenues lorsque la structure de la séquence de test est proche de celle du mot de passe du locuteur (condition **MDP**).

La base de données MyIdea, à partir de laquelle ont été validés nos travaux, a été enregistrée dans des conditions de studio qui se traduisent, d'un point de vue audio, par l'absence de bruit extérieur. L'influence du module de détection *parole/non-parole* dans cet environnement n'est certainement pas aussi importante que dans un environnement bruyant. La vérification des comportements observés pour la base MyIdea nécessiterait des expériences réalisées en milieu bruyant. Encore une fois, les données d'évaluation nous font défaut. Nous retiendrons tout de même les bonnes performances de notre approche contrainte par un alignement phonétique en l'absence d'un module de détection *parole/non-parole* .

9.3 Retour sur la structuration temporelle des vidéo

Une étude des traitements vidéo existants dans l'état-de-l'art a été présentée dans le chapitre 4. Il apparaît que deux techniques majeures permettent de caractériser la structure temporelle du flux vidéo. Cette section est consacrée à une analyse de ces méthodes dans le contexte qui est le nôtre.

9.3.1 Approches Markoviennes

Les modèles Markoviens ont été présentés de façon détaillée dans la section 8.1.1. Leur capacité à modéliser la structure temporelle du signal audio de parole a été démontrée dans ce même chapitre. Différentes utilisations de ces modèles peuvent permettre de caractériser la structure temporelle issue d'un signal vidéo. Le formalisme des visèmes (cf. section 4.2.2), par exemple, est particulièrement adapté à la tâche qui nous préoccupe.

Quelle que soit l'approche retenue, les modèles HMMs nécessitent une paramétrisation plus ou moins complexe du signal de parole vidéo. Cette paramétrisation, même si elle peut prendre de nombreuses formes (ACP, LDA calcul de coefficients DCT, flot optique, extraction de paramètres dynamiques) , implique un traitement coûteux du flux vidéo, qui est ajouté au coût de la modélisation Markovienne. Ce coût ne nous paraît pas respecter les contraintes fixées dans nos travaux.

9.3.2 Approches morphologiques

Les approches morphologiques permettent d'extraire une information dont la corrélation avec le flux audio a été prouvée (Eveno et Besacier, 2005), (Chetty et Wagner, 2004a), (Goecke, 2005). Ces méthodes mesurent, sur le visage, des distances entre points caractéristiques. C'est l'évolution de ces distances au cours du temps qui caractérise le mouvement des articulateurs.
La méthode la plus courante consiste à détecter et suivre le contour de la bouche afin d'en mesurer l'ouverture et éventuellement d'autres caractéristiques (Goecke et al., 2000), (Lievin et Luthon, 1998).
Ces algorithmes, qui arrivent à maturité, nécessitent un processus complexe et très coûteux, puisqu'il s'agit d'abord de détecter les points caractéristiques avant de les suivre (Eveno et al., 2004) (Wiskott et al., 1997) à travers la séquence vidéo.
La chaîne de traitements nécessaire à ces opérations ne correspond pas aux contraintes calculatoires imposées par le contexte de nos travaux.

9.4 Calcul d'une synchronisation vidéo dans le cadre de nos contraintes

Les variations rapides du signal vidéo, ajoutées au faible taux d'échantillonnage des signaux, rendent difficile une caractérisation simple du mouvement des articulateurs. Les mouvements de tête des locuteurs, dans le champ de cadrage de la vidéo, compliquent encore la tâche et justifient la réalisation d'un suivi de visage au cours du temps. Nous avons cependant choisi de rechercher une information dans l'image entière, sans pratiquer de détection, de manière à limiter le coût calculatoire. Nous sommes toutefois conscients des difficultés qui peuvent apparaître dans le cas où le

fond visuel n'est plus un plan fixe.

Notre méthode de caractérisation du mouvement lié à la production de parole repose sur l'hypothèse selon laquelle les variations vidéo, dues au mouvement des articulateurs, ont une fréquence plus élevée que les variations dues au mouvement de la tête. Nous supposons ainsi que les variations vidéo entre deux images, dues au mouvement de la tête, sont négligeables devant les variations entraînées par le mouvement des articulateurs.

La quantité de mouvement présente dans le flux vidéo est estimée par une simple différence entre images successives. Le flux vidéo est tout d'abord soumis à un changement d'espace colorimétrique pour obtenir une représentation YUV. Une soustraction des composantes Y est effectuée au sein de chaque couple d'images successives. De cette soustraction résulte pour chaque couple d'images successives une *image différence*. La somme des valeurs absolues de la composante Y de chaque pixel de cette *image différence* fournit une valeur caractérisant la quantité de mouvement présente entre ces deux images. La succession de ces valeurs calculées sur l'ensemble de la vidéo forme un signal temporel discret, lié au mouvement présent dans la vidéo. Ce processus est décrit par l'équation 9.1 :

$$S(n) = \sum_{l=0}^{L} \sum_{h=0}^{H} |I^{n}_{(l,h)} - I^{n+1}_{(l,h)}| \tag{9.1}$$

où L et H sont respectivement la largeur et la hauteur des images, I^n et I^{n+1} sont deux images consécutives et $I^n_{(l,h)}$ est la composante Y du pixel (w, h) de l'image I^n. S est le signal temporel discret qui résulte de ce processus.

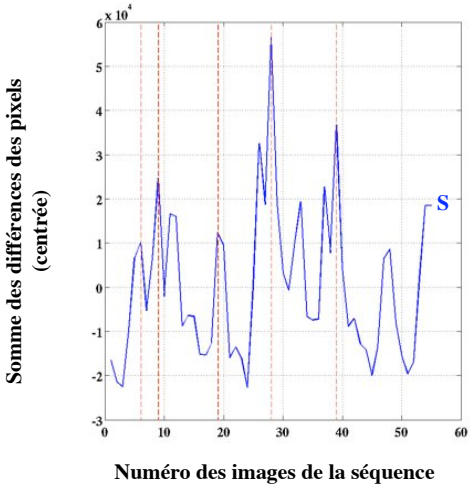


FIG. 9.8: *Somme des différences entre les pixels de deux images consécutives de la séquence vidéo.*

Les maxima locaux du signal S sont détectés par application d'une fenêtre glissante.

Les coordonnées temporelles de ces maxima seront utilisées comme points de synchronisation dans le décodage Viterbi. La figure 9.21 représente le signal S calculé pour l'une des séquences vidéo de la base de données MyIdea. Les segments verticaux superposés à ce signal correspondent aux points de synchronisation calculés.

9.5 Validation expérimentale

Une première expérience, identique à celle présentée dans la section 9.2.2, permet d'évaluer les performances obtenues lorsque l'apprentissage et le test sont contraints par la synchronisation vidéo présentée précédemment. Le tableau 9.5 présente les résultats obtenus, ainsi que ceux des systèmes de référence, auxquelles ils doivent être comparés : GMM/UBM de référence, approche structurale non-contrainte et approche structurale contrainte par un alignement phonétique.

	Configurations			
	GMM-UBM	Aucune segmentation	Segmentation phonétique	Segmentation vidéo
TOUS	3,22	3,17	3,17	3,77
MDP	4,00	4,62	4,56	5,11
FAUX	2,46	0,94	0,62	1,28

TAB. 9.5: *Incidence d'une contrainte externe provenant d'un alignement vidéo sur les performances en vérification du locuteur de l'approche structurale. Résultats exprimés en terme de taux d'égales erreurs (EER).*

Les taux d'égales erreurs qui résultent de cette expérience sont nettement supérieurs à ceux des trois systèmes de référence dans toutes les conditions de tests. La fusion des scores dépendants et indépendants du texte ne modifie pas cet état de fait puisque même dans la condition **FAUX**, le taux d'égale erreurs reste supérieur de $+36\%$ relatif à celui obtenu pour un système structural non synchronisé.

Afin de déterminer si la synchronisation issue du flux vidéo contient une information relative à la structure du signal de parole, nous réalisons l'expérience suivante :

- l'apprentissage des modèles de locuteur dépendants du texte synchronisé par la segmentation vidéo n'est pas modifié ;
- lors de la phase de test en revanche, le décodage des séquences de tests audio (clients et imposteurs) est synchronisé par une information vidéo tirée aléatoirement parmi celles des autres locuteurs de la base prononçant un énoncé différent.

Les résultats de cette expérience sont présentés dans le tableau 9.13.

Les taux d'erreurs résultant de la synchronisation aléatoire de notre système sont inférieurs à ceux obtenus par une synchronisation utilisant le flux vidéo correspondant

	Configurations		
	GMM-UBM	Segmentation vidéo	Segmentation vidéo aléatoire
TOUS	3,22	3,77	3,56
MDP	4,00	5,11	4,94
FAUX	2,46	1,28	1,06

TAB. 9.6: *Incidence d'une contrainte externe provenant d'un alignement vidéo aléatoire sur les performances en vérification du locuteur de l'approche structurale. Résultats exprimés en terme de taux d'égales erreurs (EER)*

au flux audio considéré. Ce constat peut avoir deux explications :

- soit le signal temporel représentant la différence de la somme des pixels entre images ne contient pas d'information ;
- soit ce même signal est trop bruité pour permettre d'obtenir cette information de façon aussi simple.

Un test de corrélation COIA, réalisé entre les signaux temporels vidéo et les vecteurs acoustiques du signal de parole n'a d'ailleurs montré aucune corrélation entre ces signaux. Cette expérience montre que notre synchronisation, calculée à partir du flux vidéo, ne contient pas d'information caractéristique du signal de parole.

Conclusion

Nous avons présenté dans ce chapitre une nouvelle méthode permettant d'intégrer une information temporelle au sein d'un système acoustique. Cette approche repose sur la synchronisation du processus acoustique par une information provenant d'un processus externe. Ce procédé a été développé afin d'améliorer la modélisation des cellules acoustiques composant les mots de passe mais surtout de permettre une meilleure discrimination des locuteurs ne prononçant pas le mot de passe choisi par le client.

Notre méthode a été validée par une série d'expériences utilisant une synchronisation provenant d'un alignement phonétique. Les résultats que nous avons présenté sont prometteurs. Ils montrent que si la modélisation des cellules acoustiques ne bénéficie pas de la contrainte temporelle ajoutée, cette contrainte permet de réduire le taux d'égales erreurs d'environ 20% dans le cas où les imposteurs ignorent le mot de passe des clients tout en conservant des performances équivalentes au système GMM/UBM lorsque les imposteurs ont connaissance du mot de passe des clients.

La tentative d'extraction d'information à partir du flux vidéo, que nous avons présentée dans ce chapitre, nous amène à conclure qu'il est difficile d'extraire une information fiable du flux vidéo, sans recourir à des méthodes complexes. Ce résultat est d'ailleurs en accord avec l'analyse des méthodes états-de-l'art présentée en début de

chapitre. Cette difficulté s'explique par un bruit non négligeable dans le signal vidéo. Ce bruit peut provenir des mouvements du visage du locuteur dans le champ de la caméra et de la rapidité de ces mouvements, mais également des conditions d'illumination du sujet. Même si les enregistrements ont été effectués sous un éclairage contrôlé, ce problème ne peut être complètement écarté.

Ainsi la dégradation des résultats, observée lors de l'ajout d'une information issue de la vidéo, s'explique, selon nous, par la présence de bruit qui occulte l'information utile. Nous continuons cependant à penser qu'il est possible d'extraire une information pertinente à partir du flux vidéo et qu'une telle information pourrait améliorer les performances observées dans le chapitre 9. Nous pensons en effet que la vidéo peut amener, dans le décodage acoustique, une information différente de celle du flux audio. L'extraction d'une telle information est cependant peu compatible avec les contraintes applicatives qui ont été définies pour nos travaux et nécessiterait éventuellement de redéfinir un contexte adapté. L'utilisation d'une segmentation provenant de la modalité vidéo offre l'avantage lutter contre les play-backs (Eveno et Besacier, 2005).

Conclusions et perspectives

CETTE thèse s'inscrit dans une volonté d'améliorer les performances des systèmes d'authentification biométrique embarqués et, plus particulièrement, les systèmes de reconnaissance du locuteur. L'approche proposée s'accommode des ressources limitées disponibles dans ce contexte applicatif ainsi que des contraintes ergonomiques imposées. Les données disponibles, pour l'enrôlement des utilisateurs, sont ainsi très limitées et la durée des séquences de test n'excède pas quelques secondes. Aussi, nous avons proposé de compenser le manque de données d'apprentissage ainsi que la courte durée des séquences de tests par :

- l'utilisation de mots de passe personnels, dont la structure temporelle est utilisée, en plus de l'information acoustique, pour assurer un haut niveau de performance. Cette structure est incorporée à une représentation hiérarchique des modèles acoustiques permettant de minimiser les besoins en termes de données d'apprentissage comme en termes de ressources calculatoires ;
- l'ajout, au sein du système acoustique, d'une information provenant d'un processus externe, qui renforce la structure temporelle des modèles de mots de passe.

Influence des mots de passe sur l'approche GMM/UBM

Les différentes tâches utilisant le signal de parole, décrites dans la partie II, montrent que ce signal contient une information structurale qui est souvent exploitée en reconnaissance de la parole mais qui peut également être utile en reconnaissance automatique du locuteur.

L'approche GMM/UBM, état-de-l'art, est à la base de nos travaux. Nous l'avons également utilisée comme système de référence tout au long de cette thèse. Pour cela, nous avons mené une étude approfondie des comportements de l'approche GMM/UBM dans le contexte de la reconnaissance du locuteur et plus particulièrement lorsque les ressources et données biométriques sont limitées. Nous avons montré que les performances de ce système sont fortement liées au ratio *quantité de données / nombre de distribution des modèles GMMs*. Le nombre de distributions des modèles GMMs doit être suffisant pour permettre une modélisation précise des caractéristiques des locuteurs. Par exemple, l'augmentation du nombre de distributions, de 16 à 512, permet de faire chuter le taux d'égales erreurs de $7,01\%$ à $3,06\%$ lorsque les modèles sont appris avec une dizaine de secondes de parole. La quantité de données d'apprentissage doit cependant être assez importante pour permettre une bonne estimation des paramètres du modèle. Pour des modèles GMMs à 512 distributions, le taux d'égales erreurs peut décroître de 60% relatifs selon qu'ils sont appris avec 12 ou 14 secondes de parole, passant ainsi de $3,06\%$ à $1,31\%$.

Nous avons également voulu estimer l'influence, sur des systèmes GMM/UBM, d'énoncés contraints syntaxiquement, même si l'approche GMM/UBM ne prend pas intrinsèquement en considération les informations temporelles. Pour cela nous avons évalué leurs performances dans le cadre de la reconnaissance de la parole ainsi qu'en reconnaissance du locuteur dépendante du texte.

Pour une tâche de détection de mot de passe dépendante du locuteur, et non de reconnaissance de locuteur, notre système GMM/UBM de référence obtient un taux d'égales erreurs de 31,4% sur la base de données MyIdea avec 10 énoncés différents. Bien que ce taux de reconnaissance de mots soit nettement inférieur aux approches classiques de reconnaissance de la parole, les résultats démontrent que notre système GMM/UBM est capable de discriminer une dizaine d'énoncés différents.
L'influence du texte sur les performances de l'approche GMM/UBM a été évaluée dans le cadre de la reconnaissance du locuteur dépendante du texte. Pour des modèles appris avec une seule répétition des mots de passe, le taux d'égales erreurs qui est de 3,68% lorsque les imposteurs connaissent les mots de passe des clients, chute jusqu'à 2,11% lorsque les mêmes imposteurs ignorent ces mots de passe. Ces résultats montrent que l'utilisation de mots de passe personnels améliore les performances des systèmes GMM/UBM. L'écart observé entre les deux conditions de test s'accroît lorsque le modèle GMM de locuteur est appris avec deux exemples du mot de passe. Le taux d'égales erreurs passe de 0,56% lorsque les imposteurs ignorent le mot de passe à 2% lorsqu'ils en ont connaissance.
Bien qu'ils n'exploitent pas la structure temporelle des séquences acoustiques, nous avons montré que les systèmes GMM/UBM affichent un comportement différent selon le texte prononcé, dès lors que la phase d'apprentissage est contrainte en termes de contenu lexical.

Mots de passe et information structurale

Nous avons proposé de structurer l'information acoustique pour tirer parti de l'organisation temporelle de mots de passe personnels. Cette information structurelle est exploitée à travers des modèles de Markov semi-continus qui prennent en considération l'information structurelle, tout en respectant les contraintes ergonomiques et matérielles de l'environnement embarqué. Nous avons développé une architecture hiérarchique à trois niveaux qui permet une spécialisation progressive des modèles acoustiques qui la composent. L'organisation de cette architecture est la suivante :

1^{er} **niveau** un modèle du monde modélise l'ensemble de l'espace acoustique ;

2^{e} **niveau** une représentation du locuteur, indépendante du texte, est apprise avec l'ensemble des données disponibles pour ce locuteur ;

3^{e} **niveau** un modèle de Markov semi-continu, dérivé du modèle du deuxième niveau, permet de modéliser deux informations : l'une spécifique au locuteur et l'autre relative à la structure du mot de passe choisi par ce locuteur.

Cette architecture permet d'obtenir, à moindre coût, un modèle intégrant des données spécifiques au locuteur et à la structure temporelle de son mot de passe. Les deux premiers niveaux de notre architecture reprennent le paradigme GMM/UBM qui est à la base des principales méthodes état-de-l'art de ces dix dernières années en reconnaissance du locuteur. La représentation indépendante du texte des locuteurs bénéficie ainsi des avantages des approches génératives, qui intègrent naturellement la variabilité acoustique intra-locuteur. Le modèle dépendant du texte, le troisième niveau

de notre architecture, est directement obtenu en dérivant les densités de probabilité d'émission des états des modèles de Markov semi-continus depuis le GMM indépendant du texte du niveau précédent. Chacun des états du modèle SCHMM (modèle de mot de passe) est composé des distributions Gaussiennes du modèle GMM indépendant du texte, pour lesquelles un nouveau paramètre de poids est appris.

La configuration des modèles SCHMMs doit être définie pour représenter au mieux la structure temporelle des mots de passe à modéliser. Une étude expérimentale nous a permis de déterminer les principales caractéristiques de cette structure : le nombre et la dimension des états du SCHMM, la stratégie et le critère d'adaptation à utiliser pour leur apprentissage ou encore les transitions entre les états.
Cette étude a confirmé nos conclusions quant à l'importance du rapport entre la quantité de données d'apprentissage des modèles GMMs et le nombre de distributions qui les composent. Un nombre trop faible de distributions par état du modèle SCHMM ne permet pas de modéliser l'information acoustique correspondante. L'augmentation du nombre de distributions par état s'accompagne d'une diminution du taux d'erreurs jusqu'à atteindre une configuration optimale. Dans le cas de modèles SCHMMs à 20 états, le taux d'erreurs diminue de $4,51\%$ à $3,17\%$ lorsque le nombre de distributions augmente de 32 à 256. Une fois cette configuration optimale atteinte, l'augmentation du nombre de distributions par état du modèle SCHMM entraîne une dégradation des performances. Le taux d'erreur atteint, dans la même configuration que précédemment, $3,89\%$ lorsque le nombre de distributions augmente jusqu'à 1024. Cette baisse de performances est due à l'augmentation du ratio entre le nombre de paramètres à estimer lors de l'apprentissage des modèles et la quantité de données disponible.
La stratégie que nous avons développée pour obtenir une segmentation en cellules acoustiques pertinente repose sur une approche itérative. Chaque itération de ce processus fournit une segmentation du flux acoustique, utilisée pour l'apprentissage des densité de probabilité d'émission. Un décodage Viterbi permet d'aligner la séquence d'entraînement avec le nouveau modèle et de calculer une nouvelle segmentation. Bien que la convergence de cette stratégie n'ait pas été établie, cette stratégie a démontré expérimentalement son efficacité, notamment par sa cohérence avec les résultats issus d'un étage de VAD.
Le critère d'adaptation des paramètres de poids des états des SCHMMs a été déterminé expérimentalement. Nous avons contraint le critère du maximum a posteriori, utilisé pour l'adaptation des modèles indépendants du texte, afin d'accorder moins d'importance aux données d'apprentissage de ces états du fait de la très faible quantité de données disponible. Cette contrainte permet de réduire le taux d'égales erreurs de 20% par rapport à une adaptation MAP standard.
Une étude portant sur le choix des transitions entre états au sein des modèles SCHMMs ainsi que sur la structure de ces modèles a montré que les modèles SCHMMs *Gauche-Droite* sont les plus adaptés à notre tâche.
L'estimation des probabilités de transition entre états, calculées à partir des alignements des séquences d'entraînement sur les modèles SCHMMs, s'est révélée moins pertinente que l'utilisation de transitions équiprobables. Nous pensons que ce constat est dû au manque d'exemples de mots de passe pour l'apprentissage.

Les résultats de ces expériences ont montré que notre approche, utilisant une modélisation des mots de passe par des SCHMMs, permet une réelle prise en compte de la structure temporelle des mots de passe. Dans le cas où les imposteurs ne connaissent pas le mot de passe des clients, le taux d'égales erreurs obtenu passe de 2,46% pour l'approche GMM/UBM de référence à 0,94% pour notre approche utilisant des modèles SCHMMs. L'information structurelle se combine donc à l'information acoustique pour discriminer les imposteurs des clients. L'un des inconvénients de cette approche apparaît dans le cas où les imposteurs prononcent le mot de passe des clients. Dans cette configuration, l'information structurale exploitée par notre approche semble masquer en partie l'information spécifique au locuteur. Les performances de notre approche sont alors inférieures à celle du système GMM/UBM de référence (taux d'égales erreurs de 4,62% contre 4% pour notre référence).
Dans notre architecture acoustique hiérarchique, la spécialisation progressive des modèles acoustiques nous permet de disposer, à moindre coût, de deux approches, dépendante et indépendante du texte. Le surcoût du calcul de deux scores, dépendant et indépendant du texte, est rendu négligeable par le partage des paramètres des modèles statistiques entre les différentes couches de notre architecture.
La fusion de ces deux scores par une somme pondérée permet à notre approche d'égaler les performances du modèle GMM/UBM dans le cas où les imposteurs connaissent le mot de passe des clients. Alors que notre approche structurale seule obtient dans cette condition un taux d'erreurs de 4,62%, une fusion avec le niveau indépendant du texte permet d'obtenir des performances équivalentes à celles de notre système de référence (4,06% contre 4%).
Les performances obtenues par cette fusion, dans le cas où les imposteurs ne disposent pas du bon mot de passe (1,11%), restent supérieures à l'approche GMM/UBM puisque nous obtenons des taux d'erreurs inférieur de 55% (en relatif) à ceux de notre système de référence (2,46%).

Ajout d'une contrainte temporelle issue d'un processus externe

Le décodage de Viterbi, utilisé pour aligner les séquences de test sur les modèles SCHMMs des mots de passe, maximise la vraisemblance entre le modèle de mot de passe et la séquence de test considérée. Cette maximisation est adaptée dans le cas d'un test client mais nuit aux performances de notre approche dans le cas de tests imposteurs. Souhaitant conserver le score optimal pour les tests clients et dégrader les scores des tests imposteurs, nous avons introduit, au cours du décodage de Viterbi, une information a priori sur la structure du signal acoustique.
Des points de synchronisation forts, calculés lors de la phase d'apprentissage, sont introduis au sein du modèle SCHMM de mot de passe. Le même procédé appliqué durant la phase de test permet de déterminer les points de synchronisation inhérents à cette séquence. Nous imposons alors, durant le décodage de Viterbi, une correspondance entre les points de synchronisation du modèle et ceux issus de la séquence de test.

Cette approche a été validée dans un premier temps grâce à l'utilisation d'une contrainte externe issue d'un alignement phonétique automatique. Ce procédé impose des points de synchronisation forts, correspondant à des frontières inter-mots.
Les tests effectués sur la base de données audio-vidéo MyIdea ont permis de montrer que la mise en place de cette contrainte améliore les performances du système de vérification du locuteur dans le cas où les imposteurs ne connaissent pas le mot de passe des clients. La contrainte temporelle permet de réduire les taux d'égales erreurs obtenus par l'approche GMM/UBM (2,46%) et par notre approche structurale (1,11%) de respectivement 73% et 20% relatifs. Nous avons ainsi montré que ce procédé permet de dégrader les scores des tests imposteurs lorsque ceux-ci ne prononcent pas le bon énoncé et que l'intégration de la contrainte externe améliore la prise en compte de la structure temporelle des mots de passe. Le taux d'erreurs chute ainsi de 0,94% à 0,62% lors de la synchronisation du processus acoustique.
En revanche, cette contrainte n'agit pas dans le cas où les imposteurs prononcent le mot de passe des clients. Comme précédemment, l'information relative à la structure temporelle du mot de passe masque l'information spécifique au locuteur.
La fusion des scores dépendants et indépendants du texte permet, à nouveau, de remédier à cet inconvénient et d'égaler les performances de l'approche GMM/UBM dans le cas où les imposteurs connaissent le mot de passe des clients.

Notre approche audio assistée par une contrainte externe surpasse l'approche GMM/UBM lorsque le contenu lexical prononcé par les imposteurs diffère du mot de passe des clients, tout en offrant des performances équivalentes lorsque les imposteurs connaissent les mots de passe des clients.

L'analyse de l'état-de-l'art que nous avons présentée nous laisse penser que la vidéo offre un fort potentiel pour l'extraction d'une information structurelle corrélée au flux audio. Les points de synchronisation utilisés pour contraindre le décodage acoustique peuvent alors être issus du signal vidéo. Nous avons cependant expliqué qu'il est très difficile d'extraire du flux vidéo une information pertinente permettant une synchronisation efficace du processus acoustique, tout en respectant la limitation des ressources imposée dans le contexte des systèmes embarqués.
Nous avons proposé dans ce document de tirer parti de la quantité de mouvement présente entre les images successives du flux vidéo pour estimer des points de synchronisation, basés sur les maxima de la quantité de mouvement. Une approche globale a été privilégiée afin d'éviter l'usage d'algorithmes de détection et suivi des visages, qui entraîne un surcoût calculatoire conséquent. Nous avons proposé une mesure simple de la quantité de mouvement au sein du flux vidéo, reposant sur la différence des pixels de deux images successives. Ce processus n'a cependant pas permis d'extraire une information pertinente à cause, probablement, d'un faible rapport signal à bruit.

L'usage de la modalité vidéo semble donc difficilement conciliable avec la limitation

des ressources imposée par la contexte embarqué. Cependant, l'usage d'une modalité supplémentaire permettrait d'utiliser les différentes informations structurelles pour déjouer d'éventuelles impostures par play-back. Ce travail ouvre ainsi de nombreuses perspectives relatives à l'utilisation d'information structurelle dans le cadre de la vérification du locuteur et aux approches de reconnaissance du locuteur assistée par la modalité vidéo.

Perspectives

Comme nous l'avons indiqué dans ce document, la base de données MyIdea, utilisée pour la validation de nos travaux reste limitée. Il serait donc pertinent de confirmer l'ensemble de nos résultats sur un corpus plus large et davantage représentatif des conditions embarquées lorsqu'un tel corpus sera disponible.

Améliorations du système proposé

Parmi les modifications qui peuvent être apportées à notre approche, l'apprentissage des modèles acoustiques qui composent les trois niveaux de notre architecture peut être amélioré.

Le modèle du monde (UBM, appris à partir de centaines d'heures de parole) a pris une importance considérable dans les systèmes qui constituent l'état-de-l'art actuel en reconnaissance du locuteur. Son rôle structurant, dont nous avons fait mention au cours de cette étude, soulève de nombreuses questions. Compte tenu des fortes disparités comportementales observées entre les différents modèles et locuteurs, il pourrait être intéressant d'accroître le rôle structurant de ce modèle en tenant compte du type de données utilisées pour son apprentissage. Il nous paraît judicieux d'accorder une importance différente aux données d'apprentissage selon leur provenance et leur spécificité. Ainsi nous proposons de prendre en compte, lors de l'apprentissage du modèle du monde, une information relative au son prononcé afin de s'assurer que le modèle du monde contienne des informations spécifiques à chaque son.
Ce travail peut être étendu pour garantir la qualité des modèles de mots de passe appris pour chaque client. Considérant que les sons modélisés par le modèle du monde ont été répertoriés, de façon explicite ou non, le mot de passe prononcé par un client durant la phase de test est analysé afin de s'assurer que la suite de sons qu'il contient est présente au sein du modèle du monde. Cette précaution permet de garantir la qualité de la modélisation de ce mot de passe. Dans le cas où certains sons sont absents du modèle du monde, il sera conseillé au client de choisir un nouveau mot de passe.

Tout au long de nos travaux, nous avons montré l'importance du ratio *quantité de données d'apprentissage / taille des modèles*. L'architecture acoustique à trois niveaux que nous avons proposée peut être améliorée pour permettre d'introduire la scalabilité des

modèles GMMs de probabilités d'émission, qui composent les nœuds de notre architecture. En effet, la modélisation de l'espace acoustique est d'autant meilleure que le modèle UBM contient de distributions Gaussiennes, alors que nous avons montré que les modèles de locuteur dépendants du texte, et plus encore, les probabilités d'émission des états des modèles SCHMMs, doivent respecter un ratio *quantité de données/nombre de distributions des modèles*.
Le modèle du monde peut être utilisé afin de créer des modèles de locuteur scalables, dont le nombre de distributions Gaussiennes varie selon le contenu lexical utilisé à l'enrôlement. Nous proposons ainsi de ne conserver dans le modèle du locuteur que les distributions présentant une forte vraisemblance avec les données d'apprentissage. L'utilisation de mots de passe ou d'un prompteur permet, durant la phase de test, de s'assurer que le contenu lexical correspond au modèle utilisé. Une telle approche, utilisée dans un contexte embarqué client/serveur, permettrait d'utiliser des modèles de taille réduite lors d'une authentification en ligne, directement sur le système embarqué, et des modèles de taille plus importante sur le serveur, pour un niveau de sécurité plus élevé.

Le pouvoir discriminant de notre approche peut également être accru par l'utilisation de modèles d'états plus discriminants. Considérant que l'information indépendante du texte relative au locuteur est modélisée par le modèle GMM du second niveau de notre architecture, dériver les états du modèle SCHMM par un critère d'adaptation discriminant peut permettre, comme c'est le cas en reconnaissance de la parole, d'améliorer la prise en compte de la structure temporelle du signal acoustique. L'approche MMIE peut convenir à cette tâche.

Concernant l'algorithme d'apprentissage des modèles SCHMMs, proposé dans nos travaux, nous avons remarqué qu'il est peu probable que le découpage initial en segments de même longueur permette d'obtenir une segmentation optimale en cellules acoustiques. Une meilleure segmentation permettrait sans doute une meilleure caractérisation de la structure temporelle des mots de passe. Une information plus détaillée peut être obtenue par la prise en compte, lors de l'alignement phonétique, de la totalité des phonèmes et pas seulement des limites inter-mots, comme nous l'avons réalisé dans ce document. Il est également envisageable d'utiliser une approche de type GLR (*Generalized Likelihood Ratio*) ou BIC (*Bayesian Information Criterion*) pour détecter les irrégularités au sein des flux audio ou vidéo.

L'estimation des probabilités de transition entre les états des SCHMMs présente un fort potentiel pour améliorer la représentation de la structure temporelle des mots de passe. Le signal de synchronisation intégré au décodage acoustique peut être remplacé par un processus continu utilisant des modèles de durées et permettant une prise en compte plus détaillée de la structure temporelle des mots de passe. La granularité finale est cependant limitée par les contraintes propres à la modalité dont serait extrait ce signal. Dans le cas de la vidéo, par exemple, la limite est fixée par le taux d'échantillonnage plus faible que celui du signal audio.

Enfin, d'autres architectures acoustiques peuvent être développées pour intégrer des états modélisant explicitement les vecteurs étiquetés *non-parole* afin d'obtenir des alignements temporels plus pertinents.
Affranchis des contraintes liées aux ressources des systèmes embarqués, l'utilisation de deux modèles SCHMMs appris, l'un avec le *flux brut*, l'autre avec le *flux de parole* (cf. section 8.3.1), peut permettre d'améliorer l'apprentissage des modèles, via une amélioration de l'alignement effectué durant la phase d'enrôlement.

Extensions bi-modales de notre approche

Nous proposons ici différentes extensions de nos travaux à la biométrie bi-modale dans le cas où l'application visée n'impose plus le respect des contraintes de ressources et de données spécifiques à l'embarqué.

Nous avons fait référence, dans ce document, à des travaux ayant trait à la synchronisation des modalités audio et vidéo en parole. Cependant, dans la littérature, peu d'études sont consacrées à la caractérisation des silences et pauses. Une caractérisation robuste de ces événements à partir du flux vidéo pourrait apporter beaucoup à la reconnaissance de la parole, à la segmentation en locuteur ou, dans notre cas, à la reconnaissance du locuteur. Cette caractérisation peut être menée en analysant les corrélations existant entre différents types de paramètres audio et vidéo.

La suppression des contraintes calculatoires liées aux systèmes embarqués permet de remplacer le processus vidéo que nous avons proposé par un système biométrique uni-modal tirant partie de la dynamique du flux vidéo, proche de ceux que nous avons présenté.
Cette approche, permettant d'exploiter les modalités audio et vidéo selon le contexte applicatif pour garantir la robustesse du système biométrique, autoriserait l'exploitation de différents modes de fonctionnement :

- un mode de reconnaissance du locuteur (audio) tant que le rapport signal sur bruit ne dépasse pas un seuil fixé ou que la modalité vidéo est inexploitable ;
- un mode de reconnaissance de l'identité par l'utilisation de la vidéo uniquement dans le cas où la modalité audio est inutilisable et que la luminosité le permet ;
- un mode de reconnaissance du locuteur assisté par la vidéo, selon le fonctionnement décrit dans ce document, pour le cas où le bruit acoustique est faible et que la luminosité le permet ;
- un mode symétrique au précédent, dans lequel la modalité vidéo est assistée par la composante acoustique du système, dans le cas où le bruit acoustique est élevé et que la modalité vidéo est exploitable.

Dans cette perspective, il s'agit d'optimiser un processus de décision permettant de passer d'un mode de fonctionnement à un autre selon la qualité des signaux et les conditions environnementales.

Dans le cadre de l'apprentissage des modèles de locuteur, nous pensons qu'une information vidéo pertinente peut être combinée à une détection d'activité vocale afin d'apprendre différents modèles d'un même locuteur, chaque modèle correspondant à un couple (classe audio ; classe vidéo).
De nombreuses techniques de paramétrisation d'images existent dans l'état de l'art. Les paramétrisations du flux vidéo sont moins nombreuses et se heurtent principalement à la dimension importante du signal. Le domaine de la compression vidéo peut, selon nous, être utilisé pour fournir des paramétrisations propres à diminuer la dimension des données, à extraire l'information pertinente et à éliminer le bruit présent dans les vidéos. Des codeurs vidéo existent et peuvent être utilisés ou modifiés pour extraire des informations exploitables en biométrie vidéo. Ces traitements vidéo, qui nécessitent une puissance de calcul élevée et s'accommodent encore mal des contraintes liées à notre contexte embarqué, peuvent être envisagés dans un contexte moins contraint.

Extension à d'autres modalités

Enfin d'autres couples de modalités biométriques émergentes présentent une forte corrélation dans le domaine temporel et pourraient tirer parti de l'approche bi-modale que nous avons proposée. Nous pensons particulièrement aux modalités qui analysent la démarche (*gaite*) ou le contact du pied sur le sol (*footstep*). En *footstep*, les systèmes état-de-l'art utilisent des modèles de Markov cachés. Il est facilement envisageable de remplacer le flux de données audio de notre système par les mesures de la force de réaction du sol, mesurée par un capteur à deux dimensions au cours du temps. Les cellules acoustiques de notre approche sont alors remplacées par les différentes configurations du pied sur le sol tandis que l'enchaînement des pas mesuré par la modalité *gates* peut être utilisée pour synchroniser le processus *footstep*.

Reinforced Temporal Structure of Acoustic Models for Speaker Recognition.

Embedded speaker recognition in mobile devices may involve several ergonomic constraints and a limited amount of computing resources. For example both the enrolment and the test have to be done using short audio sequences. Even if they have proved their efficiency in more classical contexts, GMM/UBM based systems show their limits in such situations, with good accuracy demanding a relatively large quantity of speech data, but with negligible harnessing of linguistic content.

This thesis deals with this problem and proposes to take into account the linguistic nature of the speech material inside the GMM/UBM framework by using client-customised utterances. Furthermore, the acoustic structure is then reinforced with new temporal information.

Experiments on the MyIdea database are performed when impostors know the client utterance and also when they do not, highlighting the potential of this new approach.
The whole approach allows a gain of up to 65% in terms of Equal Error Rate (EER) over a basic GMM/UBM system when impostors do not know the client utterance. Performance is equivalent to the GMM/UBM baseline system in other configurations.

1 Introduction

The efficiency of speaker recognition systems in a realistic application can be influenced by several constraints. For example, an application which should be immediately usable will strongly limit the enrolment material and hence could lead to poor recognition accuracy. Given that limited data can greatly reduce recognition performance, ergonomic constraints may also impact performance negatively by imposing short test sequences. Some other constraints can be memory and computational resource limitation or the use in variable environments. Embedded systems might well present these conditions.
State-of-the-art speaker recognition engines tend to be assessed on text-independent inputs and often follow the GMM/UBM (Gaussian Mixture Model/ Universal Background Model) paradigm (Bimbot et al., 2004). This solution gives a high level of performance as shown during the NIST evaluations (Przybocki et al., 2007). Unfortunately, the GMM/UBM depends strongly on the quantity of training data available to enrol a speaker in contrast to the context considered here which involves speech material of relatively short duration. A solution to this problem is to increase the amount of information taken into account by the system by including text dependencies, like in a user-customised utterance scenario (BenZeghiba et Bourlard, 2006). In this case, the Temporal Structure Information (TSI) gathered from the utterance can help to compensate for the short duration of the audio sequences. In order to model the TSI of speech while achieving statistical modelling, a word recognition system could be combined with a speaker recognition system(Hebert et Heck, 2003), (Navratil et al., 2000).

To satisfy application ergonomic constraints and to allow the speaker to chose his/her own customised-utterance, the system should accept all kinds of utterances, especially short ones, and also be language-independent. Adding language options when using a phoneme-based word recognition system would seem viable with an appropriate choice of phonemes covering the languages. However, this solution could be expensive in terms of storage and computational cost.
Furthermore, an embedded system could be confronted with strongly variable environments. Due to this constraint, the acoustic modelling used in the recognition system has to be adapted to the environment and the computational cost of the adaptation has to follow the targeted context resource constraints. HMM adaptation does not seem well suited as it normally requires a relatively large amount of training data.
The solution proposed in this thesis associates the well known advantages of a GMM-based statistical acoustic model with an original architecture able to deal with the application context constraints and to incorporate external temporal information. It uses the GMM/UBM paradigm for the general acoustic space modelling and its text-independent speaker recognition capabilities. It also involves an HMM/Viterbi approach in order to incorporate the text-dependent and TSI aspects using a Semi-Continuous HMM (SCHMM) (Young, 1992). Such a combined system was originally proposed in (Bonastre et al., 2003) for speaker recognition and extended to word recognition in (Lévy et al., 2006).

The bi-modal aspect of speech (audio-video) can be viewed as a possible way to increase robustness and performance of speaker recognition systems. Two main approaches are possible to take into account this bi-modal aspect of speech. Generally, this problem is viewed as a fusion process between the two modalities. Early fusion at the data level is difficult due to the different nature of the parameters and their asynchronism. Several works were proposed, mainly in speech recognition (Chibelushi et al., 1997) (Faraj et Bigun, 2007), and show a performance improvement only when noisy audio data are used. Fusion at the score level is more often proposed due to its simplicity (Cetingul et al., 2006), but such a fusion process does not take advantage of the temporal joint information and it is still costly in terms of computational resources (separate systems are needed). Finally, an interesting alternative to a fusion process consists in a joint decoding of both modalities. However the asynchronism aspect of audio-video modalities leads to complex algorithms like in (Bengio, 2003b) and (Chetty et Wagner, 2005).
In this paper we propose a new approach for speaker verification. The speech modality is reinforced by an additional temporal information coming for example from the video stream. Thanks to the specific aspects of our system, this joint-decoding shows an acceptable level of complexity. In order to reinforce the relaxed synchronisation between states and frames due to the SCHMM structure of the TSI modelling, we propose to embed external information during the audio decoding by adding further time-constraints gathered for example from a video synchronisation process.

The specific three-stage architecture is described in Section 2 as well as the way of reinforcing the TSI with an external synchronisation. The experimental protocol as well as the audio-video MyIdea database are described in Section 3. The ability of GMM/UBM systems to exploit temporal structure of speech is studied in Section 4. We propose in Section 5 an extension of the GMM/UBM paradigm which takes into account the temporal structure of speech. The temporal structure of user-customised passwords is then reinforced by an external synchronisation process and the synchronisation of the acoustic process by video information is discussed. Section 6 summarizes the benefits of this approach and presents possible future work directions.

2 Approach overview

The proposed system combines a statistical representation of the acoustic space and a precise modelling of the TSI. Based on a semi-continuous hidden Markov model (SCHMM), it operates a three-stage acoustic modelling architecture. In order to involve the TSI with respect to training data and resource limits, a common GMM which represents the acoustic space is derived to obtain the SCHMM state probability functions. External information is then used to reinforce the temporal structure of the password model. The three stages of the proposed hierarchical architecture illustrated in Figure 9.9, denoted EBD, for Embedded LIA_SpkDET (Bonastre et al., 2008) system are described below.

Universal Background Model : the top layer is the least specialised one and is a classical UBM. It aims to model the general acoustic space.

Speaker-dependent GMM model : the middle layer contains the text-independent characteristics of each speaker. These text-independent speaker models are obtained by a classical GMM/UBM adaptation method : each speaker model is derived from the UBM following the MAP criterion and using the EM algorithm.

Password SCHMM model : the previous text-independent speaker model is used to obtain a left-right SCHMM with the goal of harnessing the TSI of the utterance chosen by this speaker. Using an iterative Viterbi decoding process, each state of the SCHMM is trained from a part of that utterance. During the test, Viterbi decoding is again performed with this SCHMM. Details of the train and test processes are given below.

External information : the goal here is to use further information to assist in the overall verification task by adding constraining components. This constraint is computed during the training phase and used to constrain both the training and test process. This information is here labelled external to reflect that it is aimed to come from an independent process.

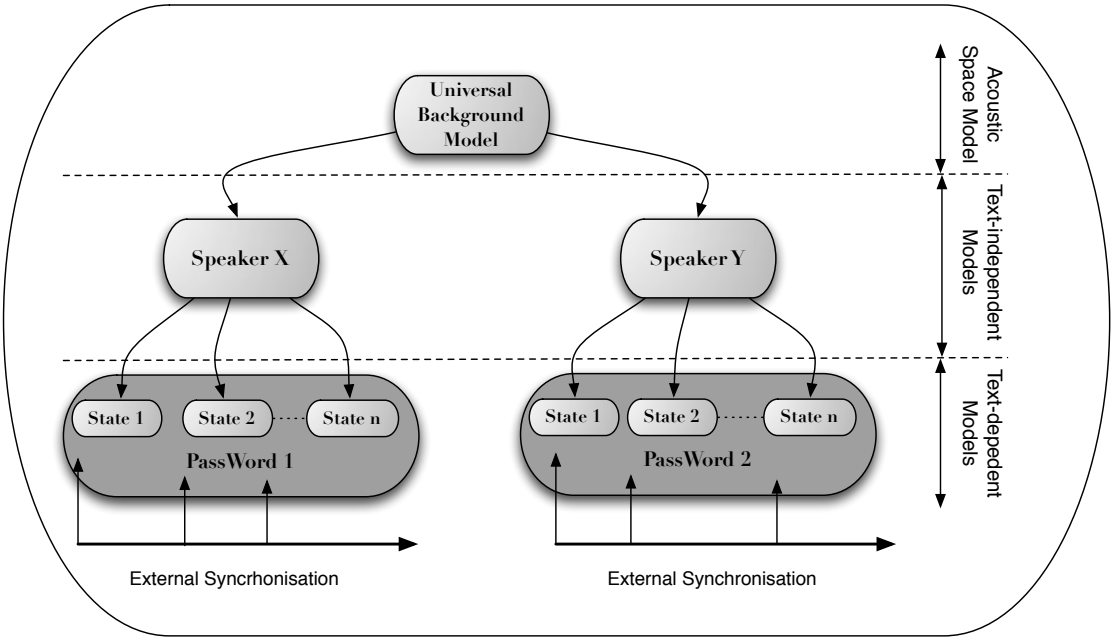


FIG. *9.9:* *General view of the EBD model architecture.*

3 Corpus and protocol

3.1 The MyIdea database

Experiments are performed on the BIOMET part of the MyIdea database. This database contains audio-video records of 30 male speakers. In this subpart of MyIdea, 25 sentences were recorded in 3 sessions for each speaker. Twelve of these sentences are the same for all the speakers, ten short (about 3 seconds) and 2 long sentences (about 6 seconds). Other occurrences are speaker or session dependent. The three sessions were recorded under controlled acoustic and lighting conditions. However, MyIdea presents several drawbacks for our work. The recordings were not done in a real environment. The sentence duration variability is limited (2 or 3 seconds for the utterance occurrences), the sentences are too long for a real password dedicated system, and the number of speakers is small for a speaker verification experiment.

3.2 Experimental protocol

The 30 male speakers are separated into two groups - A and B - of i15 speakers each. Each group is successively used as the Client-set with the others used to train the UBM. The UBM is trained using the whole recorded material of the 15 speakers of the UBM set.
When using the A-group data set to train the UBM model, the speakers from the B-group are used for enrolment and tests. Due to the small number of speakers available,

a jacknifing process is used by training a client model for each available speaker session. Each of the 15 speakers of this client set is successively considered as a client for which the 14 other speakers of the client set are impostors. Two conditions are defined :

1-occ in this condition, each client's text-independent GMM model is derived from the UBM by using two long sentences and one occurrence of the selected short sentence (around 8 seconds of speech). The password-dependent model is trained with the same short sentence occurrence (around 2 seconds of speech). With the jacknifing process, 900 utterance models are trained (10 short sentences, 3 sessions and 30 clients).

2-occ this condition is the same as above except that an additional occurrence of the selected short sentence (which is the password) is used to train both the text-independent and the utterance models. The number of utterance models is still the same as above.

1-occ-Random in this condition, each client's text-independent GMM model is derived from the UBM by using two long sentences, one occurrence of the selected password and a random additional short sentence, different from the speaker password. The password-dependent model is trained with the only available password occurrence.

The number of target trials is condition-dependent. The short sentences not used for utterance training are compared to the client model. 1,800 client tests are performed in the *1-occ* and *1-occ-Random* conditions (2 test occurrences for each of the 900 utterances) while 900 client tests are performed in the *2-occ* condition (1 test occurrence for each of the 900 utterances).
Three configurations of impostor tests are proposed. The speaker and utterance models are compared to the 14 impostors who are the remaining speakers of the same group.

- **UNKNOWN configuration** the linguistic content of the test occurrences is different from the training material of client models. Each speaker model is compared to three randomly selected short sentences (one per session) out of the 9 remaining sentences of each of the 14 impostor speakers. 37,800 impostor tests are performed in this configuration.
- **KNOWN configuration** the linguistic content of the test sequences is the same as the occurrences used to train the client models. Each utterance model is compared to three randomly selected sentences from each of the 14 other speakers of the Client-set. 37,800 impostor tests are performed in this configuration.
- **ALL configuration** the tests are from both the KNOWN and the UNKNOWN configurations. The number of impostor tests in this configuration is 75,600.

3.3 System Configuration

Mel-scaled frequency cepstral coefficients (MFCC) are used, computed every 10ms. An energy labelling is applied to separate the speech frames from the non-speech frames. Acoustic feature frames are 32-dimension vectors, with 15 cepstral coefficients, the log-energy and the corresponding Δ coefficients.

4 Baseline System

GMM do not model the temporal structure of speech and are mainly dedicated to the text-independent speaker verification task. As our approach is based on the GMM/UBM paradigm and will be compared in this paper to a state-of-the-art GMM/UBM system, in terms of performance, we propose to evaluate the performance of GMM/UBM systems for text-dependent speaker verification The first part of this section is dedicated to the configuration of GMM/UBM systems in our specific embedded context and the second part of this study evaluates the ability of GMM/UBM systems to take advantage of the text-dependency constraint.

4.1 Ratio between quantity of data and model size

Dimension of GMMs has a significant effect in terms of accuracy and resources consumption which should be determined, particularly when considering an embedded context. (Mason et al., 2005) shows that the ratio *quantity of training data / number of Gaussian distributions* strongly affects GMM/UBM system performance.
Considering our applicative constraints (restricted training data and short test utterances), the aims of this section is to determined the most suitable configuration in terms of number of Gaussian distributions per GMM.

The same experiment is performed with different sizes of GMM (from 16 to 2048 Gaussians per model) for the 1-occ configuration. Results of this experiments are given in Table 9.7.

	Number of Gaussian distributions per GMM							
	16	32	64	128	256	512	1024	2048
Equal Error Rate (%)	7.01	5.28	4.33	3.67	3.22	3.06	3.16	3.17

TAB. 9.7: *Equal Error Rate of a GMM/UBM state-of-the-art system for different sizes of models. The number of Gaussian distributions grows from 16 to 2048. Impostors pronounce the client password as well as other sentences (1-occ ALL configuration).*

The lowest Equal Error Rate (EER) is provided by GMMs with 512 Gaussian distributions. The performance of the GMM/UBM system drops drastically when decreasing the number of distributions per model. This phenomenon is ordinarily explained by a number of distributions too small to precisely model the density of probability of a speaker.
Increasing the number of distributions beyond 512 provokes a weak increase of the EER probably due to the lack of training data.

4.2 Influence of text-dependency for GMM/UBM systems

Most of the time, the GMM/UBM paradigm is used for text-independent speaker recognition. As such a system is considered as a reference in our study, we aim to evaluate the effect of text-dependency on the performance of GMM/UBM systems.
The same GMM/UBM system (512 Gaussian distributions per model) is used in three different configurations : ALL, KNOWN and UNKNOWN. Results are given in Table 9.8.

Test Condition	Equal Error Rate (%)
KNOWN	3.68
UNKNOWN	2.11
ALL	3.06

TAB. *9.8:* *Performance of a GMM/UBM system for different text-dependency configurations.*

GMM do not intrinsically take into account the temporal structure of speech. However, the performance of our system strongly varied due to the lexical content of the training and test utterances. Our reference GMM/UBM system takes unfair advantage of the lexical variability between the client and impostor utterances in the UNKNOWN conditions and provides 2.11% of EER instead of 3.68 % when the lexical content of those utterances is the same.
This gap of performance could be explained by the lack of training data. As the speaker models are adapted from only a few seconds of speech, they could not be considered as exhaustive acoustic models of the clients but mainly model phonemes belonging to the pronounced sentences.
The difference between KNOWN and UNKNOWN conditions should probably decrease when improving the quantity and variability of the training data.

4.3 Performance of GMM/UBM with small quantity of data

According to previous results, increasing the quantity of training data should improve performance of GMM/UBM systems for all the test conditions and should reduce the performance gap between them. A batch of experiments are performed to verify this assumption and evaluate the influence of the lexical content on this improvement. Our reference GMM/UBM system is tested in three conditions :

1-occ considered as the reference condition ;

2-occ in this configuration, another utterance of the speaker password is added to the training data. The quantity of training data is increased but the lexical content remains the same ;

1-occ-Random in this condition, a short sentence, different from the speaker password, is used to train the text-independent model. The quantity as well as the lexical variability of training data are increased.

		Test Condition		
		1-occ	1-occ-Random	2-occ
Equal Error Rate (%)	ALL	3.06	2.33	1.31
	KNOWN	3.68	2.89	2.00
	UNKNOWN	2.s11	1.78	0.56

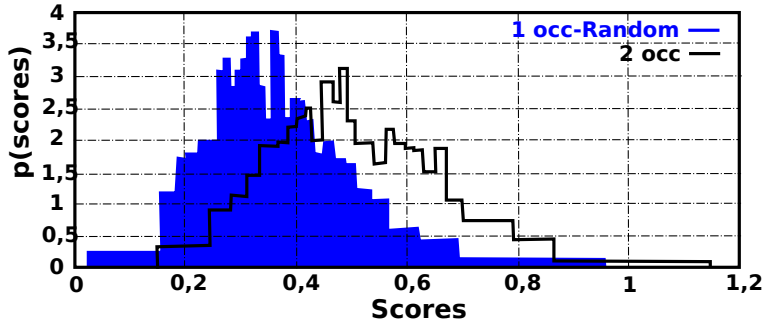
TAB. 9.9: *Influence of the lexical content for GMM speaker model training in text-dependent configuration.*

Increasing the quantity of data increases the speaker model quality whatever the lexical content of the additional data is.
For a text-dependent task, the best configuration is to add the same lexical content as the one used during the testing phase. The UNKNOWN configuration results show that adding the same lexical content increases the discriminative power of the GMM/UBM paradigm.
In the KNOWN configuration, it is difficult to know if the better performance of the 2-occ configuration comes from better scores obtained by client test utterances or by worse results for the impostor tests. Figures 9.10(a) and 9.10(b) show distributions of client and impostor scores in 1-occ-Random and 2-occ configuration for the ALL test condition.

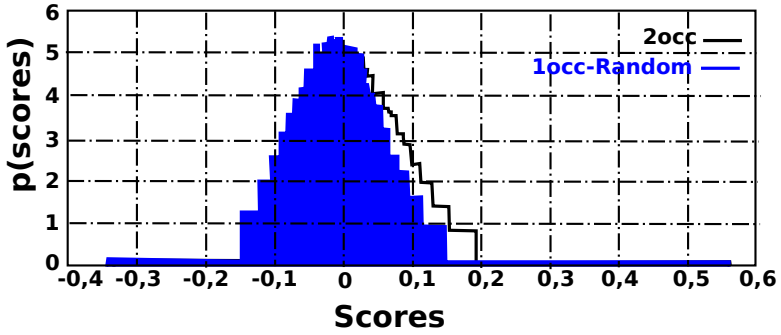
Figure 9.10(b) shows that the variation of the impostor scores distribution related to the lexical content of the additional utterance is extremely limited. This weak variation shows that the more different the training occurrences are, the best the models are trained.
Figure 9.10(a) shows that client scores are dependent on the training data. When the additional sentence is the same as the test utterance, client scores are higher than when a different sentence is used. In this context, our GMM/UBM system takes advantage of the phonetic content of the training utterances.

5 Extensions of the GMM/UBM paradigm

As explain in Section 2 our method is based on a three level architecture to model the acoustic structure of speaker specific password utterances. The two first layers of this architecture are based on the GMM/UBM paradigm studied in Section 4. The third layer is thought to model the temporal structure of user-customised passwords as well as the speaker acoustic specificity. An external information is then added to this three stages architecture in order to reinforce the temporal structure of the password models. These two points are described in this part.



(a) Client scores distributions



(b) Impostor scores distributions

FIG. *9.10:* *Distribution of impostor and client scores depending on training data for 1-occ-Random and 2-occ conditions in ALL condition.*

5.1 Temporal structure modelling

Contrary to vectorial approaches, generative methods are naturally able to deal with intra-speaker variability. Hidden Markov Models (HMM) which belong to this category are heart of most of the state-of-the-art speech recognition and isolated words recognition approaches. Nevertheless, HMM models adaptation requires a relatively large amount of training data and involves a high computational cost which does not match with the targeted resources constraint of embedded environment.
Semi-Continuous Hidden Markov Models (SCHMM) were introduced in (Young, 1992). The emission probability function associated with each of the SCHMM states is a GMM derived from the corresponding text-independent speaker model. The transformation function works only on the weights of the GMMs, the other parameters are directly taken from this middle level model. This allows to model the temporal structure of speech utterances such as continuous HMM while requiring minimal resources.

Iterative training process

The EBD model is trained in three steps, each corresponding to one level of the architecture. The UBM is firstly trained to model the largest part of the acoustic space. It is built off line using a suitably large amount of representative data. It is trained with a classical EM/ML algorithm. The training of the text-independent speaker models consists in adapting the UBM with the available data pronounced by the client. The model is obtained by adapting the UBM using the EM algorithm with the MAP criterion.

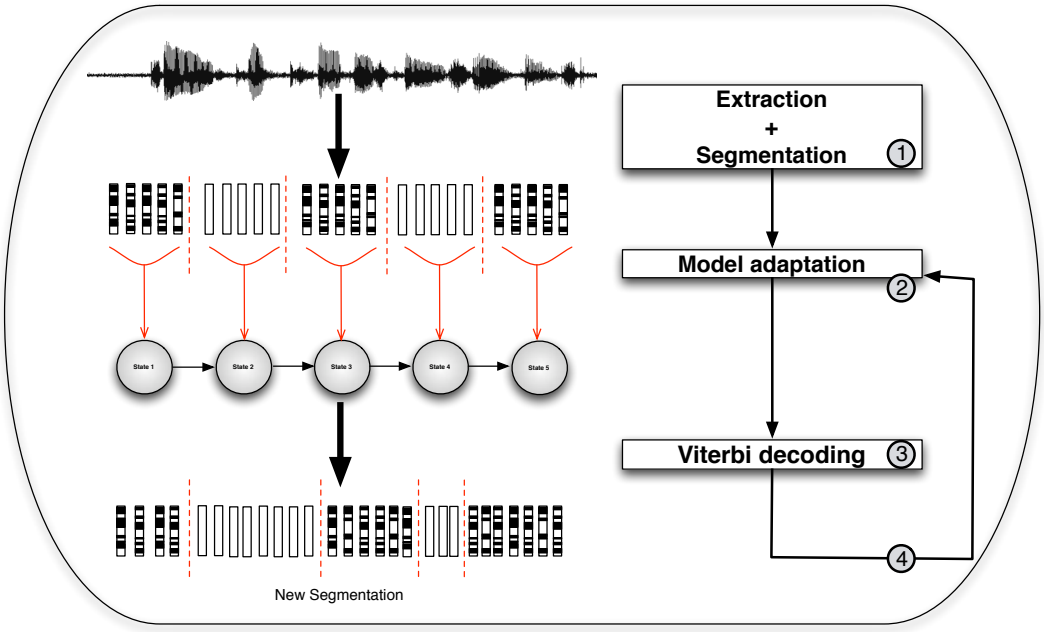


FIG. *9.11:* *Iterative training process.*

The third layer text-dependent model results from the four steps iterative training process shown by Figure 9.11.

1. In order to initialise the password SCHMM model, the password sequence is cut into S segments $\{seg_i\}$ of the same length.
2. Each state i of the SCHMM is adapted from the speaker text-independent speaker model using the speech-labelled frames of seg_i. An EM/MAP algorithm is applied on the weight parameters.
3. Then a new segmentation is achieved by a classical Viterbi decoding.
4. The new segmentation provided by the Viterbi decoding is used to adapt the state models.

As all parameters except weights are tied between the states of the SCHMM and the text-independent model, the log-likelihood for an input frame is only computed for

each Gaussian component of the text-independent model. Then the log-likelihood of this frame with each state of the SCHMM involves a weighted sum which is negligible compared to the full log-likelihood computation.

Structure of the SCHMM

Place of the non-speech frames in the password temporal structure

State-of-the-art speaker recognition approaches include a Voice Activity Detection (VAD) stage. This VAD is used to label *speech* the acoustic frames which contain a useful information for the speaker recognition task. The remaining frames are labelled *non-speech* as a complement. Only *speech* frames are used in the rest of the process and this selection strongly increases speaker recognition performance as shown in (Besacier et al., 2000) and (Scheffer, 2006). However, considering *non-speech* frames repartition and duration could be worthwhile in order to discriminate user-customised password structures. Using the whole acoustic signal, without any selection, we consider that *non-speech* segments are part of the temporal structure information and that they should be modeled in the acoustic model as a specific part of the password temporal structure.

The alignment resulting from the four step process described above is compared to a VAD stage. An example of the two segmentations is given in Figure 9.12.

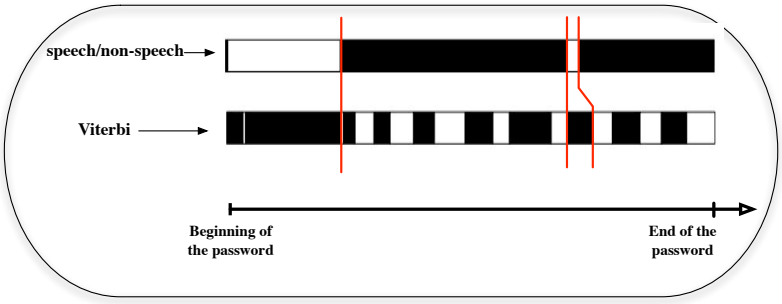


FIG. 9.12: *Segmentation speech/non-speech resulting from the voice activity detection stage and from the iterative training process*

In this example, the VAD determines 5 segments (3 speech and 2 non-speech). The segmentation provided by our system (based on a password model using 20 states per SCHMM) shows similitude with the VAD result. It seems that the iterative training process partially matches with the result of the VAD. In this case, the training process allocates dissimilar states to the *speech* segments and to the *non-speech* segments.
A complete evaluation performed with the 1-occ-ALL configuration shows that the use of a VAD decreases the EER from 3.89% (with no selection) to 3.17% by using only speech labelled frames.
The Viterbi may provide bad segmentation mixing speech and non-speech frames in a

same state that would affect performance of the EBD system.

Ratio training data quantity/ number of Gaussian distributions
The number of states used in the SCHMM to model password occurrences determines the quantity of data available to adapt each of the states. Increasing the granularity of the password model involves an equivalent reduction of training data quantity per state.
The ratio *quantity of training data/number of Gaussian distribution per GMM model* has been shown to be very important for GMM/UBM approach in section 4.1. An estimation of the necessary granularity is given by state-of-the-art phonetic speech recognition systems. As the quantity of data is strongly limited in our context, it seems that we can not use the same granularity as in speech recognition approaches.
Experiments are performed to fixe the optimal SCHMM state number. Figure 9.13 shows the evolution of EER related to the number of distribution of the state GMM models and to the number of states in SCHMM.

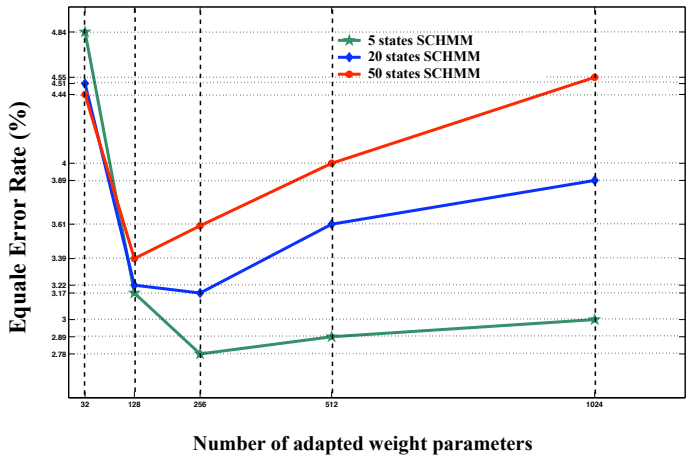


FIG. 9.13: *Equal Error Rate evolution (%) depending on the GMM model dimension for different numbers of SCHMM initial states*

The EER function, depending of the number of distribution per state is convex. The best ratio is obtained when the number of distributions is sufficient to model the temporal structure of the password. However, as the number of distributions increases, the quantity of training data do not suffice to adapt GMM state models and performance drops.
According to those results forthcoming experiments will be performed with 256 Gaussian distributions per GMM model.

Adaptation of the weight parameters for SCHMM states
Each state of SCHMM models is a mixture of Gaussian characterised by a vector of weight parameters. It was previously shown that reducing the dimension of GMM state models spoil performance of the speaker recognition system. Reducing the number of weight parameters which characterise states of SCHMMs allows to greatly reduce required resources while keeping a satisfactory speaker model.
Experiments are performed in order to evaluate to effect of reducing the number of adapted weight parameters per state. Non-adapted weight parameters are normalised to assure the sum of the weight to be equal to one.

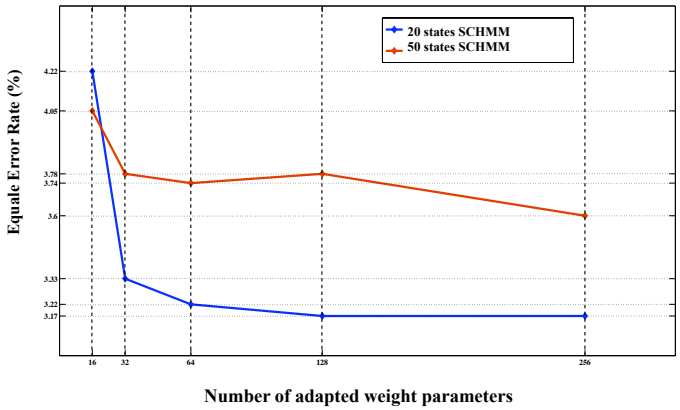


FIG. 9.14: *performance of the EBd system for different numbers of adapted weight parameters.*

Figure 9.14 shows the degradation of performance for two size of SCHMM models (20 and 50 states and 256 Gaussian per GMM model) when varying the number of adapted weight parameters per state. Adapting a small part of the weight parameters strongly degrades performance of our approach. With 20 states per SCHMM for example, the EER falls down from 3.17% to 4.22% However, the performance still equivalent while a quarter of the parameters are at least adapted. Forthcoming experiments are realised adapting all the weight parameters.

Scoring

The scoring process developed for EBD system is equivalent in terms of computation to a classical GMM/UBM system and produces two scores. The first is obtained with only the text-independent speaker model. It corresponds to the GMM/UBM paradigm. The second score is computed with the SCHMM model and takes advantages of the TSI of the user-customised password. This score is text and speaker dependent. Both of the scores are normalised using the log-likelihood of the UBM. They are then combined to give a final score for the decision stage. An empirically-tuned weighted

linear combination is used.

During a test, the score between the input signal and an utterance SCHMM model is derived from the corresponding Viterbi path. All the input frames are used during this Viterbi decoding phase. The log-likelihood of a frame sequence could be expressed as a sum of two log-likelihood accumulations, one using speech-labelled frames and the other using non-speech-labelled frames, as shown in Equation 9.2.

$$log\ p(X|\lambda) = log\ p(X_{speech}|\lambda) + log\ p(X_{non-Speech}|\lambda) \tag{9.2}$$

The final speaker matching score corresponds to the log-likelihood computed with the $log\ p(X_{speech}|\lambda)$ only.
As for the training, the log-likelihood for an input frame is only computed for each Gaussian component of the text-independent model. The computation of the log-likelihood of this frame with each state of the SCHMM which involves a linear combination is negligible.

Test condition	Equal Error Rate (%)		
	GMM/UBM	Fusion	Text-dependent
UNKNOWN	2.46	1.11	0.94
KNOWN	4.00	4.06	4.62
ALL	3.22	2.83	3.17

TAB. *9.10:* *performance of the different layers of the EBD system : GMM/UBM (text-independent), text-dependent as well as a score fusion of those two stages.*

Experiments are conducted to assess the contributions coming from the Temporal Structure Information (TSI). The GMM/UBM is regarded as the baseline. The experimental results presented in Table 9.10, expressed in terms of equal error rates, show performance of the EBD system depending of the nature of the impostor tests. performance of the baseline GMM in the same conditions is provided for comparison. It is important to note that the GMM/UBM system reflects the non-structural layer of the EBD system.

The first row in Table 9.10 shows the results when the impostors do not know the speaker utterances. The EBD system takes advantages of the TSI and the Error rate falls from 2.46 down to 0.94 when text-dependent approach is used. A loss of performance is observed when the impostors know the client password. It seems that, in speaker matching scores, the TSI is dominated by the utterance content information rather than the speaker specific information, *i.e.* the system recognises the utterance instead of the speaker. Indeed, the text-independent system performs better than the text-dependent one. Fusion of the two scores allowed to get the best of both approaches in both KNOWN and UNKNOWN conditions. Moreover, this fusion outperforms text-dependent and text-independent approaches in the ALL condition.

5.2 External Synchronisation

This section describes the integration of an external synchronisation within the Viterbi decoding. This synchronisation signal is used to reinforced the temporal structure of the password models in order to compensate the lack of training data and the short duration of test utterances due to the targeted context.
Synchronisation points are generated off line from an external source. These points are used during both the training and test phase to strongly constrain the Viterbi decoding. This constraint is obtained by allowing or forbidding transitions of the SCHMM corresponding to the synchronisation points (labelled *S* in Figure 9.15). Other transitions, labelled *W* are computed from the Audio and not modified when adding the external segmentation.

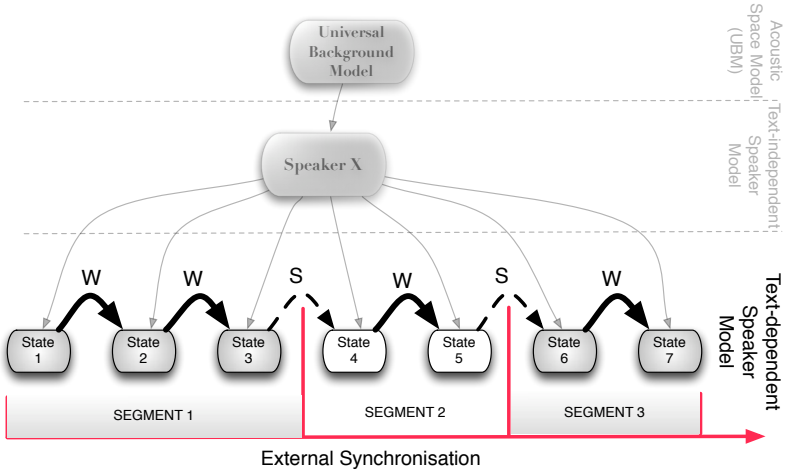


FIG. *9.15:* *Use of an external Segmentation in the bottom layer of the EBD system to constrain the Viterbi decoding and reinforce the TSI in the utterance SCHMMs. The S labelled transitions are constrained by the external segmentation as the W labelled ones still unchanged*

The number of synchronisation points depends on both the speaker and the utterance he pronounces.

The effect of the external constraint expected during both the training and testing phases are described below. Validation is performed by using a phonetic-learnt synchronisation signal. Finally, the use of a video-learnt information within the targeted context is discussed.

During the training phase

Initialisation

Training process described in Section 5.1 is modified to integrate the external synchronisation. Initialisation of the process could be decomposed in three steps shown by

Figure 9.16 :

1. temporal constraint gives a first segmentation of the acoustic password utterance ;
2. states of the SCHMM are spread between the previously determined acoustic segments. Sub-segments are attributed in order to adapt each of the states ;
3. two types of transitions are fixed :

   S transitions the value of these transitions is directly determined by the external information and could be 0 or 1.

   W transitions all these transitions are equiprobable.

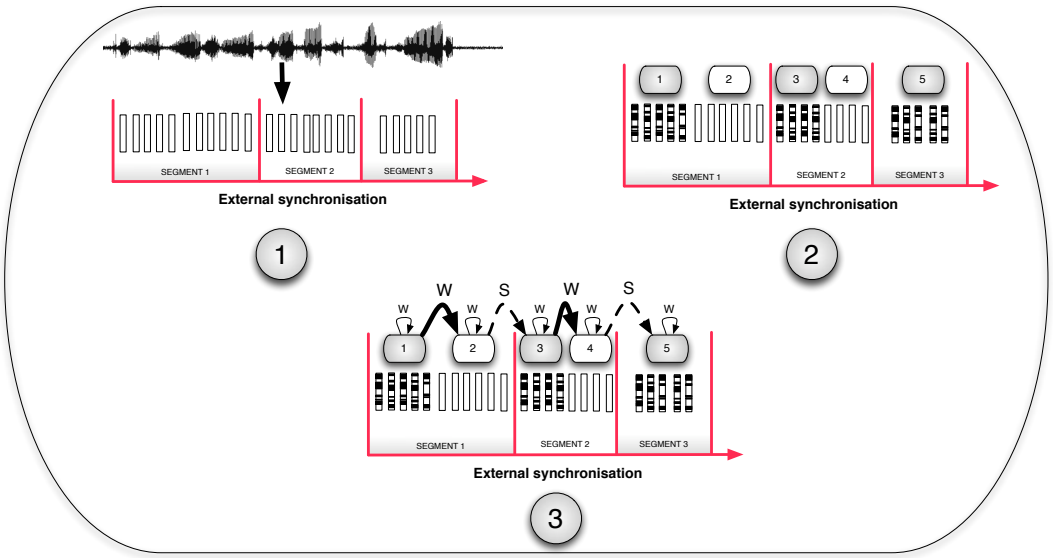


FIG. 9.16: *Illustration of the initialisation three steps constrained process of SCHMM training.*

Iterations

The iterative process is the same as in Section 5.1 excepted that during the Viterbi decoding, performed to obtain a new segmentation, values of S transitions of the SCHMM model vary in accordance with the external synchronisation.

As the external synchronisation constraint is used to reinforced the temporal structure of the password models, we venture the hypothesis that adding temporal information during the training process will improve the adaptation of the probability density of SCHMM states.

During the testing phase

The Viterbi decoding used to score a test utterance in our approach is an optimal algorithm. This algorithm maximises the resulting score which corresponds to the best

matching alignment. Decision module in speaker recognition involves to use a fix threshold. Maximising scores in this context increases the probability to go past this threshold whether in respect of the client scores but even of the impostor scores.

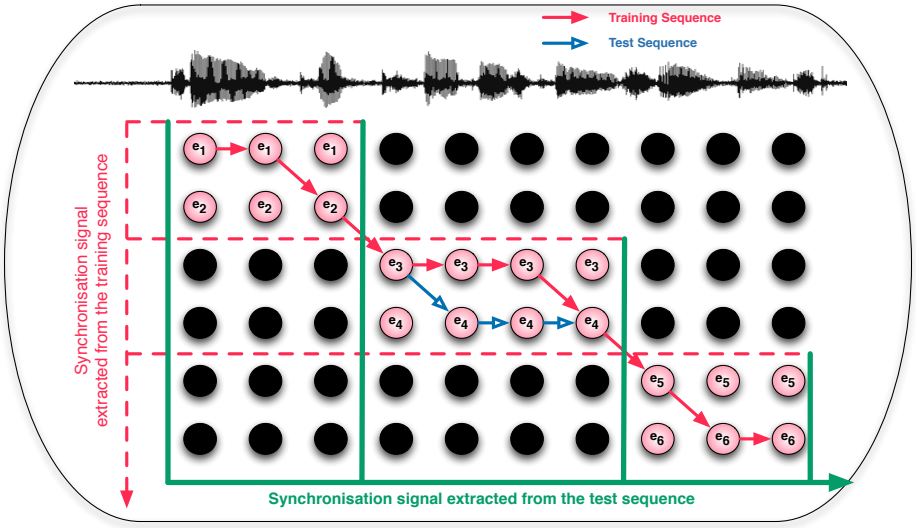


FIG. 9.17: *Alignment of client training and test utterances with and without external synchronisation*

The Viterbi decoding is modified such as in the training phase : S transition values vary according to the external synchronisation computed with the test utterances. The temporal constraint added during the Viterbi decoding forbid the Viterbi path to go across certain areas of the lattice. These areas are illustrated on Figures 9.17 and 9.18 by filled states. Synchronising the Viterbi decoding with an external signal aims to add an a priori information on the nature of the test utterance and expected behaviour is described below.

Influence of the external synchronisation on the client tests
Temporal structure of target utterances are supposed to be very similar to the train one and the test-utterance path should then belong to the permitted area (Figure 9.17). As a consequence, the external constraint is not supposed to affect the Viterbi decoding which still optimal.

Influence of the external synchronisation on the impostor tests
Temporal structure of impostor utterances are supposed to be different to the client train utterance and the test-utterance path should then go outside to the permitted area (Figure 9.18). As a consequence, the external constraint force the Viterbi path to stay in the permitted area. In pursuance thereof, it makes the decoding sub-optimal and decreases impostors scores.

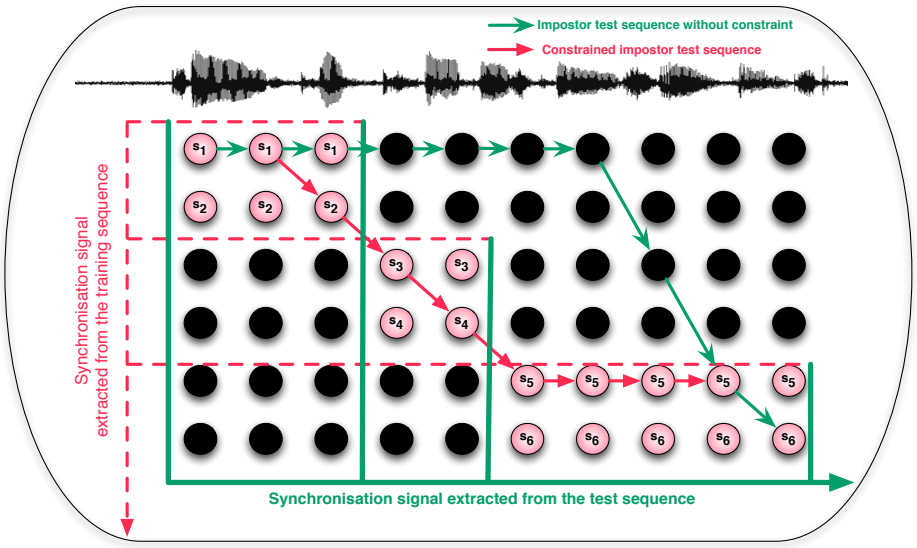


FIG. 9.18: *Alignment of an impostor test sequence with or without external synchronisation.*

Synchronisation by a phonemic constraint

In order to evaluate the effect of an external synchronisation, a new experiment is performed. The external information which is here a lexical constraint is computed using the *LIA SPEERAL Toolkit* (Bürki et al., 2008). Results of this experiment are presented in Table 9.11.

The first column is the original GMM baseline and the next three columns give the EBD results as in Table 9.10, with the use of the external information and then with a random synchronisation signal.
In the KNOWN condition where impostors pronounce the client password utterances, the synchronisation constraint has no effect on the EER. On the other hand, when impostors do not know the client password, adding the external constraint allowed a relative gain of 34% relative (from 0.94 to 0.62%) in terms of EER.
Experiments performed with a random synchronisation show that the gain observed with the phonetic constraint is related to the correlation between acoustic and phonetic signals.

It is shown in Section 5.1 that fusing the scores of the text-dependent and text-independent layers of the EBD system strongly improves performance. The integration of the synchronisation constraint in the structural approach should improve performance of the fused system due to the complementarity of both score information. Table 9.12 shows results of score fusion for three systems : the reference GMM/UBM,

	Configurations			
	GMM-UBM	No Synchronisation	phonetic synchronisation	Random synchronisation
ALL	3.22	3.17	3.17	3.78
KNOWN	4.00	4.62	4.56	5.56
UNKNOWN	2.46	0.94	0.62	1.38

TAB. *9.11:* *EER (%) of GMM/UBM compared to the EBD system with 20 states in ALL condition when using or not a synchronisation coming from a phonetic alignment. Scores provided for the EBD system result from the third layer of the architecture.*

the EBD structural system without any external constraint and the EBD system with a phonetic-learnt synchronisation.

	Configurations		
	GMM-UBM	No Synchronisation	phonetic synchronisation
ALL	3.22	2.83	2.83
KNOWN	4.00	4.06	4.07
UNKNOWN	2.46	1.11	0.89

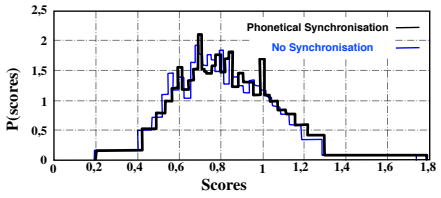
TAB. *9.12:* *EER (%) of GMM/UBM compared to the EBD system in ALL condition when using or not a synchronisation coming from a phonetic alignment.*

Fusion of both text-dependent and text-independent scores only decreases error rates in UNKNOWN condition. Weighted sum could be unsuitable for those type of information or information could be redundant. Indeed, the external synchronisation results from a phonetic alignment and could be too much correlated to the SCHMM structure.

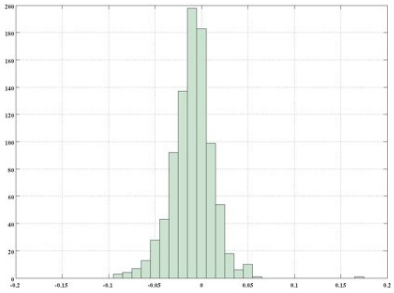
Assumptions enunciated in section 5.2 have found no answer in the previous results. A deeper analyse of the evolution of the client and impostor score is presented in the following paragraphs.

Influence of the external synchronisation on the client tests
Figures 9.19(a) and 9.19(b) show the evolution of client scores when introducing the phonetic synchronisation constraint. Distributions of client score with and without external information are equivalent and do not allowed us to conclude.

Influence of the external synchronisation on impostor tests
The integration of an external source of information in the Viterbi decoding is supposed to degrade the impostor scores. Figures 9.20(a) and 9.20(b) show the evolution of impostor scores distributions in the UNKNOWN condition (*i.e.* when impostor do not



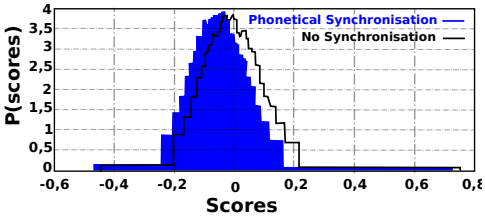
(a) Client scores distributions.



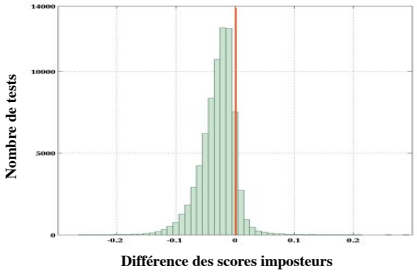
(b) Distribution of the client scores differences.

FIG. *9.19:* *Evolution of the of client text-dependent scores distributions with or without external synchronisation constraint.*
Figure (a) shows the scores distributions with or without the synchronisation constraint.
Figure (b) shows the difference between scores computed with a synchronisation constraint and the scores computed without this information.

know the client passwords).



(a) Impostor scores distributions in



(b) Difference des scores imposteurs

FIG. *9.20:* *Evolution of the of impostor text-dependent scores distributions in UNKNOWN condition with or without external synchronisation constraint.*
Figure (a) shows the impostor scores distributions with or without the synchronisation constraint.
Figure (b) shows the difference between impostor scores computed with a synchronisation constraint and the scores computed without this information.

Figure 9.20(a) confirms that impostor scores decrease significantly when adding an external source of information. Figure 9.20(b) corroborates this conclusion as the observed score differences are mostly negative. The same experiment performed with impostor knowing the client password shows that the effect of the external information is less significant in this configuration due to the similar temporal structure of the password utterance. This effect is not sufficient to discriminate speakers by the way they pronounce a lexical content.

Video Learnt Synchronisation

Description of the information extraction
Synchronisation points are now extracted from a very simple video processing. The video stream is first pre-processed to obtain a black-and-white sequence which is the Y component of the sequence resulting of an RGB to YCbCr transformation.
A mono-dimensional temporal signal is issued from this black-and-white video stream in order to estimate the quantity of change between successive frames. Subtractions are processed between the pixels of one image and thus of the following one. The absolute values of pixel subtractions are summed to obtain a value of the discrete temporal signal S. This computational process is described by Equation 9.3.

$$S(n) = \sum_{w=0}^{W} \sum_{h=0}^{H} abs(I^{n}_{(w,h)} - I^{n+1}_{(w,h)}) \tag{9.3}$$

where W and H are respectively the width and the height of the video images, I^n and I^{n+1} are two consecutive images of the video stream and $I^n_{(w,h)}$ is the value of the pixel (w,h) of the image I^n. $S(n)$ is the discrete temporal signal from which the synchronisation points are extracted.
Local maxima of the signal S are found by applying a sliding window algorithm. These local maxima are stored to become the video-learnt synchronisation points. Figure 9.21 is an example of the S signal taken from the MyIdea database. Red markers indicate the selected synchronisation points.

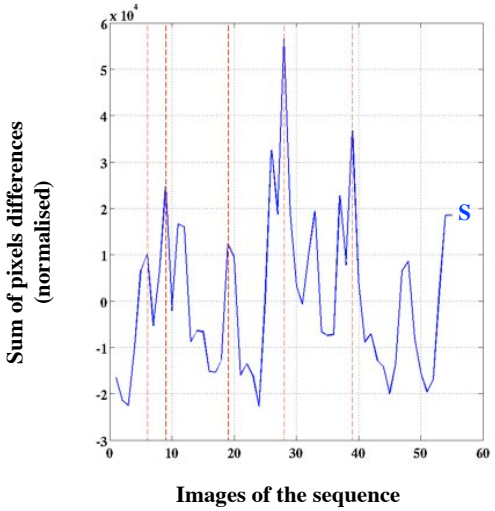


FIG. 9.21: *Sum of pixel differences between two consecutive images of a video stream.*

Experiments
We wish to evaluate the influence of the video-learnt synchronisation on the EBD system overall performance. Same experiments as in the previous section are performed.

The experimental results presented in Table 9.13, expressed in terms of equal error rates (EER), show the influence of a video-learnt synchronisation on performance of the EBD system. performance of the baseline GMM in the same conditions are provided for comparison.

	Configurations		
	GMM-UBM	Video segmentation	Random segmentation
ALL	3.22	3.77	3.56
KNOWN	4.00	5.11	4.94
UNKNOWN	2.46	1.28	1.06

TAB. *9.13:* *EER (%) of GMM/UBM compared to the EBD system in ALL condition when using or not a synchronisation coming from a video-learnt alignment.*

When video-learnt synchronisation is used error rates rise from 3.22% up to 3.77% and 4.00% to 5.11% for the ALL and KNOWN conditions respectively. Moreover, the use of a random synchronisation signal outperforms the video-learnt segmentation.

In conclusion, it seems that the temporal signal extracted from the video stream does not contain any information or that it is too noisy to be used this way. A Co-Inertia Analysis (Doledec et Chessel, 1994) test between both the acoustic and the video signal does not bring to light any correlation. According to the previous results, it seems difficult to extract a useable synchronisation signal from the video stream while respecting ergonomic and material constraints. Nevertheless, we keep thinking that such a signal could be extracted in this context and that the integration of a complementary information coming from the video stream should improve performance of our approach.

6 Conclusion and Future Works

The approach proposed in this paper is designed for embedded applications. It takes advantages of a GMM/UBM text-independent approach and the HMM/Viterbi speech-recognition power in order to compensate the lack of data. In addition we propose to reinforce the Temporal Structure Information modelling by a synchronisation issued from an external stream.

The EBD approach is based on the GMM/UBM paradigm which is also used as reference system in our study. Experiments performed to evaluate the influence of the ratio *quantity of data / dimension of GMM models* show that this parameter strongly affects recognition performance. The number of Gaussian distributions per model has to

be high enough to model speaker characteristics. At the same time, the quantity of available data limits the number of parameters which could be correctly estimated. In the targeted context, the optimal number of distribution is 512 per GMM. Decreasing this number increases the equal error rate from 3.06% to 7.01% when increasing it gently increases the EER from 3.06% to 3.17%.
Others experiments show that GMM/UBM systems are sensitive to lexical-constrain when the quantity of training data is limited. The GMM/UBM does not take advantages from the temporal structure information of the password utterances. However, GMM/UBM systems take benefit from the similarity between lexical content of the training data and the test utterance. Experiments performed in the text-dependent speaker recognition task show that the EER which is 3.68%, when impostors pronounce the same lexical content than client, drop to 2.11% when impostors pronounce a different lexical content.

The main expected advantage of the EBD system compared to a classical UBM/GMM is to incorporate the password-based information like the password itself and its relative TSI. This point was evaluated by the experiments presented in section 5.1.
These experiments show that in the same way as for the GMM/UBM system, the ratio *quantity of data / dimension of GMM models* is critical for the training of the density of probability of SCHMM models. The size of GMM models should be sufficient to model the password utterance without needing too much training data.
The use of the temporal information in a speaker recognition system allows to improve performance particularly when a relatively small quantity of speech data is available for training and test. performance of our approach are equivalent to the GMM/UBM baseline system when not considering the linguistic content (example of EER in KNOWN condition, GMM : 4.00%, EBD 4.06%) whereas the proposed approach outperforms the GMM/UBM when impostors do not know the client utterance (EER in UNKNOWN condition, GMM : 2.46%, EBD : 1.11%). The three stages hierarchical architecture of the EBD approach produces two scores, dependent and independent of the password. The scoring process is equivalent in terms of computation to a classical GMM/UBM system. The system resulting from this fusion keep the best of both tetxt-dependent and text-independent approaches.

Synchronisation points extracted from an external process are introduced into the Viterbi decoding. The temporal constraint aims to reinforce the temporal structure of password models, to increase performance and to thwart replay attacks.
This approach has been first validated by using a phonetic segmentation. The external-learnt synchronisation leads to a gain in all situations, when impostors know or not the client-utterance, and outperforms or is equivalent to the GMM/UBM baseline in all situations. The gain is particularly significant when impostors do not know the client password. In this condition equal error rates obtained by the GMM/UBM system (2.46%) and the EBD system without any additional constraint (1.11%) drop of respectively 64% and 20% relative.

The phonetic segmentation has been replaced by a video-learnt information coming

from a very simple video process in order to deal with embedded constraints. the use of a video-learnt constraint was supposed to improve performance of the EBD system by adding a complementary source of information. Experiments performed show that this synchronisation signal does not contain any information, or more presumably that the signal to noise ratio is too low.

Future work will focus on improving the discriminant power of the SCHMM models. The Maximum Mutual Information Estimation criteria, for example, could replace the Maximum A Posteriori criteria to adapt state models.
The discrete synchronisation process could also be replaced by a continuous one in order to model the temporal structure information with more accuracy.
In a multimodal scheme, it could be worth caracterising silences with the video stream to carry this information forward in the acoustic processing.
Relaxing the ergonomic and material constraints could allow to extract a more accuracy synchronisation from the video stream and to improve performance by including a complementary information.
It is also possible to invert audio and video streams in the proposed approach. The role of audio and video in the recognition process could then be determined by the environment.

Annexes

Annexe A

Base de données MyIdea

Présentation générale

MyIdea est une base de données multi-modale créée au Département d'Informatique de l'Université de Fribourg en Suisse.
Les modalités présentes dans cette base de données sont au nombre de 7 :

- image,
- voix,
- empreintes digitales,
- signature,
- dynamique de l'écriture,
- empreinte de la main,
- géométrie de la main.

En plus de ces modalités isolées, MyIdea intègre des enregistrements synchronisés de deux couples de modalités :

- audio/vidéo,
- écriture/voix.

MyIdea se présente comme une extension de différentes bases de données existantes : BANCA (Bailly-Bailliere et al., 2003), BIOMET (Garcia-Salicetti et al., 2003), XM2VTSDB (Messer et al., 1999), Mcyt (Ortega-Garcia et al., 2003) et IAM (Marti et Bunke, 2002).

Nous présentons, dans la partie suivante, les enregistrements conjoints des modalités audio et vidéo.

Modalités audio-vidéo

La description de la partie audio-vidéo de la base de données MyIdea proposée ici ne présente que la partie validée des enregistrements existant. D'autres enregistrements existent mais ne sont pas, à l'heure actuelle, validés.
Les enregistrements audio-vidéo de MyIdea correspondent aux protocoles de deux bases existantes : BANCA et BIOMET. Seule la partie enregistrée selon le protocole de BIOMET a été validée et utilisée pour nos travaux. C'est cette partie que nous décrivons ci-dessous.

Sujets

La partie audio-vidéo de la base de données MyIdea utilisée dans nos travaux contient les enregistrements de 30 locuteurs hommes d'âge variable.

Conditions d'enregistrement

- Les enregistrements sont en langue française.
- Le texte prononcé est lu.
- Les enregistrements sont effectués dans un studio à l'intérieur duquel conditions sonores et lumineuses sont contrôlées, c-à-d que les conditions d'éclairage des sujets sont optimales et qu'aucun bruit n'est présent sur la bande sonore.

Contenu

Chaque sujet de la base de données a participé à 3 sessions d'enregistrement. Au cours de chacune de ces sessions, il prononce un certains nombre de contenus :

Compter 0 1 2 3 4 5 6 7 8 9

Décompter 9 8 7 6 5 4 3 2 1 0

Phrase longue 1 Alors que Monsieur Gorbatchev regagnait Moscou au terme d'un difficile voyage en Lituanie, une partie du Caucase s'est embrasée.

Phrase longue 2 Chaque jour ils reçoivent, dans la bonne humeur, la visite du commissaire des renseignements généraux, qui suit de loin l'opération.

Oui/Non

10 Phrases fixes Ces 10 phrases sont prononcées lors de chaque session par tous les locuteurs.

- Il se garantira du froid avec un bon capuchon.
- Annie s'ennuie loin de mes parents.
- Les deux camions se sont heurtés de face.
- Un loup s'est jeté immédiatement sur la petite chèvre.
- Dés que le tambour bat, les gens accourent.

– Mon père m'a donné l'autorisation.
– Vous poussez des cris de colère.
– Ce petit canard apprend à nager.
– La voiture s'est arrêtée au feu rouge.
– La vaisselle propre est mise sur l'évier.

4 phrases "mot de passe" ces phrases sont les mêmes à chaque session mais elles sont spécifiques à chaque locuteur. Elles sont limitées entre 25 et 40 caractères. Les 3 phrases suivantes sont citées à titre d'exemple :
– Plusieurs façades ont été endommagées.
– Le premier traité antitabac entre en vigueur.
– Le chameau est loin de son abris.

4 phrases "mot de passe" d'imposteurs Ces phrases correspondent aux phrases "mot de passe" de 4 autres clients.

5 phrases aléatoires Ces 5 phrases diffèrent pour chaque session et chaque locuteur. Elles sont générées d'après le contenu de la base de données BREF (Larnel et al., 1991). La longueur de ces phrase est limitée entre 50 et 90 caractères, elles ne contiennent pas de citation et le sens de ces phrases a été vérifié par des opérateurs humains.

Après chaque enregistrement, les sujets réalisent différents mouvements de tête : gauche-droite, 15 degrés haut-bas et mouvement let du corps pas rotation de la chaise, ce qui permet d'obtenir une vue de leur profiles gauche et droit.

Le protocole d'impostures n'est pas utilisable en raison de problèmes survenus lors de l'enregistrement.

Protocole expérimental

Les 30 hommes sont séparés en deux groupes A et B, de chacun 15 locuteurs. Chaque groupe est successivement considéré comme groupe de clients tandis que l'ensemble des enregistrements de l'autre groupe est utilisé pour entraîner le modèle du monde.

Dans la suite, nous nommerons :

phrases longues les phrases longues de type 1 ou 2 de la base de données MyIdea de durée $\sim$ 5 secondes,

phrases courtes les 10 phrases fixes de durée $\sim$ 2 secondes

Considérons pour la suite que le groupe A est le groupe clients et le groupe B est utilisé pour entraîner le modèle du monde (UBM). En raison du faible nombre de locuteurs, nous utilisons un protocole tournant. Chaque locuteur du groupe A est successivement considéré comme le client alors que les 14 autres locuteurs représentent

les imposteurs qui tentent d'usurper son identité. Un protocole symétrique est utilisé lorsque le groupe B est le groupe clients et le groupe A le groupe UBM.

Phase d'enrôlement

Trois conditions d'entraînement sont définies :

Configuration 1-occ Chaque locuteur dispose des 2 phrases longues et d'une occurrence d'une phrase courte.

Le modèle indépendant du texte est appris avec les 2 phrases longues et la phrase courte disponibles, ce qui constitue environ 12 secondes de parole.

Le modèle de mot de passe est appris avec une seule occurrence du mot de passe choisi (la phrase courte disponible), soit environ 2 secondes de parole.

Cette configuration constitue notre référence en termes de données d'apprentissage.

Configuration 2-occ Chaque locuteur dispose des 2 phrases longues et de 2 occurrences d'une même phrase courte.

Le modèle indépendant du texte est appris avec les 2 phrases longues et les 2 occurrences de la phrase courte disponibles, ce qui constitue environ 14 secondes de parole.

Le modèle de mot de passe est appris avec les 2 occurrences du mot de passe choisi, soit environ 4 secondes de parole.

Cette configuration a pour but d'évaluer l'effet de l'augmentation de la quantité de données d'apprentissage (par rapport à la configuration **1-occ**) pour les deux modèles du locuteur, puisque le contenu linguistique ajouté est une occurrence du mot de passe.

Configuration 1-occ + aléatoire Chaque locuteur dispose des 2 phrases longues, d'une occurrence d'une phrase courte qui constitue son mot de passe et d'une occurrence d'une phrase courte supplémentaire, différente de son mot de passe.

Le modèle indépendant du texte est appris avec les 2 phrases longues et les 2 phrases courtes disponibles, ce qui constitue environ 14 secondes de parole.

Le modèle de mot de passe est appris avec une seule occurrence du mot de passe choisi, soit environ 2 secondes de parole.

Cette configuration a pour but d'évaluer l'effet de l'augmentation de la quantité de données d'apprentissage (par rapport à la configuration **1-occ**) pour le seul modèle indépendant du texte puisque le contenu linguistique ajouté est une phrase courte différente du mot de passe choisi.

Phase de test

Tests clients

Le nombre de tests clients dépend de la condition d'entraînement choisie. Seules les occurrences des phrases fixes qui ne sont pas utilisées pour l'entraînement des modèles sont testées. De ce fait, 1 800 tests clients sont effectués pour la condition 1-occ (900×2 sessions) alors que la condition 2 fournit seulement 900 tests clients.

Tests imposteurs

Trois configurations de tests imposteurs sont possibles. Dans ces trois configurations, le modèle de chacun des 15 locuteurs du groupe clients est confronté à des phrases fixes prononcées par chacun des 14 autres locuteurs du même groupe.

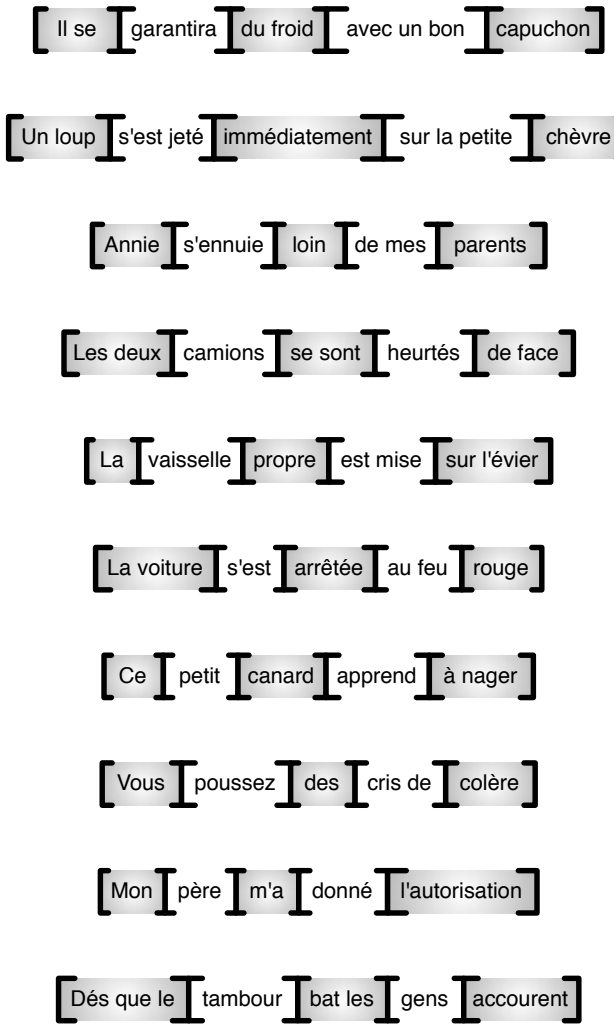
configuration : MDP le contenu linguistique est le même celui utilisé pour entraîner les modèles du client. Chaque modèle de client est comparé à 3 occurrences de la même phrase fixe choisies au hasard parmi les réalisations des 14 imposteurs du groupe. Chacune de ces trois phrases provient d'une session différente. Cette configuration simule le cas où les imposteurs connaissent le mot de passe des clients.

configuration : FAUX le contenu linguistique de la séquence de test est différent du mot de passe du client. Chaque modèle de client est comparé à 3 phrases fixes différentes de celle utilisée pour l'enrôlement du client. Chacune de ces 3 séquences de test est choisie au hasard parmi les 9 phrases fixes restantes et les 14 imposteurs disponibles. Chacune de ces trois phrases provient d'une session différente. Cette configuration simule le cas où les imposteurs ne connaissent pas le mot de passe des clients.

configuration : TOUS cette configuration regroupe tous les tests des deux configurations précédentes (**MDP** et **FAUX**).

Les configurations de tests imposteurs **MDP** et **FAUX** comptent chacune le même nombre de tests : 37 800 tests imposteurs.
Dans la configuration **TOUS**, le nombre de tests est doublé : 75 600 tests imposteurs.

Segmentation phonétique



Annexe B

Algorithme d'Espérance Maximisation

La présentation de l'algorithme EM qui suit s'inspire de (Acero, 1990).

Inégalité de Jensen

L'inégalité de Jensen s'applique aux fonctions convexes, elle s'applique également aux fonctions concaves, comme le logarithme, si elle est inversée. C'est cette forme qui est utilisée dans notre démonstration de l'algorithme EM.

La fonction logarithme ln vérifie :

$$x - 1 \geq \ln x, \forall x \in \mathbb{R}_+^* \tag{B.1}$$

Soit X une variable aléatoire et f_X et g_X deux fonctions de densités de probabilité (probability density function - pdf). Alors :

$$\int f_X(x)dx \; = \; \int g_X(x)dx \; = \; 1 \tag{B.2}$$

et

$$f_X(x) \geq 0 \text{ et } g_X(x) \geq 0 \tag{B.3}$$

Or par nature, f et g sont définies sur $\mathbb{R}$ et admettent une primitive sur $\mathbb{R}$, alors par linéarité, la fonction $f - g$ admet une primitive sur $\mathbb{R}$ et vérifie :

$$\int \left[g_X(x) - f_X(x) \right] dx = 0 \tag{B.4}$$

Pour tout x tel que $f_X(x) \neq 0$. L'équation B.4 permet d'écrire :

$$\int f_X(x) \left[\frac{g_X(x)}{f_X(x)} - 1 \right] dx = 0 \tag{B.5}$$

De plus, f_X et g_X étant des pdf :

$$\frac{g_X(x)}{f_X(x)} \geq 0 \tag{B.6}$$

D'après les équations B.1 et B.6 :

$$\frac{g_X(x)}{f_X(x)} - 1 \geq \ln \frac{g_X(x)}{f_X(x)} \tag{B.7}$$

La multiplication par un terme positif des deux termes de l'inégalité conserve le sens de celle-ci, et comme indiqué dans B.3, $f_X(x) \geq 0$. L'inéquation B.7 devient :

$$f_X(x) \left[\frac{g_X(x)}{f_X(x)} - 1 \right] \geq f_X(x) \ln \frac{g_X(x)}{f_X(x)} \tag{B.8}$$

Après intégration de l'équation B.7

$$\int f_X(x) \left[\frac{g_X(x)}{f_X(x)} - 1 \right] \geq \int f_X(x) \ln \frac{g_X(x)}{f_X(x)} \tag{B.9}$$

d'après les équations B.8 et B.5 :

$$0 \geq \int f_X(x) \ln \frac{g_X(x)}{f_X(x)} \tag{B.10}$$

d'où

$$0 \geq \int f_X(x) \left[\ln g_X(x) - \ln f_X(x) \right] \tag{B.11}$$

Par linéarité,

$$0 \geq \int f_X(x) \ln g_X(x) - \int f_X(x) \ln f_X(x) \tag{B.12}$$

et finalement :

$$\boxed{\int f_X(x) \ln f_X(x) \geq \int f_X(x) \ln g_X(x)} \tag{B.13}$$

C'est cette forme de l'inégalité de Jensen que nous utilisons dans notre démonstration de l'algorithme EM.

Forme générale de l'algorithme EM

Soit une variable aléatoire X régie par une loi de probabilité f_X dépendant d'un vecteur de paramètres Θ. L'algorithme EM permet d'estimer le paramètre Θ et donc la distribution de probabilité qui génère X.

Cet algorithme procède de façon itérative et assure la convergence vers un maximum local selon le critère du maximum de vraisemblance. Pour estimer ce paramètre, supposons que nous disposons d'un certain nombre, T, d'observations des réalisations

de la variable X, $\mathcal{X} = \{x(1), x(2), ..., x(T)\}$. Si de plus, ces données sont indépendantes entre elles, alors :

$$f(\mathcal{X}|\Theta) = \prod_{t=1}^{T} f_X(x(t)|\Theta) \tag{B.14}$$

où $f(X|\Theta)$ est un estimateur de la vraisemblance du vecteur de paramètres Θ conditionné par les données $\mathcal{X}$. L'estimateur de Θ obtenu par la méthode du maximum de vraisemblance est :

$$\hat{\Theta} = \underset{\Theta}{argmax} f(X|\Theta) \tag{B.15}$$

Lorsque la fonction de vraisemblance $f(X|\Theta)$ est compliquée, l'algorithme EM permet d'estimer Θ en supposant l'existence de paramètres cachés et inconnus. Les données $\mathcal{X}$ ont été observées. Supposons l'existence des données cachées $\mathcal{Y}$ et notons $\mathcal{Z}$ la réunion de ces données : $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. Nous avons :

$$f_Z(z|\Theta) = f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta) \tag{B.16}$$

et

$$f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta) = f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta).f_X(\mathcal{X}|\Theta) \tag{B.17}$$

où $f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta)$ est la probabilité conjointe de X et Y, $f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta)$ est la probabilité de la variable cachée Y sachant les données observées X égales à $\mathcal{X}$, et $f_X(\mathcal{X}|\Theta)$ est la probabilité de la variable observée X. Le logarithme de l'équation B.17 s'écrit :

$$\ln f_X(\mathcal{X}|\Theta) = \ln f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta) - \ln f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta) \tag{B.18}$$

On calcule maintenant l'espérance selon $\mathcal{Y}$ des termes de l'équation B.18 conditionnée par $X = \mathcal{X}$ et le vecteur de paramètres Θ'

$$\ln f_X(\mathcal{X}|\Theta) = E_Y\{\ln f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta)|X = \mathcal{X}, \Theta'\} - E_Y\{\ln f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta)|X = \mathcal{X}, \Theta'\} \tag{B.19}$$

Pour plus de lisibilité nous utiliserons les notations :

$$\begin{aligned} L(\Theta) &= \ln f_X(\mathcal{X}|\Theta) \\ U(\Theta, \Theta') &= E_Y\{\ln f_{XY}(\mathcal{X}, \mathcal{Y}|\Theta)|X = \mathcal{X}, \Theta'\} \\ V(\Theta, \Theta') &= E_Y\{\ln f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta)|X = \mathcal{X}, \Theta'\} \end{aligned}$$

L'équation B.19 prend la forme suivante :

$$L(\Theta) = U(\Theta, \Theta') - V(\Theta, \Theta') \tag{B.20}$$

Les termes $V(\Theta', \Theta')$ et $V(\Theta, \Theta')$ s'écrivent :

$$V(\Theta', \Theta') = \int_{\mathcal{Y}} \underbrace{f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta')}_{f(x)} \underbrace{\ln f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta')}_{lnf(x)} d\mathcal{Y} \tag{B.21}$$

$$V(\Theta,\Theta') = \int_{\mathcal{Y}} \underbrace{f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta')}_{f(x)} \underbrace{\ln f_{Y/X=\mathcal{X}}(\mathcal{Y}|\Theta)}_{lng(x)} d\mathcal{Y} \tag{B.22}$$

L'inégalité de Jensen B.13 permet alors d'écrire :

$$V(\Theta',\Theta') \geq V(\Theta,\Theta') \tag{B.23}$$

S'il est possible de trouver une valeur du vecteur Θ qui permette de maximiser $U(\Theta,\Theta')$ de la façon suivante :

$$U(\Theta,\Theta') \geq U(\Theta',\Theta') \tag{B.24}$$

alors d'après les équations B.20, B.23 et B.24 cette valeur de Θ permet de maximiser la valeur de $L(\Theta)$.

$$L(\Theta) > L(\Theta') \tag{B.25}$$

Le problème qui était jusque là la maximisation $L(\Theta)$ est remplacé par celui, plus simple, de la maximisation de la fonction $U(\Theta,\Theta')$. L'algorithme EM démarre avec une valeur du vecteur de paramètres $\hat{\Theta}^{(0)}$ et on note $\hat{\Theta}^{(n)}$ la valeur de ce vecteur après n itérations. Chaque itération de l'algorithme est décomposable en deux étapes :

1. Estimation de $U(\Theta,\hat{\Theta}^{(n)})$
2. Maximisation $U(\Theta,\hat{\Theta}^{(n)})$, la valeur de Θ qui maximise $U(\Theta,\hat{\Theta}^{(n)})$ devient $\hat{\Theta}^{(n+1)}$

Application aux mixtures de Gaussiennes

Considérons maintenant le cas où la variable aléatoire X suit une loi de probabilité dont la densité est un mélange de Gaussiennes (GMM). Dans ce cas, l'estimation du vecteur de paramètres Θ va être grandement simplifiée.
La variable cachée Y renseigne sur la Gaussienne du mélange qui a généré chaque observation de la variable X. C'est pourquoi cette variable cachée peut être facilement estimée.

La vraisemblance d'une réalisation x_i avec le GMM à N distributions et de paramètres Θ s'écrit :

$$\begin{aligned} f(x_i|\Theta) &= \sum_{j=1}^{N} f(x_i,j|\Theta) \\ &= \sum_{j=1}^{N} f(x_i|j,\Theta).f(j|\Theta) \end{aligned} \tag{B.26}$$

On introduit alors la variable $\beta_{i,j}$ telle que :

$$\beta_{i,j} = \begin{cases} 1 & \text{si } x_i \text{ est émis par j} \\ 0 & \text{sinon} \end{cases} \tag{B.27}$$

Cette variable permet d'exprimer la vraisemblance des données complètes $\mathcal{Z}$ de la façon suivante :

$$f(X, Y|\Theta) = \prod_{i=1}^{T}\prod_{j=1}^{N}[f(j|\Theta).f(x_i|j,\Theta)]^{\beta_{i,j}} \tag{B.28}$$

Le logarithme de l'équation B.28 donne :

$$\ln f(X, Y|\Theta) = \sum_{i=1}^{T}\sum_{j=1}^{N}\beta_{i,j}\left[\ln f(j|\Theta) + \ln f(x_i|j,\Theta)\right] \tag{B.29}$$

La fonction auxiliaire U peut alors s'écrire :

$$\begin{aligned} U(\Theta, \Theta^{(n)}) &= E_Y\{\ln f(\mathcal{X}, \mathcal{Y}|\Theta)|X = \mathcal{X}, \Theta^{(n)}\} \\ &= E_Y\left[\sum_{i=1}^{T}\sum_{j=1}^{N}\beta_{i,j}.\ln f(j|x_i,\Theta) + \beta_{i,j}.\ln f(x_i|\Theta)|\mathcal{X}, \Theta^{(n)}\right] \\ &= \sum_{i=1}^{T}\sum_{j=1}^{N} E_Y\left[\beta_{i,j}|\mathcal{X}, \Theta^{(n)}\right].\ln f(j|x_i,\Theta) + E_Y\left[\beta_{i,j}|\mathcal{X}, \Theta^{(n)}\right].\ln f(x_i|\Theta) \end{aligned} \tag{B.30}$$

Étant donné l'expression de la fonction auxiliaire de l'équation B.30, les deux étapes de l'algorithme EM consistent à

Étape E

estimer la quantité :

$$\begin{aligned} E_Y\left[\beta_{i,j}|\mathcal{X}, \Theta^{(n)}\right] &= 1 \cdot f(\beta_{i,j} = 1|\mathcal{X}, \Theta^{(n)}) + 0 \cdot f(\beta_{i,j} = 0|\mathcal{X}, \Theta^{(n)}) \\ &= f(j|x_i, \Theta^{(n)}) = \frac{f(x_i|j, \Theta^{(n)})\, f(j|\Theta^{(n)})}{f(x_i|\Theta^{(n)})} \end{aligned} \tag{B.31}$$

Étape M

trouver le paramètre ★ (μ_j, σ_j ou w_j) en résolvant :

$$\frac{\partial U}{\partial \bigstar} = 0 \tag{B.32}$$

Les formules finales d'estimation des paramètres $\hat{\mu}_j$, $\hat{\sigma}_j^2$ et $\hat{w}_j$ sont :

$$\hat{\mu}_j = \frac{\sum_{i=1}^{T} x_i f(j|x_i, \Theta^{(n)}) \frac{f(x_i|j,\Theta)}{f(x_i|\Theta)}}{\sum_{i=1}^{T} f(j|x_i, \Theta^{(n)}) \frac{f(x_i|j,\Theta)}{f(x_i|\Theta)}} \tag{B.33}$$

$$(\hat{\sigma}_j)^2 = \frac{\sum_{i=1}^{T} (x_i - \hat{\mu}_j)^2 f(j|x_i, \Theta^{(n)}) \frac{f(x_i|j,\Theta)}{f(x_i|\Theta)}}{\sum_{i=1}^{T} f(j|x_i, \Theta^{(n)}) \frac{f(x_i|j,\Theta)}{f(x_i|\Theta)}} \tag{B.34}$$

$$\hat{w}_j = \frac{\sum_{i=1}^{T} w_j f(j|x_i, \Theta^{(n)}) \frac{f(x_i|j,\Theta)}{f(x_i|\Theta)}}{\sum_{k=1}^{N} \sum_{j=1}^{T} w_k f(k|x_i, \Theta^{(n)} \frac{f(x_i|k,\Theta)}{f(x_i|\Theta)}} \tag{B.35}$$

Remarque : dans l'équation B.35, le dénominateur est ajouté pour rendre la somme des poids des Gaussiennes du GMM égale à 1.

Illustration

Dans un premier temps, considérons que les données observées nous ont permis d'estimer un premier jeu de paramètres correspondant à deux distributions Gausiennes A et B. Cette situation est représentée par la figure B.1.

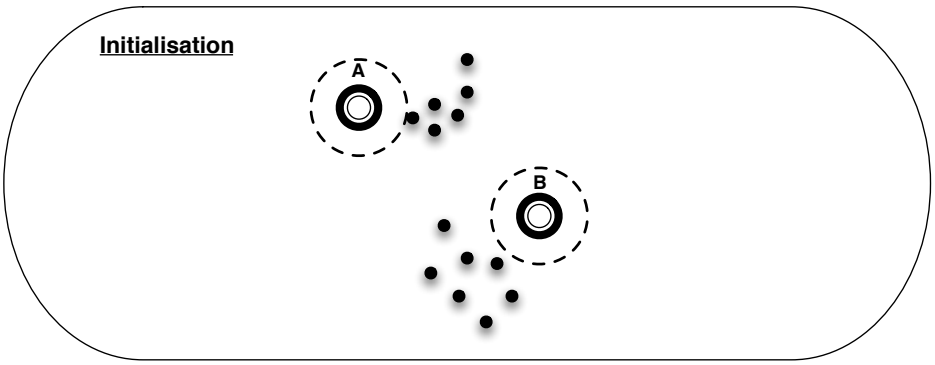


FIG. *B.1:* *Représentation d'un couple de distributions Gaussiennes obtenu par apprentissage sur un jeu de données observées. Ces Gaussiennes peuvent être les distributions choisies pour initialiser l'algorithme EM ou être le résultat obtenu après quelques itérations d'EM.*

Lors de la phase d'estimation, la probabilité pour chaque observation d'être générée

par l'une ou l'autre des Gaussiennes est calculée. Dans cet exemple (figure B.2), chaque observation permet de calculer un couple de valeurs $(p(A), p(B))$ qui sont les probabilités que ce vecteur soir généré par les sus-dites Gaussiennes.

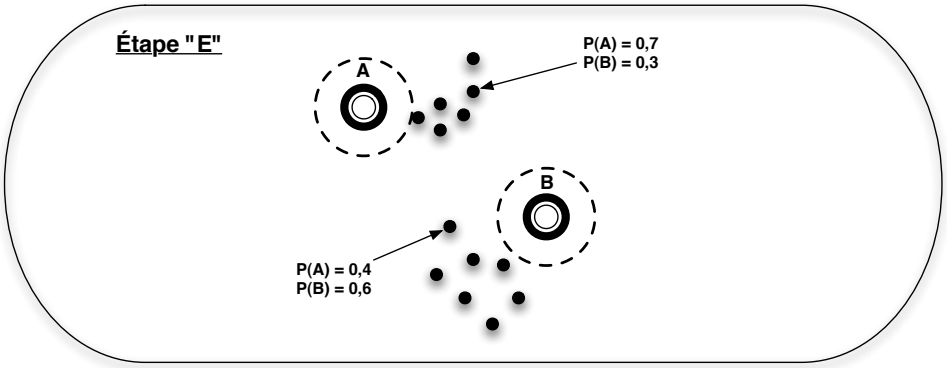


FIG. *B.2:* *Illustration de l'étape d'estimation de l'algorithme EM. Pour chaque observation, la probabilité que celle-ci soit émise par chacune des Gaussiennes (paramètres caché) est calculée.*

Ces estimations sont utilisées durant l'étape de maximisation afin d'estimer les nouveaux paramètres des deux Gaussiennes A et B qui sont alors modifiés. Dans cet exemple, les flèches de la figure B.3 représentent le *déplacement* des moyennes des distributions.

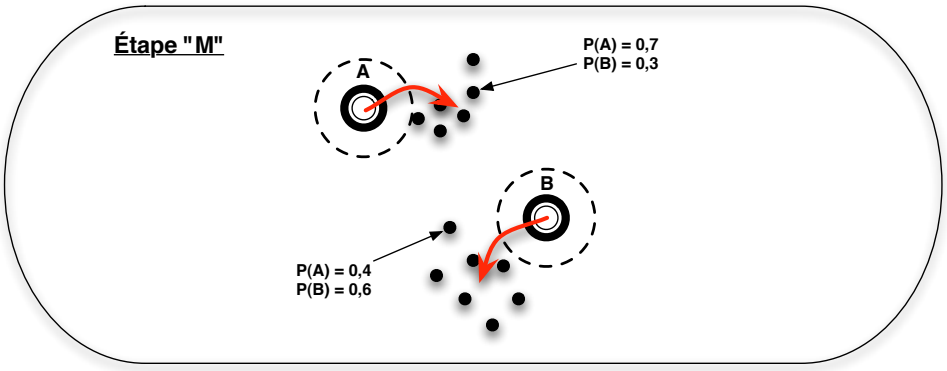


FIG. *B.3:* *Illustration de l'étape de maximisation de l'algorithme EM. Les nouveaux paramètres des Gaussiennes sont estimés d'après les probabilités calculées à l'étape "E"*

Annexe C

Algorithmes *Forward, Backward* et *Forward-Backward*

Algorithme *Forward*

L'algorithme *Forward* est utilisé pour estimer la probabilité d'émission d'une séquence d'observations $\mathcal{O}$ par un HMM de paramètres Λ. Cet algorithme de programmation dynamique permet de réduire la complexité originale de ce calcul, de l'ordre de $2^T \times N^T$, jusqu'à une complexité inférieure, de l'ordre de $N^2 \times T$ où T est le nombre d'observations et N le nombre d'états.

L'algorithme *Forward* se décompose en trois étapes :

1. Initialisation :

$$\alpha_1(i) = \pi_i f_i(o_1) \quad 1 \leq i \leq N \tag{C.1}$$

2. Itération :

$$\alpha_{t+1}(j) = f_j(o_{t+1}) \cdot \sum_{i=1}^{N} \alpha_t(i)\tau_{ij} \quad 1 \leq j \leq N \text{ et } 1 \leq t \leq T-1 \tag{C.2}$$

3. Conclusion :

$$P(O|\Lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{C.3}$$

$\alpha_t(i)$ est la probabilité d'observer la séquence $\{o_0, o_1, ..., o_t\}$ et l'état i au temps t.

Algorithme *Backward*

L'algorithme *Backward* est le pendant de l'algorithme *Forward*. Cet algorithme comporte également trois étapes qui sont :

1. Initialisation :

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \tag{C.4}$$

2. Itération :

$$\beta_t(i) = \sum_{j=1}^{N} \tau_{ij} f_{j(o_{t+1}} \beta_{t+1}(j) \quad 1 \leq 1 \leq N \text{ et } t = T-1, T-2, ..., 1 \tag{C.5}$$

3. Conclusion :

$$P(O|\Lambda) = \sum_{i=1}^{N} \pi_i f_i(o_1) \beta_1(i) \tag{C.6}$$

Algorithme de *Baum-Welch*, dit *Forward-Backward*

L'algorithme de *Baum-Welch* (Baum et al., 1970) est un algorithme d'entraînement itératif qui permet de maximiser la probabilité de génération d'une séquence $\mathcal{O} = \{o_0, o_1, ..., o_T\}$ par un modèle de Markov caché de paramètres Λ. C'est une forme généralisée de l'algorithme *EM*. Comme l'algorithme *EM*, l'algorithme de Baum-Welch converge vers un maximum local.

On associe aux symboles, états et transitions, le nombre de fois où ils sont utilisés pour toutes les séquences et tous les chemins qui sont susceptibles de générer la séquence donnée, pondéré par la probabilité du chemin.

Soit une séquence d'observations $\mathcal{O} = \{o_0, o_1, ..., o_T\}$. L'algorithme de *Baum-Welch* permet de trouver Λ qui maximise la probabilité $P(O|\Lambda)$.

Pour ce faire, deux variables auxiliaires sont introduites en plus des variables *forward* et *backward*. La première, $\xi_t(i,j)$ est la probabilité de transiter de l'état i vers l'état j à l'instant $t+1$ sachant le modèle Λ et les observations $\mathcal{O}$.

$$\xi_t(i,j) = P(a_t = S_i, a_{t+1} = S_j|(O), \Lambda) \tag{C.7}$$

Cette valeur peut être exprimée en fonction des variables *forward* et *backward* définies précédemment sous la forme :

$$\begin{aligned}\xi_t(i,j) &= \frac{\alpha_t(i)\tau_{i,j}f_j(o_{t+1})\beta_{t+1}(j)}{P(\mathcal{O}|\Lambda)} \\ &= \frac{\alpha_t(i)\tau_{i,j}f_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{i,j}\tau_{i,j}f_j(o_{t+1})\beta_{t+1}(j)}\end{aligned}$$

La deuxième variable auxiliaire introduite est la probabilité d'être dans l'état i à l'instant t sachant le modèle Λ et les observations $\mathcal{O}$.

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{C.8}$$

La valeur $\sum_{t=1}^{T}\gamma_t(i)$ est alors l'espérance du nombre de transitions à partir de l'état i.

Les probabilités ré-estimées des transitions entre les états du modèles HMM sont :

$$\hat{\tau}_{i,j} = \frac{\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i)} \tag{C.9}$$

La probabilité d'émettre un symbole k, associée à l'état j est ré-estimée par :

$$\hat{f}_j(k) = \frac{\sum_{t=1_{o_t=k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{C.10}$$

La probabilité Π_i d'être dans l'état i à l'instant initial est :

$$\hat{\Pi}_i = \gamma_1(i) \tag{C.11}$$

Comme l'algorithme *EM*, l'algorithme *forward-backward* assure :

$$P(\mathcal{O}|\hat{\Lambda}_n) \leqslant P(\mathcal{O}|\hat{\Lambda}_{n+1}) \tag{C.12}$$

Annexe D

Le projet BIOBIMO

Présentation

Le projet BIOBIMO [1] (BIOmétrie BImodale sur MObile) est un projet supporté par l'Agence Nationale de la Recherche (ANR) dans le cadre de l'appel à projets 2005 du RNRT (Réseau National de Recherche en Télécommunications). La description suivante est inspirée de l'annexe technique du projet BIOBIMO (Consortium BIOBIMO, 2005).

Objectifs

De nos jours, les terminaux mobiles, tels que PDA ou téléphone portable, sont de plus en plus utilisés. Que ce soit dans le domaine privé ou professionnel, ils fournissent de nombreux services, tels que le rechargement du téléphone, l'accès à Internet, les transactions bancaires, la consultation de comptes. La plupart de ces services nécessitent une vérification de l'identité du client, généralement réalisée en demandant un code PIN ou un mot de passe à l'utilisateur. Ces moyens classiques de vérification d'identité, basés sur la possession d'un secret, présentent de nombreuses faiblesses : ils peuvent facilement être oubliés, découverts ou volés. La multiplication des secrets accentue également la fragilité de ces méthodes, notamment en encourageant les clients à écrire ceux-ci, ou à utiliser le même secret pour plusieurs applications. De ce fait, il apparaît nécessaire en terme de sécurité de faire appel à la biométrie pour déterminer l'identité d'une personne. Actuellement de nombreux terminaux mobiles sont équipés d'un microphone ainsi que d'une caméra.
Ainsi, l'objectif de BIOBIMO est de faire de l'authentification biométrique bimodale, basée sur la voix et le visage, sur des terminaux mobiles. Ce problème d'authentification biométrique est un problème complexe et faire appel à deux modalités non-intrusives, rend le système plus acceptable d'un point de vue utilisateur. Le projet BIOBIMO portera sur la biométrie bimodale conjointe (visage/voix) dans des conditions réelles et

[1] http://biobimo.eurecom.fr/

variées. En effet, la plupart des systèmes biométriques multimodaux effectuent une simple fusion des résultats (scores ou seuil de décisions) obtenus pour chaque modalité sans tenir compte des corrélations temporelles entre les différentes modalités. Au niveau de la communauté, peu de recherches ont été entreprises sur l'utilisation conjointe de deux modalités pour réaliser une authentification.
Les différents thèmes abordés dans ce projet sont : la localisation du visage robuste aux changements d'illumination ou de pose, la segmentation de la parole en environnement bruité, la reconnaissance du visage par le biais de la vidéo (aspect dynamique), la reconnaissance du locuteur et l'authentification bimodale conjointe.
BIOBIMO se propose de développer une application biométrique embarquée sur téléphone mobile. Celle-ci doit donc être adaptée aux ressources disponibles en terme de mémoire, de puissance de calcul, etc.

Les objectifs scientifiques recouvrent les thèmes suivants :
- Localisation robuste du visage (face aux changements d'illumination et de pose) et reconnaissance robuste face aux attaques par « doublage » (« replay attacks »). Face à ces attaques utilisant les images fixes de visage, nous tenterons d'améliorer l'authentification en utilisant l'aspect dynamique de la vidéo, et plus particulièrement les expressions faciales.
- Segmentation de la parole en environnement bruité et vérification du locuteur.
- Authentification bimodale (visage/voix) conjointe tenant compte des corrélations temporelles entre les modalités.

La mise en œuvre des techniques biométriques présente encore des verrous technologiques importants, en particulier au niveau des perturbations dues à l'environnement (bruit, éclairage,...) et au niveau de l'acquisition des références. Ces différents points amènent de grands écarts en terme de performance entre des conditions d'évaluation « type laboratoire » et un environnement réel. La prise en compte du contexte complet d'exploitation (scénario de mise en œuvre) est une nécessité pour répondre à ces questions, ou pour en mesurer l'influence.

Marché et intérêt du projet

La vérification biométrique est un marché qui reçoit actuellement un intérêt particulier aussi bien de la part des industriels que des laboratoires de recherche. Plusieurs produits sont apparus sur le marché, reposant le plus souvent sur les empreintes digitales ou l'iris, qui se montrent fiables mais qui restent mal acceptés par les utilisateurs du fait de leur caractère intrusif. Alors que les modalités visage et parole, de par leur absence de contact et leur discrétion, possèdent un certain avantage par rapport aux autres modalités biométriques. En environnement contraint, les performances de systèmes biométriques basés sur le visage ou la voix sont bonnes mais elles se dégradent fortement dans des conditions réelles. En effet, la reconnaissance de visage montre une grande fragilité aux changements d'éclairage, de pose. Quant à la vérification du locuteur, elle s'accommode encore assez mal d'une prise de son en milieu bruité C'est pourquoi, faire appel à la multimodalité est un axe de recherche intéressant afin d'amé-

liorer la robustesse.

Partenariat

Le consortium associe 3 partenaires complémentaires :

EURECOM, GET , Sophia Antipolis (Jean-Luc Dugelay, Caroline Mallauran).
EURECOM est un laboratoire académique avec des connaissances solides en traitement d'image et en biométrie. Le laboratoire est notamment un spécialiste reconnu en reconnaissance des visages 2D et en 3D.

LIA (EA 931, FRE 2487), Université d'Avignon (Jean-François Bonastre, Corinne Fredouille).
Ce laboratoire académique possède une expertise reconnue en reconnaissance du locuteur (nombreuses publications, nombreuses thèses soutenues ou en cours, participation constante aux campagnes d'évaluations internationales du domaine, organisation du workshop RLA2C 1998, coorganisation des workshops Speaker Odyssey 2001/04/06, du workshop MMUA 2006, fondateur du groupe de travail SPLC de l'ISCA, membre du projet COST 275, membre du NOE BIOSECURE, direction du projet BIO_MUL, direction du projet ALIZE).

E2V-Grenoble Saint Egrève `http://www.e2v.com` E2V-Grenoble mettra en œuvre son expertise de la biométrie et des ressources dans l'algorithme de reconnaissance de personnes et le portage de ces algorithmes sur un processeur ARM.

Bibliographie personnelle

Conférences internationales

A. Larcher, J-F Bonastre et J.S.D. Mason, 2008. From GMM to HMM for Embedded Password-Based Speaker Recognition. Dans les actes de *European Signal and Image Processing Conference (EUSIPCO)*, Lausanne (Suisse).

A. Larcher, J-F Bonastre et J.S.D. Mason, 2008. Reinforced Temporal Structure Information For Embedded Utterance-Based Speaker Recognition. *International Conference on Speech Communication and Technology (Interspeech)*, Brisbane (Australie).

Workshops internationaux

J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, et J. S.D. Mason, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. Dans les actes de *Odyssey Conference - The Speaker and Language Recognition Workshop*. http ://mistral.univavignon.fr

D. Matrouf, J-F Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren et F. Huenupan, 2008. LIA GMM-SVM system description : NIST SRE 2008. Dans les actes de *NIST Speaker Recognition Evaluation Workshop*, Montreal (Canada).

A. Larcher, J-F Bonastre et J.S.D. Mason, 2008. Short utterance-based video aided speaker recognition. Dans les actes de *IEEE International workshop on Multimedia Signal Processing (MMSP)*, 897-901, Cairns (Australie).

Conférences nationales

A. Larcher, J-F Bonastre et J.S.D. Mason, 2008. Utilisation de la structure de mots de passe personnalisés pour la reconnaissance de locuteurs embarquée. Dans les actes de *Journees d'Études sur le Parole (JEP)*, Avignon (France).

S. Meigner, T. Merlin, C. Lévy, A. Larcher, E. Charton, J-F Bonastre, L. Besacier, J. Farinas et B. Ravera, 2008. Mistral : Plate-forme open source d'authentification biométrique. Dans les actes de *Journees d'Études sur le Parole (JEP)*, Avignon (France).

Liste des illustrations

Liste des tableaux

Glossaire

ACP Analyse en Composantes Principales. 72
ADN Acide Désoxyribo-Nucléique. 26
ANR Agence Nationale de la Recherche. 15

CANCOR CANonical CORrelation. 84
CHMM Coupled Hidden Markov Model. 92
CMAP Constrained Maximum A Posteriori. 151
COIA CO-Interia Analysis. 84

DCT Discrete Cosine Transform. 51
DET Detection Error Tradeoff. 38
DTW Dynamic Time Warping. 66

EBGM Elastic Bunch Graph Matching. 71
EER Equal Error Rate. 40
EGM Elastic Graph Matching. 71
EM Algorithme d'Espérance Maximisation. 227

FA Fausses Acceptations. 37
FFT Fast Fourier Transform. 51
FHMM Fused Hidden Markov Model. 92
FR Faux Rejets. 37

GDW Gaussian Dynamic Warping. 67
GLDS Generalized Linear Discriminant Sequence Kernel. 60
GMM Gaussian Mixture Model. 112

HMM Hidden Markov Model. 132

LDA Linear Discriminant Analysis. 75
LIA Laboratoire d'Informatique d'Avignon. 62

LPC Linear Predictive Coefficient. 50

LPCC Linear Prediction Cepstral Coefficients. 50

MAP Maximum A Posteriori. 116

MFCC Mel Frequency Cepstral Coefficient. 51

MLE Maximum Likelihood Estimation. 136

MMIE Maximum Mutual Information Estimation. 136

MSSO Méthodes Statistiques du Second Ordre. 55

NAP Nuisance Attribute Prediction. 60

NIST National Institute of Standards and Technology. 61

NIST-SRE National Institute of Standards and Technology - Speaker Recognition Evaluation. 55

PDA Personal Digital Assistant. 45

PLP Perceptual Linear Predictive. 51

RAL Reconnaissance Automatique du Locuteur. 15

RAP Reconnaissance Automatique de la Parole. 17

RASTA RelAtive SpecTrAl. 51

ROC Receiver Operating Characteristic. 38

ROI Region Of Interest. 78

SCHMM Semi-Conituous Hidden Markov Model. 136

SMM Segment Mixture Model. 55

SVM Support Vector Machine. 58

TRAPS TempoRAl PatternS. 53

UBM Universal Background Model. 114

VAD Voice Activity Detection. 53

VQ Vector Quantization. 57

VTLN Vocal Track Length Normalisation. 51

Bibliographie

(Abry et Lallouache, 1991) C. Abry et M. T. Lallouache, 1991. Audibility and stability of articulatory movements : Deciphering two experiments on anticipatory rounding in French. Dans les actes de *International Congress of Phonetic Sciences (ICPhS)*, Volume 1, 220–225. 84

(Acero, 1990) A. Acero, 1990. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Thèse de Doctorat, Department of Electrical and Computer Engineering Carnegie Mellon University, Pittsburgh, Pennsylvania (USA). 227

(Achermann et Bunke, 1996) B. Achermann et H. Bunke, 1996. Combination of classifiers on the decision level for face recognition. Rapport technique, Institut d'Informatique et de Mathématiques Appliqués de l'Université de Bern. 89

(Adami, 2007) A. G. Adami, 2007. Modeling prosodic differences for speaker recognition. *Speech Communication 49*(4), 277–291. 48

(Adjoudani et Benoît, 1995) A. Adjoudani et C. Benoît, 1995. Audio-Visual Speech Recognition Compared Across Two Architectures. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 85

(Andre-Obrecht et Jacob, 1997) R. Andre-Obrecht et B. Jacob, 1997. Direct identification vs. correlated models to process acoustic and articulatory informations in automatic speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2, Munich (Germany), 999–1002. 84

(Aubert, 2002) X. Aubert, 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer speech & language 16*(1), 89–114. 48

(Auckenthaler et al., 2000) R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. Score Normalization for Text-Independent Speaker Verification System. *Digital Signal Processing 1*(10), 42–54. 40

(Bahl et al., 1986) L. Bahl, P. Brown, P. De Souza, et R. Mercer, 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 11. 136, 150, 159

(Bailly-Bailliere et al., 2003) E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, et al., 2003. The BANCA database and evaluation protocol. *Lecture Notes in Computer Science (LNCS) 2688*, 625–638. 103, 221

(Barker et Berthommier, 1999) J. P. Barker et F. Berthommier, 1999. Estimation of Speech Acoustics from Visual Speech Features : A Comparison of Linear and Non-Linear Models. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*, Santa Cruz (USA), 112–117. ISCA. 83, 162

(Baum et al., 1970) L. E. Baum, T. Petrie, G. Soules, et N. Weiss, 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*(1), 164–171. 132, 136, 237

(Belhumeur et al., 1997) P. N. Belhumeur, J. P. Hespanha, et D. J. Kriegman, 1997. Eigenfaces vs. Fisherfaces : recognition using class specific linear projection. *IEEE transactions on Pattern Analysis and Machine intelligence 19*(7), 711–720. 75

(Ben-Yacoub et al., 1999) S. Ben-Yacoub, Y. Abdeljaoued, et E. Mayoraz, 1999. Fusion of face and speech data for person identity verification. *IEEE Transactions on Neural Networks 10*(5), 1065–1074. 89

(Bengio, 2003a) S. Bengio, 2003a. *Audio-and Video-Based Biometric Person Authentication*, Volume 2688/2003, Chapter Multimodal Authentication Using Asynchronous HMMs, 1056. Springer. 83

(Bengio, 2003b) S. Bengio, 2003b. Multimodal authentication using asynchronous HMMs. Dans Springer-Verlag (Ed.), *International Conference of Audio and Video-Based Person Authentication, AVBPA*, Guildford (UK), 770–777. 85, 93, 196

(Bengio, 2004) S. Bengio, 2004. Multimodal speech processing using asynchronous Hidden Markov Models. *Information Fusion 5*(2), 81–89. 84

(Benoît et al., 1991) C. Benoît, T. Lallouache, T. Mohamedi, A. Tseva, et C. Abry, 1991. Nineteen ($\pm$Two) French Visemes for Visual Speech Synthesis. Dans les actes de *The ESCA Workshop on Speech Synthesis*. ISCA. 80

(BenZeghiba et Bourlard, 2006) M. F. BenZeghiba et H. Bourlard, 2006. User-customized password speaker verification using multiple reference and background models. *Speech Communication 48*(9), 1200–1213. 63, 65, 195

(Besacier et al., 2000) L. Besacier, J.-F. Bonastre, et C. Fredouille, 2000. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication 31*(2-3), 89–106. 53, 157, 205

(Bimbot et al., 2004) F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, et D. A. Reynolds, 2004. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing 4*, 430–451. 50, 53, 112, 195

(Bimbot et al., 1995) F. Bimbot, I. Magrin-Chagnolleau, et L. Mathan, 1995. Second-order statistical measures for text-independent speaker identification. *Speech Communication 17*(1-2), 177–192. 55, 112

(Binnie et al., 1974) C. Binnie, A. Montgomery, et P. Jackson, 1974. Auditory and visual contributions to the perception of consonants. *Journal of Speech, Language and Hearing Research 17*(4), 619. 79

(Boehringer et al., 2006) S. Boehringer, T. Vollmar, C. Tasse, R. Wurtz, G. Gillessen-Kaesbach, B. Horsthemke, et D. Wieczorek, 2006. Syndrome identification based on 2D analysis software. *European Journal of Human Genetics 14*, 1082–1089. 71

(Boite et al., 2000) R. Boite, H. Bourlard, et T. Dutoit, 2000. *Traitement de la parole*. PPUR presses polytechniques. 50

(Bonastre et al., 2003) J.-F. Bonastre, P. Morin, et J.-C. Junqua, 2003. Gaussian Dynamic Warping (GDW) Method Applied to Text-Dependent Speaker Detection and Verification. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Geneva (Switzerland). 68, 98, 137, 196

(Bonastre et al., 2008) J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, et J. S. Mason, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. Dans les actes de *Speaker and Language Recognition Workshop (IEEE Odyssey)*. http ://mistral.univ-avignon.fr/. 99, 197

(Bregler et Konig, 1994) C. Bregler et Y. Konig, 1994. Eigenlips for robust speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2, Adelaïde (Australia), 669–672. 74
english

(Burget et al., 2007) L. Burget, P. Matejka, P. Schwarz, O. Glembek, et J. Cernocky, 2007. Analysis of feature extraction and channel compensation in GMM speaker recognition system. *IEEE Transactions on Audio, Speech, and Language Processing 15*(7), 1979–1986. 119

(Bürki et al., 2008) A. Bürki, C. Gendrot, G. Gravier, G. Linares, et C. Fougeron, 2008. Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa. *TAL*. 167, 212

(Campbell et al., 2006a) W. Campbell, J. Campbell, D. Reynolds, E. Singer, et P. Torres-Carrasquillo, 2006a. Support vector machines for speaker and language recognition. Dans les actes de *Computer Speech & Language*, Volume 20, 210–229. Elsevier. 60

(Campbell et al., 2006b) W. Campbell, D. Sturim, et D. Reynolds, 2006b. Support Vector Machines Using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters 13*(5), 308. 60

(Carey et Parris, 1992) M. Carey et E. Parris, 1992. Speaker verification using connected words. *Proc. Institute of Acoustics 14*(6), 96–100. 114

(Cetingul et al., 2006) H. E. Cetingul, E. Erzin, Y. Yemze., et A. M. Tekalp, 2006. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal processing 86*(12), 3549–3558. 196

(Charlet, 1997) D. Charlet, 1997. *Authentification vocale du locuteur ‡ travers le rÈseau tÈlÈphonique*. Thèse de Doctorat, ENST. 65

(Chellappa et al., 1995) R. Chellappa, C. L. Wilson, et S. Sirohey, 1995. Human and machine recognition of faces, a survey. *Proceedings of the IEEE 83*(5), 705–741. 70

(Chen et al., 2001) L.-F. Chen, H.-Y. M. Liao, et J.-C. Lin, 2001. Person identification using facial motion. Dans les actes de *IEEE International Conference on Image Processing*, Volume 2, Singapore, 677–680. 77

(Chetty et Wagner, 2004a) G. Chetty et M. Wagner, 2004a. Automated lip feature extraction for liveness verification in audio-video authentication. Dans les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*, Jeju (Korea), 17–22. 175

(Chetty et Wagner, 2004b) G. Chetty et M. Wagner, 2004b. "Liveness" verification in Audio-Video Authentication. Dans les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*, Jeju (Korea). 94

(Chetty et Wagner, 2005) G. Chetty et M. Wagner, 2005. Liveness detection using cross-modal correlations in face-voice person authentication. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, 2181–2184. 93, 196

(Chibelushi et al., 1997) C. C. Chibelushi, J. S. D. Mason, et F. Deravi, 1997. Feature-level data fusion for bimodal person recognition. Dans les actes de *Sixth International Conference on Image Processing and Its Applications*. 196

(Consortium BIOBIMO, 2005) Consortium BIOBIMO, 2005. Projet BIOBIMO, Annexe Technique. Rapport technique, Université d'Avignon et des Pays de Vaucluse. 239

(Cunado et al., 1997) D. Cunado, M. S. Nixon, et J. N. Carter, 1997. *Using gait as a biometric, via phase-weighted magnitude spectra*. Springer. 26

(Daugman, 2000) J. Daugman, 2000. Biometric decision landscape. Rapport technique, University of Cambridge, Computer Laboratory. 89

(Dean et al., 2008a) D. Dean, P. Lucey, S. Sridharan, et T. Wark, 2008a. Fused-HMM Adaptation of Synchronous HMMs for Audio-Visual Speech Recognition. *Digital Signal Processing 1051-2004*. 83

(Dean et al., 2008b) D. Dean, S. Sridharan, et P. Lucey, 2008b. Cascading Appearance-Based Features for Visual Speaker Verification. Dans les actes de *International Conference on Speech Communication and Technology*. 78

(Dean et al., 2006) D. Dean, S. Sridharan, et T. Wark, 2006. Audio-visual speaker verification using continuous fused HMMs. Dans les actes de *HCSNet workshop on Use of*

vision in human-computer interaction, Volume 56, 87–92. Australian Computer Society, Inc. Darlinghurst, Australia. 92

(Delacretaz et Hennebert, 1998) D. Delacretaz et J. Hennebert, 1998. Text-prompted speaker verification experiments with phonemespecific MLPs. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2. 63

(Deleglise et al., 1996) P. Deleglise, A. Rogozan, et M. Alissali, 1996. Asynchronous integration of audio and visual sources in bimodal automatic speech recognition. Dans les actes de *European Signal and Image Processing Conference (EUSIPCO)*, 10–13. 84

(Doddington et al., 2000) G. R. Doddington, M. A. Przybockib, A. F. Martin, et D. A. Reynolds, 2000. The NIST speaker recognition evaluation - Overview, Methodology, Systems, Results, Perspective. *Speech communication 31*(2-3), 225–254. 55, 94

(Doledec et Chessel, 1994) S. Doledec et D. Chessel, 1994. Co-inertia analysis : An alternative method for studying species-environment relationships. *Freshwater biology 31*, 277–294. 84, 87, 216

(Dong et al., 2005) L. Dong, S. W. Foo, et Y. Lian, 2005. A two-channel training algorithm for Hidden Markov Model and its application to lip reading. *EURASIP Journal on Applied Signal Processing 2005*(9), 1382–1399. 79

(Dongmei et al., 2002) J. Dongmei, X. Lei, Z. Rongchun, W. Verhelst, I. Ravyse, et H. Sahli, 2002. Acoustic viseme modelling for speech driven animation : a case study. Dans les actes de *MPCA*, Volume 1, Leuven. 79

(Eide et Gish, 1996) E. Eide et H. Gish, 1996. A parametric approach to vocal tract length normalization. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Atlanta (USA). 51

(Enqing et al., 2002) D. Enqing, L. Guizhong, Z. Yatong, et Z. Xiaodi, 2002. Applying support vector machines to voice activity detection. Dans les actes de *International Conference on Signal Processing*, Volume 2. 54

(Erzin et al., 2005) E. Erzin, Y. Yemez, et A. M. Tekalp, 2005. Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia 7*(5), 840–852. 103

(Eveno et Besacier, 2005) N. Eveno et L. Besacier, 2005. A Speaker Independent "Liveness" Test for Audio-Visual Biometrics. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa (Portugale). 84, 162, 168, 175, 179

(Eveno et al., 2004) N. Eveno, A. Caplier, et P.-Y. Coulon, 2004. Accurate and quasi-automatic lip tracking. *IEEE Transactions on Circuits and Systems for Video Technology 14*, 706–715. 88, 93, 175

(Faraj et Bigun, 2007) M.-I. Faraj et J. Bigun, 2007. Audio-visual person authentication using lip-motion from orientation maps. *Pattern recognition letters 28*, 1368–1382. 87, 196

(Fauve et al., 2008) B. Fauve, H. Bredin, W. Karam, F. Verdet, A. Mayoue, G. Chollet, J. Hennebert, R. Lewis, J. Mason, C. Mokbel, et D. Petrovska, 2008. Some results from the biosecure talking face evaluation campaign. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 4137–4140. 80

(Fauve et al., 2007) B. Fauve, D. Matrouf, N. Scheffer, J. Bonastre, et J. Mason, 2007. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Transactions on Audio, Speech, and Language Processing 15*(7), 1960–1968. 60, 99

(Ferrer et al., 2007) L. Ferrer, E. Shriberg, S. Kajarekar, et K. Sonrnez, 2007. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 4. 49

(Fine et al., 2001) S. Fine, J. Navratil, et R. A. Gopinath, 2001. A Hybrid Gmm/Svm Approach To Speaker Identification. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1, Salt Lake City (USA), 417–420. 58

(Fisher, 1968) C. G. Fisher, 1968. Confusions among visually perceived consonants. *Journal of Speech, Language and Hearing Research 11*(4), 796. 79

(Foo et al., 2003) S. Foo, Y. Lian, et L. Dong, 2003. A two-channel training algorithm for hidden Markov model to identify visual speech elements. Dans les actes de *International Symposium on Circuits and Systems*, Volume 2. ISCA. 79, 80

(Forney, 1973) G. D. Forney, 1973. The Viterbi Algorithm. *Proceedings of the IEEE 61*(3), 268–278. 133

(Fox et Reilly, 2003) N. Fox et R. B. Reilly, 2003. Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features. Dans les actes de *International Conference of Audio and Video-Based Person Authentication, AVBPA*, Guildford (UK). 86, 87

(Fox et al., 2005) N. A. Fox, B. A. O'Mullane, et R. B. Reilly, 2005. The Realistic Multimodal VALID database and Visual Speaker Identification Comparison Experiments. Dans les actes de *International Conference of Audio and Video-Based Person Authentication, AVBPA*, New York (US). 88, 103

(Fredouille, 2000) C. Fredouille, 2000. *Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances*. Thèse de Doctorat, Université d'Avignon, Avignon, FRANCE. 51

(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 les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Volume 2, Budapest (Hungary), 983–986. 40

(Furui, 1981) S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing] 29*(2), 254–272. 50, 51, 52

(Furui, 1986) S. Furui, 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing 34*(1), 52–59. 48

(Gagnon et al., 2001) L. Gagnon, P. Stubley, et G. Mailhot, 2001. Password-Dependent Speaker Verification Using Quantized Acoustic Trajectories. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1, Salt Lake City (USA), 449–452. 63

(Garcia-Salicetti et al., 2003) S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Jardins, J. Lunter, Y. Ni, et D. Petrovska-Delacretaz, 2003. BIOMET : A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities. *Lecture Notes in Computer Science 2688/2003*, 845–853. 103, 221

(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 les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2, Adelaide (Australia), 291–298. 116

(Georghiades et al., 2001) A. S. Georghiades, P. N. Belhumeur, et D. J. Kriegman, 2001. From Few to Many : Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE transactions on Pattern Analysis and Machine intelligence 6*, 643–660. 73

(Gittins, 1985) R. Gittins, 1985. *Canonical Analysis : A Review with Applications in Ecology*. Springer-Verlag. 88

(Goecke, 2005) R. Goecke, 2005. 3D Lip Tracking and Co-Inertia Analysis for Improved Robustness of Audio-Video Automatic Speech Recognition. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*, Vancouver Island (Canada). 88, 175

(Goecke et Asthana, 2008) R. Goecke et A. Asthana, 2008. A Comparative Study of 2D and 3D Lip Tracking Methods for AV ASR. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*. 70

(Goecke et Millar, 2003) R. Goecke et B. Millar, 2003. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*, St Jorioz (France). 83, 84, 162

(Goecke et al., 2000) R. Goecke, B. Millar, A. Zelinsky, et J. Robert-Ribes, 2000. Automatic extraction of lip feature points. Dans les actes de *Australian Conference on Robotics and Automation, ACRA*, Melbourne (Australia), 31–36. 103, 175

(Goff et Benoit, 1996) B. L. Goff et C. Benoit, 1996. A text-to-audiovisual-speech synthesizer for french. Dans les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*. ISCA. 79

(Hansen et Bria, 1990) J. Hansen et O. Bria, 1990. Lombard effect compensation for robust automatic speech recognition in noise. Dans les actes de *First International Conference on Spoken Language Processing*. ISCA. 94

(Haton et al., 2006) J.-P. Haton, C. Cerisara, D. Fohr, Y. Laprie, et K. Smaili, 2006. *Reconnaissance automatique de la parole, du signal ‡ son inteprÈtation*. Dunod. 44

(Hazen, 2006) T. J. Hazen, 2006. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 14*(3), 1082–1089. 93

(Hazen et al., 2004) T. J. Hazen, K. Saenko, C.-H. La, et J. R. Glass, 2004. A segment-based audio-visual speech recognizer : data collection, development, and initial experiments. Dans les actes de *International Conference on Multimodal Interfaces*, 235–242. 93

(Hebert et Heck, 2003) M. Hebert et L. P. Heck, 2003. Phonetic class-based speaker verification. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, 1665–1668. 195

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustic Society of America 87*, 1738–1752. 51

(Hermansky et al., 1991) H. Hermansky, N. Morgan, A. Bayya, et P. Kohn, 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 51

(Hermansky et al., 1992) H. Hermansky, N. Morgan, A. Bayya, et P. Kohn, 1992. RASTA-PLP speech analysis technique. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 51

(Hermansky et Sharma, 1999) H. Hermansky et S. Sharma, 1999. Temporal patterns (TRAPs) in ASR of noisy speech. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 53

(Higgins et al., 1991) A. Higgins, L. Bahler, et J. Porter, 1991. Speaker verification using randomized phrase prompting. Dans les actes de *Digital Signal Processing*, Volume 1, 89–106. 40, 114

(Hjelmas et Low, 2001) E. Hjelmas et B. K. Low, 2001. Face detection : A survey. *Computer Vision and Image Understanding 83*(3), 236–274. 70

(Ho et al., 1992) T. K. Ho, J. J. Hull, et S. N. Srihari, 1992. Combination of decisions by multiple classifiers. Dans les actes de *Structured Document Image Analysis*, 188–202. Springer-Verlag. 89

(Hollien et al., 1974) H. Hollien, W. Majewski, et P. Hollien, 1974. Perceptual identification of voices under normal, stress, and disguised speaking conditions. *The Journal of the Acoustical Society of America 56*, S53. 28

(Hotelling, 1936) H. Hotelling, 1936. Relations between two sets of variates. *Biometrika 28*(3-4), 321–377. 84

(Huang et Trivedi, 2002) K. Huang et M. Trivedi, 2002. Streaming face recognition using multicamera video arrays. Dans les actes de *International Conference on Pattern recognition*, Volume 16, 213–216. 74

(Ikedo, 1998) J. Ikedo, 1998. Voice activity detection using neural network. *IEICE Transactions on Communications 81*(12), 2509–2513. 54

(Israel et al., 2005) S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, et B. K. Wiederhold, 2005. ECG to identify individuals. *Pattern Recognition 38*(1), 133–142. 27

(Iyengar et al., 1995) S. Iyengar, L. Prasad, et H. Min, 1995. *Advances in Distributed Sensor Technology*. Prentice-Hall. 85

(Jaakkola et Haussler, 1999) T. S. Jaakkola et D. Haussler, 1999. Exploiting generative models in discriminative classifiers. *Advances in Neural Informations Processing Systems 11*, 487–493. 60

(Jain et al., 2004) A. K. Jain, A. Ross, et S. Prabhakar, 2004. An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on 14*(1), 4–20. 35

(Jain et al., 1999) D. A. K. Jain, R. Bolle, et S. Pankanti, 1999. *Biometrics : Personal Identification in Networked Society*. Kluwer Academic Publishers. 25

(Jourlin, 1998) P. Jourlin, 1998. *Approche Bimodale du Traitement Automatique de la Parole : application à la Reconnaissance du Message et du Locuteur*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse. 93

(Jourlin et al., 1997) P. Jourlin, J. Luettin, D. Genoud, et H. Wassner, 1997. Acoustic-labial speaker verification. *Pattern Recognition Letters 18*(9), 853–858. 83, 85

(Kenny et al., 2005) P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel, 2005. Factor analysis simplified. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 74, 120

(Kenny et Dumouchel, 2004) P. Kenny et P. Dumouchel, 2004. Disentangling speaker and channel effects in speaker verification. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 37–40. 119

(Kenny et al., 2003) P. Kenny, M. Mihoubi, et P. Dumouchel, 2003. New MAP estimators for speaker recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 119

(Kittler et al., 1998) J. Kittler, M. Hatef, R. P. Duin, et J. Matas, 1998. On combining classifiers. *IEEE transactions on Pattern Analysis and Machine intelligence 20*(3), 226–239. 88

(Kuhn et al., 1998) R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, et M. Contolini, 1998. Eigenvoices for speaker adaptation. Dans les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*, Sydney (Australia), 1771–1774. 74, 118

(Lades et al., 1993) M. Lades, J. C. Vorbruggen, Joachim, J. Lange, C. von der Malsburg, R. P. Wurtz, et W. Konen, 1993. Distortion invariant object recognition in the dynamic linkarchitecture. *Computers, IEEE Transactions on 42*(3), 300–311. 71

(Larnel et al., 1991) L. Larnel, J. Gauvain, et M. Eskenazi, 1991. BREF, a Large Vocabulary Spoken Corpus for French. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 223

(Leggetter et Woodland, 1995) C. J. Leggetter et P. C. Woodland, 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language 9*(2), 171–185. 67

(Lévy et al., 2006) C. Lévy, G. Linares, P. Nocera, et J.-F. Bonastre, 2006. *Mobile Phone Embedded Digit-Recognition*, Chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2, 71–84. Springer Sciences. 98, 137, 196

(Li et Porter, 1998) K.-P. Li et J. E. Porter, 1998. Normalizations and selection of speech segments for speaker recognition scoring. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1, New York (USA), 595–598. 40

(Lievin et Luthon, 1998) M. Lievin et F. Luthon, 1998. Lip Features Automatic Extraction. Dans les actes de *IEEE International Conference on Image Processing*, Volume 3, Chicago (USA), 168–172. 175

(Liu et Chen, 2003) X. Liu et T. Chen, 2003. Video-Based Face Recognition Using Adaptive Hidden Markov Models. Dans les actes de *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 1, Madison (USA), 340–345. 79

(Louradour, 2007) J. Louradour, 2007. *Noyaux de séquences pour la vérification du locuteur par Machines à Vecteurs de Support*. Thèse de Doctorat, Université Toulouse III - Paul Sabatier. 60, 67

(Louradour et Daoudi, 2005) J. Louradour et K. Daoudi, 2005. SVM speaker verification using a new sequence kernel. Dans les actes de *European Signal and Image Processing Conference (EUSIPCO)*. 60

(Luettin et Thacker, 1997) J. Luettin et N. A. Thacker, 1997. Speechreading using probabilistic models. *Computer Vision and Image Understanding 65*(2), 163–178. 45

(Lyngso et Pedersen, 2001) R. B. Lyngso et C. N. Pedersen, 2001. Complexity of comparing hidden Markov models. 416–428. Springer. 67

(Lyngso et al., 1999) R. B. Lyngso, C. N. Pedersen, et H. Nielsen, 1999. Metrics and similarity measures for hidden Markov models. Dans les actes de *International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 178–186. 67

(Mahalanobis, 1936) P. C. Mahalanobis, 1936. *On the generalized distance in statistics*, Volume 2. 74

(Maison et al., 1999) B. Maison, C. Neti, A. Senior, I. Center, et Y. Heights, 1999. Audio-visual speaker recognition for video broadcast news : some fusion techniques. Dans les actes de *IEEE International workshop on Multimedia Signal Processing*, 161–167. 89

(Mami et Charlet, 2004) Y. Mami et D. Charlet, 2004. Représentation compacte des locuteurs par distribution sur les modèles d'ancrage. Dans les actes de *Proceedings of journees d'etudes sur le parole, JEP*, Fes (Maroc). 58

(Marcel et del R. Millan., 2007) S. Marcel et J. del R. Millan., 2007. Person Authentication using Brainwaves (EEG) and Maximum A Posteriori Model Adaptation. *IEEE transactions on Pattern Analysis and Machine intelligence 29(4)*, 743–748. 27

(Mariethoz, 2006) J. Mariethoz, 2006. *Algorithmes d'apprentissage discriminants en vérification du locuteur*. Thèse de Doctorat, Lyon II Lumière. 40

(Marti et Bunke, 2002) U. Marti et H. Bunke, 2002. The IAM-database : an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition 5*(1), 39–46. 104, 221

(Martin et Przybocki, 2000) A. Martin et M. Przybocki, 2000. The NIST 1999 speaker recognition evaluation - An overview. *Digital Signal Processing 10*(1-3), 1–18. 39

(Martin et al., 1997) A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, et M. A. Przybocki, 1997. The DET Curve in Assessment of Detection Task Performance. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 38

(Martinez et Kak, 2001) A. M. Martinez et A. C. Kak, 2001. PCA versus LDA. *IEEE transactions on Pattern Analysis and Machine intelligence 23*(2), 228–233. 75

(Mason et al., 1989) J. S. Mason, J. Oglesby, et L.-Q. Xu, 1989. Codebooks to optimise speaker recognition. Dans les actes de *First European Conference on Speech Communication and Technology*. ISCA. 57

(Mason et al., 2005) J. S. D. Mason, N. W. D. Evans, R. Stapert, et R. Auckenthaler, 2005. Data-model relationship in text-independent speaker recognition. *EURASIP Journal on Applied Signal Processing 4*, 471–481. 121, 200

(Matrouf et al., 2008) D. Matrouf, J.-F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren, et F. Huenupan, 2008. LIA GMM-SVM system description : NIST SRE08. Dans les actes de *NIST Speaker Recognition Evaluation Workshop*, Montreal (Canada). 29, 60, 61

(Matrouf et al., 2007) D. Matrouf, N. Scheffer, B. Fauve, et J.-F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *International Conference on Speech Communication and Technology*. 62

(Matsui et Furui, 1993) T. Matsui et S. Furui, 1993. Concatenated phoneme models for text-variable speaker recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2, Minneapolis (USA), 391–394. 40, 63, 65

(Matta, 2008) F. Matta, 2008. *Video person recognition strategies using head motion and facial appearance*. Thèse de Doctorat, University of Nice Sophia-Antipolis (UNSA). 77

(Matta et Dugelay, 2006) F. Matta et J.-L. Dugelay, 2006. Person recognition using human head motion information. Dans les actes de *Conference on Articulated Motion and Deformable Objects*, Andratx (Spain), 326–335. 78

(Matthews et al., 1998) I. Matthews, T. Cootes, S. Cox, R. Harvey, et J. A. Bangham, 1998. Lipreading using shape, shading and scale. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*, Sydney, Australia, 73–78. ISCA. 85

(McCool et Marcel, 2009) C. McCool et S. Marcel, 2009. Parts-based face verification using local frequency bands. Dans les actes de *IEEE IAPR International Conference on Biometrics (ICB)*. 76

(Meigner et al., 2008) S. Meigner, T. Merlin, C. Lévy, A. Larcher, E. Charton, J.-F. Bonastre, L. Besacier, J. Farinas, et B. Ravera, 2008. Mistral : Plate-forme open source d'authentification biométrique. Dans les actes de *Proceedings of journees d'etudes sur le parole, JEP*. 99

(Merlin et al., 1999) T. Merlin, J.-F. Bonastre, et C. Fredouille, 1999. Non directly acoustic process for costless speaker recognition and indexation. Dans les actes de *International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communications*, Neuchâtel (Switzerland). 58

(Messer et al., 1999) K. Messer, J. Matas, J. Kittler, J. Luettin, et G. Maitre, 1999. XM2VTSDB : The Extended M2VTS Database. Dans les actes de *International Conference of Audio and Video-Based Person Authentication, AVBPA*, Volume 626. 103, 221

(Minematsu et al., 2003) N. Minematsu, K. Yamauchi, et K. Hirose, 2003. Automatic estimation of perceptual age using speaker modeling techniques. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 48

(Moghaddam et Pentland, 1997) B. Moghaddam et A. Pentland, 1997. Probabilistic visual learning for object representation. *IEEE transactions on Pattern Analysis and Machine intelligence 19*, 696–710. 74

(Monrose et Rubin, 2000) F. Monrose et A. D. Rubin, 2000. Keystroke dynamics as a biometric for authentication. *Future Gener Comput Syst 16*(4), 351–359. 26

(Morishima et al., 2002) S. Morishima, S. Ogata, K. Murai, et S. Nakamura, 2002. Audio-visual speech translation with automatic lip synchronization and face tracking based on 3-D head model. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 2. 79

(Myers et al., 1980) C. Myers, L. R. Rabiner, et A. E. Rosenberg, 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-28*(6), 623–635. 66

(Nakagawa et al., 2004) S. Nakagawa, Z. Wei, et M. Takahashi, 2004. Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 5, Montreal (Canada). 65

(Navratil et al., 2000) J. Navratil, U. V. Chaudhari, et S. H. Maes, 2000. A speech biometrics system with multigrained speaker modeling. Dans les actes de *Conference for Natural Speech Processing*. 195

(Nefian et al., 2003) A. V. Nefian, L. H. Liang, T. Fu, et X. X. Liu, 2003. *A Bayesian Approach to Audio-Visual Speaker Identification*, Volume 2688-1. Springer Berlin / Heidelberg. 92

(Oglesby, 1995) J. Oglesby, 1995. What's in a number ? Moving beyond the equal error rate. *Speech Communication 17*(1-2), 193–208. 38

(Orr et Abowd, 2000) R. J. Orr et G. D. Abowd, 2000. The smart floor : a mechanism for natural user identification and tracking. Dans les actes de *Conference on Human Factors in Computing Systems*, New York (USA), 275–276. ACM. 26

(Ortega-Garcia et al., 2003) J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J. Igarza, C. Vivaracho, et al., 2003. MCYT baseline corpus : a bimodal biometric database. *IEEE Proceedings on Vision, Image and Signal Processing 150*(6), 395–401. 104, 221

(Patterson et Gowdy, 2003) E. Patterson et J. Gowdy, 2003. An audio-visual approach to simultaneous-speaker speech recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 5. 87

(Phillips et al., 2005) P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, et W. Worek, 2005. Overview of the face recognition grand challenge. Dans les actes de *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 1. 80

(Phillips et al., 2006) P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, W. Worek, N. I. of Standards, et T. (US, 2006. Preliminary Face Recognition Grand Challange Results. Dans les actes de *International Conference on Automatic Face and Gesture Recognition (FGR)*. US Dept. of Commerce, National Institute of Standards and Technology. 29, 80

(Potamianos et al., 2001) G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, et A. Verma, 2001. A cascade visual front end for speaker independent automatic speechreading. *International Journal of Speech Technology 4*(3), 193–208. 78

(Pouchoulin et al., 2007) G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, et J. Revis, 2007. Characterization of the pathological voices (dysphonia) in the frequency space. Dans les actes de *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, Volume 16, 6–10. 48

(Preti, 2008) A. Preti, 2008. *Surveillance de rÈseaux professionnels de communication par la reconnaissance du locuteur*. Thèse de Doctorat, UniversitÈ d'Avignon et des Pays de Vaucluse. 54

(Przybocki et al., 2007) M. A. Przybocki, A. F. Martin, et A. N. Le, 2007. NIST speaker recognition evaluations utilizing the mixer corpora - 2004, 2005, 2006. *IEEE Transactions on Audio, Speech and Language Processing 15*(7), 1951–1959. 195

(Quenot, 1992) G. M. Quenot, 1992. The orthogonal algorithm for optical flow detection using dynamic programming. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 3. 76

(Rabiner, 1989) L. R. Rabiner, 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE 77*(2), 257–286. 63, 64

(Rabiner et al., 1978) L. R. Rabiner, A. E. Rosenberg, et S. E. Levinson, 1978. Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*(6), 575–582. 66

(Reynolds, 1994) D. A. Reynolds, 1994. Experimental evaluation of features for robust speaker identification. *Speech and Audio Processing, IEEE Transactions on 2*(4), 639–643. 50

(Reynolds, 1995) D. A. Reynolds, 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication 17*(1-2), 91–108. 114

(Reynolds, 1996) D. A. Reynolds, 1996. The effects of handset variability on speaker recognitionperformance : experiments on the Switchboard corpus. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 40

(Reynolds et al., 2000) D. A. Reynolds, T. F. Quatieri, et R. B. Dunn, 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing 10*, 19–41. 112, 116

(Reynolds et Rose, 1995) D. A. Reynolds et R. C. Rose, 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Acoustics, Speech and Signal Processing* *3*(1), 72–83. 55, 112

(Robert-Ribes, 1995) J. Robert-Ribes, 1995. *Modèles d'intégration audiovisuelle de signaux linguistiques*. Thèse de Doctorat, Institut National Polytechnique de Grenoble (INPG). 85

(Rodriguez et al., 2007) R. Rodriguez, N. Evans, R. Lewis, B. Fauve, et J. Mason, 2007. An experimental study on the feasibility of footsteps as a biometric. Dans les actes de *European Signal and Image Processing Conference (EUSIPCO)*. 26

(Rodríguez et al., 2008) R. V. Rodríguez, R. P. Lewis, J. S. Mason, et N. W. Evans, 2008. Footstep recognition for a samrt home environment. *International Journal of Smart Home* *2*(2), 95–110. 26

(Rogozan et Deléglise, 1998) A. Rogozan et P. Deléglise, 1998. Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication* *26*(1-2), 149–161. 84

(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 les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*, 599–602. 114

(Rosenberg et al., 1991) A. E. Rosenberg, C. Lee, et S. Gokcen, 1991. Connected word talker verification using whole word Hidden Markov Models. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 381–384. 64 anglais

(Rouas et al., 2005) J.-L. Rouas, J. Farinas, F. Pellegrino, et R. AndrÈ-Obrecht, 2005. Rhythmic unit extraction and modelling for automatic language identification . *Speech Communication* *47*(4), 436–456. ?OLDEditeur(Speech Communication, Elsevier). 48

(Saeed et al., 2006) U. Saeed, F. Matta, et J.-L. Dugelay, 2006. Person recognition based on head and mouth dynamics. Dans les actes de *IEEE International workshop on Multimedia Signal Processing*, Victoria (Canada). 78

(Sáenz-Lechón et al., 2006) N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, et P. Gómez-Vilda, 2006. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control* *1*(2), 120–128. 48

(Sanderson, 2008) C. Sanderson, 2008. *Biometric Person Recognition : Face, Speech and Fusion*. VDM-Verlag. 103

(Sanderson et Paliwal, 2004) C. Sanderson et K. K. Paliwal, 2004. Identity verification using speech and face information. *Digital Signal Processing* *14*(5), 449–480. 85

(Sanderson et al., 2005) C. Sanderson, M. Saban, et Y. Gao, 2005. On local features for GMM based face verification. Dans les actes de *International Conference on Information Technology and Applications*, 638–643. 75

(Satoh, 2000) S. Satoh, 2000. Comparative evaluation of face sequence matching for content-based video access. Dans les actes de *IEEE International Conference on Automatic Face and Gesture Recognition*, 163–168. 74, 75

(Scheffer, 2006) N. Scheffer, 2006. *Structuration de l'espace acoustique par le modèles générique pour la vérification du locuteurs*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse. 53, 205

(Schwarz et al., 2006) P. Schwarz, P. Matejka, et J. Cernocky, 2006. Hierarchical structures of neural networks for phoneme recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 53

(Schwerdt et Crowley, 2000) K. Schwerdt et J. L. Crowley, 2000. Robust face tracking using color. Dans les actes de *IEEE International Conference on Automatic Face and Gesture Recognition*, 90–95. 70

(Sharma et Mammone, 1996) M. Sharma et R. Mammone, 1996. Subword based text dependent speaker verification system with user selectable password. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Atlanta (USA). 63

(Shimodaira et al., 2001) H. Shimodaira, K. ichi Noma, M. Nakai, et S. Sagayama, 2001. Dynamic Time-Alignment Kernel in Support Vector Machine. Dans les actes de *Neural Information Processing Systems*, Vancouver. 67

(Singer et al., 2003) E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, et D. Reynolds, 2003. Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, 1345–1348. ISCA. 48

(Siracusa et Fisher, 2007) M. R. Siracusa et J. W. Fisher, 2007. Dynamic dependency tests for audio visual speaker association. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Hawaï (USA). 83, 162

(Sönmez et al., 1997) M. Sönmez, L. Heck, M. Weintraub, et E. Shriberg, 1997. A lognormal tied mixture model of pitch for prosody based speaker recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA. 49

(Soong et al., 1985) F. K. P. Soong, A. Rosenberg, L. Rabiner, et B. Juang, 1985. A Vector Quantization Approach to Speaker Recognition. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 10, Tampa (USA), 387–390. 57, 75

(Soong et Rosenberg, 1988) F. K. P. Soong et A. E. Rosenberg, 1988. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing] 36*(6), 871–879. 52

(Starpert et Mason, 2001) R. P. Starpert et J. S. Mason, 2001. A segmental mixture model for speaker recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Volume 4, Aalborg (Denmark), 2509–2512. 55

(Steffens et al., 1998) J. Steffens, E. Elagin, et H. Neven, 1998. PersonSpotter - Fast and Robust System for Human Detection, Tracking and Recognition. Dans les actes de *International Conference on Automatic Face and Gesture Recognition*, 516–521. 70, 72

(Stolcke et al., 2005) A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, et A. Venkataraman, 2005. MLLR Transforms as Features in Speaker Recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa (Portugale), 2425–2428. 67

(Sumby et Pollack, 1954) W. Sumby et I. Pollack, 1954. Visual contribution to speech intelligibility in noise. *Journal of Acoustic Society of America 26*, 212. 79

(Tan et al., 2006) X. Tan, S. Chen, Z. Zhou, et F. Zhang, 2006. Face recognition from a single image per person : A survey. *Pattern Recognition 39*(9), 1725–1745. 29

(Technical Specification Digital, European Telecommunications Standards Institute (ETSI), 2005) Technical Specification Digital, European Telecommunications Standards Institute (ETSI), 2005. Speech Processing, Transmission and Quality aspects (STQ) ; Distributed speech recognition ; Extended advanced front-end feature extraction algorithm ; Compression algorithms ; Back-end speech reconstruction algorithm. [ES 202 212 v1.1.2 (2005-11)]. 54

(Teoh et al., 2004) A. Teoh, S. A. Samad, et A. Hussain, 2004. Nearest Neighbourhood Classifiers in Biometric Fusion. *International Journal of the computer, ther internet and management 2 12*(1), Teoh04. 89

(Teunen et al., 2000) R. Teunen, B. Shahshahani, et L. Heck, 2000. A model-based transformational approach to robust speaker recognition. Dans les actes de *Proceedings International Conference on Spoken Language Processing, ICSLP*. ISCA. 119

(Torres et Vilà, 2002) L. Torres et J. Vilà, 2002. Automatic face recognition for video indexing applications. *Pattern Recognition 35*(3), 615–625. 74

(Tremain, 1982) T. Tremain, 1982. The government standard linear predictive coding algorithm : LPC-10. *Speech Technology 1*(2), 40–49. 50

(Turk et Pentland, 1991a) M. A. Turk et A. P. Pentland, 1991a. Eigenfaces for Face Detection/Recognition. *Journal of Cognitive Neuroscience 3*, 71–86. 72, 118

(Turk et Pentland, 1991b) M. A. Turk et A. P. Pentland, 1991b. Face recognition using eigenfaces. Dans les actes de *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaï (USA), 586–591. 72

(Vapnik, 1998) V. Vapnik, 1998. *Statistical learning theory*. John Wiley & Sons. 59

(Verlinde et Cholet, 1999) P. Verlinde et G. Cholet, 1999. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application. Dans les actes de *International Conference of Audio and Video-Based Person Authentication, AVBPA*, 188–193. Citeseer. 85, 89

(Verlinde et al., 2000) P. Verlinde, G. Chollet, et M. Acheroy, 2000. Multi-modal identity verification using expert fusion. *Information Fusion 1*(1), 17–33. 88

(Verma et al., 2003) A. Verma, N. Rajput, et L. Subramaniam, 2003. Using viseme based acoustic models for speech driven lip synthesis. Dans les actes de *International Conference on Multimedia and Expo*, Volume 3. 80

(Viterbi, 1967) A. J. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269. 132, 133

(Wan et Campbell, 2000) V. Wan et W. M. Campbell, 2000. Support Vector Machines for Speaker Verification and Identification. Dans les actes de *IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, Volume 2, Sydney (Australia), 775–784. 58

(Wan et Carmichael, 2005) V. Wan et J. Carmichael, 2005. Polynomial Dynamic Time Warping Kernel Support Vector Machines for Dysarthric Speech Recognition with Sparse Training Data. Dans les actes de *International Conference on Speech Communication and Technology*, Lisboa. 67

(Wang et al., 2008) Y. Wang, Z. Wu, L. Cai, et H. M. Meng, 2008. Modeling the Synchrony between Audio and Visual Modalities for Speaker Identification. Dans les actes de *Phonetic Conference of China and the International Symposium on Phonetic Frontiers*, Beijing, China. 93

(Wark et Sridharan, 2001) T. Wark et S. Sridharan, 2001. Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification. *Digital Signal Processing 11*(3), 169–186. 89

(Wiskott et al., 1997) L. Wiskott, J.-M. Fellous, N. Krüger, et C. von der Malsburg, 1997. Face Recognition by Elastic Bunch Graph Matching. *IEEE transactions on Pattern Analysis and Machine intelligence 19*, 775–779. 71, 175

(Yang et al., 2002) M.-H. Yang, D. J. Kriegman, et N. Ahuja, 2002. Detecting Faces in Images : A Survey. *IEEE transactions on Pattern Analysis and Machine intelligence 24*(1), 34–58. 70

(Yehia et al., 1997) H. Yehia, P. Rubin, et E. Vatiokis-Bateson, 1997. Quantitative association of orofacial and vocal-tract shapes. Dans les actes de *International Conference on Audio-Visual Speech Processing, AVSP*, Rhodes (Greece). 83, 162

(Young, 1992) S. J. Young, 1992. The general use of tying in phoneme-based HMM speech recognisers. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1, San Francisco (USA), 569–572. 136, 196, 203

(Zavaliagkos, 1995) G. Zavaliagkos, 1995. *Maximum A Posteriori Adaptation Techniques For Speech Recognition*. Thèse de Doctorat, Northeastern University. 118

(Zhang et al., 2002) X. Zhang, R. M. Mersereau, et M. A. Clements, 2002. Audio-Visual Speech Recognition by Speechreading. Dans les actes de *International Conference on Digital Signal Processing*, Volume 2, Island of Santorini (Thera), 1069–1072. 83

(Zhao et al., 2003) W. Zhao, R. Chellappa, P. Phillips, et A. Rosenfeld, 2003. Face recognition : A literature survey. *ACM Computing Surveys (CSUR) 35*(4), 399–458. 70

(Zissman et Singer, 1994) M. A. Zissman et E. Singer, 1994. Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. Dans les actes de *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Volume 1. 48