HAL
open science

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

Anthony Larcher

To cite this version:

Anthony Larcher. Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée. Autre [cs.OH]. Université d'Avignon, 2009. Français. NNT : 2009AVIG0170 . tel-00453645

HAL Id: tel-00453645

https://theses.hal.science/tel-00453645v1

Submitted on 5 Feb 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 :

Mme	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

Résumé **9**

Abstract **11**

Introduction **15**

I Introduction à la biométrie **21**

1 De l'individu à la biométrie **23**
1.1 Un individu - une identité 24
1.2 Les biométries 25
1.3 Biométrie et systèmes automatiques 27
1.4 Applications et tâches biométriques 30

2 Description générale des systèmes de vérification biométrique d'identité **33**
Introduction 34
2.1 Structure de la phase d'enrôlement 34
2.2 Structure de la phase de test 35
2.3 Quel résultat ? 37

II La parole en biométrie **41**

Introduction **43**

3 Vérification automatique du locuteur **47**
Introduction 48
3.1 Extraction d'information du signal de parole 49
3.2 Vérification du locuteur non-structurale 54
3.3 Vérification du locuteur structurale 62
Conclusion 68

4 Reconnaissance visuelle de personnes **69**
Introduction 70

4.1 La vidéo, un signal à 2+1 dimensions 71
4.2 La vidéo, un signal temporel 76
Conclusion 80

5 Authentification bi-modale audio-visuelle 81
Introduction 82
5.1 Audio et Vidéo, un lien étroit 83
5.2 Bi-Modalité et fusion 85
5.3 Traitement conjoint des modalités audio et vidéo 91
Conclusion 94

III Vérification du locuteur et synchronisation contrainte 95

Introduction 97

6 Corpus et protocole expérimental 101
Introduction 102
6.1 Contraintes fixées 102
6.2 Bases de données existantes 102
6.3 La base de données MyIdea 104
6.4 Protocole expérimental 106
Conclusion 110

7 Représentation des locuteurs 111
7.1 Le paradigme GMM/UBM 112
7.2 Place des modèles de locuteurs dans l'espace acoustique 115
7.3 Performances des systèmes GMM/UBM 120
Conclusion 126

8 Structuration temporelle de la séquence acoustique 129
Introduction 130
8.1 Modélisation des mots de passe 131
8.2 Apprentissage itératif des modèles de mot de passe 138
8.3 Améliorations dues à la structuration du modèle acoustique 142
8.4 Exploiter pleinement l'architecture à trois niveaux 155
Conclusion 158

9 Renforcement de la structure temporelle par une contrainte de synchronisation 161
Introduction 162
9.1 Intégration d'une information temporelle externe 163
9.2 Validation expérimentale avec un alignement phonétique 167
9.3 Retour sur la structuration temporelle des vidéo 174
9.4 Calcul d'une synchronisation vidéo dans le cadre de nos contraintes . . 175
9.5 Validation expérimentale 177
Conclusion 178

Conclusion et perspectives **183**

Long Abstract **195**
1 Introduction 195
2 Approach overview 197
3 Corpus and protocol 198
4 Baseline System 200
5 Extensions of the GMM/UBM paradigm 202
6 Conclusion and Future Works 216

Annexes **221**

A Base de données MyIdea **221**

B Algorithme d'Espérance Maximisation **227**

C Algorithmes *Forward*, *Backward* et *Forward-Backward* **235**

D Le projet BIOBIMO **239**

Bibliographie personnelle **243**

Liste des illustrations **247**

Liste des tableaux **250**

Glossaire **252**

Bibliographie **253**

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

1.1 Un individu - une identité 24
1.1.1 Une identité pour reconnaitre l'individu 24
1.1.2 Différentes définitions de l'identité 24
1.2 Les biométries 25
1.2.1 La biométrie morphologique 25
1.2.2 La biométrie comportementale 26
1.2.3 Les biométries mixtes 27
1.3 Biométrie et systèmes automatiques 27
1.4 Applications et tâches biométriques 30
1.4.1 Identification 30
1.4.2 Vérification d'identité 31

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 ».