Vérification de l'identité d'un visage parlant. Apport de la mesure de synchronie audiovisuelle face aux tentatives délibérées d'imposture.

Hervé Bredin

20 novembre 2007

Pour Mum – un petit peu, mais pas trop !
Pour Dad – finalement convaincu ?
Pour Carine – parce que.

Résumé

La *sécurité* des personnes, des biens ou des informations est l'une des préoccupations majeures de nos sociétés actuelles. L'*authentification de l'identité des personnes* est l'un des moyens permettant de s'en assurer. La principale faille des moyens actuels de vérification d'identité est qu'ils sont liés à ce qu'une personne *possède* (un passeport, un badge magnétique, etc.) et/ou ce qu'elle *sait* (un code PIN de carte bancaire, un mot de passe, etc.). Or, un badge peut être volé, un mot de passe deviné ou cassé par force algorithmique brute. La *biométrie* est le domaine technologique traitant de la *vérification d'identité* et/ou de l'*identification* de personnes par leurs caractéristiques physiques individuelles, pouvant être morphologiques ou morpho-comportementales. Elle apparaît comme une solution évidente au problème soulevé précédemment : l'identité d'une personne est alors liée à *ce qu'elle est* et non plus à ce qu'elle possède ou sait.

En plus d'être une des modalités biométriques les moins intrusives et donc plus facilement acceptée par le grand public, la vérification d'identité basée sur les visages parlants est intrinsèquement multimodale : elle regroupe à la fois la reconnaissance du visage, la vérification du locuteur et une troisième modalité relevant de la synchronie entre la voix et le mouvement des lèvres.

La première partie de notre travail est l'occasion de faire un tour d'horizon de la littérature portant sur la biométrie par visage parlant et nous soulevons le fait que les protocoles d'évaluation classiquement utilisés ne tiennent pas compte des tentatives délibérées d'imposture. Pour cela, nous confrontons un système de référence (basé sur la fusion des scores de vérification du locuteur et du visage) à quatre types d'imposture délibérée de type *rejeu* et mettons ainsi en évidence les faiblesses des systèmes actuels.

Dans la seconde partie, nous proposons d'étudier la synchronie audiovisuelle entre le mouvement des lèvres acquis par la caméra et la voix acquise par le microphone de façon à rendre le système de référence robuste aux attaques. Plusieurs nouvelles mesures de synchronie basées sur l'analyse de corrélation canonique et l'analyse de co-inertie sont présentées et évaluées sur la tâche de détection d'asynchronie. Les bonnes performances obtenues par la mesure de synchronie basée sur un modèle dépendant du client nous encouragent ensuite à proposer une nouvelle modalité biométrique basée sur la synchronie audiovisuelle. Ses performances sont comparées à celle des modalités *locuteur* et *visage* et sa robustesse intrinsèque aux attaques de type *rejeu* est mise en évidence. La complémentarité entre le système de référence et la nouvelle modalité *synchronie* est soulignée et des stratégies de fusion originales sont finalement mises en place de façon à trouver un compromis entre les performances brutes du premier et la robustesse de la seconde.

Abstract

Authenticating people is a means to ensure the safety of people, goods or sensitive information, which is one of the major concerns of our modern societies. The main weakness of current authentication systems is that they rely on what a person owns (a passport, a magnetic card, etc.) and/or what he/she knows (a PIN number, a password, etc.). Still, a card can be stolen and a password broken.

Biometrics is the technological field dealing with authentication and/or identification of people using their physical characteristics, including morphological and behavioral measurements. This happens to be an obvious solution to the issue previously highlighted : the identity of a person is then related to who he/she is and no longer to what he/she owns or knows.

Not only is *talking face* one of the less intrusive biometric modality, it is also intrinsically multimodal : it includes both speaker and face verification, and a third modality related to audiovisual speech synchrony between the voice and lip motion.

In the first part of this document, we overview the literature about the talking-face biometric modality and we underline that deliberate impostor attacks are often forgotten in the development process of talking-face authentication algorithms. We simulate four kinds of deliberate impostor replay attacks in order to uncover the main weakness of classical systems based on the fusion of speaker and face verification scores.

In the second part, we propose to study the audiovisual synchrony between voice and lip motion as a way of making a classical *speaker+face* robust to attacks. Several novel synchrony measures based on canonical correlation analysis and co-inertia analysis are introduced and tested on the asynchrony detection task. The promising results that we obtained with a client-dependent synchrony measure led us to introduce a novel biometric modality based on audiovisual synchrony. Though it is not as efficient as speaker and face verification, this new modality is intrinsically robust to deliberate impostor attacks. We finally point out the complementarity between the *speaker+face* reference system and the synchrony modality and introduce novel fusion strategies that allow to achieve a good compromise between the efficiency of the former and the robustness of the latter.

Remerciements

Je tiens tout d'abord à remercier Mme Sylvie Lelandais-Bonade et M. Gérard Bailly pour avoir accepté de rapporter sur mes travaux. Je remercie par la même occasion Mmes Régine André-Obrecht et Delphine Charlet, ainsi que MM. Farzin Deravi et Chafic Mokbel pour avoir accepté de participer à mon jury de thèse.

Je remercie chaleureusement mon directeur de thèse, M. Gérard Chollet, qui m'a fait confiance tout au long de ces trois années de thèse. Il a su créer des conditions de travail idéales, alliant une très grande liberté et la participation à de nombreux projets à l'échelle nationale ou européenne. Il n'a jamais rechigné à me laisser partir à l'autre bout de la planète pour présenter mes travaux : mes différents voyages en Australie, en Chine, en Inde et même à Hawaii (pour ne citer que les destinations les plus exotiques) en sont la preuve éclatante. Pour tout cela, je le remercie vivement.

Ces trois années auraient été bien longues sans les différents collègues avec qui j'ai partagé mon bureau. Merci en particulier à Leila et Rémi que j'ai côtoyés pendant la plus grande partie de mes trois années de thèse. J'ai comme l'impression que vous allez me manquer ! J'ajouterai une petite pensée pour Patricia et Catherine qui m'ont grandement facilité la vie à maintes reprises.

Ce rapport de thèse ne serait pas ce qu'il est aujourd'hui sans les nombreuses (et fastidieuses, si si !) relectures qui en ont été faites. Merci à Marc et Gérard pour leurs remarques sages et avisées. Un très grand merci à Rémi pour avoir épluché avec le plus grand soin les pages qui suivent. Mille mercis à Émilie, Fanny et Nicolas pour avoir osé se plonger dans ce charabia à la recherche de coquilles et autres croustillantes formulations dont j'ai le secret.

Je finirai par un petit mot à l'attention de ma petite Lili qui a su être (très) patiente quand je n'y voyais plus très clair. Merci, merci, merci…

Table des matières

Table des figures

Liste des tableaux

Acronymes

AHMM Modèle de Markov caché asynchrone – *Asynchronous Hidden Markov Model*

CANCOR Analyse de corrélation canonique – *Canonical Correlation Analysis*

CoIA Analyse de co-inertie – *Co-Inertia Analysis*

DCF Fonction de coût de détection – *Detection Cost Function*

DCT Transformée en cosinus discrète – *Discrete Cosine Transform*

DET Courbe de détection – *Detection Error Tradeoff*

DFFS Distance à l'espace de visage – *Distance From Face Space*

EER Taux d'égale erreur – *Equal Error Rate*

EM *Expectation Maximization*

FA Fausse Acceptation

FAR Taux de fausse acceptation – *False Acceptance Rate*

fps Images par seconde – *frame per second*

FR Faux Rejet

FRR Taux de faux rejet – *False Rejection Rate*

GMM Modèle de mélange de gaussiennes – *Gaussian Mixture Model*

HMM Modèle de Markov caché – *Hidden Markov Model*

ICA Analyse en composantes indépendantes – *Independent Component Analysis*

LPC *Linear-Predictive Coding*

LSF *Line Spectral Frequencies*

MAP Maximum A Posteriori

MFCC *Mel-Frequency Cepstral Coefficients*

MLP Réseau de neurones multi-couches – *Multiple Layer Perceptron*

NN Réseau de neurones – *Neural Network*

PCA Analyse en composantes principales – *Principal Components Analysis*

PIN Numéro d'identification personnel – *Personal Identification Number*

RMS Valeur efficace – *Root Mean Square*

ROI Région d'intérêt – *Region Of Interest*

SIFT *Scale Invariant Feature Transform*

SVM Machine à vecteur de support – *Support Vector Machine*

UBM Modèle du monde – *Universal Background Model*

WER Taux d'erreur pondéré – *Weighted Error Rate*

Introduction générale

La *sécurité* des personnes, des biens ou des informations est l'une des préoccupations majeures de nos sociétés actuelles. L'*authentification de l'identité des personnes* permet de s'en assurer. Ainsi, une personne désirant traverser une frontière sensible se verra systématiquement demander de décliner et prouver son identité à l'aide de son passeport par exemple ; une autre voulant accéder à un service bancaire sur l'Internet devra la plupart du temps saisir un nom d'utilisateur et le mot de passe correspondant. La grande faiblesse des moyens actuels de vérification d'identité apparaît clairement ici : l'identité d'une personne est directement liée à *ce qu'elle possède* (un passeport, un badge magnétique, etc.) et/ou *ce qu'elle sait* (un code PIN de carte bancaire, un mot de passe, etc.). Or, un badge peut être volé, un mot de passe deviné ou cassé par force algorithmique brute : ceci menant à l'*usurpation d'identité*.

La *biométrie* est le domaine technologique traitant de la *vérification d'identité* et/ou de l'*identification* de personnes par leurs caractéristiques physiques individuelles, pouvant être morphologiques ou morpho-comportementales. Elle apparaît comme une solution évidente au problème soulevé précédement : l'identité d'une personne est liée à *ce qu'elle est* et non plus à ce qu'elle possède ou sait.

Modalités biométriques Les modalités biométriques *morphologiques* les plus courantes sont obtenues à partir de plusieurs parties du corps humain, telles que l'oeil (pour l'iris et la rétine), la main (pour les empreintes digitales et palmaires ou encore la forme de la main) ou le visage. Cette liste peut être allongée par des modalités moins répandues (voire exotiques) telles que la forme de l'oreille, les vaisseaux sanguins de la main, etc.

FIG. 1 – Modalités biométriques morphologiques

Comme leur nom l'indique, les modalités biométriques *morpho-comportementales* sont liées autant à

la morphologie humaine qu'à la dynamique du comportement. Nous pouvons citer des modalités telles que la voix, la dynamique de la signature, la démarche ou la dynamique de la frappe sur un clavier. À titre d'exemple, les caractéristiques physiques de la voix sont à la fois guidées par le comportement et la morphologie du conduit vocal du locuteur. Il en est de même pour la démarche qui ne saurait être complètement décorrélée de la morphologie du marcheur.

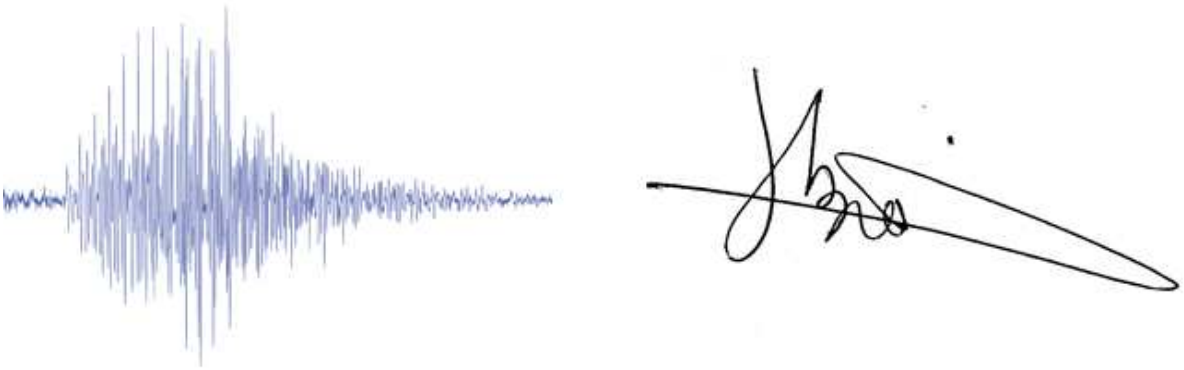


FIG. 2 – Modalités biométriques morpho-comportementales

Vérification biométrique d'identité

Quelle que soit la nature de la modalité utilisée, les systèmes biométriques partagent tous une structure de base commune.

Enrôlement La première étape indispensable à l'utilisation d'un système biométrique par une personne λ est son enrôlement. Il s'agit du processus pendant lequel un échantillon biométrique de la personne est acquis, à partir duquel un modèle λ d'identité de la personne est obtenu et stocké (par exemple, sur un serveur central ou sur une carte à puce que seule la personne λ possède). L'acquisition de l'échantillon biométrique est effectuée de différentes façons selon la modalité : à l'aide d'un appareil photo pour la reconnaissance du visage, un microphone pour la vérification du locuteur ou encore une tablette graphique pour la signature. Cette étape d'enrôlement est résumée schématiquement dans la figure 3.

Vérification d'identité vs. identification Deux applications de la biométrie sont alors envisageables : la vérification d'identité et l'identification. Souvent confondus dans la littérature, ces deux termes n'en décrivent pas moins deux applications différentes :

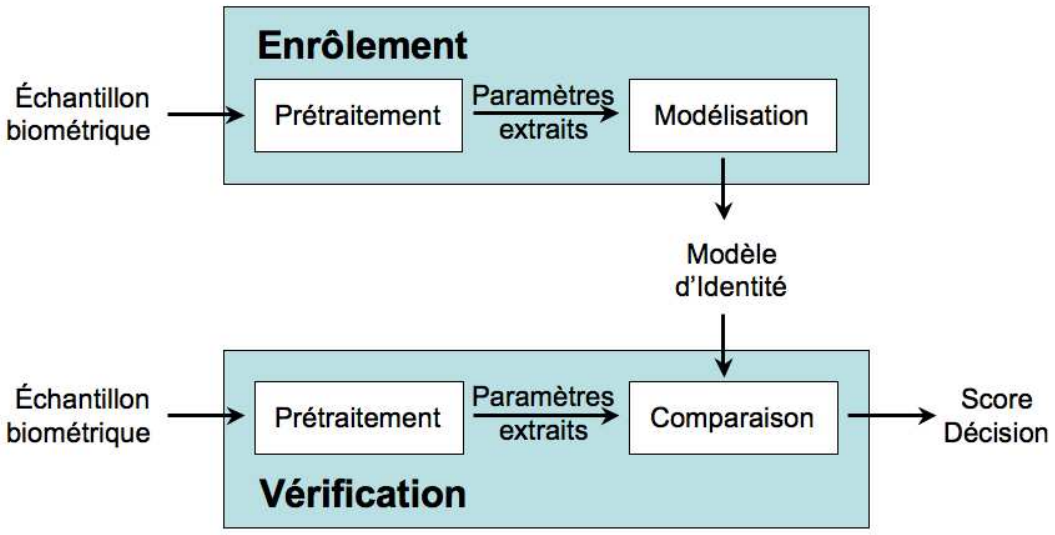


FIG. 3 – Système de vérification biométrique d'identité

– La vérification d'identité consiste à décider si l'identité ϵ, clamée par une personne λ, est correcte. Il s'agit donc de comparer les données biométriques de la personne λ au modèle constitué lors de l'enrôlement de la personne ϵ et de fournir une décision du type *accepté* (si $\lambda = \epsilon$) ou *refusé* (si $\lambda \neq \epsilon$). Une seule comparaison [λ vs. ϵ] est effectuée à chaque accès.
– Quant à l'identification, elle consiste à déterminer si une personne λ est enregistrée dans le système et, le cas échéant, quelle est son identité. La décision attendue est du type *identité* = λ *: accepté* ou *identité inconnue : refusé*. Si N personnes sont enregistrées dans le système, N comparaisons sont effectuées à chaque accès [λ vs. ϵ_i], pour $i \in \{1 \dots N\}$.

Nos travaux se limiteront au cadre de la vérification d'identité. L'étape de vérification est résumée schématiquement et mise en relation avec l'étape d'enrôlement dans la figure 3.

Client et imposteur Au moment du test, deux situations peuvent se produire :

– On parle d'accès légitime ou accès *client* lorsque une personne λ clame sa propre identité λ auprès du système de vérification biométrique.
– On parle d'accès illégitime ou accès *imposteur* lorsqu'une personne λ clame une identité ϵ différente de la sienne ($\epsilon \neq \lambda$). La personne ϵ est la cible de l'imposteur λ.

Quelle modalité pour quelle application ?

Toutes les modalités ne sont pas équivalentes et leur utilisation varie selon l'application visée et les performances souhaitées. Ainsi, un système biométrique destiné à gérer l'accès à une zone contenant des informations très sensibles sera différent d'un système biométrique dit de *confort*, permettant par exemple de jouer dans un salon la musique préférée de la personne reconnue. Plusieurs critères permettent de choisir celle qui est la plus adaptée.

Le critère le plus évident est la **performance**. Certaines modalités sont beaucoup plus performantes que d'autres et obtiennent de très faibles taux d'erreur (citons l'iris et les empreintes digitales, voir le tableau 1).

Modalité	**Ordre de grandeur**
Iris	0.1 %
Empreinte digitale	1 %
Voix	5 %
Visage	10 %

TAB. 1 – Comparaison des taux d'égale erreur – d'après [Ross *et al.*, 2006]

Le **passage à l'échelle** d'un système (*scalability* dans la littérature anglophone) est un point qui doit être mentionné. Ainsi, en fonction de l'application visée et du nombre de personnes enregistrées dans la base de données du système biométrique, les performances de ce dernier peuvent être dégradées aussi bien du point de vue des taux d'erreur que de la rapidité d'exécution. Selon que l'on se pose le problème de la *vérification d'identité* – "la personne correspond-elle au modèle de l'identité clamée ?" – ou de l'*identification de personnes* – "quel modèle y correspond le mieux ?" –, le nombre de comparaisons entre données biométriques et modèles varie énormément. Il est alors important d'associer cette variation à celui du temps d'attente effectif avant la prise de décision finale lorsque le système est mis en application.

Le **coût** de la mise en place d'un système biométrique dépend beaucoup de celui du capteur associé. En effet, un capteur pour la modalité *visage* peut être très bon marché ; en témoigne la multiplication des téléphones portables munis d'un appareil photo. Déjà des téléphones sont proposés avec un système de vérification du visage pour accéder aux fonctionnalités du téléphone. À l'opposé, les capteurs d'iris, beaucoup plus coûteux, ne sont installés que pour protéger l'accès à des lieux ou données dont la sécurité est très sensible.

Le niveau d'**acceptabilité** des modalités par le grand public varie selon les modalités. Ainsi, la captation de l'image de l'iris peut effrayer certaines personnes – "cela va-t-il abîmer mes yeux ?" – et les questions d'hygiène peuvent survenir au moment de passer son doigt sur un capteur d'empreinte digitale – "qui est passé avant moi ?" : il s'agit de modalités dites *intrusives*. Ainsi, la coopération de la personne à identifier est souvent indispensable au bon déroulement du processus de vérification d'identité [Bolle et Pankanti, 1998]. Cependant, certaines modalités (telles que le visage, la voix ou la démarche) peuvent mener à une vérification biométrique *à l'insu* de la personne, résolvant ainsi le problème de la coopération de la personne mais soulevant aussi quelques questions éthiques.

La première étape dite d'*enrôlement* d'une personne consiste à acquérir un ou des échantillon(s) de la modalité de la personne de façon à construire un modèle qui lui sera associé. Cette étape peut parfois échouer (*failure to enrol*, dans la littérature anglophone). À titre d'exemple, environ 2% de la population testée n'a pas pu s'enregistrer en utilisant la modalité *empreinte digitale* dans les travaux reportés dans [Fairhurst *et al.*, 2004]. Il convient donc de considérer le critère d'**universalité** de la modalité biométrique.

Multi-modalité Bien que déjà très performants pris séparément, les systèmes *mono-modaux* (i.e. ne faisant appel qu'à une seule modalité) peuvent mener à un système *multi-modal* encore meilleur lorsqu'ils sont fusionnés [Ross *et al.*, 2006]. Cette amélioration est sensible au niveau du taux d'erreur mais aussi au niveau de l'universalité, l'utilisation de plusieurs modalités limitant de façon drastique l'échec de l'acquisition d'échantillons au moment de l'enrôlement et/ou du test [Jain et Ross, 2002].

Visage parlant

L'échantillon biométrique disponible pour la vérification d'identité par la modalité **visage parlant** est un enregistrement audiovisuel de la personne parlant face à la caméra. En plus d'être l'une des modalités les moins intrusives et donc plus facilement acceptée par le grand public [Bolle et Pankanti, 1998], la vérification d'identité basée sur les visages parlants est intrinsèquement multi-modale : elle inclut en particulier la modalité *visage* et la modalité *voix*. Son coût est, en outre, très faible : une simple *webcam* équipée d'un microphone suffit pour acquérir les échantillons biométriques. La modalité *visage parlant* apparaît donc comme un très bon compromis entre tous les critères définis précédemment.

Plan du document

La première partie de notre exposé sera l'occasion de faire un tour d'horizon de la littérature portant sur la biométrie par visage parlant (chapitre 1). Les protocoles d'évaluation utilisés pour reporter les performances de nos différents algorithmes seront l'objet du chapitre 2. Nous décrirons ensuite le système classique que nous avons développé, basé sur la fusion des modalités *voix* et *visage*, et évaluerons ses performances (chapitre 3). Notre première contribution originale consistera à définir des tentatives délibérées d'imposture (où l'imposteur a acquis au préalable une photographie du visage et un enregistrement de la voix de sa cible) afin de mettre en évidence la principale faiblesse du système initial *voix+visage* (chapitre 4).

Dans la deuxième partie, nous proposerons un moyen de rendre le système de base robuste à ces attaques élaborées. La solution proposée repose sur l'étude de la synchronie audiovisuelle entre la voix et le mouvement des lèvres. Une revue de la littérature du domaine sera l'objet du chapitre 5. Une nouvelle mesure de la synchronie audiovisuelle (et ses quatre variantes) sera introduite et évaluée au chapitre 6. L'étude plus approfondie de sa variante dépendante du client nous mènera à la définition et l'évaluation d'une troisième modalité (après la *voix* et le *visage*) relevant de la synchronie de la parole audiovisuelle (chapitre 7). Dans le dernier chapitre (8), deux nouvelles stratégies de fusion des modalités *voix*, *visage* et *synchronie* seront introduites. Nous montrerons alors comment l'utilisation de cette nouvelle modalité *synchronie* permet d'augmenter la robustesse du système *voix+visage* initial face aux tentatives délibérées d'imposture.

Avertissement

La notion de synchronie audiovisuelle développée dans ce rapport peut prêter à confusion. Là où un lecteur averti pourrait s'attendre à une approche locale visant à détecter des incohérences entre événements acoustiques et visuels, les approches développées ici visent à évaluer un degré global de cohérence entre les flux acoustiques et visuels.

L'évaluation, sur la tâche de détection d'asynchronie, des différentes mesures que nous proposons a pour unique objectif la sélection de la meilleure mesure qui sera utilisée par la suite dans le cadre de la vérification d'identité. Il apparaît clairement que les méthodes locales basées sur la détection d'événements ont leur mot à dire pour la tâche de détection d'asynchronie. En effet, une seule incohérence (au niveau d'une plosive par exemple) suffit à une méthode locale pour détecter des flux asynchrones, alors que cette incohérence sera noyée au milieu du flux pour une mesure globale.

Cependant, l'objectif final est d'obtenir, sur la tâche de vérification d'identité, un degré de vraisemblance plutôt qu'une décision binaire synchrone/asynchrone. Les mesures globales prennent alors tout leur sens : ce sont elles que nous étudierons dans la deuxième partie de ce document.

Première partie

Vérification audiovisuelle de l’identité

Chapitre 1

Tour d'horizon

La vérification d'identité basée sur les visages parlants est souvent introduite dans la littérature sous la dénomination *biométrie audiovisuelle* [Aleksic et Katsaggelos, 2006] ; la plupart des travaux existants ne considérant un visage parlant que comme la fusion des deux modalités *audio* (vérification du locuteur) et *visuelle* (reconnaissance du visage). L'objet de ce chapitre n'est pas d'entrer dans les détails des processus de vérification du locuteur et du visage et ni dans ceux du processus de fusion : ils ont déjà été largement étudiés dans la littérature [Furui, 1997, Reynolds, 2002, Zhao *et al.*, 2003, Li et Jain, 2005, Ross *et al.*, 2006]. Il s'agit plutôt de détailler la spécificité de la modalité visage parlant.

1.1 Vérification du visage à partir d'une séquence vidéo

Là où un algorithme de vérification du visage *classique* ne dispose que d'une image (ou d'un petit nombre d'images) du visage, la vérification du visage parlant repose sur une séquence vidéo constituée d'un grand nombre de trames. Tout algorithme de vérification du visage (qu'il utilise une seule image ou une séquence vidéo) est généralement constitué de trois modules bien distincts : un module de détection du visage, un autre module d'extraction de caractéristiques et un dernier module qui est chargé de la comparaison entre ces caractéristiques. Chacun de ces trois modules peut bénéficier de l'information supplémentaire apportée par une séquence vidéo.

Détection du visage L'information de mouvement est particulièrement utile à la détection du visage. Par exemple, *Turk et al.* calculaient la différence entre les niveaux de gris des pixels de deux trames successives, afin de réduire la région de recherche du visage [Turk et Pentland, 1991b]. Cependant, une telle approche est très sensible au mouvement qui peut se produire derrière la personne dont on cherche à vérifier l'identité.

La combinaison de l'information de couleur et de mouvement permet de déterminer une zone de recherche de visage plus robuste [Choudhury *et al.*, 1999]. Cependant, l'étape finale de détection du visage de ce type d'approches repose tout de même sur les méthodes classiques de détection du visage [Yang *et al.*, 2002]. L'apport le plus évident de l'utilisation des séquences vidéo réside dans le fait qu'il est possible de mettre en place des techniques de suivi du visage détecté sur la première trame ou sur une autre trame où la détection est plus facile. Plusieurs techniques sont élaborées qui se rapprochent généralement de l'approche CAMshift [Bradski, 1998] ou des modèles actifs d'apparence [Zhou *et al.*, 2004].

Extraction des caractéristiques Les premiers travaux de vérification du visage bénéficiant des séquences vidéo considèrent ces dernières comme un ensemble d'images indépendantes les unes des autres. Les caractéristiques sont alors extraites d'une ou plusieurs trames (choisies aléatoirement) de la séquence ; comme s'il s'agissait de vérification du visage à partir d'une image fixe [Zhao *et al.*, 2003]. Une amélioration consiste à ne conserver que les *meilleures* trames selon un critère défini au préalable. En effet, une rotation du visage, de mauvaises conditions d'éclairage ou une expression du visage peuvent entraîner une dégradation des performances du système. Par exemple, la distance à l'espace de visage (DFFS, pour *Distance From Face Space* en anglais) peut être utilisée comme une mesure du caractère *normal* d'un visage détecté et ainsi permettre de rejeter les trames éventuelles sources d'erreur [Turk et Pentland, 1991b]. Comme le résume la figure 1.1, DFFS est la distance entre un visage et sa projection sur l'espace de visage obtenu par analyse en composantes principales. Plus récemment, les changements de pose du visage tout au long de la séquence vidéo

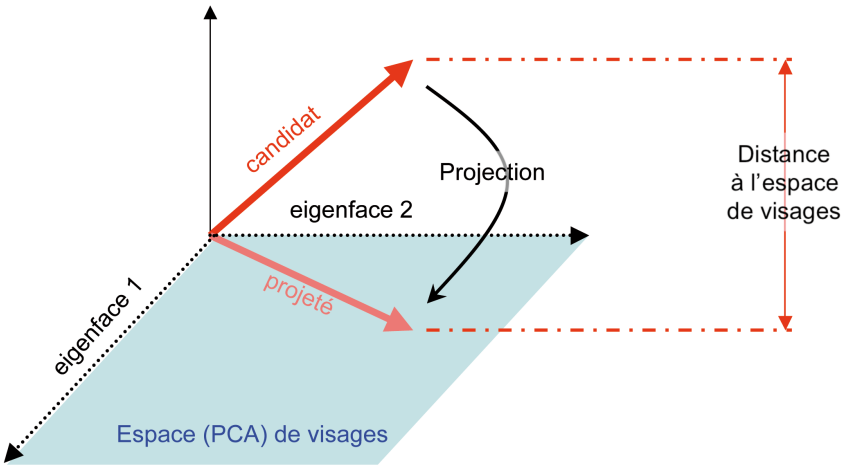


FIG. 1.1 – Distance à l'espace de visage

sont utilisés afin de mener à bien une reconstruction en trois dimensions du visage [Chowdhury *et al.*, 2002] et ainsi procéder à la vérification 3D du visage [Zhao *et al.*, 2003]. Un modèle 2D générique est proposé dans [Choudhury *et al.*, 1999] afin d'estimer la pose du visage et de reconstruire artificiellement une vue frontale du visage pour chacune des trames de la séquence vidéo. Enfin, l'information contenue dans les mouvements du visage apporte une dimension dynamique aux caractéristiques qui peuvent être extraites. Dans [Saeed *et al.*, 2006] par exemple, l'orientation du visage, les clignements des yeux et l'ouverture de la bouche sont autant de caractéristiques dynamiques du visage utiles à la vérification.

Modèle et comparaison Le grand nombre de caractéristiques extraites grâce à l'abondance de trames dans les séquences vidéos a donné naissance à de nouvelles approches de modélisation et comparaison. Les algorithmes *classiques* de vérification du visage à partir d'images fixes n'utilisent généralement pas de modèle : les caractéristiques de l'image testée sont directement comparées aux caractéristiques issues de l'image utilisée pour l'enrôlement. Ce principe peut être directement étendu aux séquences vidéo en appliquant un processus de vote : les n trames de la séquence de test sont comparées, une à une, aux m trames de la séquence d'enrôlement. Chacune des $n \times m$ comparaisons fournit une décision (acceptation ou rejet) et un vote à la majorité permet la décision finale. D'autres schémas de vote (tels que le minimum, le maximum ou la moyenne des distances) sont aussi parfois envisagés. [Krueger et Zhou, 2002] montrent que les performances sont améliorées avec l'utilisation d'un plus grand nombre d'échantillons. Là où la modélisation statistique est souvent impossible avec un seul échantillon, il est possible d'entraîner des modèles à partir d'un ensemble d'échantillons d'une même personne. Ainsi, dans [Bicego *et al.*, 2005], une machine à vecteurs de support à une classe (*one-class SVM*) est apprise à partir de l'ensemble des trames de la séquence d'enrôlement. Par analogie avec les approches classiques en vérification du locuteur, nous avons proposé de modéliser le visage d'une personne à l'aide d'un modèle de mélange de gaussiennes [Bredin *et al.*, 2006a, Bredin *et al.*, 2006b]. Dans [Bicego *et al.*, 2006], des modèles de Markov cachés pseudo-hiérarchiques sont proposés, où le nombre d'états est déterminé automatiquement en fonction des mouvements du visage.

1.2 Détection d'attaques

Comme on l'a écrit précédemment, la plupart des systèmes biométriques basés sur les visages parlants n'est basée que sur la fusion des scores produits par deux algorithmes de vérification du locuteur et de vérification du visage. Ainsi, si aucune vérification de la présence effective d'une personne réelle devant la caméra n'est réalisée, un tel système est directement menacé par un imposteur montrant une photographie du visage de sa cible et rejouant une enregistrement audio de sa voix. Ce type d'attaque (de type *rejeu* ou *replay*

attacks dans la littérature anglophone) n'est que très rarement pris en compte, alors même qu'il constitue l'une de ses plus dangereuses menaces.

Mot clé aléatoire La première parade contre ce type d'attaques peut tout aussi bien être implémentée dans le cadre de la seule vérification du locuteur. Elle consiste à demander de prononcer un mot clé (ou phrase) aléatoire différent à chaque accès (un système de transcription automatique de la parole se chargeant de vérifier l'exactitude de la phrase réellement prononcée). Cette méthode simple permet d'empêcher l'utilisation d'un enregistrement préalable de la voix de la cible. Toutefois, un système de synthèse de parole par concaténation pourrait aisément tromper cette parade.

Analyse du mouvement Une autre solution (spécifique à la vérification du visage à partir de séquences vidéo) consiste à analyser le mouvement du visage et des parties du visage afin de vérifier qu'il ne s'agisse pas d'un faux (une photographie, par exemple). Dans [Kollreider *et al.*, 2005], les mouvements de plusieurs parties du visage (nez, oreille, yeux, ...) sont comparés et, selon qu'ils soient proches les uns des autres ou non, l'accès est refusé (les mouvements des différentes parties du visage sont presque identiques dans le cas d'une image) ou accepté. Cependant, il existe aujourd'hui de nombreux outils permettant d'animer artificiellement une photographie contre lesquels ces techniques sont inefficaces.

Mesure de synchronie Une troisième solution, qui tire profit du caractère bimodal d'un visage parlant, consiste à mesurer le degré de synchronie entre la voix acquise par le microphone et le mouvement des lèvres de la personne devant la caméra. Seul un petit nombre de travaux porte sur la question spécifique de la détection d'asynchronie pour la biométrie basée sur les visages parlants. *Chetty et al.* proposent un modèle de mélange de gaussiennes dans l'espace des paramètres acoustiques et visuels concaténés [Chetty et Wagner, 2004]. Au moment du test, la mesure de synchronie est donnée par la moyenne des vraisemblances des paramètres audiovisuels de la séquence de test par rapport à ce modèle de synchronie. *Eveno et al.* proposent une mesure basée sur la corrélation entre paramètres acoustiques et visuels [Eveno et Besacier, 2005a, Eveno et Besacier, 2005b] à l'aide de l'analyse de corrélation canonique [Weenink, 2003] et l'analyse de co-inertie [Dolédec et Chessel, 1994].

1.3 Parole audiovisuelle

La dernière spécificité de la modalité *visage parlant* est le fait que le signal de parole y est audiovisuel. En effet, le mouvement des lèvres peut être utilisé comme une source complémentaire d'information au signal de parole acoustique. La fusion des signaux de parole acoustique et visuel tombe classiquement dans

l'une de ces trois catégories : la fusion au niveau des scores, la fusion au niveau des paramètres et la fusion au niveau des modèles [Chibelushi *et al.*, 2002].

Fusion au niveau des scores La grande majorité des systèmes audiovisuels de vérification du locuteur est basée sur la fusion des scores de deux systèmes de vérification du locuteur : l'un basé sur le signal acoustique seul, et l'autre basé sur le seul signal visuel. Nous n'entrerons pas dans les détails du premier. Des modèles de Markov cachés (HMM, pour *Hidden Markov Model*) dépendant du locuteur sont entraînés à l'aide de paramètres liés à la forme des lèvres dans [Jourlin *et al.*, 1997], à l'aide de paramètres de type *eigenlips* (zone de la bouche transformée par analyse en composantes principales) dans [Dean *et al.*, 2005] et des coefficients DCT (transformée en cosinus discrète) de la zone de la bouche dans [Sargin *et al.*, 2006]. Dans [Fox *et al.*, 2007], les auteurs concluent cependant que l'utilisation de modèles de mélange de gaussiennes serait suffisante puisque les meilleures performances sont obtenues avec des HMM à un seul état. Tous ces travaux tirent la même conclusion selon laquelle la fusion des deux scores monomodaux (acoustique et visuel) est un moyen simple et efficace d'améliorer les performances globales de la vérification d'identité, et tout particulièrement en milieu bruité.

Fusion au niveau des paramètres La fusion au niveau des paramètres consiste en la combinaison de deux (ou plus) vecteurs de paramètres monomodaux afin de former un unique vecteur de paramètres multimodal, utilisé en entrée d'un système de vérification. Dans le cas de la parole audiovisuelle, les fréquences d'échantillonnage diffèrent entre les deux modalités acoustique et visuelle. Typiquement, 100 vecteurs de paramètres acoustiques sont extraits chaque seconde alors que seulement 25 (ou 30) trames vidéo sont disponibles pendant la même période. Afin d'équilibrer les fréquences d'échantillonnage, une solution consiste à interpoler linéairement les vecteurs de paramètres visuels [Sargin *et al.*, 2006, Bredin *et al.*, 2006a]. Une autre solution consiste à sous-échantillonner les vecteurs de paramètres acoustiques [Arsic *et al.*, 2006]. Les vecteurs concaténés sont fournis en entrée d'un réseau de neurones (MLP, pour *Multiple Layer Perceptron*) dans [Chibelushi *et al.*, 1997a] et d'un GMM dans [Arsic *et al.*, 2006]. Le fléau de la dimension (*curse of dimensionality* dans la littérature anglophone) est évoqué dans [Chibelushi *et al.*, 1997a]. Une solution consiste à appliquer une analyse en composantes principales ou une analyse discriminante linéaire pour réduire la dimension des paramètres audiovisuels dans le but d'obtenir consécutivement une meilleure modélisation. Dans [Sargin *et al.*, 2006], une analyse de corrélation canonique permet d'extraire des paramètres acoustiques et visuels à dimensions réduites avec une corrélation maximisée, utilisés par la suite pour entraîner un unique HMM audiovisuel. Comme dans le cas de la fusion au niveau des scores, la fusion au niveau des paramètres est surtout efficace dans le cas d'un environnement acoustique bruité.

Fusion au niveau des modèles Dans le cadre de la fusion au niveau des modèles, les modèles sont intrinsèquement conçus de façon à tenir compte du caractère bimodal de la parole audiovisuelle. Par exemple, les HMM couplés peuvent être décrits comme deux HMM parallèles dont les probabilités de transition dépendent des états de chacun d'eux. Ils ont été appliqués à des paramètres acoustiques (MFCC) et visuels (*eigenlips*) transformés par LDA dans [Nefian et Liang, 2003]. Les HMM-produits permettent de tenir compte de l'asynchronie entre paramètres acoustiques et visuels [Lucey *et al.*, 2005] : une transition acoustique ne correspond pas forcément à une transition visuelle. [André-Obrecht *et al.*, 1997] proposent l'utilisation de deux HMM corrélés : un HMM maître traitant le flux visuel et un HMM esclave qui en dépend, traitant du flux acoustique. Enfin, les HMM asynchrones proposés dans [Bengio, 2003] modélisent la différence des fréquences d'échantillonnage acoustique et visuelle, en introduisant la probabilité d'existence d'un vecteur de paramètres visuels à un temps donné.

Chapitre 2

Évaluation

Dans ce chapitre, nous définissons les mesures qui seront utilisées pour réaliser l'évaluation objective des performances obtenues par nos différentes propositions. Nous présenterons ensuite la base de données BANCA à partir de laquelle nous avons mené nos expériences ainsi que les protocoles d'évaluation associés.

2.1 Mesures de performance

De façon à comparer objectivement deux systèmes, il convient d'introduire des grandeurs mathématiques liées aux erreurs qu'ils peuvent commettre : c'est l'objet de ce paragraphe.

2.1.1 Faux rejet et fausse acceptation

Un système de vérification d'identité biométrique peut faire deux types d'erreur. Une fausse acceptation (FA) se produit lorsqu'un imposteur λ (clamant l'identité de sa cible $\epsilon \neq \lambda$) n'est pas rejeté par le système et un faux rejet (FR) se produit lorsqu'un client ϵ (clamant sa propre identité ϵ) est rejeté par le système. Ces deux types d'erreur dépendent du seuil de décision θ auquel est comparé le score issu du processus de comparaison. Une valeur élevée de θ tendra à rendre l'accès plus difficile, en augmentant le nombre de faux rejets NFR et diminuant le nombre de fausses acceptations NFA. Inversement, une valeur faible de θ tendra à faciliter l'accès et donc à augmenter le nombre de fausses acceptations NFA.

En pratique, les taux de fausse acceptation (FAR, pour *False Acceptance Rate*) et de faux rejet (FRR, pour *False Rejection Rate*) sont mesurés expérimentalement à partir d'un corpus de test en comptant le nombre

de fausses acceptations et de faux rejets :

$$\mathrm{FAR}(\theta) = \frac{\mathrm{NFA}(\theta)}{\mathrm{NI}} \tag{2.1}$$

$$\mathrm{FRR}(\theta) = \frac{\mathrm{NFR}(\theta)}{\mathrm{NC}} \tag{2.2}$$

où θ est le seuil de décision, NI et NC sont les nombres d'accès *imposteur* et *client* respectivement dans le corpus de test.

2.1.2 Courbe DET et EER

Les taux de faux rejet et de fausse acceptation étant tous les deux fonctions du seuil de décision θ, il est possible de représenter les performances d'un système en traçant la valeur de FRR(θ) en fonction de FAR(θ) pour θ variant de $-\infty$ à $+\infty$, comme illustré dans la figure 2.1. En utilisant une échelle logarithmique, nous obtenons une courbe de détection (DET, pour *Detection Error Tradeoff*) introduite dans [Martin *et al.*, 1997].

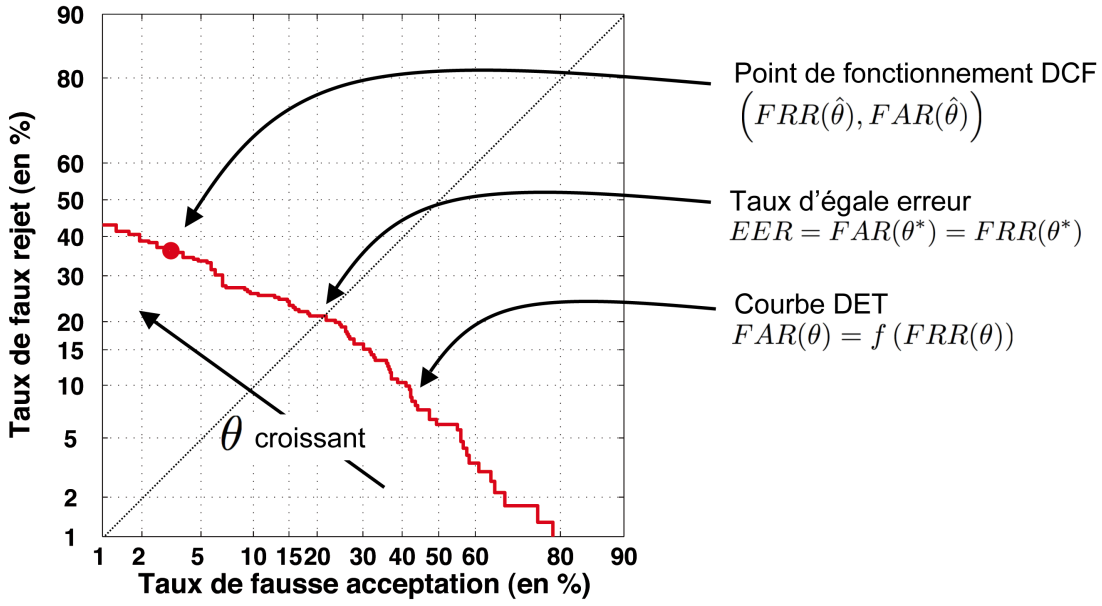


FIG. 2.1 – Courbe DET

Une mesure de performance découlant de la courbe DET est le taux d'égale erreur (EER, pour *Equal Error Rate*), qui correspond au point particulier de la courbe DET défini par le seuil θ^* (situé à l'intersection de la courbe DET et de la droite FAR=FRR) vérifiant l'équation (2.3) :

$$\text{EER} = \text{FAR}(\theta^*) = \text{FRR}(\theta^*) \tag{2.3}$$

2.1.3 DCF

Bien que la courbe DET et le taux d'égale erreur fournissent un bon moyen de comparer les performances de différents systèmes en phase de développement, ils ne permettent pas d'évaluer les performances de ces systèmes en situation réelle de fonctionnement. En effet, en situation réelle, le seuil de décision θ a été fixé une fois pour toute à partir d'un ensemble de développement, optimisé pour une application donnée, et le corpus de test est inconnu.

On définit la fonction de coût de détection (DCF, pour *Detection Cost Function*) comme la somme, pondérée par les coûts C_a et C_r, des taux de fausse acceptation FAR et faux rejet FRR [Martin et Przybocki, 2000] :

$$\text{DCF}(\hat{\theta}) = C_a \cdot \text{FAR}(\hat{\theta}) + C_r \cdot \text{FRR}(\hat{\theta}) \tag{2.4}$$

où le seuil de décision $\hat{\theta}$ a été optimisé au préalable par minimisation du DCF sur l'ensemble de développement. Dans notre cas particulier où l'objectif principal est la robustesse aux tentatives d'imposture, on convient des coûts $C_a = 0.99$ et $C_r = 0.10$ [Martin et Przybocki, 2000] : il est ainsi plus coûteux pour le système de faire une erreur de type *fausse acceptation* que de rejeter une personne dont la demande d'accès était légitime.

$$\text{DCF}(\hat{\theta}) = 0.99 \cdot \text{FAR}(\hat{\theta}) + 0.10 \cdot \text{FRR}(\hat{\theta}) \tag{2.5}$$

Variante Une variante du DCF est le taux d'erreur pondéré (WER, pour *Weighted Error Rate*) qui est défini de façon analogue à la fonction de coût de détection, par l'équation (2.6) :

$$\text{WER}_r(\hat{\theta}) = \frac{r \cdot \text{FAR}(\hat{\theta}) + \text{FRR}(\hat{\theta})}{r+1} \tag{2.6}$$

où r décrit le coût d'une fausse acceptation vis-à-vis d'un faux rejet. Typiquement, trois valeurs de r peuvent être choisies : $r = 0.1$, $r = 1$ et $r = 10$.

2.1.4 Comment s'assurer qu'un système est meilleur qu'un autre ?

S'assurer que la différence de performance entre deux systèmes est statistiquement significative est une question trop souvent passée sous silence dans la littérature. Une différence de 10% de la valeur d'un taux de fausse acceptation estimée à partir de 5 accès *imposteur* n'est sans doute pas statistiquement significative alors qu'une différence de 0.5% estimée à partir de 100000 accès l'est peut-être... [Guyon *et al.*, 1998, Bengio, 2003].

Modélisation statistique des taux d'erreur (d'après [Bengio, 2003]) Sous l'hypothèse que les accès au système sont indépendants, les décisions binaires prises par le système sont elle aussi indépendantes. Il est donc raisonnable de supposer que la variable aléatoire $\mathbf{X}$ représentant le nombre d'erreurs suit une loi binomiale $\mathcal{B}(n,p)$ où n est le nombre de tests et p est le taux d'erreur. En outre, il est connu qu'une distribution binomiale $\mathcal{B}(n,p)$ peut être approximée par une distribution normale $\mathcal{N}\left(\mu,\sigma^2\right)$ avec

$$\mu = np \text{ et } \sigma^2 = np(1-p) \tag{2.7}$$

lorsque n est *suffisamment* grand. Enfin, si $\mathbf{X} \sim \mathcal{N}(np, np(1-p))$, alors la distribution du taux d'erreur $\mathbf{Y} = \frac{\mathbf{X}}{n}$ est aussi une distribution normale :

$$\mathbf{Y} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{2.8}$$

Il est alors possible de calculer un intervalle de confiance autour de l'estimation p du taux d'erreur en déterminant les bornes $\{p-\beta, p+\beta\}$ telles que :

$$P(p-\beta < \mathbf{Y} < p+\beta) = \delta \tag{2.9}$$

où δ est la mesure de confiance (classiquement $\delta = 95\%$) en l'estimation p du taux d'erreur. Dans le cadre d'une distribution normale, la valeur de β peut facilement être obtenue à l'aide de la table de la loi normale.

Intervalle de confiance sur FAR, FRR et DCF En appliquant la méthode au nombre de fausses acceptations ($X = \text{NFA}$), on obtient $p = \text{FAR}$, $n = \text{NI}$ et l'intervalle de confiance $\text{CI}(\text{FAR})$ sur le taux de fausse acceptation via l'équation (2.10) :

$$\text{CI}(\text{FAR}) \quad = \quad \text{FAR} \pm \alpha \cdot \sqrt{\frac{1}{\text{NI}} \cdot \text{FAR}(1-\text{FAR})} \tag{2.10}$$

où $\alpha = 1.960$ décrit un intervalle de confiance à 95% et $\alpha = 2.576$ un intervalle de confiance à 99%. Par analogie, on obtient pour les faux rejets :

$$\text{CI(FRR)} \quad = \quad \text{FRR} \pm \alpha \cdot \sqrt{\frac{1}{\text{NC}} \cdot \text{FRR}(1 - \text{FRR})} \tag{2.11}$$

Sous l'hypothèse que les accès *client* et *imposteur* sont indépendants, les taux d'erreur FAR et FRR le sont aussi. Or, la somme de deux lois normales indépendantes est aussi une loi normale. Par conséquent, l'intervalle de confiance CI(DCF) de la fonction de coût de détection est obtenu via l'équation (2.12) :

$$\text{CI(DCF)} \quad = \quad \text{DCF} \pm \alpha \cdot \sqrt{\frac{0.99^2}{\text{NI}} \cdot \text{FAR}(1 - \text{FAR}) + \frac{0.10^2}{\text{NC}} \cdot \text{FRR}(1 - \text{FRR})} \tag{2.12}$$

Comparaison On considère deux systèmes S_1 et S_2 dont les performances sur l'ensemble de test sont DCF_1 et DCF_2 avec $DCF_2 < DCF_1$. Le système S_2 est significativement meilleur que le système S_1 si la valeur de DCF_2 est à l'extérieur de l'intervalle de confiance $\text{CI}(DCF_1)$. Dans le cas contraire, aucune conclusion statistiquement significative ne peut être déduite quant au meilleur des deux systèmes.

Remarque Les équations (2.10) et (2.11) reposent toutes deux sur l'hypothèse selon laquelle "*n est suffisamment grand*" ($n =$ NI ou NC). Dans le cas des attaques de type *Big Brother* et *Crazy Talk* définies au chapitre 4, cette hypothèse n'est pas vérifiée (NI $= 52$). Il est alors possible de montrer que l'intervalle de confiance à $100 \cdot (1 - \delta)\,\%$ sur FAR est défini par l'équation suivante :

$$\text{FAR} \in \left[\frac{\chi^2_{\frac{\delta}{2}}\,(2 \cdot \text{NFA})}{2 \cdot \text{NI}}, \frac{\chi^2_{1-\frac{\delta}{2}}\,(2 \cdot \text{NFA} + 2)}{2 \cdot \text{NI}} \right] \tag{2.13}$$

où $\chi^2\,(n)$ est la loi du Khi-Deux à n degrés de liberté [Saporta, 1978]. Dans le cas d'un intervalle de confiance à 95%, $\delta = 0.05$.

2.2 Base de données

Plusieurs bases contenant des données permettant l'évaluation de systèmes de vérification de l'identité des visages parlants sont disponibles. Citons les bases de données BT-DAVID [BT-DAVID, 1996], XM2VTSDB [Messer *et al.*, 1999], CUAVE [Patterson *et al.*, 2002], BANCA [Bailly-Baillière *et al.*, 2003], Biomet [Garcia-Salicetti *et al.*, 2003], MyIDEA [Dumas *et al.*, 2005] et SecurePhone [Morris *et al.*, 2006].

La base de données BANCA (*Biometric Access control for Networked and e-Commerce Applications*) est une base de données audiovisuelles destinée à l'aide au développement et à l'évaluation de systèmes de vérification d'identité [Bailly-Baillière *et al.*, 2003]. Les séquences audiovisuelles ont été acquises dans quatre langues européennes : une séquence de chiffres suivie des nom et adresse de la personne ont été enregistrés pour chaque accès. Nous nous concentrons cependant sur la seule partie en langue anglaise, dont la constitution est résumée dans le schéma de la figure 2.2 et détaillée ci-dessous.

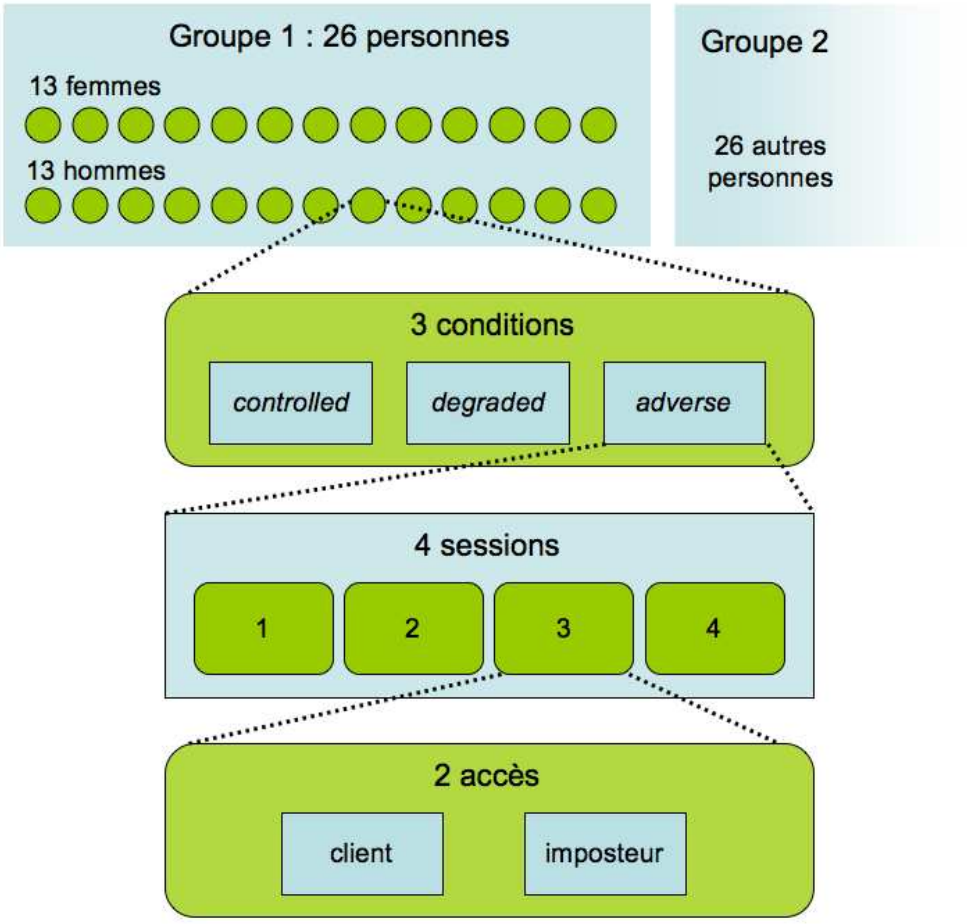


FIG. 2.2 – Description de la base BANCA

Elle est divisée en **deux groupes disjoints** de personnes, appelés G1 et G2 par la suite. Chaque groupe est constitué de 26 personnes réparties équitablement par genre : 13 femmes et 13 hommes. **Trois conditions** différentes d'enregistrement, illustrées dans la figure 2.3, ont été appliquées. Dans la condition *controlled*, la personne apparaît face à la caméra sur un fond bleu fixe et une caméra DV est utilisée pour l'acquisition. Dans la condition *degraded*, l'enregistrement a lieu dans un bureau à l'aide d'une webcam de moins bonne qualité. Enfin, les enregistrements de la condition *adverse* ont lieu dans un réfectoire universitaire, où d'autres personnes peuvent circuler en arrière-plan.

FIG. 2.3 – Exemple des conditions *controlled*, *degraded* et *adverse*

Dans chacune des trois conditions, chaque personne a participé à **quatre sessions** espacées dans le temps, numérotées de 1 à 4 pour la condition *controlled*, de 5 à 8 pour la condition *degraded* et de 9 à 12 pour la condition *adverse*. Chaque session est elle-même constituée de **deux séquences** audiovisuelles. Dans la première, la personne prononce, face à la caméra, son nom et son adresse : on parle d'accès *client*. Dans la seconde, elle prononce le nom et l'adresse d'une autre personne : on parle d'accès *imposteur*, puisqu'elle prétend être une autre personne.

Un troisième groupe, appelé *world model*, regroupe une trentaine de personnes ayant chacune enregistré deux séquences. Les 60 séquences résultantes (20 en conditions *controlled*, 20 *degraded* et 20 *adverse*) sont disponibles pour le développement des différents algorithmes.

2.3 Protocoles d'évaluation

La base de données BANCA étant constituée de deux groupes disjoints G1 et G2, il est prévu d'utiliser l'un d'eux comme ensemble de développement et l'autre comme ensemble de test. L'ensemble de déve-

loppement permet le réglage des différents modules du système de vérification d'identité. Par exemple, le seuil de décision $\hat{\theta}$ utilisé pour le calcul du DCF, les poids de fusion définis à la page 64 et la mesure de confiance définie à la page 122 sont tous réglés à partir de l'ensemble de développement avant d'être appliqués directement sur l'ensemble de test de façon à mesurer les performances.

En pratique, les mesures DCF, FAR et FRR sont calculées à partir des deux ensembles de test G1 et G2. $\widehat{\theta_2}$ et $\widehat{\theta_1}$ étant les seuils optimisés par minimisation du DCF sur les ensembles de développement G2 et G1 respectivement, les équations (2.1) et (2.2) de la page 42 deviennent :

$$\text{FAR} = \frac{\text{NFA}_{\text{G1}}(\widehat{\theta_2}) + \text{NFA}_{\text{G2}}(\widehat{\theta_1})}{\text{NI}_{\text{G1}} + \text{NI}_{\text{G2}}} \tag{2.14}$$

$$\text{FRR} = \frac{\text{NFR}_{\text{G1}}(\widehat{\theta_2}) + \text{NFR}_{\text{G2}}(\widehat{\theta_1})}{\text{NC}_{\text{G1}} + \text{NC}_{\text{G2}}} \tag{2.15}$$

2.3.1 Protocole P

Le protocole *Pooled* est l'un des protocoles distribués avec la base de données BANCA et qui a été utilisé lors d'une compétition en 2004 [Messer *et al.*, 2004].

Enrôlement Pour chaque personne λ, la séquence audiovisuelle de l'accès *client* de la session 1 de la condition *controlled* est utilisée comme donnée d'enrôlement pour obtenir le modèle λ.

Tests *client* Pour chaque personne λ, les séquences audiovisuelles des accès *client* de λ des sessions 2 à 4 (*controlled*), 6 à 8 (*degraded*) et 10 à 12 (*adverse*) sont comparées au modèle λ. Au final, le protocole P prévoit donc 9 tests *client* par personne, soit 234 tests *client* par groupe.

Tests *imposteur* Pour chaque personne λ, toutes les séquences audiovisuelles des accès *imposteur* de λ (sessions 1 à 12) sont comparées au modèle de la personne dont λ prononce le nom et l'adresse. En entrant dans le détail de ces accès *imposteur*, on note que la personne λ est en fait comparée à chacune des 12 autres personnes du même groupe et du même sexe. Au final, le protocole P prévoit donc 12 tests *imposteur* par personne, soit 312 tests *imposteur* par groupe.

2.3.2 Protocole txtP

Le protocole P peut être considéré comme un protocole dépendant du texte. En effet, à chaque personne λ sont associés un nom et une adresse qui lui sont propres et qu'elle prononce lors de ses accès *client*. En outre, lors des accès *imposteur*, l'imposteur prononce le nom et l'adresse que sa cible utilise pour s'authentifier. On introduit donc le protocole txtP indépendant du texte, qui est une adaptation du protocole P original.

Enrôlement En ce qui concerne l'enrôlement, les protocoles P et txtP sont identiques.

Tests *client* Pour chaque personne λ, les séquences audiovisuelles des accès *imposteur* de la personne λ des sessions 1 à 12 sont comparés au modèle de la personne λ. Ainsi, le texte prononcé dans la séquence de test est toujours différent de celui prononcé dans la séquence d'enrôlement. Au final, le protocole txtP prévoit donc 12 tests *client* par personne, soit 312 tests *client* par groupe.

Tests *imposteur* Les tests *imposteur* du protocole txtP sont identiques à ceux du protocole P.

Le protocole txtP est dit indépendant du texte dans le sens où le texte prononcé lors de chaque accès *client* est différent du texte prononcé lors de la séquence audiovisuelle d'enrôlement. Il sera utilisé pour évaluer l'influence de la phrase d'authentification sur la nouvelle modalité biométrique basée sur la synchronie audiovisuelle introduite dans le chapitre 7.

2.3.3 Protocole xP

Le nombre de tests du protocole original P étant très limité, ce dernier possède un intérêt limité par les larges intervalles de confiance (définis dans le paragraphe 2.1.4) qui en découlent. Nous avons donc défini le protocole xP, comme une extension du protocole P.

Enrôlement En ce qui concerne l'enrôlement, les protocoles P et xP sont identiques.

Tests *client* Pour chaque personne λ, toutes les séquences audiovisuelles de λ (accès *client* et *imposteur*, à l'exception de la session 1) sont comparées au modèle λ. Au final, le protocole xP prévoit donc 22 tests *client* par personne, soit 572 tests *client* par groupe.

Tests *imposteur* Pour chaque personne λ, toutes les séquences audiovisuelles des autres personnes (accès *client* et *imposteur*) sont comparées au modèle λ. Au final, le protocole xP prévoit donc 600 tests *imposteur* par personne, soit 15600 tests *imposteur* par groupe.

Ce protocole étendu permet d'obtenir un nombre de scores plus important qui pourra être utilisé pour le réglage des différents paramètres que nous introduirons au fur et à mesure de notre exposé ; en particulier, les paramètres de normalisation des scores de la page 63 et la mesure de confiance définie à la page 122.

2.3.4 Protocole S

Là où le protocole P (et ses variantes) permet l'évaluation des performances d'un système de vérification d'identité, le protocole S que nous avons défini s'attaque à un problème différent, soulevé dans le chapitre 6 : la détection de l'asynchronie. Il s'agit de décider si la voix captée par le microphone et le mouvement des

lèvres acquises par la caméra ont été produits simultanément par une seule et même personne. On parle alors de séquence *synchrone*. Dans le cas contraire, la séquence est dite *asynchrone*. Deux ensembles de séquences audiovisuelles (où la personne prononce une séquence de chiffres suivie d'un nom et une adresse) sont ainsi constitués :

Séquences synchrones Toutes les séquences originales de la base BANCA sont synchrones. Aussi, les 24 séquences (12 sessions constituées d'un accès *client* et d'un accès *imposteur*) de chaque personne constituent l'ensemble des séquences synchrones : nous obtenons ainsi $\texttt{NC} = 624$ accès synchrones par groupe.

Séquences asynchrones Les séquences asynchrones sont générées artificiellement en combinant la partie audio et la partie vidéo de deux séquences différentes. La durée de la séquence finale est choisie comme étant le minimum des durées de la partie audio et de la partie vidéo. Pour chaque personne, 56 séquences asynchrones sont ainsi générées à partir d'un enregistrement de sa voix et de la partie vidéo d'une autre séquence (de cette même personne – 12 séquences –, ou non – 44 séquences). Nous obtenons ainsi $\texttt{NI} = 1456$ accès asynchrones par groupe.

Bien que provenant de deux séquences différentes, les mêmes nom et adresse sont prononcés dans les parties audio et vidéo des séquences asynchrones, rendant la tâche de détection d'asynchronie particulièrement difficile dans certains cas.

Important Les nombres `NI`, `NC`, `NFA` et `NFR` correspondant à l'évaluation sur le protocole `S` (dédié à la tâche de *détection d'asynchronie*) ont une signification différente de NI, NC, NFA et NFR obtenus à partir du protocole P et ses variantes (dédiés à la tâche de *vérification d'identité*) :
- `NI` est le nombre de séquences asynchrones ;
- `NC` est le nombre de séquences synchrones ;
- `NFA` est le nombre de séquences asynchrones faussement classées comme étant synchrones ;
- `NFR` est le nombre de séquences synchrones faussement classées comme étant asynchrones.

2.4 Base de données et protocoles additionnels

Dans le cadre du projet *Technovision IV2*, une base de données a été acquise dans le but d'organiser une campagne d'évaluation de systèmes biométriques basés (entre autres modalités) sur les visages parlants. Les expériences réalisées dans ce cadre sont reportées dans l'annexe A (page 137) où sont présentés la base de données, le protocole d'évaluation associé ainsi que les performances obtenues par nos différents systèmes.

Chapitre 3

Système initial

Afin de mettre en avant nos contributions – centrées sur l'analyse de la synchronie dans la parole audiovisuelle –, un premier système état-de-l'art de vérification d'identité basé sur les visages parlants a été développé. Il est basé sur la fusion des scores obtenus par deux sous-systèmes monomodaux de vérification du locuteur et d'authentification du visage. Ces travaux ont été en partie publiés dans l'article de conférence intitulé *The Biosecure Talking-Face Reference System* et reproduit en annexe [Bredin *et al.*, 2006a].

3.1 Vérification du locuteur

Le module de vérification du locuteur est basé sur l'approche classique par modèles de mélange de gaussiennes : la figure 3.1 résume son fonctionnement.

Détection du silence Afin de ne conserver que la partie du signal acoustique d'entrée correspondant aux plages où le locuteur est effectivement en train de parler, la première étape consiste à supprimer les plages de silence. Tout d'abord, la distribution de l'énergie du signal acoustique est modélisée par un mélange bigaussien : la gaussienne de moyenne la plus élevée étant associée à l'activité vocale, et celle de moyenne la plus faible au silence (qui n'est jamais parfait, du fait du bruit ambiant). La figure 3.2 présente le résultat de cette modélisation sur un exemple. Un seuil s est ensuite fixé à la valeur s_0 (représenté par le trait noir vertical) estimée à partir des moyenne μ et variance σ de la gaussienne correspondant à l'énergie d'activité vocale (celle la plus à droite) : $s_0 = \mu - 2\sigma$. Si l'énergie est inférieure au seuil, la fenêtre de signal correspondante est détectée comme silence. La figure 3.3 présente le résultat de la détection du silence sur le même exemple que précédemment.

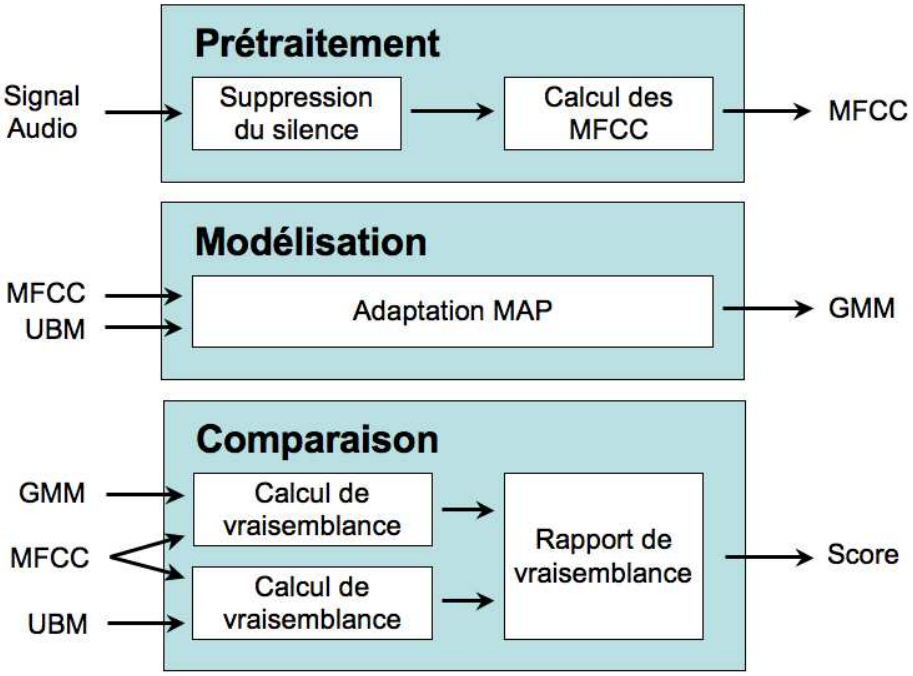


FIG. 3.1 – Détail des modules de la vérification du locuteur

Extraction des vecteurs de paramètres Un vecteur de 12 paramètres MFCC (pour *Mel-Frequency Cepstral Coefficients*), dont le processus d'extraction est schématisé dans la figure 3.4, est extrait toutes les 10 ms sur une fenêtre glissante de longueur 20 ms. Ne sont conservés que les vecteurs de paramètres correspondants aux fenêtres classées par le détecteur de silence comme *non silence*. Plusieurs jeux de paramètres peuvent être extraits selon que l'on ajoute l'énergie, les dérivés premières (appelées Δ par la suite) ou secondes ($\Delta\Delta$).

Modélisation par mélange de gaussiennes Un modèle du monde (noté *UBM* – pour *Universal Background Model* – dans la figure 3.1) est tout d'abord appris à partir d'une grande quantité de données acquises auprès d'un large échantillon de locuteurs, de façon à couvrir au maximum la variabilité des locuteurs. L'apprentissage de ce modèle de mélange de gaussiennes est réalisé par le biais de l'algorithme *EM* [Dempster *et al.*, 1977]. Une fois ce modèle disponible, il est possible de l'adapter à un locuteur particulier grâce à ses données d'enrôlement. L'approche MAP (Maximum A Posteriori) nous permet ainsi d'obtenir un modèle adapté aux données du locuteur [Reynolds *et al.*, 2000b].

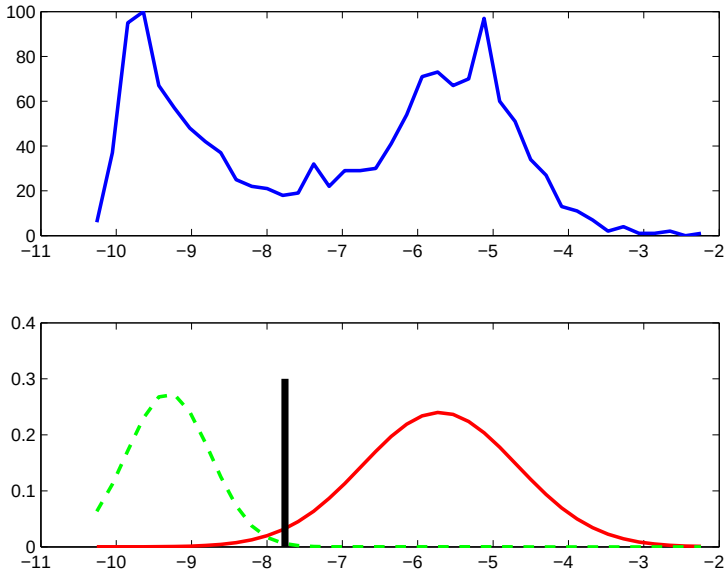


FIG. 3.2 – Modélisation bigaussienne de l'énergie. La distribution réelle de l'énergie du signal acoustique est présentée en bleu dans la courbe en haut. Sa modélisation bigaussienne est représentée dans la courbe en bas (en vert pointillé, la gaussienne associée au silence ; en rouge continu, celle associé à l'activité vocale). Le seuil (trait noir vertical) est calculé à partir de la moyenne et de la variance de la gaussienne associée à l'activité vocale.

Rapport de vraisemblance Au moment du test, il s'agit de vérifier si la personne Γ dont la voix est acquise est bien la personne λ qu'elle prétend être. Les vecteurs de paramètres extraits de la séquence Γ (MFCC) sont comparés au modèle du locuteur λ, ainsi qu'au modèle du monde Ω. Le rapport de ces vraisemblances S est finalement comparé à un seuil permettant de vérifier l'identité clamée par le locuteur. En résumé, notant x un vecteur de paramètres MFCC, on obtient :

$$S_{\text{locuteur}}(\Gamma|\lambda) = \frac{p(x|\lambda)}{p(x|\Omega)} \tag{3.1}$$

L'accès est accepté si $S_{\text{locuteur}}(\Gamma|\lambda) > \theta$ et refusé dans le cas contraire.

Performances Le système de vérification du locuteur utilisé repose sur une modélisation GMM à 256 gaussiennes, dans l'espace à 36 dimensions des MFCC auquel on a ajouté les dérivés secondes et premières (on note MFCC + Δ + $\Delta\Delta$). Le modèle du monde UBM est appris à partir des enregistrements des 30

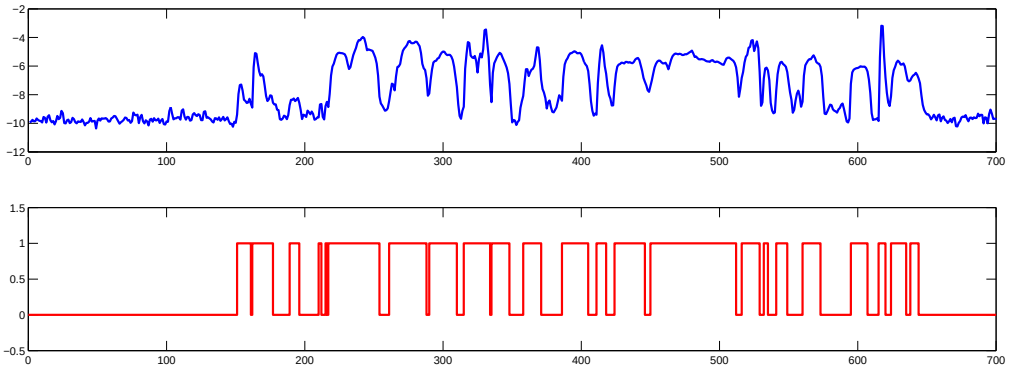


FIG. 3.3 – Détection du silence. L'évolution de l'énergie du signal acoustique est présentée en bleu, dans la courbe en haut. La courbe rouge, en bas, présente le résultat de la détection du silence (0 signifie *silence*, 1 signifie *activité vocale*).

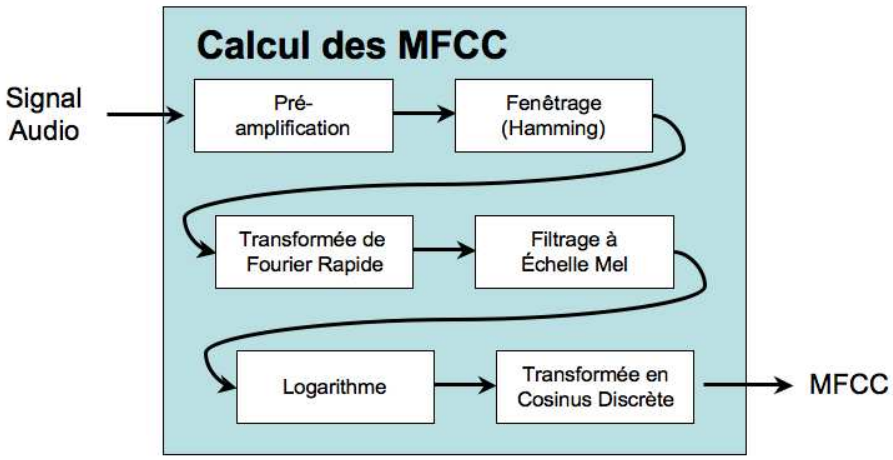


FIG. 3.4 – Extraction des MFCC

personnes issus de la partie *world model* de BANCA, qui représentent environ 10 minutes de parole. Les performances de ce système sur le protocole P sont résumées dans la figure 3.5. Nous avons, en outre, utilisé la boîte à outils BECARS [1] pour la modélisation GMM et le calcul des vraisemblances [Blouet *et al.*, 2004].

[1] `http://www.tsi.enst.fr/becars`

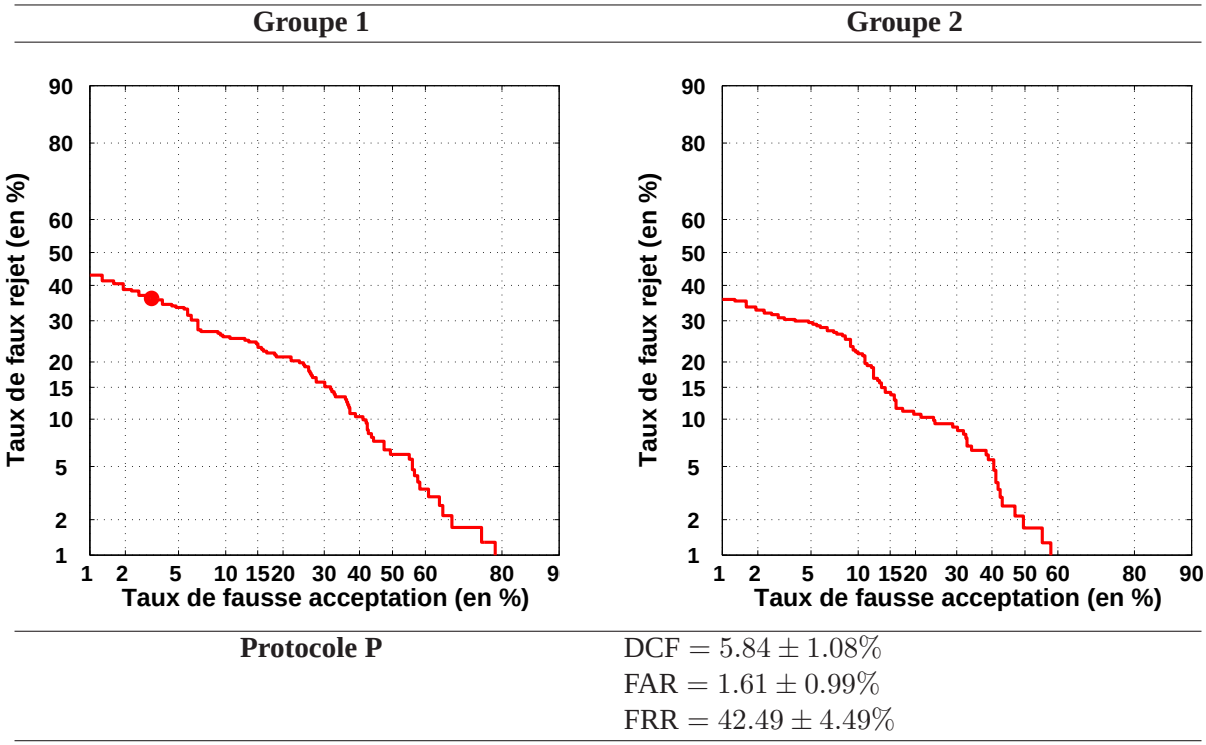


FIG. 3.5 – Performances du système de vérification du locuteur

Discussion Rappelons ici que deux canaux audio sont disponibles pour chacune des vidéos (l'un de bonne qualité et l'autre acquis avec un microphone de mauvaise qualité et très bruité). Nous avons ici choisi d'utiliser le canal de mauvaise qualité. Ceci explique en partie les performances assez éloignées de ce que l'on peut trouver dans la littérature à l'état-de-l'art. Une autre explication réside dans la petite taille de l'ensemble d'apprentissage du modèle du monde.

3.2 Vérification du visage

Le module de vérification du visage est basé sur l'approche classique des *eigenfaces* proposée par *Turk et Pentland* [Turk et Pentland, 1991a]. En outre, nous proposons d'utiliser la redondance d'information disponible dans la séquence vidéo (chacune des trames fournissant un vecteur de paramètres décrivant le visage) en sélectionnant plusieurs trames selon un critère de qualité défini par la suite. La mesure de similarité entre données d'enrôlement et données de test est basée sur l'approche classique de la distance de Mahalanobis [Mahalanobis, 1936]. La figure 3.6 décrit cette approche.

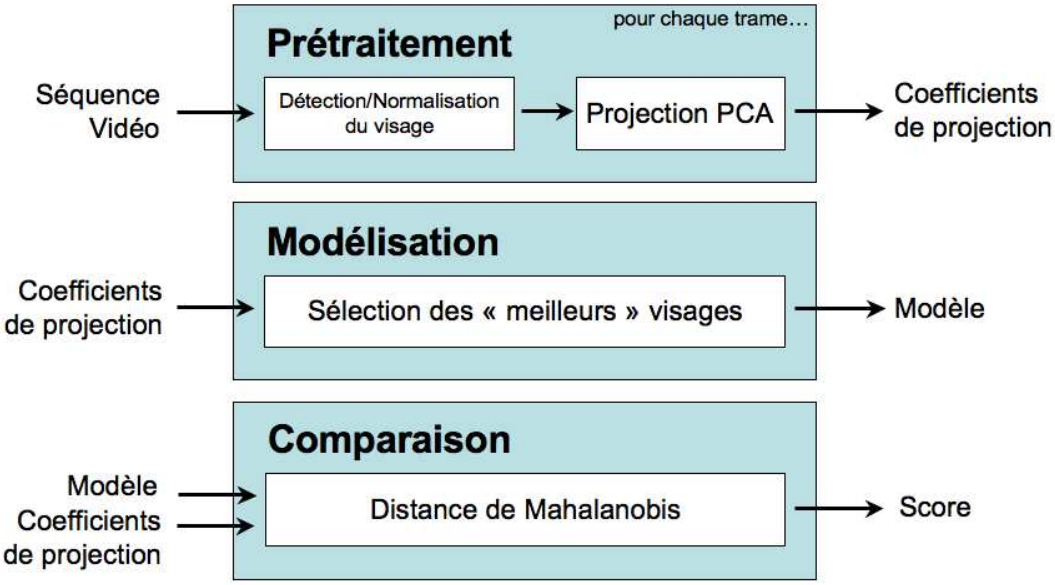


FIG. 3.6 – Détail des modules de la vérification du visage

Détection du visage La première étape indispensable à la vérification du visage est la détection de celui-ci. L'algorithme proposé par *Fasel et al.* est utilisé ici [Fasel *et al.*, 2004]. Dans chaque trame de la séquence vidéo, un détecteur de *Viola & Jones* est appliqué pour obtenir toutes les zones candidates à contenir un visage [Viola et Jones, 2002]. Afin d'améliorer la précision spatiale des visages détectés (et en prévision de l'étape suivante de normalisation), les yeux sont à leur tour détectés en appliquant le très fort *a priori* selon lequel deux yeux doivent être détectés dans la région d'intérêt. Les détails de cet algorithme sont décrits dans [Fasel *et al.*, 2004]. Nous utilisons son implémentation *open-source* proposée dans la boîte à outils *Machine Perception Toolbox* [Fasel *et al.*, 2004]. Enfin, nous faisons appel à l'*a priori* très contraignant qu'un seul visage est censé apparaître devant la caméra (la personne dont on cherche à authentifier l'identité). Par conséquent, tous les visages détectés à tort peuvent être supprimés en ne conservant que le plus grand des visages détectés.

Extraction des vecteurs de paramètres Une fois le visage et les yeux détectés, le visage est normalisé de façon à ce que les yeux soient centrés et alignés horizontalement. Un masque ovale permettant de supprimer des pixels de fond et une égalisation d'histogramme sont aussi appliqués, comme le montre la figure 3.7. Le visage de chaque trame de la vidéo peut alors être projeté dans l'espace de visage, obtenu par analyse en

FIG. 3.7 – Normalisation du visage. Le visage est normalisé en fonction de la position des yeux détectés. Un masque ellipsoïde est ajouté afin de supprimer le fond.

composantes principales suivant le principe des *eigenfaces* [Turk et Pentland, 1991a]. Cependant, la qualité de détection varie selon les images. Aussi, pour mener à bien la modélisation et/ou la reconnaissance, une sélection des meilleurs visages (selon un critère que l'on définit par la suite) est effectuée afin de ne conserver que ces vecteur de paramètres.

Sélection des meilleurs visages La distance à l'espace de visage (DFFS, pour *Distance From Face Space*) [Turk et Pentland, 1991b, Potamianos *et al.*, 2003] est utilisée comme un indicateur permettant de déterminer la vraisemblance selon laquelle une zone détectée automatiquement correspond effectivement à un visage. Plus une zone candidate est proche de son propre projeté dans l'espace de visage, plus il est probable que le visage ait été correctement détecté. La figure 3.8 illustre ce principe. Pour chaque zone candidate, un indice de confiance r est calculé à partir de la distance à l'espace de visage par inversion de celle-ci : $r = 1/\text{DFFS}$. Ainsi, plus r est grand, plus il est probable que le visage soit correctement localisé. La figure 3.9 illustre le résultat de la sélection des meilleurs visages d'une même séquence, en utilisant cet indice.

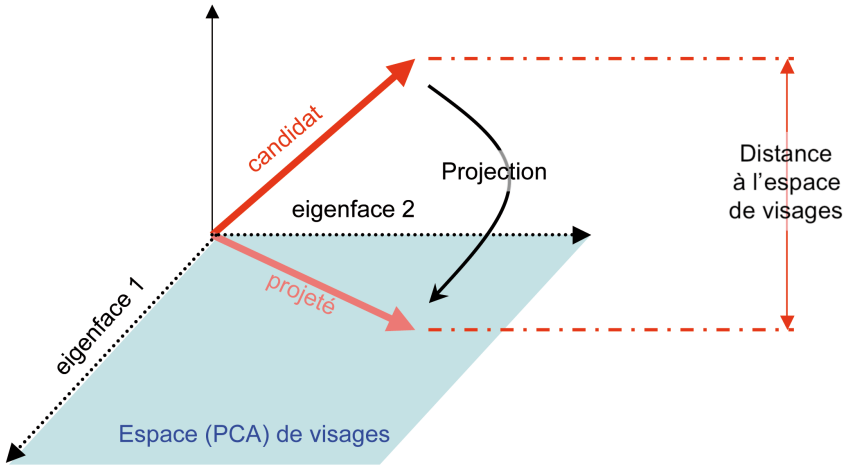


FIG. 3.8 – Distance à l'espace de visages

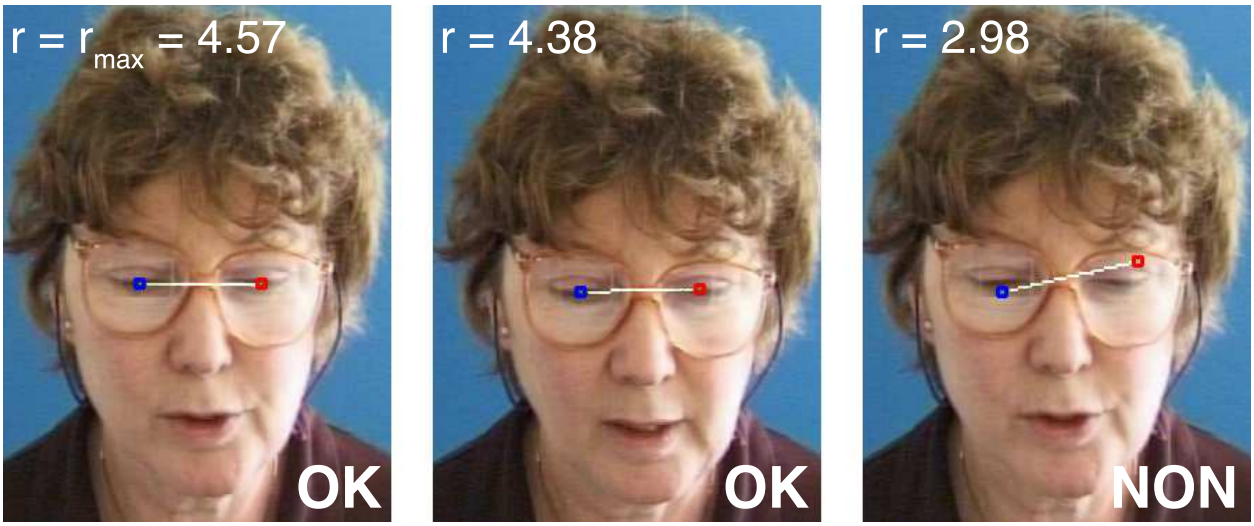


FIG. 3.9 – Sélection des meilleurs visages. Étant donnée une séquence vidéo, l'indice de confiance $r(t)$ est calculé pour chaque trame t. Le maximum $r_{\text{max}} = \max r(t)$ est choisi comme référence. Un visage t est finalement conservé si $r(t) > \alpha \cdot r_{\text{max}}$ où α est fixé à $2/3$ dans notre cas.

Distance de Mahalanobis Au moment du test, il s'agit de vérifier si la personne Γ dont le visage est acquis est bien la personne λ qu'elle prétend être. Les vecteurs de paramètres x^i extraits de la séquence de test (i variant de 1 à N) sont comparés aux vecteurs de paramètres x_λ^j extraits de la séquence d'enrôlement (j variant de 1 à N_λ) à l'aide de la distance de Mahalanobis [Mahalanobis, 1936] :

$$d(x^i, x_\lambda^j) \quad = \quad \sqrt{(x^i - x_\lambda^j)^t \Sigma_\lambda^{-1} (x^i - x_\lambda^j)} \tag{3.2}$$

où Σ_λ est la matrice de covariance des x_λ^j : il s'agit d'une distance euclidienne dans l'espace où chaque dimension est normalisée par sa variance. Ces $N \cdot N_\lambda$ distances sont alors triées dans l'ordre croissant et l'opposé de la moyenne des n plus petites distances est choisie comme la mesure de similarité $S_{\text{visage}}(\Gamma|\lambda)$ (dans notre cas, n est fixé à 10). Cette mesure est finalement comparée à un seuil permettant de vérifier l'identité clamée par la personne :

L'accès est accepté si $S_{\text{visage}}(\Gamma\|\lambda) > \theta$ et refusé dans le cas contraire.

Performances Le système de vérification du visage utilisé repose sur des vecteurs de paramètres de dimension 80 (les projections sur les 80 premières composantes principales). Ces composantes principales ont été obtenues à partir d'une base de données d'environ 2200 visages issus de plusieurs bases :

- ATT [AT&T Laboratories Cambridge, 1994] ;
- BANCA *world model* [Bailly-Baillière *et al.*, 2003] ;
- CALTECH [Weber, 1999] ;
- GeorgiaTech [Georgia Institute of Technology, 1999] ;
- Biomet [Garcia-Salicetti *et al.*, 2003].

Les performances de ce système sur le protocole P sont résumées dans la figure 3.10.

Discussion Au vu de la relative faiblesse de ce module, nous avons proposé deux pistes d'amélioration du système de vérification du visage. La première est une adaptation directe de la technique proposée par *Dehak et Chollet* [Dehak et Chollet, 2006] pour la vérification du locuteur par modèle de mélange de gaussiennes [Bredin *et al.*, 2006b]. La seconde s'attaque au problème de la paramétrisation en fusionnant deux systèmes, l'un basé sur les *eigenfaces* et l'autre sur les descripteurs SIFT [Landais *et al.*, 2007]. Ces deux publications se trouvent en annexe.

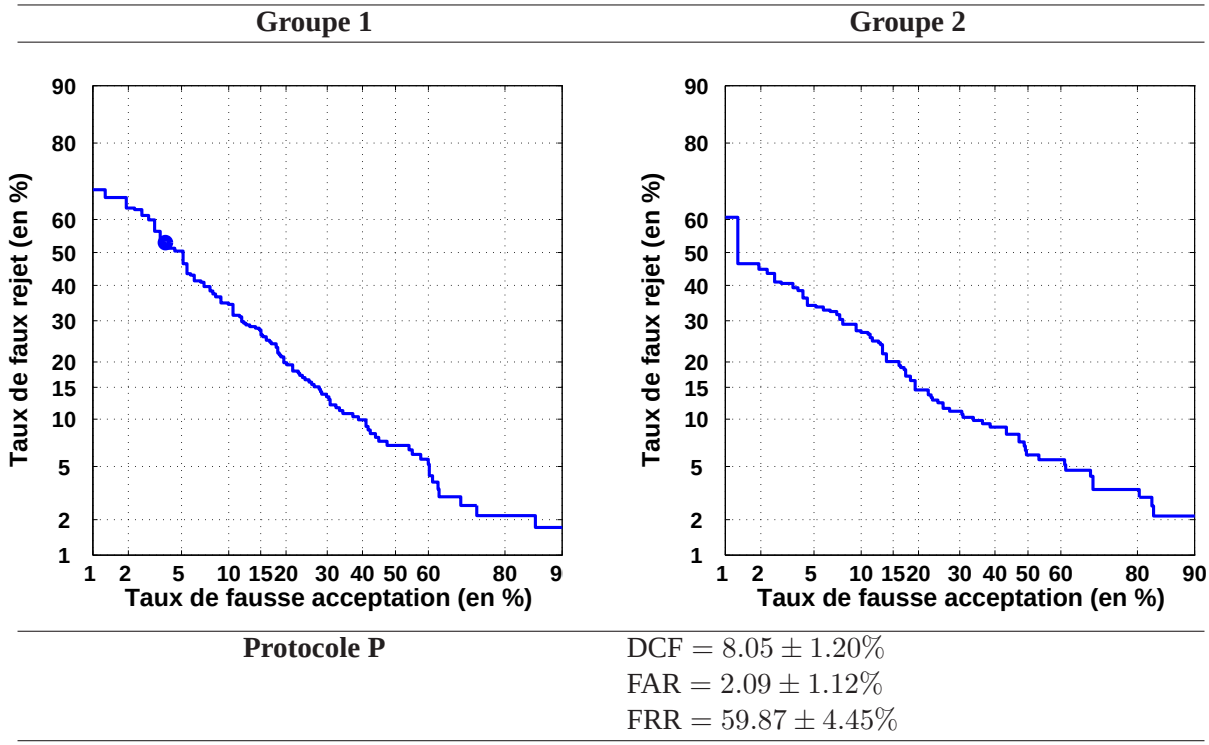


FIG. 3.10 – Performances du système de vérification du visage

3.3 Normalisation des scores

Rappelons le test menant à la décision finale des systèmes de vérification du locuteur et du visage :

L'accès est accepté si $S(\Gamma|\lambda) > \theta$ et refusé dans le cas contraire.

Le choix d'un seuil θ indépendant du client λ est un domaine de recherche à part entière, qui a été étudié en détails par la communauté des chercheurs en vérification du locuteur [Bimbot *et al.*, 2004]. La principale difficulté est issue de la grande variabilité des scores issus du module de comparaison. Cette variabilité peut provenir des conditions d'enrôlement différentes selon les clients (variabilité inter-client) ou des variabilités intra-client (dues à l'âge, l'état émotionnel ou de santé, ...) ou encore des conditions d'acquisition des données au moment du test. Une solution consiste à centrer, pour chaque client, la distribution des scores *imposteur* en appliquant une transformation σ/μ, résumée dans l'équation (3.3) :

$$S_{\text{normalisé}}(\Gamma|\lambda) = \frac{S(\Gamma|\lambda) - \mu_\lambda}{\sigma_\lambda} \tag{3.3}$$

Le lecteur intéressé pourra se référer à [Bimbot *et al.*, 2004] comme introduction aux différentes méthodes de normalisation des scores (Znorm, Hnorm, Tnorm, HTnorm, Cnorm, Dnorm, ...). Nous avons appliqué la Znorm.

3.3.1 Znorm

Au moment de l'enrôlement du client λ, son modèle est comparé à un ensemble de séquences d'imposteurs (extraites de l'ensemble de développement, typiquement) de façon à estimer la moyenne μ_λ et la variance ${\sigma_\lambda}^2$ des scores *imposteur* associés au client λ. On note Z le score issu de la Z-normalisation :

$$Z(\Gamma|\lambda) \quad = \quad \frac{S(\Gamma|\lambda) - \mu_\lambda}{\sigma_\lambda} \tag{3.4}$$

Comme l'illustre la figure 3.11, la Znorm a pour effet de centrer la distribution des scores *imposteur* (ligne fine) en en diminuant sensiblement la variance.

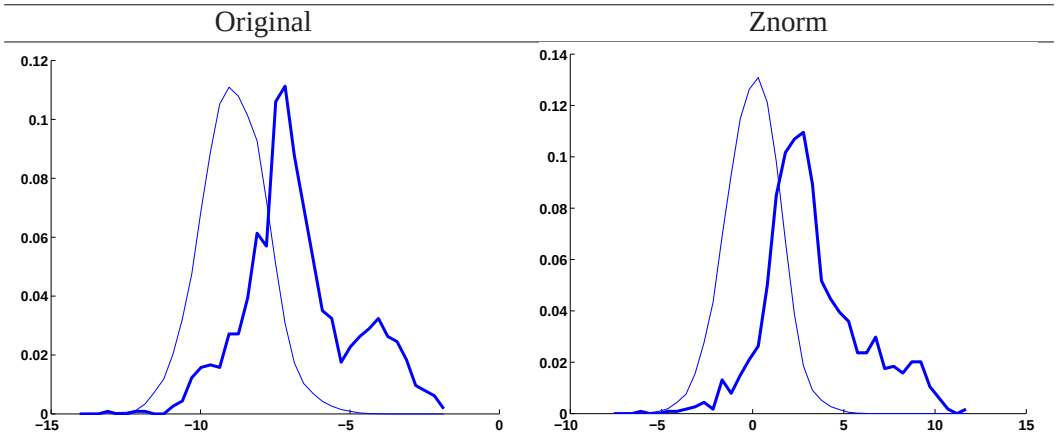


FIG. 3.11 – Effet de la Znorm sur les scores – Distribution des scores visage avant (à gauche) et après Znorm (à droite). La ligne épaisse correspond à la distribution des scores *client*, la ligne fine à celle des scores *imposteur*.

L'impact de la Znorm sur les performances des modules de vérification du locuteur et du visage est présenté dans la figure 3.12 à l'aide de courbes DET. Bien que l'amélioration apportée par la Znorm soit relativement faible, le système résultant est toujours au moins aussi bon que le système original utilisant les scores bruts.

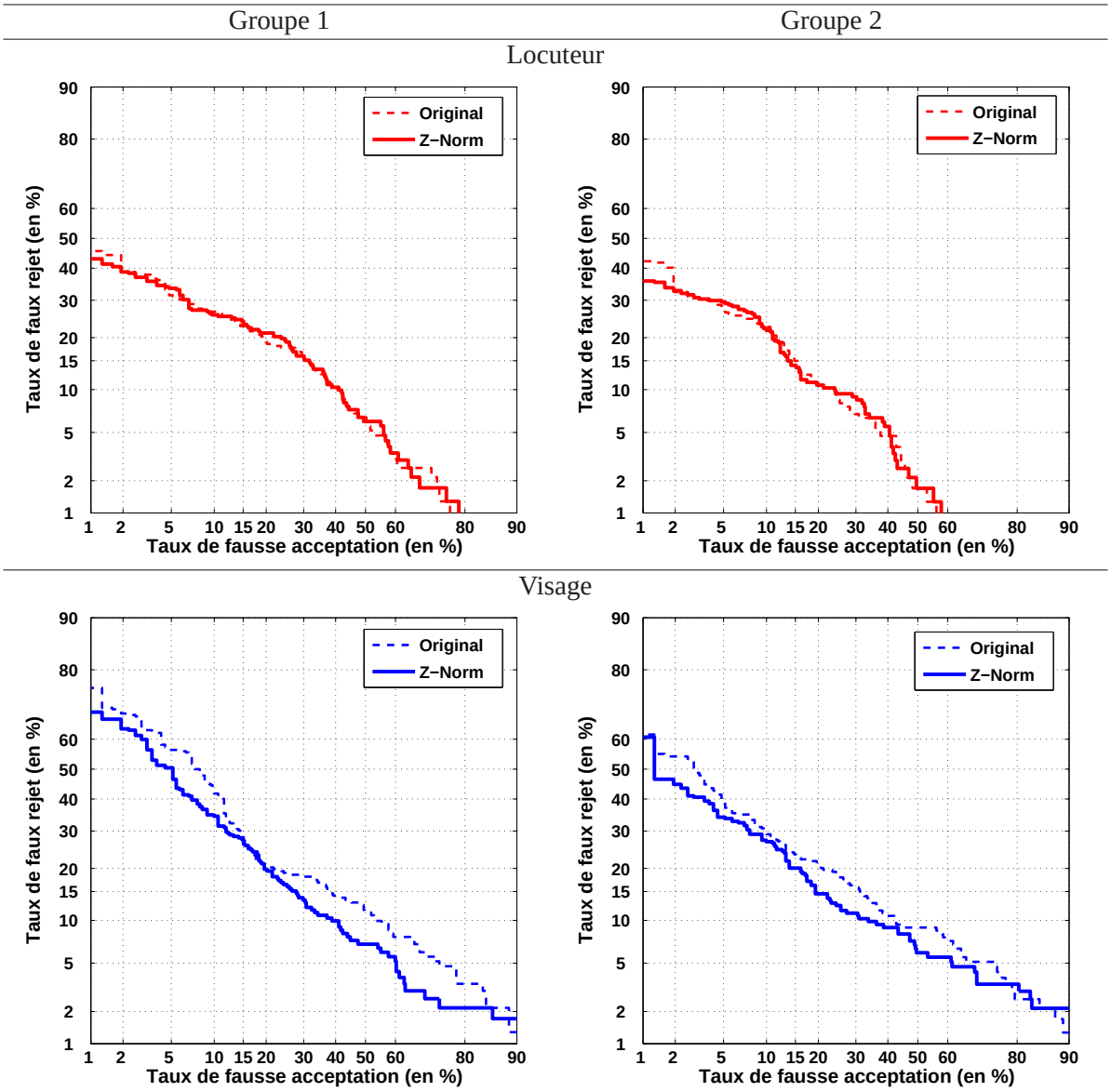


FIG. 3.12 – Effet de la Znorm sur les performances – Courbes DET avant et après Znorm, pour les modalités *voix* (en haut) et *visage* (en bas).

Remarque Les figures 3.5 et 3.10 (pages 55 et 60 respectivement) décrivant les résultats obtenus par les deux modules de vérification du locuteur et du visage tiennent compte de la Znorm.

3.3.2 Normalisation *tanh*

Le paragraphe 3.4 qui suit a pour objet la fusion au niveau des scores des deux systèmes de vérification du locuteur et du visage à l'aide d'une somme pondérée des scores S_{locuteur} et S_{visage} (ou plutôt leur version Z-normalisée Z_{locuteur} et Z_{visage}). Afin de faciliter la recherche des poids optimaux, une étape supplémentaire de normalisation des scores vise à s'assurer que les scores *locuteur* et les scores *visage* possèdent le même ordre de grandeur [Jain *et al.*, 1999, Ross *et al.*, 2006].

Les résultats de [Jain *et al.*, 2005] montrent que la normalisation *tanh* est l'une des techniques de normalisation les plus robustes et efficaces. Elle est définie par l'équation (3.5) :

$$\hat{S}(\Gamma|\lambda) = 0.5\left[1 + \tanh\left(0.01 \cdot \frac{Z(\Gamma|\lambda) - \mu_c}{\sigma_c}\right)\right] \tag{3.5}$$

où μ_c et ${\sigma_c}^2$ sont les moyenne et variance de la distribution des scores *client* estimées sur l'ensemble de développement. La figure 3.13 illustre, à droite, le résultat de cette normalisation : l'ordre de grandeur et l'amplitude de variation des scores locuteur et visage sont comparables.

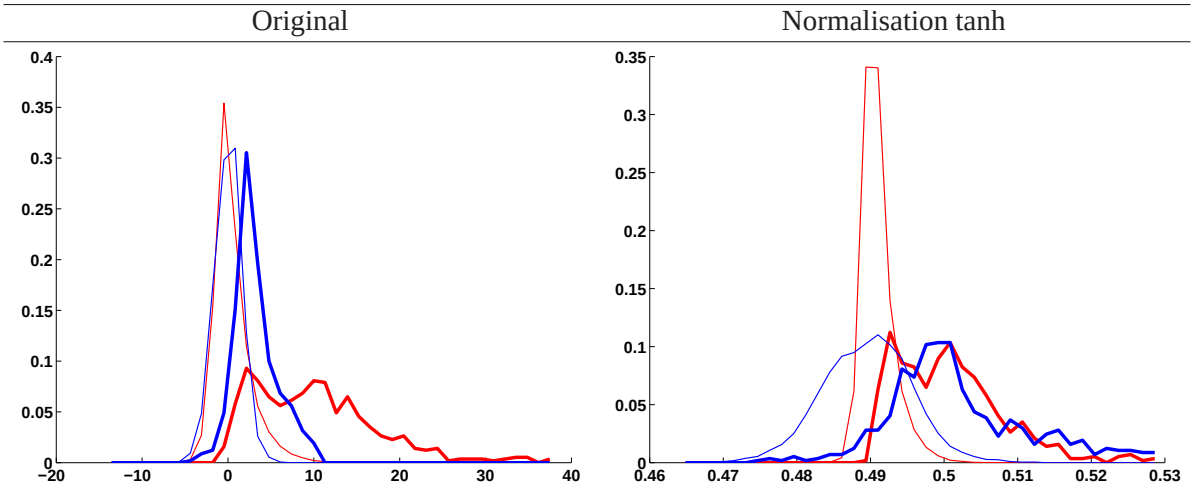


FIG. 3.13 – Effet de la normalisation *tanh* sur les scores – Distribution des scores locuteur (en rouge) et visage (en bleu) avant (à gauche) et après normalisation *tanh* (à droite). La ligne épaisse correspond à la distribution des scores *client*, la ligne fine à celle des scores *imposteur*.

3.4 Fusion des scores

À ce stade de la vérification, les deux modules de vérification du locuteur et du visage ont chacun fourni un score ($\hat{S}_{\text{locuteur}}(\Gamma|\lambda)$ et $\hat{S}_{\text{visage}}(\Gamma|\lambda)$, respectivement). L'objectif de la fusion des scores est d'obtenir, à partir de ces deux scores, un score global résultant en une performance globale meilleure que celle de chacun des deux modules. La fusion choisie, simple mais efficace [Jain *et al.*, 2005], consiste alors en une somme pondérée des scores des deux modules, comme le résume l'équation (3.6).

$$S(\Gamma|\lambda) \quad = \quad w_l \cdot \hat{S}_{\text{locuteur}}(\Gamma|\lambda) + w_v \cdot \hat{S}_{\text{visage}}(\Gamma|\lambda) \text{ avec } w_l + w_v = 1 \tag{3.6}$$

L'estimation des poids w_l et w_v se fait à l'aide de l'ensemble de développement, en minimisant la fonction de coût de détection DCF définie par l'équation (2.5) à la page 43.

L'accès est accepté si $S(\Gamma\|\lambda) > \theta$ et refusé dans le cas contraire.

Résultats L'estimation des poids w_l et w_v sur les ensembles de développement conduit à donner environ deux fois plus de poids à la modalité voix qu'à la modalité visage : $w_l = 0.66$ et $w_v = 0.34$ pour G1 et $w_l = 0.62$ et $w_v = 0.38$ pour G2. La figure 3.14 résume les performances du système basé sur la fusion *locuteur+visage*, pour le protocole P.

Discussion Au regard du tableau de la figure 3.14, les performances obtenues par le système fusionné, en termes de DCF et FAR, ne sont pas significativement différentes de celles obtenues par le meilleur des deux systèmes monomodaux, à savoir la vérification du locuteur. En revanche, l'amélioration apportée par la fusion en termes de faux rejet est statistiquement significative. À DCF constant, le système fusionné offre donc plus de confort aux clients légitimes qui se verront moins souvent refuser l'accès.

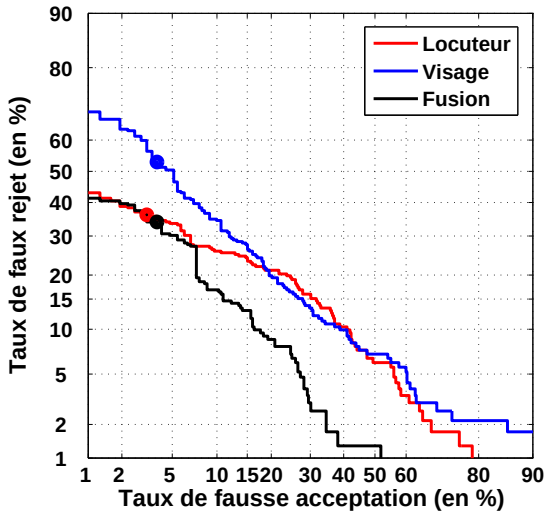
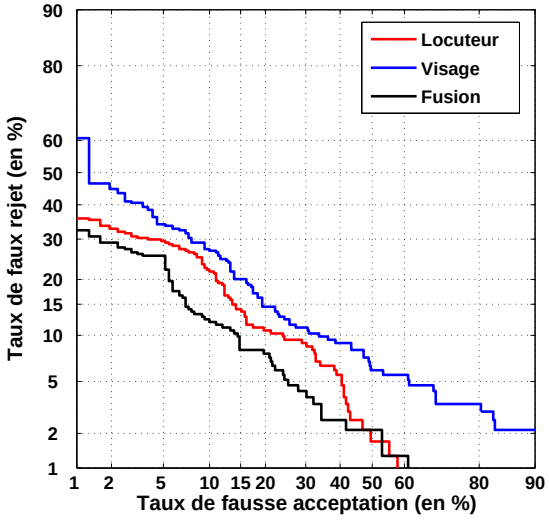
Groupe 1	Groupe 2
	
Fusion	DCF $= 5.82 \pm 1.19\%$ FAR $= 2.09 \pm 1.12\%$ FRR $= 37.55 \pm 4.40\%$
Locuteur	DCF $= 5.84 \pm 1.08\%$ FAR $= 1.61 \pm 0.99\%$ FRR $= 42.49 \pm 4.49\%$
Visage	DCF $= 8.05 \pm 1.20\%$ FAR $= 2.09 \pm 1.12\%$ FRR $= 59.87 \pm 4.45\%$

FIG. 3.14 – Performances du système locuteur+visage

Chapitre 4

Attaques

La grande majorité des systèmes de vérification d'identité basée sur les visages parlants repose uniquement sur la fusion (au niveau des scores) de deux sous-systèmes de vérification du locuteur et de reconnaissance du visage. En conséquence, un imposteur jouant un enregistrement sonore de la voix de sa cible tout en présentant une photographie de son visage devant la caméra obtiendrait l'accès : aucun des deux sous-systèmes évoqués plus haut ne permet de vérifier la présence effective d'une personne réelle devant la caméra.

Il est étonnant de constater que les bases de données et protocoles d'évaluation associés diffusés dans la communauté des chercheurs en biométrie partagent tous la même philosophie : ils sont définis de façon à évaluer les performances brutes de vérification d'identité et ne tiennent pas compte de l'éventualité d'attaques délibérées d'imposteur telles que celle que nous venons de décrire. À titre d'exemple, les accès *imposteur* du protocole P de la base BANCA ne sont considérés comme tels que du seul fait que le nom et l'adresse prononcés par la personne (l'imposteur) sont ceux d'une autre personne (la cible). Aucun réel effort n'est déployé par l'imposteur pour *ressembler* à sa cible et ainsi tromper le système. C'est à ce titre que nous parlons d'imposture aléatoire.

Par définition, un imposteur est quelqu'un qui essaie de se faire passer pour quelqu'un d'autre. Pour mettre toutes les chances de son côté, un imposteur mettra en oeuvre des techniques plus élaborées. Là où l'*imposteur aléatoire* ne possède aucune connaissance *a priori* sur sa cible autre que son nom et son adresse, l'*imposteur délibéré* aura préalablement collecté un maximum d'information sur sa cible.

Dans le cadre de la vérification d'identité biométrique d'un visage parlant, un imposteur essaiera de se procurer, à l'insu de sa cible, du matériel biométrique le représentant. Il convient de remarquer ici que la voix et le visage d'une personne ne sont pas des informations secrètes. À moins que la cible vive dans un endroit reclus tenu secret et coupé du monde, il est aisé de se procurer la photographie du visage d'une personne et/ou un enregistrement sonore de sa voix. La généralisation et la miniaturisation des appareils photo rend l'acquisition discrète de la première très facile et une simple conversation téléphonique permet d'acquérir le second. Cette spécificité de la multimodalité visage parlant est donc aussi sa plus grande faiblesse. Comparativement, il est beaucoup plus difficile (mais néanmoins réalisable) d'obtenir une image de l'iris à l'insu d'une personne.

Les tentatives délibérées d'imposture n'ont que très peu été étudiées dans la littérature. *Chetty et Wagner* simulent des attaques dans lesquelles un imposteur présente une photographie du visage de sa cible devant la caméra en construisant artificiellement des séquences audiovisuelles où la même image du visage de la cible est répétée tout au long de la séquence [Chetty et Wagner, 2004]. Bien qu'elle ait le mérite d'exposer les limites d'un système de vérification d'identité basée sur les visages parlants, cette simulation est néanmoins peu réaliste et facilement détectable. *Kollreider et al.* améliorent un peu ces simulations en ajoutant une translation horizontale et verticale à l'image répétée [Kollreider *et al.*, 2005]. Pour gagner en réalisme, *Jee et al.* ont réalisé de vraies attaques en présentant une photographie de la cible devant la caméra [Jee *et al.*, 2006]. L'exploitation des résultats est néanmoins limitée par le nombre relativement restreint des attaques (seules 10 personnes ont été photographiées).

En outre, à notre connaissance, l'impact de ces attaques sur les performances globales de vérification d'identité n'a jamais été étudié : seule l'efficacité des algorithmes de détection des attaques est reportée. Il est donc difficile d'évaluer l'apport des méthodes proposées pour la vérification d'identité.

4.1 Attaques de type rejeu

Afin de montrer les limites du système initial, nous avons simulé des attaques de type *rejeu* qui nous semblent être les attaques les plus faciles à mettre en oeuvre pour mettre à défaut un système basé sur la fusion des scores de deux sous-systèmes de vérification du locuteur et du visage. Parmi les attaques de type *rejeu*, nous distinguons les attaques suivantes[1] :

[1] Un exemple de séquence *imposteur* est proposé en ligne à l'adresse `http://www.tsi.enst.fr/~bredin/these`, section *Compléments multimédia*, pour chaque type d'attaque.

Paparazzi Dans le cadre du scénario baptisé *Paparazzi*, l'imposteur a acquis au préalable une photographie du visage de sa cible. Au moment du test, il la présente devant la caméra tout en prononçant le nom et l'adresse de sa cible. Nous avons simulé ce type d'attaque en filmant un morceau de papier bleu présenté devant la caméra, sur lequel nous avons *collé* la photographie de la cible *a posteriori*. La figure 4.1 résume schématiquement ce scénario. Cette simulation est beaucoup plus réaliste que celles proposées dans [Kollreider *et al.*, 2005] et [Jee *et al.*, 2006] puisque le mouvement de la photographie est issu d'une séquence bien réelle où le morceau de papier bleu est translaté et incliné manuellement au cours du temps. Tous les accès *imposteur* du protocole P original sont ainsi modifiés, soit un total de 312 tests *imposteur* différents par groupe.

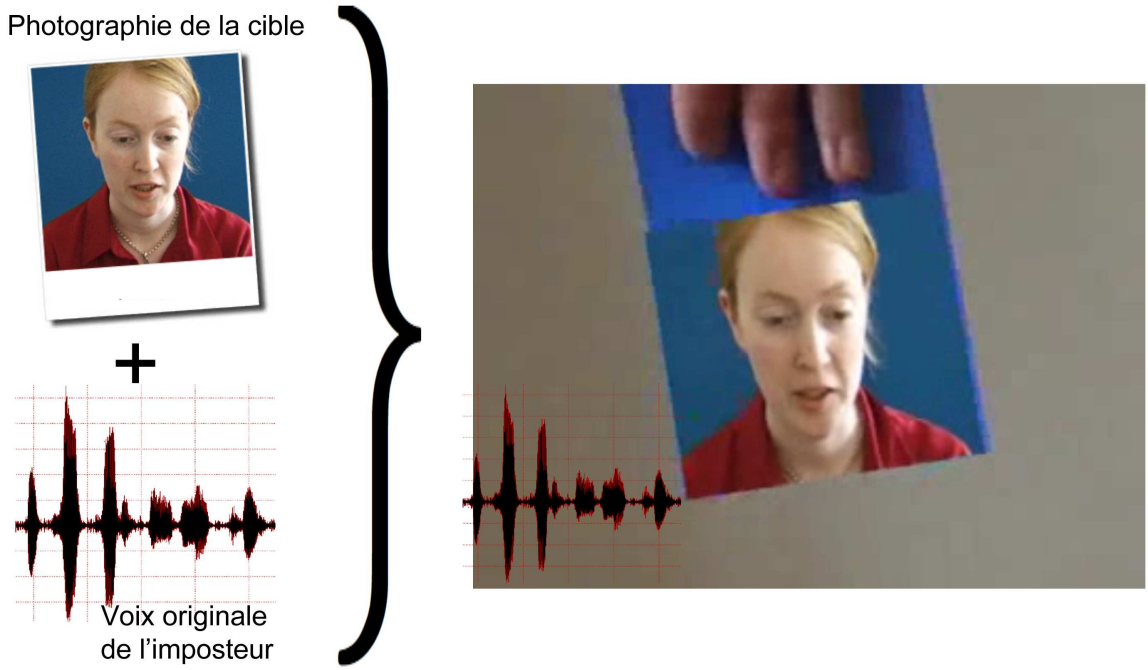


FIG. 4.1 – Attaque de type *Paparazzi*

Echelon Dans le cadre du scénario baptisé *Echelon*, l'imposteur a acquis au préalable un enregistrement de la voix de sa cible (au cours d'une conversation téléphonique, par exemple). Au moment du test, il joue cet enregistrement à l'aide d'un magnétophone et se présente devant la caméra. Nous avons simulé ce type d'attaque en remplaçant la bande sonore des tests *imposteur* par l'audio d'une séquence de la cible. La figure 4.2 résume schématiquement ce scénario. Tous les accès *imposteur* du protocole P original sont ainsi

modifiés, soit un total de 312 tests *imposteur* différents par groupe.

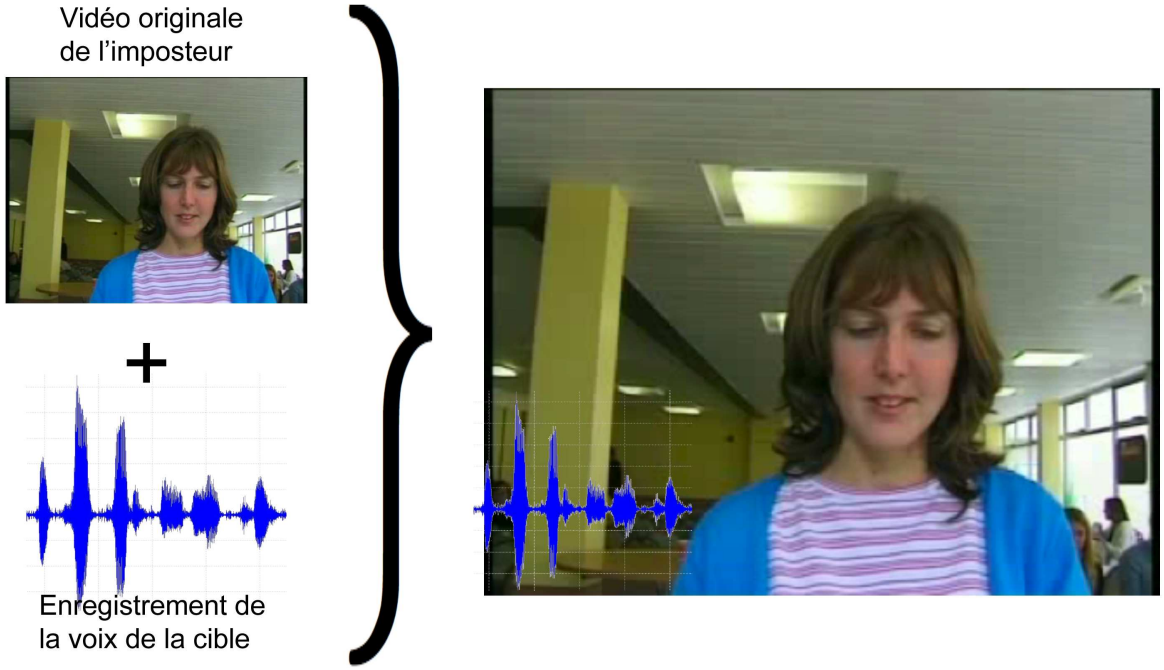


FIG. 4.2 – Attaque de type *Echelon*

Big Brother Le scénario baptisé *Big Brother* est une combinaison des deux premiers scénarii : l'imposteur a acquis à la fois une photographie du visage et un enregistrement de la voix de sa cible. Au moment du test, il présente la photographie devant la caméra tout en jouant l'enregistrement de la voix de sa cible. La figure 4.3 résume schématiquement ce scénario. Pour chaque client, une séquence *imposteur* de ce type est générée, soit un total de 26 tests *imposteur* différents par groupe.

4.2 *Crazy Talk*

Un quatrième scénario faisant appel à des techniques d'animation de visage a aussi été envisagé. Il s'agit d'animer une photographie du visage de la cible en accord avec un enregistrement audio de sa voix. Pour cela, nous avons utilisé le logiciel commercial *CrazyTalk* de la société *Reallusion*[2], appliqué aux données

[2]http ://www.reallusion.com/crazytalk/

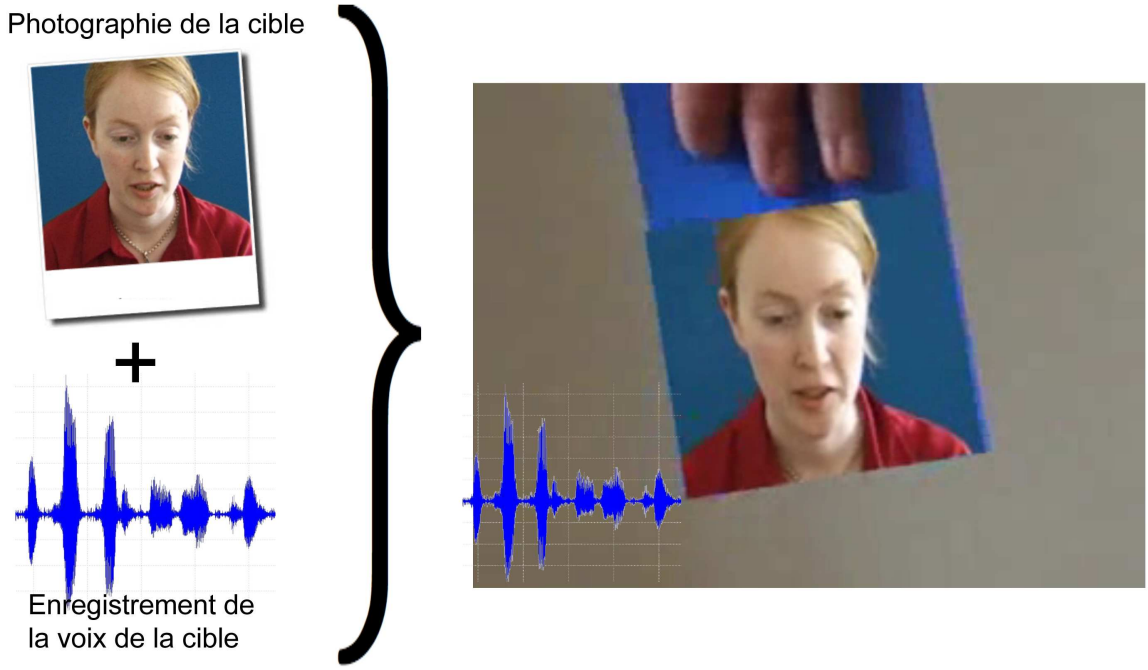


FIG. 4.3 – Attaque de type *Big Brother*

du scénario *Big Brother*. Dans le scénario *Crazy Talk*, l'imposteur possède donc une photographie du visage et un enregistrement de la voix de la cible et utilise le logiciel *CrazyTalk* pour générer une séquence audiovisuelle dans laquelle les lèvres du visage sont animées en fonction de l'enregistrement de la voix. Au delà de la seule animation des lèvres, le logiciel anime aussi d'autres parties du visage : en particulier, des clignements des yeux sont également générés, rendant la détection de l'attaque encore plus difficile. Pour chaque client, une séquence *imposteur* de ce type est générée, soit un total de 26 tests *imposteur* différents par groupe.

4.3 Évaluation

La figure 4.4 résume les performances du système de fusion *locuteur+visage* face aux différentes attaques introduites dans ce chapitre. À taux de faux rejet FRR constant (les accès *client* n'étant pas modifiés), le taux de fausse acceptation FAR passe d'environ 2% à 60% pour les attaques de type *Paparazzi* et à 76% pour les attaques de type *Echelon*. Bien pire qu'un système qui prendrait une décision aléatoire (pour lequel FAR = 50%), le système de fusion *locuteur+visage* laisse passer plus de 90% des attaques *Big Brother* et *Crazy*

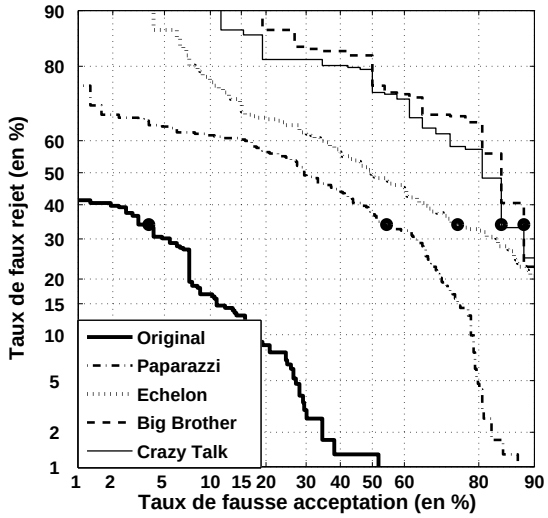
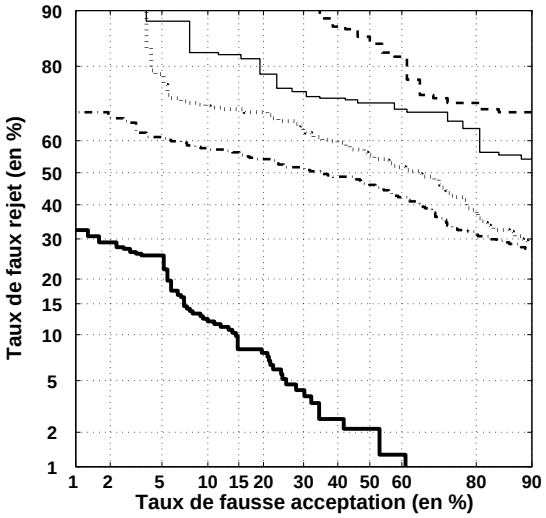
Groupe 1	Groupe 2
	
Protocole P original	DCF $= 5.82 \pm 1.19\%$ FAR $= 2.09 \pm 1.12\%$ FRR $= 37.55 \pm 4.40\%$
Attaque *Paparazzi*	DCF $= 62.71 \pm 3.84\%$ FAR $= 59.55 \pm 3.85\%$
Attaque *Echelon*	DCF $= 79.40 \pm 3.33\%$ FAR $= 76.40 \pm 3.33\%$
Attaque *Big Brother*	DCF $= 97.04\% \in [72\%, 100\%]$ FAR $= 94.23\% \in [69\%, 100\%]$
Attaque *Crazy Talk*	DCF $= 95.14\% \in [70\%, 100\%]$ FAR $= 92.31\% \in [68\%, 100\%]$

FIG. 4.4 – Performances du système *locuteur+visage* face aux attaques

Talk.

Conclusion

La grande majorité des systèmes de vérification d'identité basée sur les visages parlants repose uniquement sur la fusion au niveau des scores de deux modules de vérification du locuteur et de reconnaissance du visage. Aussi, à défaut d'un module de vérification de présence effective d'une personne bien réelle devant l'objectif de la caméra, un imposteur pourrait simplement rejouer un enregistrement sonore de la voix de sa cible tout en présentant une photographie de son visage devant la caméra. L'éventualité des attaques de type *rejeu* (ou *replay attacks* dans la littérature anglophone) a été peu fréquemment étudiée dans la littérature. Aussi, nous avons défini et simulé quatre scénarii d'attaque et évalué leur impact sur notre système de référence, reproduisant le schéma classique de fusion au niveau des scores d'un module de vérification basé sur la voix et d'un autre module basé sur le visage. Notre système de référence se trouve alors dans l'incapacité de détecter ces attaques.

L'objet de la seconde partie est donc de rendre le système de référence robuste à toutes ces attaques. Une première solution consiste à demander à la personne de prononcer une phrase aléatoire et de s'en assurer à l'aide d'un système de transcription automatique de la parole, empêchant ainsi l'utilisation d'un enregistrement sonore de la voix de la cible. Une deuxième solution consiste à étudier le mouvement du visage afin de déterminer s'il est bien réel. Dans [Kollreider *et al.*, 2005], les auteurs étudient les mouvements relatifs des différentes parties du visage (nez, oreilles, yeux, ...). Dans le cas où une photographie est présentée devant la caméra, chaque partie du visage détecté possède quasiment le même vecteur mouvement. En revanche, lorsqu'une personne bien réelle bouge devant la caméra et tourne la tête, les différentes parties du visage possèdent des vecteurs mouvements apparents différents. Dans [Jee *et al.*, 2006], les yeux sont détectés dans la séquence d'images et les variations des deux régions correspondant aux deux yeux permettent de déterminer si le visage présent devant le caméra est bien réel ou s'il s'agit d'une photographie. Cependant, la qualité d'animation de visage obtenue par un logiciel tel que *CrazyTalk* de *Reallusion* laisse présager que de telles méthodes seraient insuffisantes. Ainsi, la troisième solution, que nous avons adoptée, est basée sur l'analyse de la synchronie audiovisuelle entre la voix et le mouvement des lèvres : il s'agit de s'assurer que la voix

acquise par le microphone et le mouvement des lèvres acquises par la caméra ont été produits simultanément par une seule et même personne.

Deuxième partie

Synchronie audiovisuelle

Introduction

Le signal de parole est intrinsèquement *bimodal*. Si son traitement se limite souvent à ces caractéristiques acoustiques (transcription [Gauvain et Lamel, 2000, Young, 2001, Deng et Huang, 2004] et vérification du locuteur [Furui, 1997, Reynolds *et al.*, 2000a, Reynolds, 2002, Ben et Bimbot, 2003, Ben, 2004]), son complémentaire visuel peut être d'une grande aide, particulièrement dans des conditions acoustiques dégradées [Potamianos *et al.*, 2004].

La démonstration la plus évidente de cette *complémentarité* est donnée par la capacité qu'ont les personnes atteintes de surdité ou de problèmes d'audition à "lire sur les lèvres". En outre, dans un environnement acoustique bruité, la compréhension du signal de parole est aussi améliorée lorsque le signal visuel est disponible (i.e. lorsque les lèvres du locuteurs sont visibles). Il est par exemple plus facile de distinguer le son [*m*] du son [*n*] en voyant les lèvres. À l'opposé, il n'est pas possible de distinguer le son [*b*] du son [*p*] à la seule vue des lèvres : les signaux acoustiques correspondant étant, eux, bien distincts. Enfin, l'effet *McGurk* est une démonstration bien connue de l'intrication du signal acoustique et du signal visuel dans l'interprétation globale que l'homme a de ceux-ci. Combiner le signal acoustique correspondant au son [*ba*] au signal visuel correspondant au son [*ga*] entraîne la sensation du son [*da*]. Une vidéo de démonstration de ce phénomène peut être trouvée sur l'Internet[3].

Le signal visuel de parole correspond à l'observation des déformations et mouvements de l'appareil vocal dont résulte le signal acoustique de parole. Aussi, plus que *complémentaires*, ces deux signaux sont profondément *corrélés*, le second résultant du premier. Les travaux de *Yehia* et *Barker* ont montré qu'il était possible de partiellement déduire les signaux acoustiques de l'observation du signal visuel, et inversement [Yehia *et al.*, 1998, Barker et Berthommier, 1999b, Barker et Berthommier, 1999a].

[3] Effet *McGurk* : `http://www.media.uio.no/personer/arntm/McGurk_english.html`

Par la suite, nous qualifierons de *synchrones* deux signaux acoustique et visuel produits simultanément par une seule et même personne.

Dans le chapitre 5, nous proposons un tour d'horizon de la littérature s'intéressant au problème particulier de la synchronie audiovisuelle. La question de la paramétrisation du signal de parole audiovisuelle sera abordée ainsi que celle des différentes méthodes (le plus souvent statistiques) proposées pour évaluer le degré de synchronie entre les signaux de parole acoustique et visuel. Dans le chapitre 6, nous proposons une nouvelle mesure de synchronie et étudions son application à la détection d'asynchronie. Nous dérivons ensuite de cette mesure une nouvelle modalité biométrique et évaluons ses performances pour la vérification d'identité dans le chapitre 7. Enfin, dans le chapitre 8, nous proposons des stratégies originales de fusion de cette nouvelle modalité et du système de référence de façon à rendre ce dernier robuste aux attaques.

Chapitre 5

État de l'art

5.1 Paramétrisation de la parole

Cette section fait l'inventaire des différentes paramétrisations du signal de parole utilisées dans la littérature relative à la synchronie audiovisuelle. Toutes ces paramétrisations partagent l'objectif commun de réduire les données brutes de façon à permettre une bonne modélisation par la suite. Nous aborderons successivement la question des paramètres issus du signal de parole acoustique et ceux issus du signal de parole visuel.

5.1.1 Paramètres acoustiques

Classiquement, les vecteurs de paramètres acoustiques sont extraits du signal audio à partir d'une fenêtre temporelle glissante avec recouvrement.

Énergie brute L'amplitude du signal audio peut être utilisée telle qu'elle. Dans [Hershey et Movellan, 1999], les auteurs extraient l'énergie acoustique moyenne sur la fenêtre courante de façon à obtenir une paramétrisation mono-dimensionnelle relative à l'activité vocale. Des méthodes similaires faisant appel à la valeur efficace – Root Mean Square (RMS) en anglais – ou à la log-énergie sont aussi proposées dans les références [Barker et Berthommier, 1999b, Bredin *et al.*, 2006c].

Périodogramme Dans [Fisher *et al.*, 2001], un périodogramme du signal audio sur la plage de fréquence [0–10 kHz] est calculé sur une fenêtre glissante de durée 2/29.97 s (correspondant à la durée de 2 trames de vidéo à une fréquence de 29.97 images par seconde) et utilisé directement comme les paramètres du flux audio.

***Mel-Frequency Cepstral Coefficients* (MFCC)** Le fait que les coefficients MFCC soient très fréquemment utilisés [Slaney et Covell, 2000, Cutler et Davis, 2000, Nock *et al.*, 2002, Iyengar *et al.*, 2003] peut s'expliquer de façon pragmatique du fait qu'ils constituent la paramétrisation *état-de-l'art* de la majorité des systèmes de traitement automatique de la parole acoustique [Reynolds *et al.*, 2000b] et qu'ils ont démontré leur efficacité que ce soit en transcription de la parole ou en vérification du locuteur. La figure 3.4 à la page 54 résume les étapes de calcul des coefficients MFCC.

***Linear-Predictive Coding* (LPC) et *Line Spectral Frequencies* (LSF)** L'utilisation des LPC, ainsi que des LSF qui en dérivent [Sugamura et Itakura, 1986], a aussi été largement investiguée. *Yehia et al.* ont montré la plus grande corrélation de la géométrie du conduit vocal avec les LSF qu'avec les LPC [Yehia *et al.*, 1998].

Une comparaison de ces différents paramètres acoustiques, appliqués dans le cadre de l'opérateur linéaire *FaceSync* (voir plus bas), est rapportée dans [Slaney et Covell, 2000]. En deux mots, dans le cadre de leurs travaux, les auteurs concluent que les paramètres MFCC, LSF et LPC montrent des liens avec le signal de parole visuel plus forts que le périodogramme ou l'énergie brute. Ces résultats sont cohérents avec ceux que nous avons obtenus dans nos travaux (énergie brute dans [Bredin *et al.*, 2006c] vs. MFCC dans [Bredin et Chollet, 2007]).

5.1.2 Paramètres visuels

Dans cette section, nous appellerons région d'intérêt – *Region of Interest* (ROI) en anglais – la zone de l'image autour de la bouche. Cette région peut être beaucoup plus large que la seule zone des lèvres, jusqu'à inclure la mâchoire et les joues. Par la suite, nous ferons l'hypothèse que cette région a été détectée au préalable. La plupart des paramètres visuels proposés dans la littérature relative à la synchronie est identique à ceux utilisés dans le cadre de la transcription automatique de la parole audiovisuelle. Cependant, nous verrons que des paramétrisations bas-niveaux beaucoup plus simples spécifiques à la synchronie ont aussi été étudiées dans le cadre de l'étude de la synchronie :

Pixels Il s'agit de l'équivalent visuel de l'énergie acoustique brute. Dans [Hershey et Movellan, 1999] et [Iyengar *et al.*, 2003], l'intensité des pixels est utilisée telle qu'elle. Nos premiers travaux considéraient la somme des intensités des pixels en niveaux de gris dans la ROI, de façon à obtenir une paramétrisation mono-dimensionnelle du signal de parole visuel [Bredin *et al.*, 2006c].

L'extraction de paramètres holistiques revient à considérer la ROI comme un tout, une source d'information insécable :

Transformée en cosinus discrète *Nock et al.* applique une transformation en cosinus discrète – *Discrete Cosine Transform* (DCT) en anglais – sur la ROI, en ne conservant que les coefficients les plus énergétiques : il s'agit d'une technique classique dans le domaine de la compression d'image. À un coefficient multiplicatif de normalisation près, les coefficients DCT extraits d'une ROI de dimension $N \times N$ sont définis par l'équation (5.1) :

$$\mathrm{DCT}(u,v) = \sum_{i=1}^{N}\sum_{j=1}^{N} I(i,j)\cos\left[\frac{\pi}{N}\left(i-\frac{1}{2}\right)(u-1)\right]\cos\left[\frac{\pi}{N}\left(j-\frac{1}{2}\right)(v-1)\right] \tag{5.1}$$

où $u \in \{1 \ldots N\}$, $v \in \{1 \ldots N\}$ et $I(i,j)$ est l'intensité du pixel (i,j).

Eigenlips Des transformations linéaires tenant compte de la distribution des niveaux de gris spécifique à la ROI ont aussi été proposées. Ainsi, *Bregler et al.* projettent la ROI (représentée par un vecteur contenant la valeur de tous les pixels) sur un espace vectoriel préalablement calculé par analyse en composantes principales – Principal Components Analysis (PCA) en anglais. Les auteurs travaillent sur une ROI couvrant la zone des lèvres : à partir d'un ensemble d'apprentissage constitué de centaines d'imagettes de lèvres, des *eigenlips* sont calculés par PCA, par analogie à la méthode des *eigenfaces* [Turk et Pentland, 1991a], dans le but d'extraire les paramètres codant pour un maximum de variations de la ROI [Bregler et Konig, 1994].

Géométrie Des méthodes considèrent les lèvres comme un objet déformable dont les paramètres géométriques peuvent être extraits. Ils sont la plupart du temps basés sur des points caractéristiques qui nécessite une localisation automatique préalable. Dans [Barker et Berthommier, 1999b], deux caméras fournissent une vue frontale et de profil du visage. La localisation automatique de points caractéristiques de la ROI (tels que les commissures des lèvres, par exemple) est facilitée grâce à un maquillage adapté. *Goecke et al.* résument la forme de la bouche en quatre paramètres : la largeur et la hauteur de la bouche, la protrusion des lèvres et un dernier paramètre que les auteurs appellent un *compteur de dents* (*teeth count*, en anglais) et qui constitue une mesure de la visibilité des dents. Le modèle déformable composé de plusieurs courbes polynomiales, proposé par *Eveno et al.* dans [Eveno et Besacier, 2005b, Eveno et Besacier, 2005a] suit le contour des lèvres : la hauteur, la largeur et l'aire de la bouche en sont déduites. Enfin, le rapport entre l'ouverture et la largeur des lèvres constitue l'unique paramètre visuel dans [Chetty et Wagner, 2004].

Paramètres dynamiques *Chibelushi et al.* [Chibelushi *et al.*, 2002] soulignent que, bien qu'une importante part de l'information de parole soit dynamique, l'extraction de paramètres dynamiques est rarement pratiquée dans la littérature. Cependant, quelques travaux sur la synchronie essaient d'intégrer cette dimen-

sion du signal de parole. Ainsi, l'utilisation des dérivées temporelles est proposée dans [Fox et Reilly, 2003]. *Cutler et al.* calculent la variation temporelle totale (entre deux trames vidéo consécutives) de la valeur des pixels de la ROI, selon l'équation (5.2).

$$v_t = \sum_{i=1}^{W}\sum_{j=1}^{H} |I_t(i,j) - I_{t+1}(i,j)| \tag{5.2}$$

où $I_t(i,j)$ est l'intensité du pixel de coordonnées (i,j) de la ROI de la trame t.

5.1.3 Fréquences d'échantillonnage

Les fréquences d'échantillonnage des paramètres acoustique et visuel sont souvent très différentes. Dans le domaine de la vérification du locuteur par exemple, les MFCC peuvent être extraits toutes les 10 ms alors que les séquences vidéo sont généralement encodées à 25 fps (images par seconde) ou 29.97 fps, en fonction du *codec* utilisé. Par conséquent, il est souvent requis d'équilibrer les fréquences d'échantillonnage (sous-échantillonnage des paramètres acoustiques ou sur-échantillonnage des paramètres visuels) avant même de pouvoir évaluer la synchronie audiovisuelle. Cependant, bien que l'extraction de l'énergie acoustique brute ou le calcul du spectrogramme peuvent être effectués directement sur des fenêtres plus larges (et donc avec une fréquence d'échantillonnage proche de celle des paramètres visuels), le sous-échantillonnage des paramètres audio est connu pour dégrader les performances en traitement automatique de la parole acoustique. Aussi, le sur-échantillonnage des paramètres visuels lui est souvent préféré (par interpolation linéaire, par exemple). Il est aussi envisageable d'utiliser directement une caméra à 100 fps ou d'utiliser des paradigmes traitant directement les paramètres acoustique et visuel aux fréquences d'échantillonnage originales [Cutler et Davis, 2000, Bengio, 2003].

5.2 Sous-espaces audiovisuels

Dans cette section, nous présentons les transformations qui sont appliquées dans les espaces acoustique, visuel et/ou audiovisuel définis par les paramètres listés dans la section précédente. Ces transformations ont toujours pour but de trouver des sous-espaces dans lesquels la mesure de la synchronie audiovisuelle se trouve améliorée.

5.2.1 Analyse en composantes principales

L'analyse en composantes principales (PCA) est une transformation linéaire visant à trouver un espace de projection dans lequel l'étalement des données (leur variance) soit maximisé. La PCA permet d'obtenir une base de composantes principales, à partir des vecteurs propres de la matrice de covariance des vecteurs de paramètres issus d'un large ensemble d'apprentissage. Dans [Chibelushi *et al.*, 1997b], la PCA est appliquée dans un espace audiovisuel (créé par concaténation des paramètres acoustiques et visuels) de façon à réduire sa dimensionnalité, tout en conversant les caractéristiques contribuant le plus à sa variance.

5.2.2 Analyse en composantes indépendantes

L'analyse en composantes indépendantes – *Independent Component Analysis* (ICA) en anglais – a été introduite afin de résoudre le problème de séparation de sources [Hyvärinen, 1999]. Dans [Sodoyer *et al.*, 2003], les auteurs tiennent compte des paramètres visuels de parole afin d'améliorer la séparation de différentes sources de parole. Dans [Smaragdis et Casey, 2003], l'ICA est appliquée à un enregistrement audiovisuel d'une session de piano : la caméra fait un gros plan sur le clavier et le signal de musique est acquis à l'aide d'un microphone. L'ICA permet de découvrir clairement la correspondance entre la note acoustique et la note visuelle (le mouvement de la touche correspondante). Cependant, aucune mention de l'application de l'ICA au signal de parole audiovisuelle n'a été trouvée dans la littérature.

5.2.3 Analyse en corrélation canonique

Étant donnés deux flux de paramètres acoustiques $X \in \mathbb{R}^n$ et visuels $Y \in \mathbb{R}^m$, l'objectif de l'analyse de corrélation canonique – Canonical Correlation Analysis (CANCOR) en anglais – est de déterminer les directions $\mathbf{a} \in \mathbb{U}^n$ et $\mathbf{b} \in \mathbb{U}^m$ (avec $\mathbb{U}^d = \left\{ z \in \mathbb{R}^d | \|z\| = 1 \right\}$) telles que les projections de X et Y sur ces deux vecteurs maximisent leur corrélation (voir l'équation (5.3)).

Proposition 1 (Analyse de corrélation canonique)

$$(\boldsymbol{a}, \boldsymbol{b}) = \underset{(a,b) \in \mathbb{U}^n \times \mathbb{U}^m}{\operatorname{argmax}} \operatorname{corr}\left(a^t X, b^t Y\right) \tag{5.3}$$

Soient $\boldsymbol{a} \in \mathbb{U}^n$ et $\boldsymbol{b} \in \mathbb{U}^m$ définis par l'équation (5.3). $\boldsymbol{a}$ est le vecteur propre normé correspondant à la plus grande valeur propre λ_1 de la matrice $C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX}$ et $\boldsymbol{b}$ est le vecteur normé colinéaire à $C_{YY}^{-1} C_{YX} \boldsymbol{a}$.

Démonstration 1 *On définit*

$$\begin{aligned}\rho &= \operatorname{corr}\left(a^tX, b^tY\right) \\ &= \frac{\operatorname{cov}\left(a^tX, b^tY\right)}{\sqrt{\operatorname{cov}\left(a^tX, a^tX\right)}\sqrt{\operatorname{cov}\left(b^tY, b^tY\right)}} \\ &= \frac{a^tC_{XY}b}{\sqrt{a^tC_{XX}a}\sqrt{b^tC_{YY}b}}\end{aligned}$$

En écrivant ce problème de maximisation sous sa forme lagrangienne et en dérivant par rapport à a et b, nous obtenons les équations de l'analyse en corrélation canonique 5.4 et 5.5 (voir [Weenink, 2003] pour tous les détails) :

$$\left(C_{XY}C_{YY}^{-1}C_{XY}^t - \rho^2 C_{XX}\right)a = 0 \quad (5.4)$$

$$\left(C_{XY}^tC_{XX}^{-1}C_{XY} - \rho^2 C_{YY}\right)b = 0 \quad (5.5)$$

On multiplie à gauche (5.4) par $C_{XY}^tC_{XX}^{-1}$ pour obtenir :

$$\left(C_{XY}^tC_{XX}^{-1}C_{XY}C_{YY}^{-1}C_{XY}^t - \rho^2 C_{XY}^t\right)a = 0$$

L'introduction de $C_{YY}C_{YY}^{-1} = I$ nous permet d'obtenir :

$$\begin{aligned}\left(C_{XY}^tC_{XX}^{-1}C_{XY}C_{YY}^{-1}C_{XY}^t - \rho^2 C_{YY}C_{YY}^{-1}C_{XY}^t\right)a &= 0 \\ \left(C_{XY}^tC_{XX}^{-1}C_{XY} - \rho^2 C_{YY}\right)C_{YY}^{-1}C_{XY}^ta &= 0\end{aligned}$$

Nous avons ainsi montré que les valeurs propres des équations (5.4) et (5.5) sont les mêmes et que $b = C_{YY}^{-1}C_{XY}^ta$. On multiplie alors l'équation (5.4) par C_{XX}^1 :

$$\left(C_{XX}^{-1}C_{XY}C_{YY}^{-1}C_{XY}^t - \rho^2\right)a = 0$$

dont la solution est donnée par a vecteur propre de $C_{XX}^{-1}C_{XY}C_{YY}^{-1}C_{XY}^t$.

En triant les valeurs propres par ordre décroissant, CANCOR nous permet d'obtenir un ensemble de vecteurs orthonormaux $\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ et $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$ maximisant les corrélations $\operatorname{corr}\left(a_k^tX, b_k^tY\right)$ entre les projections de X et Y (d étant le rang de C_{XY}). En d'autres termes, CANCOR projette X et Y dans un nouveau système de coordonnées où leur corrélation est maximisée, dimension à dimension.

5.2.4 Analyse de co-inertie

L'analyse de co-inertie – Co-Inertia Analysis (CoIA) en anglais – est une transformation très proche de CANCOR. Cependant, là où CANCOR cherche à maximiser une corrélation, CoIA vise à maximiser la covariance entre les paramètres acoustiques et visuels. Elle a été utilisée par *Dolédec et Chessel* en biologie [Dolédec et Chessel, 1994] afin d'extraire les relations cachées entre les espèces et leur environnement. Son objectif est de déterminer les directions $\mathbf{a} \in \mathbb{U}^n$ et $\mathbf{b} \in \mathbb{U}^m$ telles que les projections de X et Y sur ces deux vecteurs maximisent leur covariance (voir l'équation (5.6)).

Proposition 2 (Analyse de co-inertie)

$$(\boldsymbol{a}, \boldsymbol{b}) = \underset{(a,b)\in\mathbb{U}^n\times\mathbb{U}^m}{\operatorname{argmax}} \operatorname{cov}\left(a^t X, b^t Y\right) \tag{5.6}$$

Soient $\boldsymbol{a} \in \mathbb{U}^n$ et $\boldsymbol{b} \in \mathbb{U}^m$ définis par l'équation (5.6). $\boldsymbol{a}$ est le vecteur propre normé correspondant à la plus grande valeur propre λ_1 de la matrice $C_{XY}C_{XY}^t$ et $\boldsymbol{b}$ est le vecteur normé colinéaire à $C_{XY}^t\boldsymbol{a}$.

Démonstration 2 *On note*

$$\begin{aligned} \rho &= \operatorname{cov}\left(a^t X, b^t Y\right) \\ &= a^t C_{XY} b \end{aligned} \tag{5.7}$$

Cherchant à maximiser ρ, on fait l'hypothèse que $\rho > 0$ (si $\rho < 0$, il suffit de changer a en $-a$) : il est par conséquent équivalent de maximiser ρ et ρ^2.

$$\begin{aligned} \rho^2 &= \left(a^t C_{XY} b\right)^t \left(a^t C_{XY} b\right) \\ &= \left[\left(C_{XY}^t a\right)^t b\right]^t \left[\left(C_{XY}^t a\right)^t b\right] \end{aligned}$$

Selon l'inégalité de Cauchy-Schwarz, $\rho^2 \leq \|C_{XY}^t a\| \cdot \|b\|$ avec égalité si et seulement si b peut s'écrire $\mu C_{XY}^t a$, avec $\mu \in \mathbb{R}$. Ainsi, l'équation (5.7) devient :

$$\begin{aligned} \rho &= a^t C_{XY}\left(\mu C_{XY}^t a\right) \\ &= \mu a^t \left(C_{XY}C_{XY}^t\right) a \end{aligned}$$

Puisque $\|a\| = 1$, ρ est proportionnel au quotient de Rayleigh $R\left(C_{XY}C_{XY}^t, a\right)$, qui est maximisé pour a vecteur propre de $C_{XY}C_{XY}^t$ associée à la plus grande valeur propre λ_1.

En triant les valeurs propres par ordre décroissant, CoIA nous permet d'obtenir un ensemble de vecteurs orthonormaux $\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ et $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$ maximisant les covariances $\mathrm{cov}\left(a_k^t X, b_k^t Y\right)$ entre les projections de X et Y (d étant le rang de C_{XY}). En d'autres termes, CoIA projette X et Y dans un nouveau système de coordonnées où leur covariance est maximisée, dimension à dimension.

Notations Par la suite, on notera $\mathbf{A}$ et $\mathbf{B}$ les matrices résultantes de l'analyse de corrélation canonique et/ou analyse de co-inertie dont les colonnes sont les vecteurs de projection $\mathbf{a}_k$ et $\mathbf{b}_k$:

$$\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|\ldots|\mathbf{a}_d] \text{ et } \mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\ldots|\mathbf{b}_d] \tag{5.8}$$

Remarque Des études comparatives entre CANCOR et CoIA sont proposées dans [Goecke et Millar, 2003, Eveno et Besacier, 2005b, Eveno et Besacier, 2005a]. Les auteurs de [Goecke et Millar, 2003] montrent que CoIA est plus stable que CANCOR : les résultats sont beaucoup moins sensibles au nombre d'échantillons disponibles pour l'apprentissage. En outre, le score de *liveness* proposé dans [Eveno et Besacier, 2005b, Eveno et Besacier, 2005a] (permettant de vérifier le caractère *vivant* de l'échantillon biométrique et résumé au paragraphe 5.3.3) est beaucoup plus efficace avec CoIA qu'avec CANCOR pour la tâche de détection d'asynchronie. Les auteurs de [Eveno et Besacier, 2005b] expliquent cette différence par le fait que CoIA est un compromis entre CANCOR (où la corrélation audiovisuelle est maximisée) et PCA (où seules les directions acoustiques et visuelles de plus grande variance sont conservées) et profite par conséquent des avantages de deux transformations.

5.3 Mesures

Dans cette section, nous décrivons les mesures de correspondances proposées dans la littérature pour évaluer la synchronie entre les paramètres acoustiques et visuels.

5.3.1 Corrélation

Soient X et Y deux variables aléatoires. Le carré du coefficient de corrélation de *Pearson* $R(X, Y)$ (défini dans l'équation (5.9)) décrit la portion de la variance totale de X qui peut être expliquée par une transformation linéaire de Y (et réciproquement, la mesure étant symétrique).

$$R(X, Y) = \frac{\mathrm{cov}\,(X, Y)}{\sigma_X \sigma_Y} \tag{5.9}$$

Dans [Hershey et Movellan, 1999], les auteurs calculent le coefficient R entre l'énergie acoustique X et la valeur Y des pixels de la vidéo afin de déterminer quelle zone de l'image est la plus corrélée avec l'audio. Ceci permet alors de décider quelle personne parle, parmi toutes celles apparaissant à l'écran.

5.3.2 Information mutuelle

En théorie de l'information, l'information mutuelle $MI(X,Y)$ de deux variables aléatoires X et Y mesure la dépendance mutuelle entre ces deux variables. Dans le cas où X et Y sont discrètes, MI est définie par l'équation (5.10) :

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5.10}$$

C'est une mesure non-négative ($MI(X,Y) \geq 0$) et symétrique ($MI(X,Y) = MI(Y,X)$). On peut aussi démontrer que X et Y sont indépendantes si et seulement si $MI(X,Y) = 0$. Dans le cas où X et Y sont des variables aléatoires mono-dimensionnelles normales [Hershey et Movellan, 1999], l'information mutuelle MI est lié à R via l'équation :

$$MI(X,Y) = -\frac{1}{2} \log \left(1 - R(X,Y)^2\right) \tag{5.11}$$

Dans [Hershey et Movellan, 1999, Fisher *et al.*, 2001, Nock *et al.*, 2002, Iyengar *et al.*, 2003], l'information mutuelle est utilisée afin de localiser les pixels de la vidéo qui correspondent au signal audio : le visage de la personne qui parle est la zone qui se détache clairement du reste de l'image. Cependant, il est notable que la zone de la bouche n'est pas toujours la partie du visage dont l'information mutuelle avec le signal audio est la plus élevée : les contours du visage sont parfois mis en évidence, montrant que certains mouvements globaux du visage complètent de façon synchrone le signal de parole acoustique.

Remarque Dans [Bregler et Konig, 1994], pour un signal de parole donné, l'information mutuelle entre le flux audio X (8 coefficients cepstraux) et le flux visuel Y_t (10 coefficients de type *eigenlips*) décalé dans le temps est tracée en fonction du décalage temporel t (voir la figure 5.1 tirée de l'article original). Il apparaît que l'information mutuelle atteint son maximum pour un délai du flux visuel compris entre 0 et 120 ms. Ce phénomène largement constaté dans la littérature portant sur le traitement de la parole audiovisuelle peut s'expliquer par le fait que la cause (le mouvement articulatoire) précède l'effet (le son) [Vatikiotis-Bateson *et al.*, 2006]. Cette observation a conduit les auteurs de [Eveno et Besacier, 2005a, Eveno et Besacier, 2005b] à proposer un score de *liveness* $L(X,Y)$ qui tient compte de ce délai et que nous

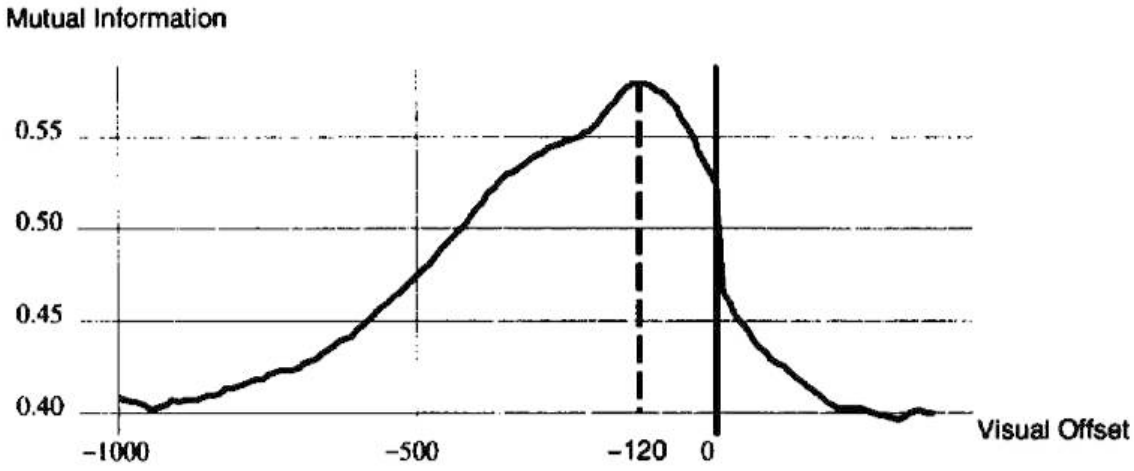


FIG. 5.1 – Mesure de l'information mutuelle (*Mutual Information*) en fonction du décalage temporel entre les flux visuel et acoustiques (*Visual Offset*). Le maximum est atteint pour un décalage de 120 ms – figure extraite de [Bregler et Konig, 1994].

décrivons dans le paragraphe suivant.

5.3.3 Mesure proposée par *Eveno et Besacier*

La paramétrisation de la parole audiovisuelle choisie par *Eveno et Besacier* est réalisée ainsi :

- 5 coefficients LPC sont extraits toutes les 40 ms et constituent ainsi le flux $X \in \mathbb{R}^5$.
- La hauteur, la largeur et l'aire de la bouche sont extraites pour chaque trame de la vidéo (toutes les 40 ms) en utilisant un outil de détection et suivi des lèvres, constituant ainsi le flux $Y \in \mathbb{R}^3$.

Les deux flux X et Y possèdent donc la même fréquence d'échantillonnage. On note Y_δ le flux visuel avec un décalage de δ trames dans le temps. L'application de l'analyse de co-inertie entre X et Y_δ permet de

déterminer les vecteurs $\mathbf{a}_\delta$ et $\mathbf{b}_\delta$ maximisant la covariance entre ${\mathbf{a}_\delta}^t X$ et ${\mathbf{b}_\delta}^t Y_\delta$. On définit alors

$$\rho_\delta = R\left({\mathbf{a}_\delta}^t X, {\mathbf{b}_\delta}^t Y_\delta\right) \quad (5.12)$$

$$\rho_{\text{ref}} = \max_{-2\leq\delta\leq 0}[\rho_\delta] \quad (5.13)$$

$$\rho_{\text{moy}} = \frac{1}{2\Delta+1}\sum_{\delta=-\Delta}^{\Delta}\rho_\delta \quad (5.14)$$

$$L(X,Y) = \frac{1}{2\Delta+1}\left(\frac{\rho_{\text{ref}}}{\rho_{\text{moy}}}-1\right)\sum_{\delta=-\Delta}^{\Delta}\mathbf{1}_{\bullet\leq\rho_{\text{ref}}}[\rho_\delta] \quad (5.15)$$

où $\Delta = 10$ (correspondant à une décalage maximum de 400 ms), $\mathbf{1}_{\bullet\leq\rho_{\text{ref}}}[\rho] = 1$ si $\rho \leq \rho_{\text{ref}}$ et 0 sinon. En résumé, plus le pic obtenu pour ρ_{ref} est marqué, plus la valeur de $L(X,Y)$ est élevée.

Cette mesure constitue l'inspiration première de nos travaux. Nous l'utiliserons, en particulier, comme mesure étalon afin de montrer les apports de nos différentes propositions.

5.3.4 Modélisation conjointe

Là où les coefficients R et MI permettent une mesure efficace de la correspondance entre deux variables aléatoires, d'autres méthodes cherchent à mesurer cette correspondance en modélisant conjointement les paramètres acoustiques et visuels.

Modèle de mélange de gaussiennes

Considérons deux variables aléatoires discrètes $X = \{x_t, t \in \mathbb{N}\}$ et $Y = \{y_t, t \in \mathbb{N}\}$ de dimensions respectives d_X et d_Y. Typiquement, X représente les paramètres acoustiques et Y les paramètres visuels [Sodoyer *et al.*, 2002, Chetty et Wagner, 2004]. On peut définir une troisième variable aléatoire discrète $Z = \{z_t, t \in \mathbb{N}\}$ de dimension d_Z où z_t est la concaténation des deux échantillons x_t et y_t, de sorte que $z_t = [x_t, y_t]$ et $d_Z = d_X + d_Y$.
Étant donné un échantillon z, le modèle de mélange de gaussiennes λ définit sa fonction de distribution de probabilité comme suit :

$$p(z|\lambda) = \sum_{i=1}^{N} w_i \mathcal{N}(z; \mu_i, \Gamma_i) \quad (5.16)$$

où $\mathcal{N}(\bullet; \mu, \Gamma)$ est la distribution normale de moyenne μ et de matrice de covariance Γ. $\lambda = \{w_i, \mu_i, \Gamma_i\}_{i\in[1,N]}$ est l'ensemble des paramètres décrivant la distribution jointe de X et Y. À partir d'un ensemble d'apprentissage d'échantillons synchrones x_t et y_t concaténés en un échantillon joint z_t, l'algorithme EM (pour Expectation-Maximization) permet l'estimation de λ. Au moment de tester la synchronie entre deux flux $X = \{x_t, t \in [1, T]\}$ et $Y = \{y_t, t \in [1, T]\}$, une mesure de correspondance $C_\lambda(X, Y)$ peut être calculée via l'équation (5.17).

$$C_\lambda(X, Y) \quad = \quad \frac{1}{T} \sum_{t=1}^{T} p([x_t, y_t]|\lambda) \tag{5.17}$$

Enfin, l'application d'un seuil θ permet de décider si les flux X et Y se correspondent (si $C_\lambda(X, Y) > \theta$) ou non (si $C_\lambda(X, Y) \leq \theta$).

Modèle de Markov caché

Le décalage temporel entre les flux acoustiques et visuels n'est pas modélisé par les GMMs, ni par les coefficients R et MI. Ainsi, *Bengio* propose un modèle de Markov caché asynchrone (AHMM) pour la reconnaissance de la parole audiovisuelle. Il fait l'hypothèse qu'à chaque instant t il existe une observation acoustique x_t et que l'observation visuelle y_t n'existe que de temps en temps. Ainsi, la différence de fréquence d'échantillonnage est directement prise en compte en introduisant la probabilité que le système émette l'observation visuelle suivante y_s au temps t. Dans [Bengio, 2003], AHMM donne de meilleurs résultats que les HMM dans la tâche de reconnaissance de la parole audiovisuelle en résolvant naturellement le problème de différence entre les fréquences d'échantillonnage.

Modèles non-paramétriques

L'utilisation des réseaux de neurones (NN) est étudiée dans [Cutler et Davis, 2000]. Étant donné un ensemble d'apprentissage de données audiovisuelles synchrones et de données asynchrones, un réseau de neurones à une couche cachée est entraîné de façon à retourner la valeur 1 quand les données en entrée sont synchrones et la valeur 0 sinon. En outre, les auteurs proposent d'utiliser une couche d'entrée au temps t de type $[X_{t-N_X}, \ldots, X_t, \ldots, X_{t+N_X}, Y_{t-N_Y}, \ldots, Y_t, \ldots, Y_{t+N_Y}]$ (au lieu de $[X_t, Y_t]$), en choisissant N_X et N_Y de sorte qu'environ 200 ms de contexte temporel soient passées en entrée. Cette proposition vise à résoudre le problème de délai entre les flux audio et visuel soulevé dans le paragraphe sur l'information mutuelle. Elle permet aussi d'ôter le besoin de sous-échantillonnage audio ou sur-échantillonnage visuel.

5.4 Applications

Mesurer la synchronie entre les flux de parole acoustique et visuel peut être d'une grande aide dans de nombreuses applications audiovisuelles et multimédia.

Localisation de source sonore La localisation de source sonore est l'application des mesures de synchronie audiovisuelle la plus citée [Barker *et al.*, 1998]. Dans [Cutler et Davis, 2000], une fenêtre glissante survole la vidéo afin de trouver la zone de la bouche qui correspond le plus probablement à la bande sonore (en utilisant un réseau de neurones). Dans [Nock *et al.*, 2002], l'information mutuelle permet de décider laquelle des quatre personnes apparaissant à l'image est la source de la voix entendue dans la bande sonore : un taux de correction de 82% est atteint (moyenne sur 1016 vidéos de test). On peut imaginer un système de visio-conférence intelligent dont la caméra zoomerait sur le locuteur courant [Yoshimi et Pingali, 2002].

Indexation de séquences audiovisuelles Dans [Iyengar *et al.*, 2003], les auteurs fusionnent les scores de trois systèmes (détection du visage, détection du silence et mesure de correspondance basée sur l'information mutuelle entre la bande sonore et la valeur des pixels) afin d'améliorer leur algorithme de détection de monologue. Des expériences réalisées dans le cadre de *TREC 2002 Video Retrieval Track* montrent une amélioration relative de 50% de la précision moyenne[1].

Post-production Lors de la post-production d'oeuvres cinématographiques, les dialogues sont souvent réenregistrés en studio. Une mesure de correspondance audiovisuelle pourrait être d'une grande aide au moment de synchroniser le nouvel enregistrement audio avec la vidéo originale. De telles mesures peuvent aussi être une façon d'évaluer la qualité d'un doublage dans une langue étrangère : la traduction choisie est-elle réaliste vis-à-vis des mouvements du visage de l'acteur ?

Autres applications Dans [Sodoyer *et al.*, 2002], la correspondance audiovisuelle est utilisée de façon à améliorer un algorithme de séparation de parole. Enfin, les auteurs de [Fisher *et al.*, 2001] élaborent des filtres pour la réduction de bruit à partir de mesure de synchronie audiovisuelle.

[1] `http://trec.nist.gov/`

Chapitre 6

Détection d'asynchronie

Contexte

Il existe relativement peu de travaux portant sur la question de la détection d'asynchronie pour la vérification d'identité. *Chetty et al.* proposent d'utiliser des modèles de mélange de gaussiennes dans un espace de paramètres constitués de la concaténation de paramètres acoustiques (les MFCC) et de paramètres visuels (*eigenlips* et mesures géométriques) [Chetty et Wagner, 2004]. Au moment du test, la vraisemblance des vecteurs de paramètres audiovisuels constitue la mesure de correspondance entre les paramètres acoustiques et visuels. Leur protocole d'évaluation mériterait cependant d'être amélioré afin de rendre les attaques plus réalistes puisqu'ils simulent des attaques de type *présentation de photographie devant la caméra* en répétant simplement la même image tout au long de la séquence vidéo. Comme nous l'avons déjà écrit au paragraphe 5.3.3 (page 90), *Eveno et Besacier* proposent une mesure de corrélation entre paramètres acoustiques et visuels du signal de parole. Elle est obtenue par analyse de corrélation canonique et/ou analyse de co-inertie de ces paramètres [Eveno et Besacier, 2005a, Eveno et Besacier, 2005b] et constitue l'inspiration première de notre travail sur la synchronie audiovisuelle.

6.1 Paramétrisation

La paramétrisation pour laquelle nous avons opté est celle classiquement utilisée dans les systèmes de reconnaissance automatique de la parole audiovisuelle [Potamianos *et al.*, 2004] : les coefficients MFCC pour la partie acoustique et les coefficients DCT de la zone de la bouche pour la partie visuelle.

6.1.1 Paramètres acoustiques X

Plusieurs jeux de paramètres acoustiques peuvent être construits à partir des coefficients MFCC présentés dans le chapitre 3, selon que l'on ajoute l'énergie du signal acoustique, les dérivées premières et secondes. Le tableau 6.1 établit un récapitulatif des 6 différents types de paramètres qui seront utilisés par la suite.

Type	**Description**	**Dimension** n
MFCC	Coefficients MFCC	12
MFCC + Δ	Ajout des dérivées premières	24
MFCC + Δ + $\Delta\Delta$	Ajout des dérivées secondes	36
MFCCE	Coefficients MFCC et énergie	13
MFCCE + Δ	Ajout des dérivées premières	26
MFCCE + Δ + $\Delta\Delta$	Ajout des dérivées secondes	39

TAB. 6.1 – Paramètres acoustiques

6.1.2 Paramètres visuels Y

Le processus d'extraction des paramètres visuels est décrit dans la figure 6.1. L'étape ① de détection du visage par localisation des yeux est celle déjà utilisée dans le système de vérification du visage présenté au chapitre 3. L'étape ② fait appel à la connaissance *a priori* de la structure géométrique du visage humain pour délimiter une zone de recherche de la bouche dont la position est déduite de la position des yeux. L'étape ③ est la détection proprement dite de la bouche à l'aide d'un détecteur de type *Viola and Jones* [Viola et Jones, 2002] : il s'agit du détecteur de bouche entraîné par *Castrillón et al.* [Castrillón Santana *et al.*, 2005], disponible librement sur l'Internet. Enfin, les étapes ④ et ⑤ consistent à extraire de la zone de la bouche (dont deux tailles sont envisageables) les 28 paramètres DCT (voir l'équation (5.1) de la page 83) correspondant aux basses fréquences spatiales, comme le montre la figure 6.2. Plusieurs jeux de paramètres visuels (rappelés dans le tableau 6.2) peuvent être construits à partir des coefficients DCT selon que l'on ajoute les dérivées premières et secondes.

Type	**Description**	**Dimension** m
DCT	Coefficients DCT	28
DCT + Δ	Ajout des dérivées premières	56
DCT + Δ + $\Delta\Delta$	Ajout des dérivées secondes	84

TAB. 6.2 – Paramètres visuels

Là où les paramètres acoustiques sont extraits toutes les 10 ms, la fréquence d'échantillonnage des

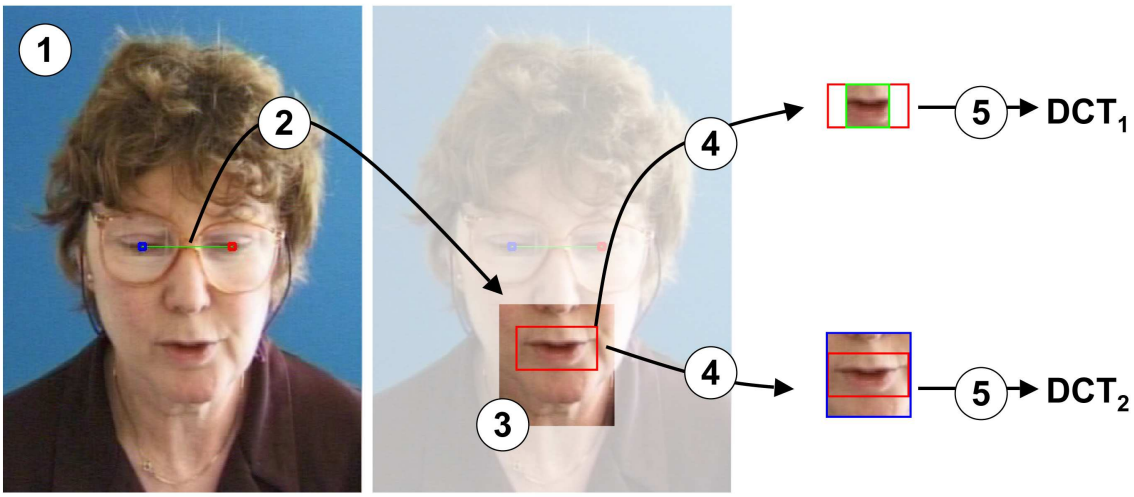


FIG. 6.1 – Extraction des paramètres visuels. ① Détection des yeux. ② Sélection de la zone d'intérêt pour la recherche de la bouche. ③ Détection de la bouche. ④ Sélection de la zone d'intérêt pour l'extraction des coefficients DCT. ⑤ Extraction des coefficients DCT.

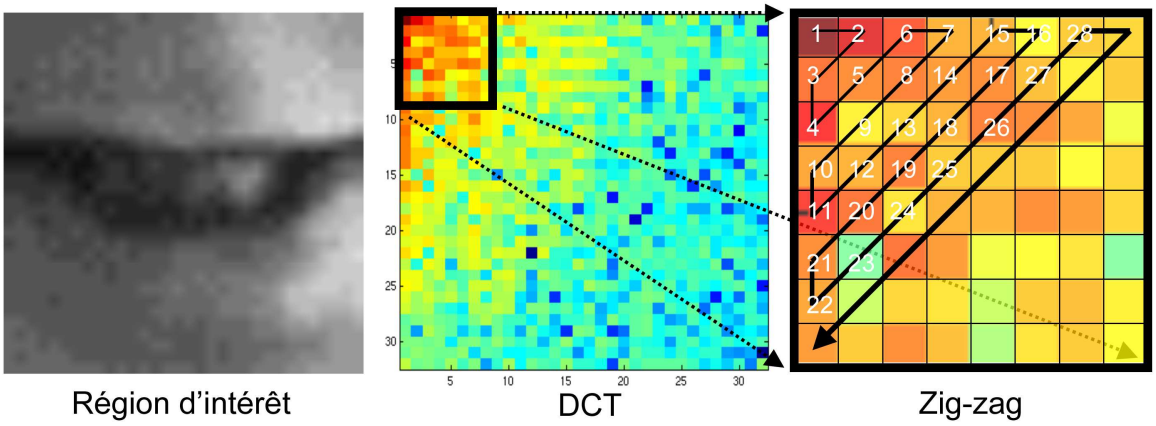


FIG. 6.2 – Extraction des 28 coefficients DCT correspondant aux basses fréquences spatiales selon le principe du zig-zag.

paramètres visuels est dépendante du nombre d'images par seconde de la séquence audiovisuelle. Dans le cas de BANCA, les paramètres visuels sont extraits toutes les 40 ms (ce qui correspond à une *frame rate* de 25 images par seconde). De façon à avoir le même nombre de paramètres acoustiques et visuels pour chaque

séquence audiovisuelle, on choisit d'effectuer une interpolation linéaire des paramètres visuels.

Remarque L'utilisation d'une paramétrisation liée à la forme des lèvres a aussi été étudiée. Un algorithme de détection et de suivi des lèvres a permis d'extraire des paramètres tels que l'aire délimitée par le contour des lèvres, la hauteur et la largeur de la bouche [Matthews et Baker, 2004]. Cependant, les premières expériences ont montré leur faiblesse et leur utilisation a donc été abandonnée [Argones-Rúa *et al.*, 2007a].

6.2 Paramètres corrélés

Étant donnés deux flux synchrones de paramètres acoustiques $X \in \mathbb{R}^n$ et visuels $Y \in \mathbb{R}^m$, CANCOR et CoIA (définies et démontrées aux pages 85 et 87 respectivement) permettent d'obtenir les matrices **A** et **B**

$$\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|\dots|\mathbf{a}_d] \text{ et } \mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\dots|\mathbf{b}_d] \tag{6.1}$$

qui, à leur tour, permettent l'extraction de paramètres acoustiques et visuels *corrélés* $\mathcal{X}$ et $\mathcal{Y}$

$$\mathcal{X} = \mathbf{A}^t X \text{ et } \mathcal{Y} = \mathbf{B}^t Y \tag{6.2}$$

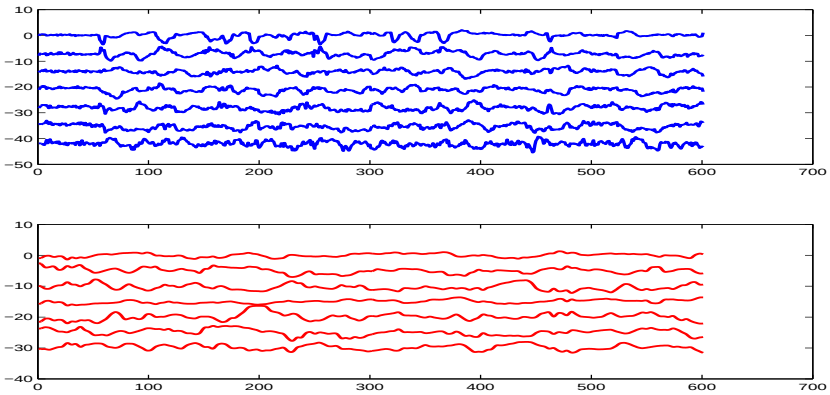
de même dimension $d = \min(n, m)$:

$$\begin{aligned} \forall k \in \{1, \dots, d\}, \quad \mathcal{X}_k &= \mathbf{a}_\mathbf{k}^t X = \sum_{i=1}^{n} \mathbf{a}_{\mathbf{k}i} X_i \\ \mathcal{Y}_k &= \mathbf{b}_\mathbf{k}^t Y = \sum_{i=1}^{m} \mathbf{b}_{\mathbf{k}i} Y_i \end{aligned} \tag{6.3}$$

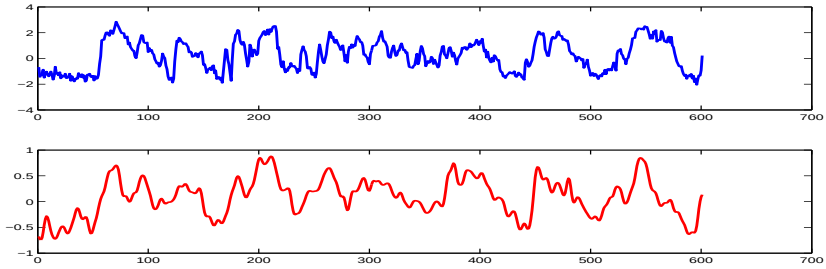
L'effet de CANCOR et CoIA sur des données réelles est illustré par la figure 6.3, qui montre des paramètres extraits d'une séquence de la base de données BANCA [Bailly-Baillière *et al.*, 2003].

Remarque En ne choisissant que les $D < d$ premières dimensions, les méthodes CANCOR et CoIA sont appliquées afin de réduire la dimension des paramètres acoustiques et visuels en limitant la perte d'information relative à leur corrélation. Cette propriété est particulièrement importante lorsque la synchronie audiovisuelle est modélisée par des outils statistiques nécessitant de grandes quantités de données d'apprentissage. Dans [Sargin *et al.*, 2006], CANCOR est utilisée pour réduire la dimension de paramètres audiovisuels en entrée d'un système de vérification du locuteur basé sur des modèles de Markov cachés (HMM, pour *Hid-*

a. Paramètres acoustiques (X_1 à X_7) et visuels (Y_1 à Y_7) originaux



b. Paramètres acoustiques $\mathcal{X}_1$ et visuels $\mathcal{Y}_1$ obtenus par CANCOR



c. Paramètres acoustiques $\mathcal{X}_1$ et visuels $\mathcal{Y}_1$ obtenus par CoIA

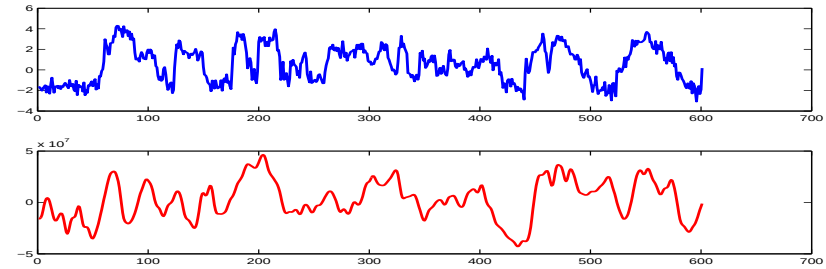


FIG. 6.3 – Évolution des paramètres acoustiques (seules les dimensions X_1 à X_7 sont représentées) et visuels (Y_1 à Y_7) avant (a) et après transformation ($\mathcal{X}_1 = \sum_{i=1}^{n} \mathbf{a_1}_i X_i$ et $\mathcal{Y}_1 = \sum_{i=1}^{m} \mathbf{b_1}_i Y_i$) par CANCOR (b) et CoIA (c). La corrélation entre X and Y est plus visible dans l'espace transformé que dans l'espace original.

den Markov models en anglais). Dans [Argones-Rúa *et al.*, 2007b, Argones-Rúa *et al.*, 2007a], nous avons proposé une modélisation statistique de la synchronie audiovisuelle à l'aide de deux HMM couplés, portant

respectivement sur des paramètres acoustiques et visuels dont la dimension est préalablement réduite par CoIA.

6.3 Mesure de synchronie

Nous introduisons dans cette section une méthode utilisant ces transformations afin de mesurer la synchronie d'une séquence audiovisuelle de test Γ dont on a extrait les paramètres acoustiques X^{Γ} et visuels Y^{Γ}.

6.3.1 Principe commun

Quatre mesures différentes sont proposées qui partagent cependant toutes un cadre commun en trois étapes, résumé schématiquement dans la figure 6.4 : après une première étape de *modélisation* de la synchronie à l'aide de CANCOR et/ou CoIA, les paramètres de la séquence de test sont *transformés* et une *mesure* de synchronie basée sur leur corrélation est finalement obtenue.

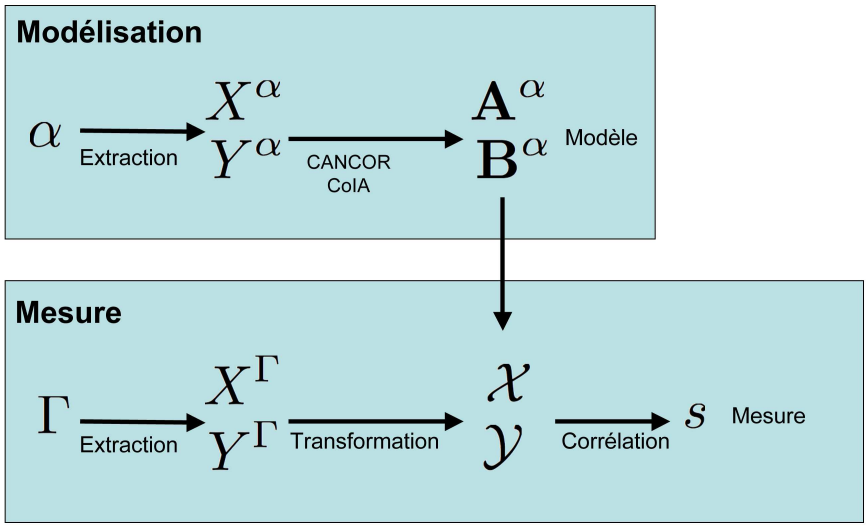


FIG. 6.4 – Mesure de synchronie

Étape 1 : Modélisation À partir de paramètres acoustiques X^{α} et visuels Y^{α} extraits des séquences issues de l'ensemble d'apprentissage α, CANCOR et/ou CoIA permettent de déduire les matrices $\mathbf{A}^{\alpha}$ et $\mathbf{B}^{\alpha}$:

$$(X^\alpha, Y^\alpha) \rightarrow (\mathbf{A}^\alpha, \mathbf{B}^\alpha) \tag{6.4}$$

Étape 2 : Transformation Les paramètres acoustiques X^Γ et visuels Y^Γ de la séquence Γ dont on cherche à mesurer la synchronie sont extraits puis transformés à l'aide des deux matrices $\mathbf{A}^\alpha$ et $\mathbf{B}^\alpha$ afin d'obtenir les paramètres acoustiques et visuels *corrélés* $\mathcal{X}$ et $\mathcal{Y}$:

$$\begin{aligned} \mathcal{X} &= \mathbf{A}^{\alpha t} X^\Gamma \\ \mathcal{Y} &= \mathbf{B}^{\alpha t} Y^\Gamma \end{aligned} \tag{6.5}$$

Étape 3 : Mesure Les corrélations entre chaque dimension de $\mathcal{X}$ et $\mathcal{Y}$ sont calculées et participent à la mesure $s_D^\alpha\left(X^\Gamma, Y^\Gamma\right)$ de synchronie entre X^Γ et Y^Γ. Plus elles sont élevées, plus le degré de synchronie est élevé :

$$\begin{aligned} s_D^\alpha\left(X^\Gamma, Y^\Gamma\right) &= \frac{1}{D}\sum_{k=1}^{D} \operatorname{corr}\left(\mathcal{X}_k, \mathcal{Y}_k\right) \\ &= \frac{1}{D}\sum_{k=1}^{D} \frac{{\mathcal{X}_k}^t \mathcal{Y}_k}{\sqrt{{\mathcal{X}_k}^t \mathcal{X}_k}\sqrt{{\mathcal{Y}_k}^t \mathcal{Y}_k}} \\ &= \frac{1}{D}\sum_{k=1}^{D} \frac{\left(\mathbf{a}_k^{\alpha t} X^\Gamma\right)^t \left(\mathbf{a}_k^{\alpha t} Y^\Gamma\right)}{\sqrt{\left(\mathbf{a}_k^{\alpha t} X^\Gamma\right)^t \left(\mathbf{a}_k^{\alpha t} X^\Gamma\right)}\sqrt{\left(\mathbf{a}_k^{\alpha t} Y^\Gamma\right)^t \left(\mathbf{a}_k^{\alpha t} Y^\Gamma\right)}} \end{aligned} \tag{6.6}$$

où

$D \leq d$ est le nombre de dimensions effectivement conservées.

6.3.2 Variantes

Synchronie Γ ($\alpha = \Gamma$) Dans le cas où l'on choisit la séquence Γ elle-même comme séquence d'apprentissage ($\alpha = \Gamma$), on parle de *synchronie* Γ. CANCOR et/ou CoIA sont directement appliquées sur la séquence audiovisuelle dont on cherche à mesurer la synchronie : il s'agit d'une mesure de la synchronie intrinsèque de la séquence Γ.

Synchronie Γ par morceau ($\alpha = \gamma$) Cette méthode est une extension de la synchronie Γ basée sur le postulat suivant :

– Si la séquence Γ est effectivement synchrone, alors chaque sous-séquence devrait suivre le même modèle de synchronie. Ainsi, un modèle de synchronie intrinsèque à une sous-séquence $\alpha \subset \Gamma$ sera aussi optimal pour toute autre sous-séquence $\gamma \subset \Gamma$;
– En revanche, si la séquence Γ n'est pas synchrone, alors un modèle de synchronie intrinsèque à une sous-séquence α ne portera que très peu d'information quant à la synchronie d'une autre sous-séquence $\gamma \subset \Gamma$ ($\alpha \cap \gamma = \varnothing$).

Notons N le nombre d'échantillons de la séquence Γ : $X^\Gamma = \{x^1, \ldots, x^N\}$ et $Y^\Gamma = \{y^1, \ldots, y^N\}$. On définit $\mathfrak{P}_\Gamma$ l'ensemble des sous-séquences α de Γ de cardinal $\lfloor N/2 \rfloor$ de façon à partitionner la séquence Γ en deux sous-séquences d'apprentissage α et de test γ de même taille (à un échantillon près) comme l'illustre la figure 6.5. La mesure de synchronie Γ par morceaux est finalement obtenue à l'aide de l'équation (6.7).

$$s_D\left(X^\Gamma, Y^\Gamma\right) = \frac{1}{\operatorname{card} \mathfrak{P}_\Gamma} \sum_{\alpha \in \mathfrak{P}_\Gamma} s_D^\alpha \left(X^\gamma, Y^\gamma\right) \tag{6.7}$$

Pour des raisons combinatoires, il n'est – en pratique – pas envisageable de sommer sur l'ensemble de toutes les partitions de Γ. Par conséquent, un petit nombre de partitions (50, dans notre cas) est tiré aléatoirement, qui participe à la mesure de synchronie Γ par morceaux.

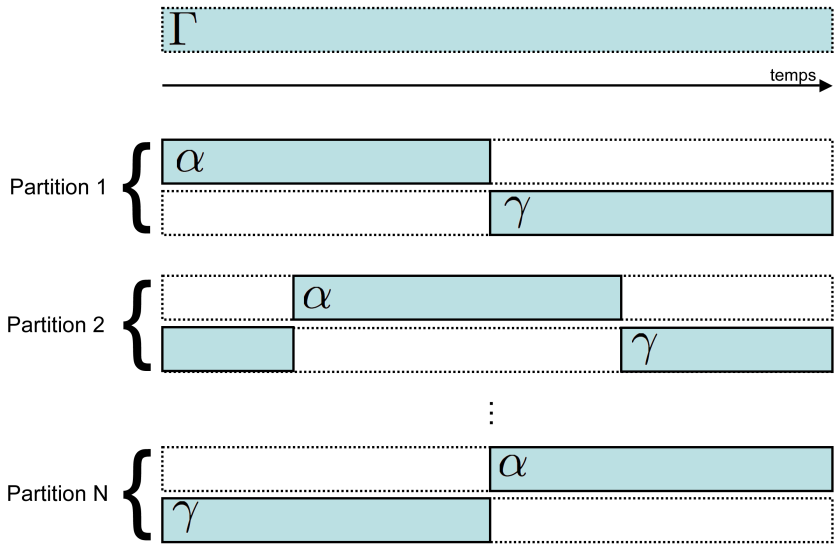


FIG. 6.5 – Partition de la séquence Γ en sous-séquence d'apprentissage α et sous-séquence de test γ

Synchronie Ω ($\alpha = \Omega$) Par analogie au modèle du monde (UBM, pour *Universal Background Model* en anglais) en vérification du locuteur, la synchronie Ω fait appel à un modèle du monde Ω. Ainsi, des paramètres acoustiques X^Ω et visuels Y^Ω sont extraits d'un ensemble de séquences audiovisuelles d'un grand nombre de personnes différentes. Ceci permet de calculer les matrices $\mathbf{A}^\Omega$ et $\mathbf{B}^\Omega$ qui décrivent des transformations maximisant *globalement* la corrélation entre les paramètres acoustiques et visuels du monde Ω. Il s'agit d'une mesure de la synchronie universelle de la séquence Γ.

Synchronie λ ($\alpha = \lambda$) La mesure de synchronie λ est une mesure dépendante de la personne. Elle repose sur le postulat que chaque personne possède sa propre façon de synchroniser sa voix et le mouvement de ses lèvres. Ainsi, des paramètres acoustiques X^λ et visuels Y^λ sont obtenus à partir d'une séquence audiovisuelle de la personne λ (la séquence d'enrôlement dans le système biométrique, typiquement) afin de calculer les matrices $\mathbf{A}^\lambda$ et $\mathbf{B}^\lambda$.

Important Ce chapitre étant dédié à la tâche de détection d'asynchronie, les performances des mesures de synchronie sont évaluées à l'aide du protocole `S`. Dans ce cadre différent de celui de la vérification d'identité, il convient de rappeler la signification des nombres `NI`, `NC`, `NFA` et `NFR` utilisés pour le calcul des valeurs de `DCF` et le tracé des courbes DET :
– `NI` est le nombre de séquences asynchrones ;
– `NC` est le nombre de séquences synchrones ;
– `NFA` est le nombre de séquences asynchrones faussement classées comme étant synchrones ;
– `NFR` est le nombre de séquences synchrones faussement classées comme étant asynchrones.

6.4 Évaluation

Un grand nombre de réglages différents peuvent influer sur les performances de ces différentes mesures de synchronie. Les figures 6.6 et 6.7 résument les nombreuses expériences menées sur le protocole `S` avec la mesure de synchronie Γ.

Pour chaque courbe, le titre indique quelle combinaison de paramètres acoustiques et visuels est utilisée. La première des trois colonnes correspond aux paramètres visuels de type DCT, la seconde à ceux de type DCT+Δ et la troisième à ceux de type DCT+Δ+$\Delta\Delta$. De façon analogue, à chaque ligne correspond un type de paramètres acoustiques ; dans l'ordre, de la première à la sixième ligne : MFCC, MFCCE, MFCC+Δ, MFCCE+Δ, MFCC+Δ+$\Delta\Delta$ et MFCCE+Δ+$\Delta\Delta$ (voir les tableaux 6.1 et 6.2).

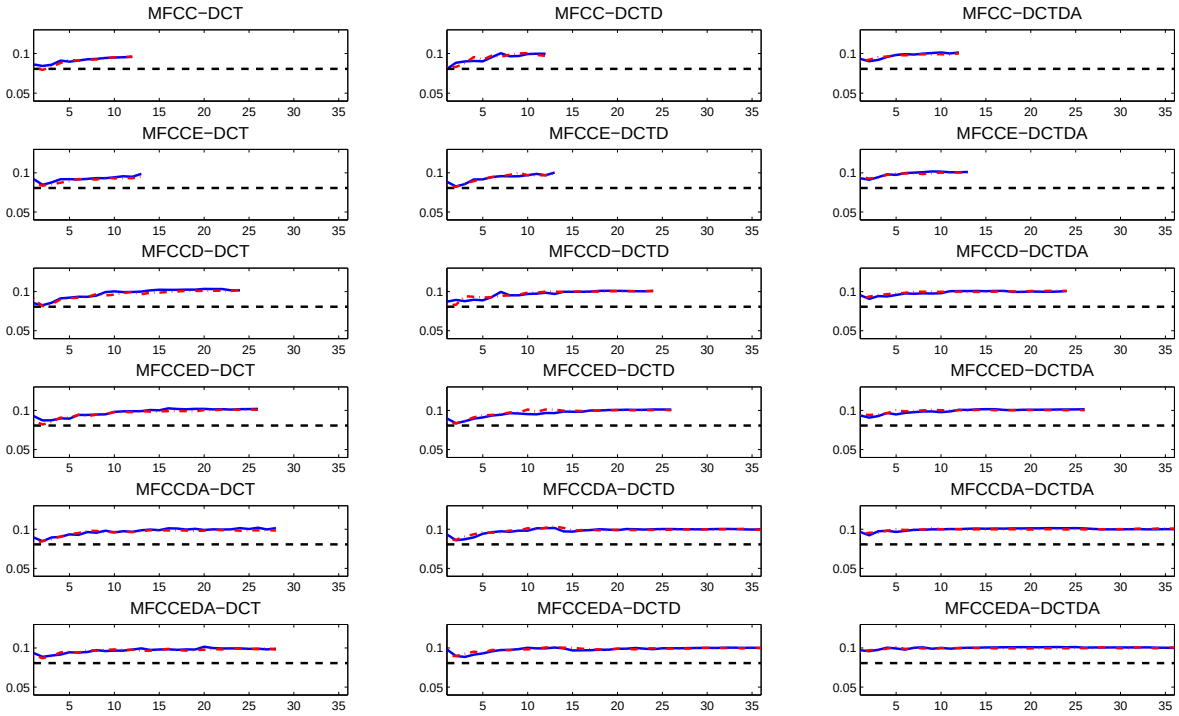


FIG. 6.6 – Performances de la synchronie CANCOR Γ sur le protocole S. Chaque courbe correspond à une combinaison MFCC/DCT (voir les tableaux 6.1 et 6.2). La valeur de DCF est tracée en fonction de D. La courbe rouge en pointillés correspond aux paramètres visuels DCT_1, celle en bleu en trait plein à ceux de type DCT_2. La ligne noire horizontale en pointillés correspond à la valeur de DCF de la meilleure mesure CANCOR Γ.

La valeur de DCF est tracée en fonction de la dimension D introduite dans l'équation (6.6). Les courbes rouges en pointillés correspondent aux paramètres visuels de type DCT_1 et les courbes bleues à ceux de type DCT_2 (voir la figure 6.8). La ligne horizontale noire en pointillés correspond à la valeur de DCF du meilleur système de chaque figure.

Taille de la région d'intérêt La première observation (surtout visible sur la figure 6.7 correspondant à CoIA) concerne la comparaison entre les paramètres DCT_1 et DCT_2. Les performances obtenues avec les paramètres DCT_1 sont toujours, sinon équivalentes, moins bonnes que celles obtenues avec les paramètres DCT_2. Il apparaît ainsi que l'information visuelle de parole n'est pas confinée dans la seule région des lèvres que décrivent les paramètres de type DCT_1 : il convient donc d'intégrer les informations contenues dans une région plus large englobant une partie de la mâchoire et des joues. Ce comportement a aussi été observé lors

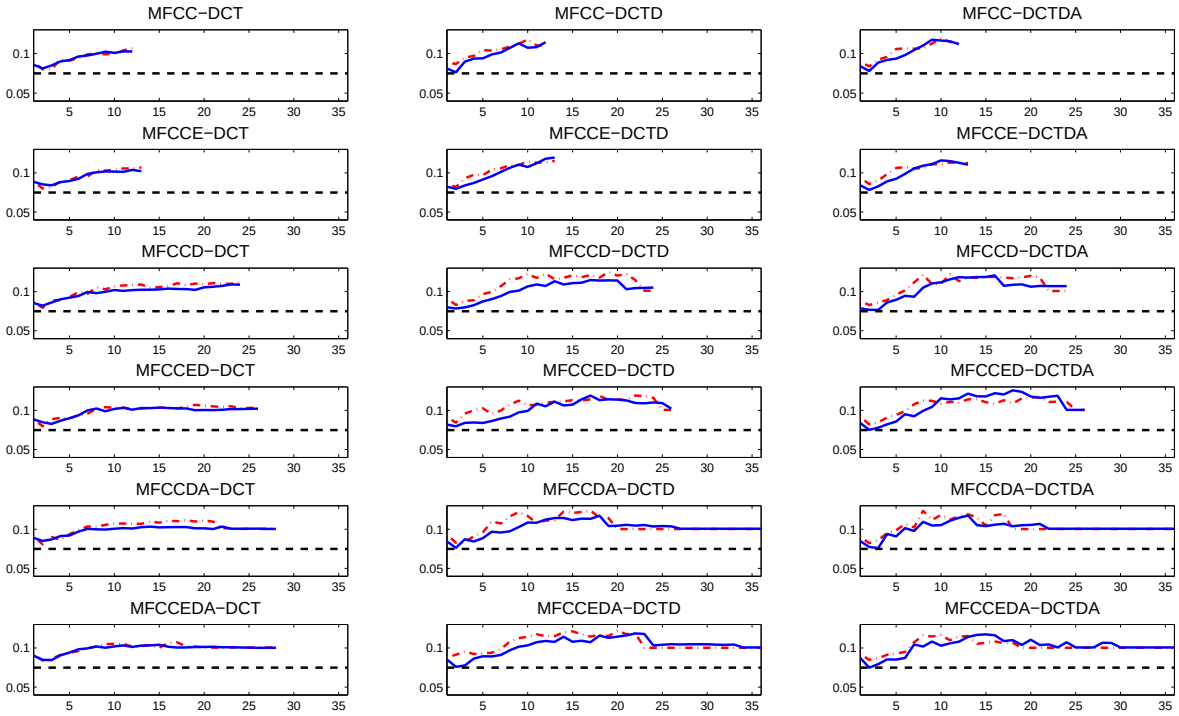


FIG. 6.7 – Performances de la synchronie CoIA Γ sur le protocole S. Chaque courbe correspond à une combinaison MFCC/DCT. La valeur de DCF est tracée en fonction de D. La courbe rouge en pointillés correspond aux paramètres visuels DCT_1, celle en bleu en trait plein à ceux de type DCT_2. La ligne noire horizontale en pointillés correspond à la valeur de DCF de la meilleure mesure CoIA Γ.

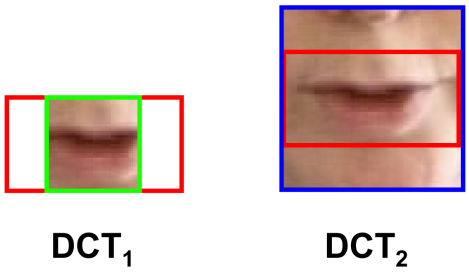


FIG. 6.8 – Taille de la région d'intérêt pour l'extraction des paramètres visuels. En rouge, la zone obtenue par l'algorithme de détection de la bouche ; en vert, la zone conservée dans le cas de DCT_1 ; en bleu, celle d'où sont extraits les coefficients DCT_2.

des expériences menées sur les mesures de synchronie Ω et λ.

Par la suite, nous ne considérerons donc que les paramètres visuels de type DCT_2.

Dérivées premières et secondes Si l'on compare les courbes bleues (correspondant aux paramètres DCT_2) des trois colonnes de la figure 6.7, l'ajout des dérivées premières des paramètres visuels apporte une petite amélioration (bien que non statistiquement significative) tandis que l'ajout complémentaire des dérivées secondes a tendance à dégrader les performances. Deux principales raisons peuvent expliquer ce phénomène. Tout d'abord, alors que le nombre d'échantillons disponibles pour l'apprentissage reste inchangé, l'ajout des dérivées secondes augmente les dimensions du modèle de synchronie (les matrices $\mathbf{A}$ et $\mathbf{B}$) ; ceci risque d'entraîner une modélisation approximative de la synchronie (le fameux fléau des dimensions). Cette remarque est d'autant plus vraie pour CANCOR qui, comme on l'a déjà mentionné, est beaucoup plus sensible à la taille de l'ensemble de l'apprentissage et nécessite généralement plus de données que CoIA pour mener correctement l'étape de modélisation. La seconde raison réside dans la méthode de calcul des dérivées secondes. Étant calculées à partir d'une fenêtre temporelle d'échantillons eux-mêmes interpolés linéairement, il est probable que les dérivées secondes ainsi estimées apportent plus de bruit que d'information pertinente.

En ce qui concerne les dérivées des paramètres acoustiques MFCC, leur influence est beaucoup moins marquée et les différences de performances observées ne permettent pas de tirer de conclusion.

Énergie acoustique La différence (en termes de `DCF`) mesurée entre les systèmes utilisant (lignes 2, 4 et 6) ou non (lignes 1, 3 et 5) l'énergie acoustique est loin d'être statistiquement significative. Pourtant, nos premiers travaux publiés dans [Bredin *et al.*, 2006c] avaient montré qu'elle est une source d'information pertinente dans la tâche de détection d'asynchronie.

Dans la suite du chapitre, la combinaison de paramètres choisie est {MFCCE+Δ, DCT+Δ}.

CANCOR vs. CoIA La figure 6.9 nous permet d'entrer dans les détails de la comparaison des comportements de CANCOR et CoIA. L'utilisation de la mesure de synchronie Γ mène systématiquement à des valeurs de `DCF` plus élevées que celles obtenues par les mesures de synchronie Ω ou λ. On remarque cependant la différence de comportement entre CANCOR et CoIA en comparant les synchronies Ω et λ. Là où la synchronie λ est bien meilleure que la synchronie Ω pour CoIA, l'inverse est constaté pour CANCOR. Ceci peut s'expliquer par le fait que l'estimation robuste des matrices $\mathbf{A}$ et $\mathbf{B}$ nécessite beaucoup plus de

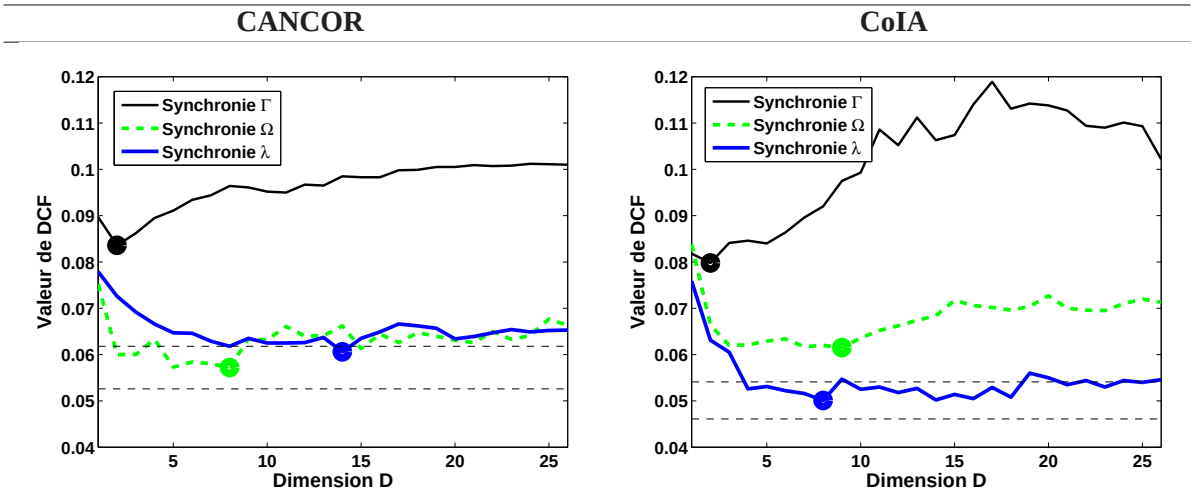


FIG. 6.9 – Comparaison des mesures basées sur CANCOR et CoIA. Les points • indiquent les meilleurs systèmes pour chaque configuration. Les lignes noires horizontales en pointillés indiquent l'intervalle de confiance à 95% sur la valeur de DCF des meilleures mesures CANCOR (Ω) et CoIA (λ).

données pour CANCOR (basée sur l'estimation des matrices de covariance C_{XY}, C_{XX} et C_{YY}) que pour CoIA (basée sur l'estimation de la seule matrice de covariance C_{XY}).

Dimension D Les courbes $\texttt{DCF} = f(D)$ pour les synchronies Γ, Ω et λ présentent des allures différentes. La mesure de synchronie Γ obtient ses meilleures performances pour des petites valeurs de D. En effet, ajouter des dimensions supplémentaires détériore irrémédiablement et très rapidement les performances. Le modèle de synchronie Γ intrinsèque d'une séquence audiovisuelle est ainsi résumé en très peu d'information, les dimensions restantes pouvant être considérées comme du bruit. La mesure de synchronie Ω obtient aussi ses meilleures performances pour des petites valeurs de D mais ajouter des dimensions supplémentaires au calcul de la mesure ne détériore que très peu les performances. Enfin, la mesure de synchronie λ nécessite un nombre plus important de dimensions pour atteindre ses meilleures performances. Il est possible d'interpréter ce comportement en considérant que les toutes premières dimensions de projection décrivent un comportement universel et les détails de la synchronie propre à chacun sont contenus dans les dimensions suivantes.

Synchronie Γ **par morceaux** Le tableau 6.3 résume en quelques chiffres les performances optimales (correspondant aux dimensions marquées d'un point • dans la figure 6.9) des différentes configurations qui partagent toutes les mêmes paramètres audiovisuels { MFCCE+Δ, DCT+Δ }. En outre, nous avons

Mesure de synchronie	Dimension D	DCF
CANCOR Γ	2	8.4 ± 0.5 %
CANCOR Γ par morceaux		7.9 ± 0.5 %
CANCOR Ω	8	5.7 ± 0.5 %
CANCOR λ	14	6.0 ± 0.4 %
CoIA Γ	2	8.0 ± 0.5 %
CoIA Γ par morceaux		7.4 ± 0.5 %
CoIA Ω	9	6.1 ± 0.5 %
CoIA λ	8	5.0 ± 0.4 %
Eveno	NA	9.7 ± 0.3 %

TAB. 6.3 – Meilleur système pour chaque mesure de synchronie sur le protocole S

reporté dans la figure 6.10 les courbes DET correspondantes. Pour des raisons de clarté, nous n'avons pas reporté celles correspondant aux mesures de synchronie Γ par morceaux. Comme le montre le tableau 6.3,

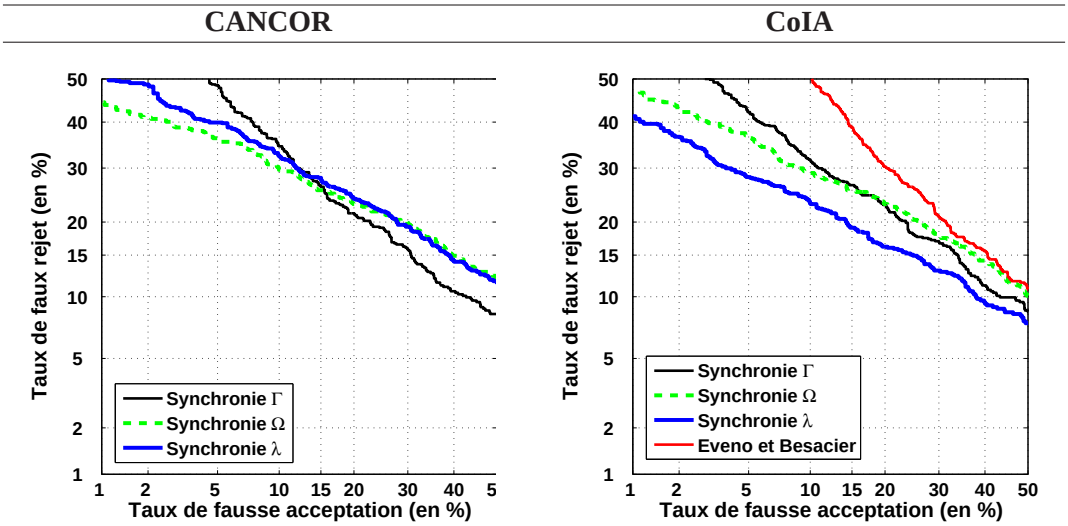


FIG. 6.10 – Courbes DET correspondant aux systèmes du tableau 6.3

l'amélioration apportée par la synchronie Γ par morceaux n'est pas significative pour le système CANCOR. Celui-ci étant déjà très limité par la quantité de données d'apprentissage disponible, la diviser par deux en appliquant la synchronie Γ par morceaux ne fait que lui rendre la tâche d'apprentissage encore plus difficile. En revanche, son application sur le système CoIA apporte une légère amélioration qu'il convient toutefois de relativiser : ses performances restent moins bonnes que la mesure de synchronie CoIA λ alors qu'elle demande pourtant environ cinquante fois plus de temps de calcul.

Comparaison avec l'existant En termes de DCF, les performances de notre implémentation de la technique proposée par *Eveno et Besacier* sont moins bonnes que chacune de nos meilleures propositions. Cependant, là où ils utilisaient des paramètres visuels liés à la forme des lèvres (hauteur, largeur et aire) et des paramètres acoustiques LPC, notre implémentation utilise les coefficients DCT et MFCC. Pouvoir tirer des conclusions définitives quant au meilleur système nécessiterait d'utiliser la même implémentation que celle décrite dans [Eveno et Besacier, 2005b].

6.5 Discussion

Nous avons proposé quatre variantes originales d'une mesure de synchronie de la parole audiovisuelle et avons évalué leurs performances pour la tâche de détection d'asynchronie.

Le système-étalon *Eveno et Besacier* permet d'obtenir une mesure du niveau de difficulté du protocole S : son taux d'égale erreur est d'environ 26%. Dans [Eveno et Besacier, 2005b], sur un protocole équivalent construit à partir de séquences de la base de données XM2VTSDB [Messer *et al.*, 1999], son taux d'égale erreur était d'environ 14%. La différence entre ces deux protocoles réside essentiellement dans les conditions d'enregistrement des séquences audiovisuelles. Là où le protocole d'*Eveno et Besacier* est basé sur des données de type *controlled* (avec fond bleu et une caméra de bonne qualité), le protocole S fait, quant à lui, appel à des données de type *controlled*, *degraded* (dans un bureau et avec une webcam) et *adverse* (la personne est debout dans un réfectoire, la tête penchée vers le bas), comme l'illustre la figure 2.3 de la page 47.

Aussi, nous avons défini un protocole $\mathtt{S}^c$ à partir du protocole S en ne conservant que les tests portant sur les séquences de type *controlled*. Les résultats obtenus par le meilleur système (CoIA λ avec $D = 8$) sur les deux protocoles sont comparés dans la figure 6.11. Les courbes DET montrent clairement que la dégradation des conditions d'enregistrement entraîne une dégradation des performances : les valeurs de DCF permettent de tirer la même conclusion : $\mathtt{DCF}\,(\mathtt{S}) = 5.0 \pm 0.4\%$ et $\mathtt{DCF}\,(\mathtt{S}^c) = 3.4 \pm 1.5\%$. Notons que pour le système *Eveno et Besacier*, $\mathtt{DCF}\,(\mathtt{S}^c) = 9.5 \pm 1.7\%$.

Le protocole S est d'autant plus difficile que des séquences "*asynchrones*" sont parfois synchrones (d'un point de vue subjectif) par le seul fait du hasard[1].

[1] Un exemple de séquence "*asynchrone*" particulièrement difficile à détecter est proposé en ligne à l'adresse `http://www.tsi.enst.fr/~bredin/these`, section *Compléments multimédia*.

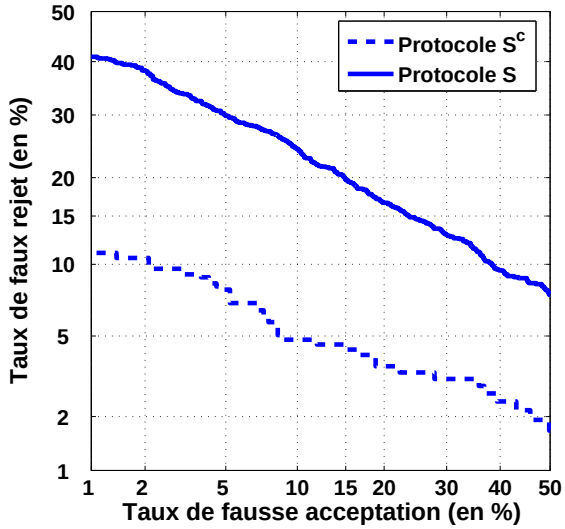


FIG. 6.11 – Performances de CoIA λ sur le protocole $\mathbb{S}^c$.

Poids des dimensions Les flux de parole acoustique X et visuel Y sont transformés en des flux $\mathcal{X}$ et $\mathcal{Y}$ de même dimension D par analyse de corrélation canonique et analyse de co-inertie. Les quatre variantes partagent la même mesure de corrélation qui affecte le même poids à chacune des D dimensions et que l'on rappelle ici :

$$s(X,Y) = \frac{1}{D}\sum_{k=1}^{D} w_k \,\mathrm{corr}\,(\mathcal{X}_k, \mathcal{Y}_k) \text{ avec } w_k = \text{constante} = 1 \tag{6.8}$$

Pourquoi ne pas pondérer différemment chacune des dimensions ? Est-il sensé de toutes leur donner le même poids ? Il serait certainement judicieux de s'intéresser à la question plus en détails. Nous allons tâcher d'y apporter une réponse préliminaire.

À gauche dans la figure 6.12, la valeur moyenne de $\mathrm{corr}\,(\mathcal{X}_k, \mathcal{Y}_k)$ estimée (pour la mesure CoIA λ) sur l'ensemble de test est reportée en fonction de la dimension k, pour les séquences synchrones (en pointillés verts) et asynchrones (en rouge). Il apparaît clairement que les premières dimensions contiennent un maximum de corrélation et la mesure définie par l'équation (6.8) donne, naturellement et malgré les apparences, plus de poids aux premières dimensions. Afin d'équilibrer l'influence de chaque dimension, nous proposons de normaliser les corrélations en introduisant les poids w_k définis par l'équation (6.9) et estimés à l'aide de

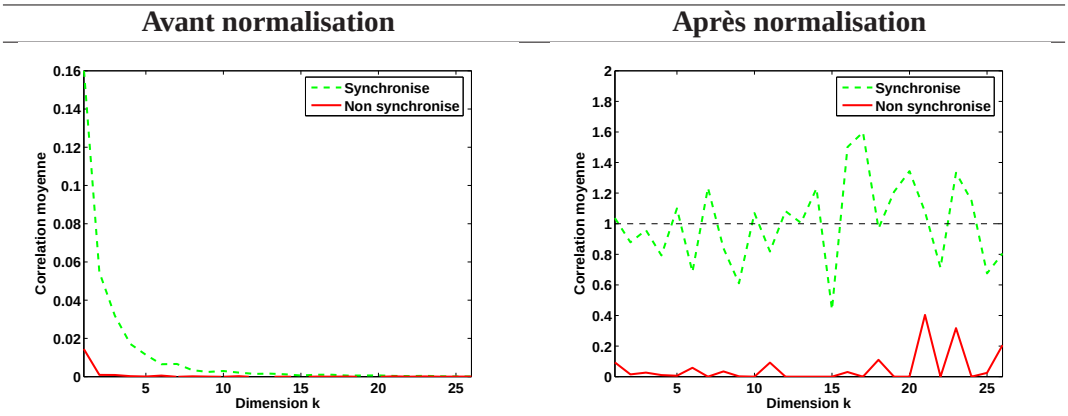


FIG. 6.12 – Effet de la normalisation sur les corrélations. La valeur moyenne de $w_k \operatorname{corr}(\mathcal{X}_k, \mathcal{Y}_k)$ estimée (pour la mesure CoIA λ) sur l'ensemble de test est reportée en fonction de la dimension k, à gauche avant normalisation ($w_k = 1$) et à droite après normalisation (voir équation (6.9)), pour les séquences synchrones (en pointillés verts) et asynchrones (en rouge).

l'ensemble (groupe G1 ou G2) de développement :

$$w_k = \mathbb{E}\left[\operatorname{corr}(\mathcal{X}_k, \mathcal{Y}_k)\right]^{-1} \tag{6.9}$$

L'effet de cette normalisation sur les corrélations est mise en évidence dans la figure 6.12 à droite. Son effet sur les performances globales (en termes de DCF) de la mesure de synchronie CoIA λ est l'objet de la figure 6.13 : elle tend à confirmer l'observation selon laquelle seules les premières dimensions apportent une réelle information, les suivantes ayant tendance à dégrader les performances. La version non-normalisée de la mesure de synchronie CoIA λ reste toutefois la meilleure.

Dépendance phonétique Une deuxième interrogation réside dans la modélisation globale de la synchronie audiovisuelle. Le mouvement conjoint des lèvres avec la voix est-il global ? Ne dépend-il pas du texte prononcé ? Ne serait-il pas préférable de modéliser la synchronie en fonction de la structure phonétique du texte prononcé ? Plusieurs pistes sont ouvertes pour essayer de répondre à ces questions. Dans [Argones-Rúa *et al.*, 2007b, Argones-Rúa *et al.*, 2007a] traitant aussi de la tâche de détection d'asynchronie, nous avons utilisé des HMM couplés à 5 états pour modéliser la synchronie, permettant ainsi de découper le signal en autant de classes "phonétiques". Aucune différence significative de performance n'a été constatée. Nous avons, en outre, implémenté une version avec fenêtre glissante de la mesure de synchronie Γ, visant à extraire localement l'information de synchronie. Les résultats ont montré une dégradation signifi-

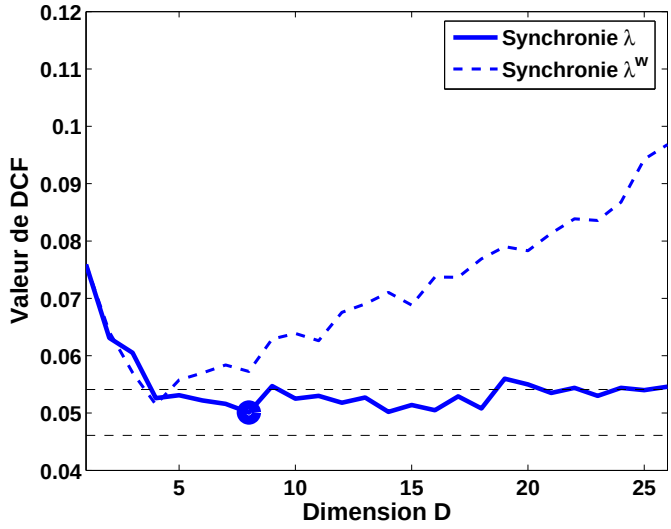


FIG. 6.13 – Comparaison des mesures de synchronie CoIA λ originale et normalisée λ^w.

cative des résultats de détection d'asynchronie, probablement du fait de la quantité trop réduite de données d'apprentissage.

Chapitre 7

Vérification d'identité

Introduction

Les bons résultats obtenus par la mesure de synchronie λ par CoIA pour la tâche de détection d'asynchronie nous ont amenés à réfléchir à son utilisation dans le cadre de la vérification d'identité. En effet, s'il est préférable d'utiliser un modèle dépendant du locuteur pour évaluer le degré de synchronie audiovisuelle, il est vraisemblable que ce modèle contienne une information relative à son identité. Nous sommes ainsi partis du postulat selon lequel chaque personne possède une façon de synchroniser sa voix et ses lèvres qui lui est propre et introduisons dans ce chapitre une troisième modalité biométrique (après la vérification du locuteur et celle du visage) liée à la synchronie audiovisuelle.

7.1 Principe de la modalité *synchronie*

Enrôlement À partir de paramètres acoustiques X^λ et visuels Y^λ extraits de la séquence d'enrôlement de la personne λ, l'application de CoIA permet de déduire les matrices $\mathbf{A}^\lambda$ et $\mathbf{B}^\lambda$.

$$\left(X^\lambda, Y^\lambda\right) \rightarrow \left(\mathbf{A}^\lambda, \mathbf{B}^\lambda\right) \tag{7.1}$$

Le couple $\left(\mathbf{A}^\lambda, \mathbf{B}^\lambda\right)$ constitue alors le modèle d'identité du client λ.

Test Les paramètres acoustiques X^Γ et visuels Y^Γ de la séquence Γ dont on cherche à déterminer si elle correspond à la personne λ sont extraits puis transformés à l'aide des deux matrices $\mathbf{A}^\lambda$ et $\mathbf{B}^\lambda$ afin d'obtenir les paramètres acoustiques et visuels *corrélés* $\mathcal{X}$ et $\mathcal{Y}$:

$$\begin{aligned} \mathcal{X} &= \mathbf{A}^{\lambda t} X^{\Gamma} \\ \mathcal{Y} &= \mathbf{B}^{\lambda t} Y^{\Gamma} \end{aligned} \tag{7.2}$$

Les corrélations entre chaque dimension de $\mathcal{X}$ et $\mathcal{Y}$ sont calculées et participent à la mesure $s_D^{\lambda}\left(X^{\Gamma}, Y^{\Gamma}\right)$ de similarité $S_{\text{synchronie}}(\Gamma|\lambda)$:

$$\begin{aligned} S_{\text{synchronie}}(\Gamma|\lambda) &= s_D^{\lambda}\left(X^{\Gamma}, Y^{\Gamma}\right) \\ &= \frac{1}{D}\sum_{k=1}^{D} \operatorname{corr}\left(\mathcal{X}_k, \mathcal{Y}_k\right) \\ &= \frac{1}{D}\sum_{k=1}^{D} \frac{{\mathcal{X}_k}^t \mathcal{Y}_k}{\sqrt{{\mathcal{X}_k}^t \mathcal{X}_k}\sqrt{{\mathcal{Y}_k}^t \mathcal{Y}_k}} \\ &= \frac{1}{D}\sum_{k=1}^{D} \frac{\left(\mathbf{a}_k^{\lambda t} X^{\Gamma}\right)^t \left(\mathbf{a}_k^{\lambda t} Y^{\Gamma}\right)}{\sqrt{\left(\mathbf{a}_k^{\lambda t} X^{\Gamma}\right)^t \left(\mathbf{a}_k^{\lambda t} X^{\Gamma}\right)}\sqrt{\left(\mathbf{a}_k^{\lambda t} Y^{\Gamma}\right)^t \left(\mathbf{a}_k^{\lambda t} Y^{\Gamma}\right)}} \end{aligned} \tag{7.3}$$

où

$D \leq d$ est le nombre de dimensions conservées dans la mesure de synchronie.

Cette mesure est finalement comparée à un seuil permettant de vérifier l'identité clamée par la personne :

L'accès est accepté si $S_{\text{synchronie}}(\Gamma|\lambda) > \theta$ et refusé dans le cas contraire.

7.2 Évaluation

La figure 7.1 résume les expériences menées sur le protocole P. Pour chaque courbe, le titre indique quelle combinaison de paramètres acoustiques et visuels est utilisée. La valeur de DCF est tracée en fonction de la dimension D.

Énergie acoustique L'observation des performances (en termes de DCF) des systèmes utilisant (lignes 2, 4 et 6) ou non (lignes 1, 3 et 5) l'énergie acoustique montre que l'ajout de l'énergie acoustique aux vecteurs de paramètres acoustiques tend à dégrader les performances. Ce comportement correspond au phénomène généralement observé en vérification du locuteur basée sur les coefficients MFCC : l'ajout de l'énergie acoustique dégrade les performances. Par la suite, on préfère donc ne pas prendre en compte cette informa-

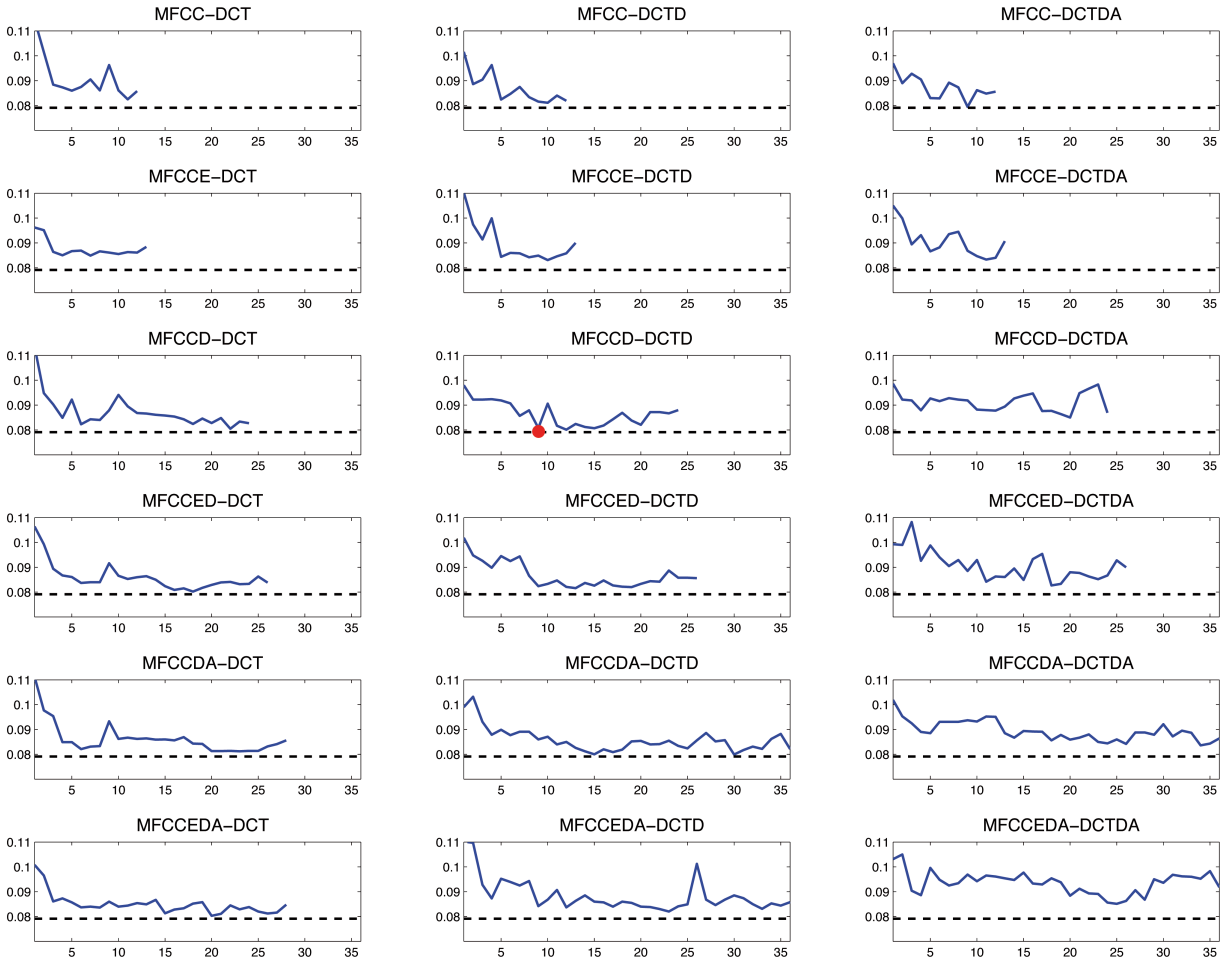


FIG. 7.1 – Performance de la modalité *synchronie* sur le protocole P. Chaque courbe représente la valeur de DCF en fonction de D et correspond à une combinaison MFCC/DCT (voir les tableaux 6.1 et 6.2 aux pages 96 et 96). La ligne horizontale noire en pointillés correspond à la valeur de DCF du meilleur système (point rouge).

tion perturbatrice.

Dérivées Comme nous l'avions constaté pour l'application de détection de synchronie, l'ajout des dérivées premières (autant acoustiques que visuelles) entraîne l'amélioration des performances. En revanche, les dérivées secondes n'apportent aucune amélioration significative supplémentaire (voire détériorent les

performances dans le cas des dérivées secondes visuelles).

Dans la suite du chapitre, la combinaison de paramètres choisie est {MFCC+Δ, DCT+Δ} avec $D = 9$.

Znorm De façon analogue aux deux systèmes basés sur les modalités *voix* et *visage*, nous appliquons une étape supplémentaire de normalisation des scores par Znorm. Les courbes DET de la figure 7.2 montrent qu'aucune amélioration significative n'est apportée par cette normalisation.

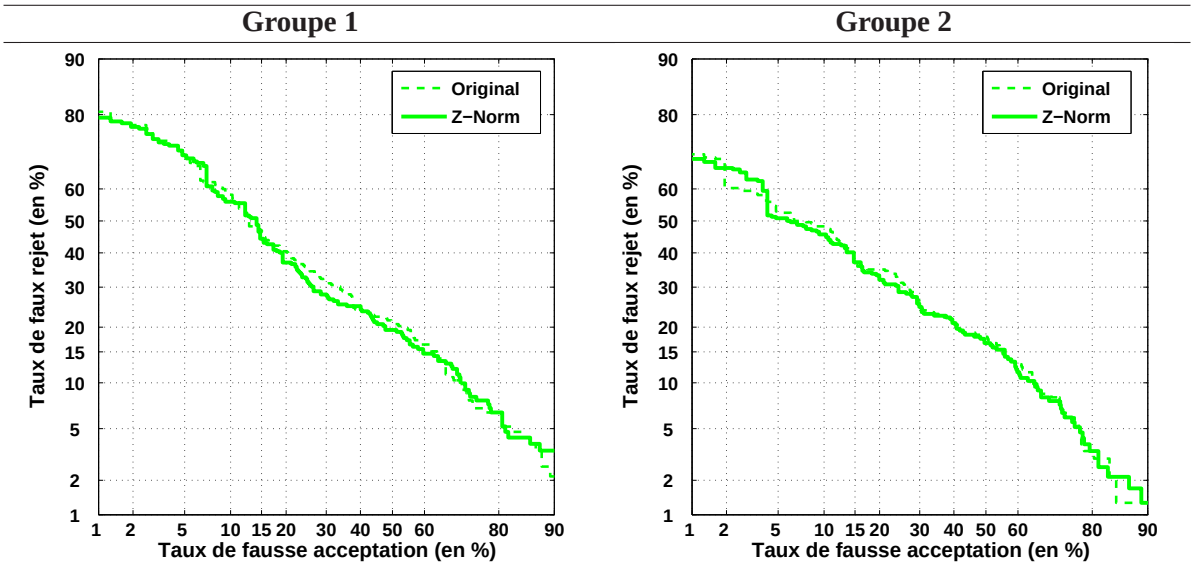
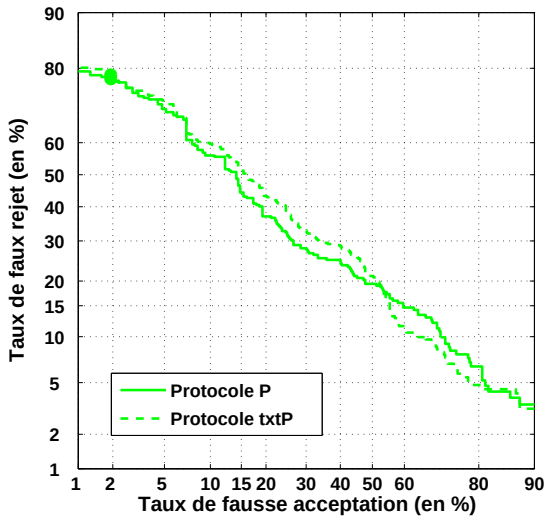
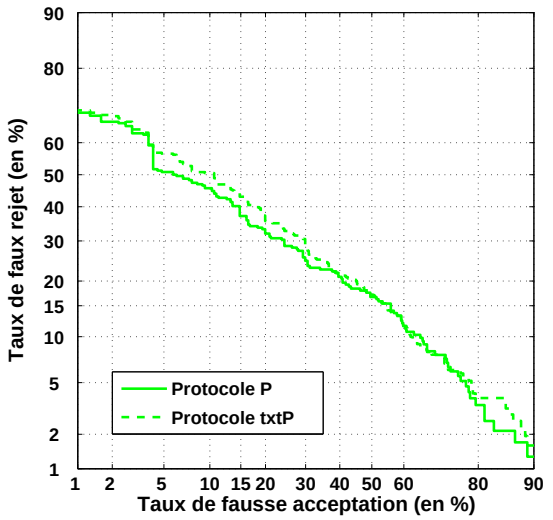


FIG. 7.2 – Influence de la Znorm sur le système basé sur la synchronie

Influence du texte prononcé Le protocole P peut être considéré comme un protocole dépendant du texte. En effet, à chaque personne λ sont associés un nom et une adresse qui lui sont propres et qu'elle prononce lors de ses accès *client*. En outre, lors des accès *imposteur*, l'imposteur prononce le nom et l'adresse que sa cible utilise pour s'authentifier. Afin d'étudier la dépendance du modèle avec la phrase d'enrôlement, nous avons introduit le protocole txtP qui est indépendant du texte dans le sens où les clients prononcent une phrase différente de celle d'enrôlement. Les courbes et chiffres de la figure 7.3 décrivent les résultats obtenus. Puisque seuls les accès *client* diffèrent entre les protocoles P et txtP, les valeurs de FAR sont les mêmes. On note une légère augmentation du FRR qui n'est cependant pas statistiquement significative (au vu des intervalles de confiance). Contrairement à la modalité *voix* dont les performances varient significativement entre les protocoles P et txtP (le FRR passe d'environ 42% à 53%), le modèle de *synchronie* créé lors de





Protocole P original	DCF $= 8.61 \pm 0.85\%$ FAR $= 0.96 \pm 0.77\%$ FRR $= 76.61 \pm 3.84\%$
Protocole txtP	DCF $= 8.72 \pm 0.83\%$ FAR $= 0.96 \pm 0.77\%$ FRR $= 77.69 \pm 3.27\%$

FIG. 7.3 – Influence du texte prononcé

l'enrôlement n'est donc pas perturbé par l'utilisation d'une phrase d'authentification différente : il est assez général pour être utilisé dans un cadre indépendant du texte. Ce résultat s'avère très satisfaisant.

Robustesse aux attaques Rappelons que cette nouvelle modalité biométrique a été introduite dans le but de parer aux attaques délibérées d'imposture. Il convient donc d'étudier ses performances face aux attaques introduites au chapitre 4 et qui constitue un réel danger pour le système de fusion *locuteur+visage*. Comme nous pouvons le constater dans la figure 7.4, la nouvelle modalité est intrinsèquement robuste aux attaques : elles sont toutes rejetées, sans exception. Cependant, ses performances brutes (sur le protocole P original, où les impostures sont aléatoires) sont beaucoup moins satisfaisantes. Là où le FRR du système de fusion *locuteur+visage* est d'environ 37%, celui de la modalité *synchronie* atteint 76%, multipliant par deux le nombre de clients faussement rejetés et potentiellement mécontents. Enfin, en termes de DCF, le système de fusion *locuteur+visage* possède des performances brutes largement meilleures (5.8% contre 8.6% pour la modalité *synchronie*).

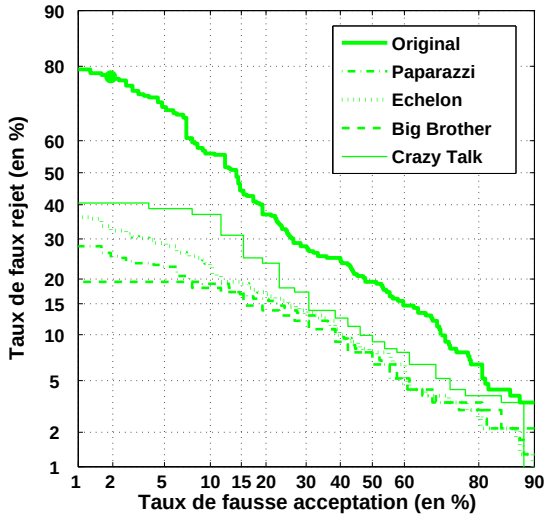
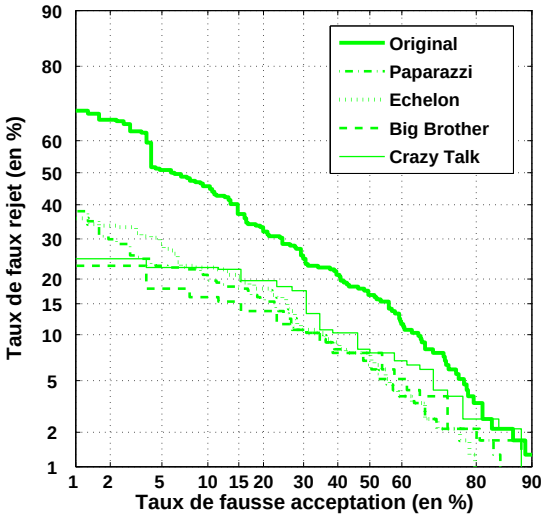
Groupe 1	Groupe 2
	
Protocole P original	DCF $= 8.61 \pm 0.85\%$ FAR $= 0.96 \pm 0.77\%$ FRR $= 76.61 \pm 3.84\%$
Attaque *Paparazzi*	DCF $= 7.66 \pm 0.38\%$ FAR $= 0\%$
Attaque *Echelon*	DCF $= 7.66 \pm 0.38\%$ FAR $= 0\%$
Attaque *Big Brother*	DCF $= 7.66\% \in [7\%, 15\%]$ FAR $= 0\% \in [0\%, 7\%]$
Attaque *Crazy Talk*	DCF $= 7.66\% \in [7\%, 15\%]$ FAR $= 0\% \in [0\%, 7\%]$

FIG. 7.4 – Performances du système basé sur la synchronie

7.3 Discussion

Nous avons proposé dans ce chapitre une méthode originale de vérification d'identité basée sur l'analyse de la synchronie audiovisuelle. Contrairement au système de fusion *locuteur+visage*, cette nouvelle modalité est intrinsèquement robuste aux attaques délibérées d'imposture introduites au chapitre 4. En outre, contrairement au module de vérification du locuteur, les performances de la modalité *synchronie* ne sont pas dégradées dans un contexte de vérification indépendante du texte.

Notons que même dans le cas des attaques *Crazy Talk* où les lèvres et la voix sont effectivement synchronisées, la modalité *synchronie* est en mesure de rejeter les imposteurs. Cette observation confirme le fait qu'il y a une réelle information d'identité dans la façon qu'a une personne de synchroniser sa voix et ses lèvres et que le logiciel *Crazy Talk* n'est pas en mesure de reproduire.

Les performances relativement faibles de la modalité *synchronie* sur le protocole P original (DCF $= 8.6\%$ à comparer à DCF $= 5.8\%$ pour le système de fusion *locuteur+visage*) s'expliquent en partie par les erreurs de segmentation des lèvres. En effet, en nous penchant sur la distribution des scores *client* issus du test sur le groupe G1, nous avons extrait les deux personnes pour lesquelles les scores étaient les plus faibles. Nous avons ensuite visualisé le résultat de la détection du visage et de la bouche sur leurs séquences d'enrôlement et présentons deux résultats typiques dans la figure 7.5. De nombreuses erreurs de segmentation surviennent

FIG. 7.5 – Erreur de détection de la bouche résultant en un mauvais modèle

tout au long de ces deux séquences. La mauvaise qualité des modèles de synchronie résultant explique alors pourquoi ces deux clients sont faussement rejetés au moment du test. La première piste d'amélioration de cette modalité réside donc dans le perfectionnement du module de segmentation du visage et de la bouche.

Malgré toutes ces propriétés très prometteuses, les performances brutes (sur le protocole P original) relativement faibles ne permettent pas la mise en place d'un système biométrique basé sur cette seule modalité *synchronie*. Il convient de tirer profit de sa complémentarité avec le système de fusion *locuteur+visage* : c'est l'objet du chapitre suivant.

Chapitre 8

Fusion robuste

Introduction

Nous avons jusqu'ici décrit et étudié deux systèmes de vérification de l'identité des visages parlants : le premier est un système classique basé sur la fusion des deux modalités *locuteur* et *visage*, le second repose quant à lui sur la modalité *synchronie* que nous avons introduite dans le chapitre précédent. Alors qu'il possède les meilleures performances brutes de vérification, le premier système est néanmoins très peu robuste aux attaques décrites au chapitre 4. À l'inverse, la modalité *synchronie* a des performances brutes relativement faibles mais est intrinsèquement robuste aux impostures délibérées. L'objectif de ce chapitre est de tirer profit de cette complémentarité en fusionnant ces deux systèmes de façon à obtenir un système final à la fois robuste aux attaques et obtenant des performances brutes satisfaisantes.

8.1 Stratégies de fusion

Chacune des trois modalités *locuteur*, *visage* et *synchronie* fournit un score : S_{locuteur}, S_{visage} et $S_{\text{synchronie}}$ respectivement. Trois stratégies de fusion de ces scores sont proposées et évaluées relativement à leurs performances brutes (sur le protocole P original) et à leur robustesse aux attaques.

Remarque Pour plus de lisibilité, nous noterons par la suite S_l, S_v et S_s les scores normalisés (voir le paragraphe 3.3.2 à propos de la normalisation *tanh*) issus respectivement de la vérification du locuteur, du visage et de la synchronie.

8.1.1 Fusion naïve

La première stratégie de fusion consiste en une extension de la stratégie de fusion des deux modalités locuteur et visage (présentée au paragraphe 3.4, page 64) aux trois modalités locuteur, visage et synchronie. Comme le résume l'équation (8.1), il s'agit de la somme pondérée des trois scores S_l, S_v et S_s.

$$S_1 = w_l S_l + w_v S_v + w_s S_s \text{ avec } w_l + w_v + w_s = 1 \tag{8.1}$$

L'estimation des poids optimaux w_l, w_v et w_s est réalisée en minimisant le taux d'erreur sur l'ensemble de développement (G1 quand le système est testé sur G2, et réciproquement).

8.1.2 Fusion robuste

Comme nous le verrons par la suite, l'apport de cette première stratégie de fusion par rapport au système de fusion *locuteur+visage* est nul. Aussi, nous proposons deux nouvelles stratégies de fusion tirant mieux profit des spécificités de chacun des systèmes en termes de robustesse aux attaques et de performances brutes.

Mesure de confiance

Comme on peut le constater dans la figure 8.1 à gauche, les scores de la modalité *synchronie* obtenus par les imposteurs (aléatoires ou délibérés, en rouge et vert) sont, en moyenne, plus faibles que ceux obtenus lors des accès *client* (en bleu).

La différence entre la distribution des scores *imposteur* aléatoires et délibérés s'explique par le fait que les lèvres et la voix ne sont pas synchrones pour les secondes (à l'exception des attaques *Crazy Talk* où elles sont artificiellement synchronisées) alors qu'elles le sont dans le cas d'imposteurs aléatoires. Néanmoins, les scores *client* sont, en moyenne, plus élevés que les scores *imposteur* aléatoires puisque la mesure de synchronie utilisée est la mesure λ dépendante de l'identité du client. Nous proposons donc de définir une mesure de confiance α en le système de fusion *locuteur+visage* initial, fonction du score S_s fourni par la modalité *synchronie* :

$$\alpha\left(S_s\right) = p\left(s \leq S_s | \text{ accès client}\right) \tag{8.2}$$

La mesure α correspond à la fonction de répartition des scores *client* de la modalité *synchronie*. Cette

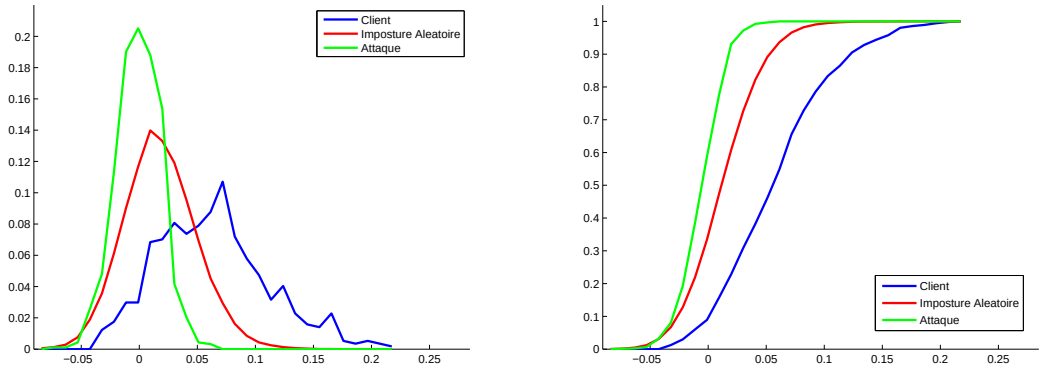


FIG. 8.1 – À gauche, distribution des scores de la modalité *synchronie* pour les accès *client*, *imposteur* aléatoires et les attaques. À droite, les fonctions de répartitions correspondantes.

fonction est estimée à partir des accès *client* de l'ensemble d'apprentissage. Son allure est dessinée dans la figure 8.1 à droite en bleu. Il s'agit d'une mesure de confiance en le système de fusion *locuteur+visage* dans le sens où sa valeur est plus élevée si la mesure de synchronie est grande (i.e. s'il est plus probable qu'aucune attaque n'ait lieu, auquel cas on peut faire confiance au système de fusion initial) et plus faible si la mesure de synchronie est petite (i.e. s'il est probable que le système soit attaqué par un imposteur –aléatoire ou délibéré–, auquel cas il est préférable de considérer avec précaution le score issu du système de fusion initial).

Pénalisation

La première stratégie de fusion robuste consiste à pénaliser les accès dont la mesure de confiance est faible. L'équation (8.3) résume ce processus de pénalisation :

$$S_2 = \alpha\left(S_s\right) S_1 \tag{8.3}$$

Le score S_1 défini par l'équation (8.1) est ainsi modulé par la mesure de confiance qui varie entre 0 (lorsque la mesure de synchronie est minimale) et 1 (lorsque elle est maximale).

Somme pondérée adaptative

La seconde stratégie de fusion robuste vise à profiter de la complémentarité entre les performances brutes de la première stratégie de fusion et la robustesse aux attaques de la modalité *synchronie*. Alors que la première stratégie de fusion est très sensible aux attaques délibérées d'imposture mais possède les meilleures performances brutes, la modalité *synchronie* est très robuste aux attaques mais possède des performances brutes limitées. Aussi, on propose de réaliser une somme pondérée de ces deux systèmes en fixant les poids en fonction de la mesure de confiance :

$$S_3 = \alpha\left(S_s\right) S_1 + \left[1 - \alpha\left(S_s\right)\right] S_s \tag{8.4}$$

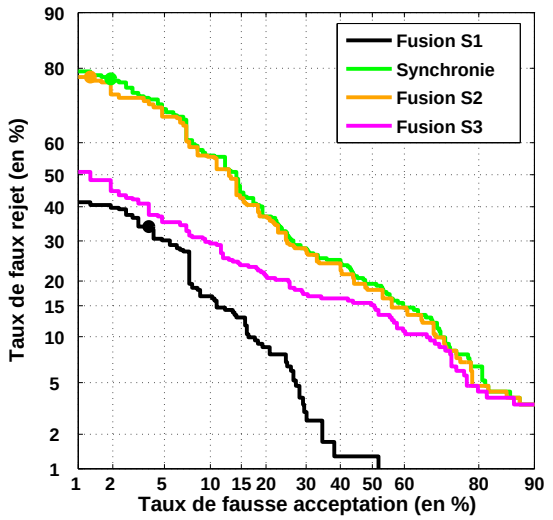
Comme le montre l'équation (8.4), cette dernière stratégie est basée sur une somme pondérée adaptative des scores normalisés. Un poids plus important est donné à la modalité *synchronie* quand la mesure de synchronie est faible. Réciproquement, son poids est réduit quand la mesure de synchronie est élevée et que l'on peut avoir confiance en la stratégie de fusion initiale.

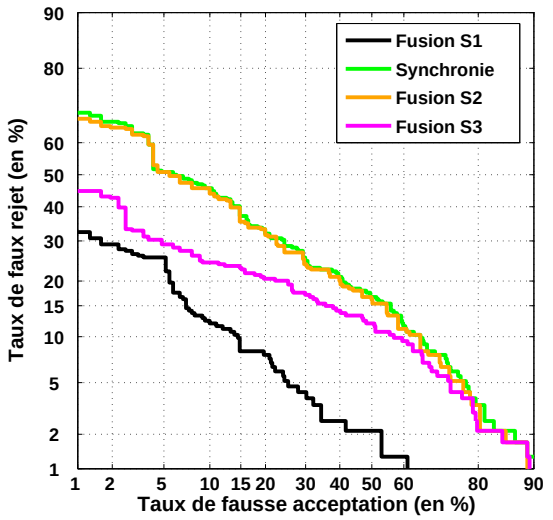
8.2 Évaluation

Fusion naïve L'apprentissage des poids w_l, w_v et w_s sur les ensembles de développement G1 et G2 met en évidence la faiblesse principale de la modalité *synchronie* : $w_l = 0.66$, $w_v = 0.34$ et $w_s = 0.00$ pour G1 et $w_l = 0.62$, $w_v = 0.38$ et $w_s = 0.00$ pour G2. En d'autres termes, ses mauvaises performances brutes ont tendance à dégrader les performances du système de fusion *locuteur+visage* initial : un poids nul lui est donc affecté et le système de fusion S_1 est identique au système de fusion initial.

Performance brute La figure 8.2 résume les performances des différentes stratégies de fusion sur le protocole P original. En termes de DCF, la fusion naïve donne les meilleures performances, très similaires à celles obtenues par la somme pondérée adaptative. En ce qui concerne la stratégie de pénalisation, ses performances sont identiques à celle de la modalité *synchronie*. Les stratégies de pénalisation et somme pondérée adaptative rendent l'accès plus difficile pour les imposteurs (FAR très faibles) comme pour les clients (FRR beaucoup plus élevés, passant de 38% à 54% et 76% respectivement).

Robustesse aux attaques Les figures 8.3 et 8.4 mettent en évidence la robustesse des stratégies de pénalisation et somme pondérée adaptative aux attaques délibérées d'imposture. Les courbes correspondant au système de fusion naïve ne sont pas répétées ici : puisque $w_s = 0$, le système de fusion naïve est le même





Somme pondérée S_1	DCF $= 5.82 \pm 1.19\%$ FAR $= 2.09 \pm 1.12\%$ FRR $= 37.55 \pm 4.40\%$
Synchronie	DCF $= 8.61 \pm 0.85\%$ FAR $= 0.96 \pm 0.77\%$ FRR $= 76.61 \pm 3.84\%$
Pénalisation S_2	DCF $= 8.21 \pm 0.73\%$ FAR $= 0.64 \pm 0.63\%$ FRR $= 75.75 \pm 3.89\%$
Somme pondérée adaptative S_3	DCF $= 6.00 \pm 0.77\%$ FAR $= 0.64 \pm 0.62\%$ FRR $= 53.65 \pm 4.53\%$

FIG. 8.2 – Performances des systèmes de fusion sur le protocole P original

que le système de fusion de référence *locuteur+visage*. Là encore, le système de fusion par pénalisation obtient des performances identiques (selon l'intervalle de confiance à 95%) à celle de la modalité *synchronie*. Il rejette la totalité des attaques délibérées d'imposture mais entraîne aussi un taux de faux rejet très élevé. Le système de fusion par somme pondérée adaptative apparaît comme un bon compromis entre performance brute et robustesse. En termes de DCF, il obtient, sur le protocole P, des performances brutes similaires au système de référence. Il est aussi très robuste face aux attaques : il est meilleur que la modalité *synchronie* pour les attaques *Paparazzi*, obtient des performances similaires pour les attaques *Echelon* et *Big Brother* et est légèrement moins efficace face à l'attaque *Crazy Talk* (mais la différence n'est pas statistiquement

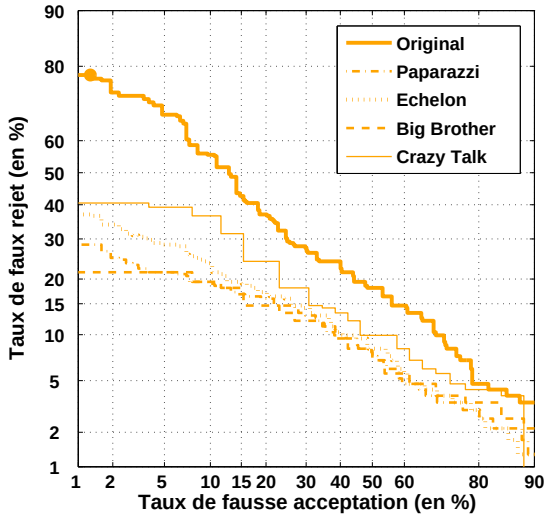
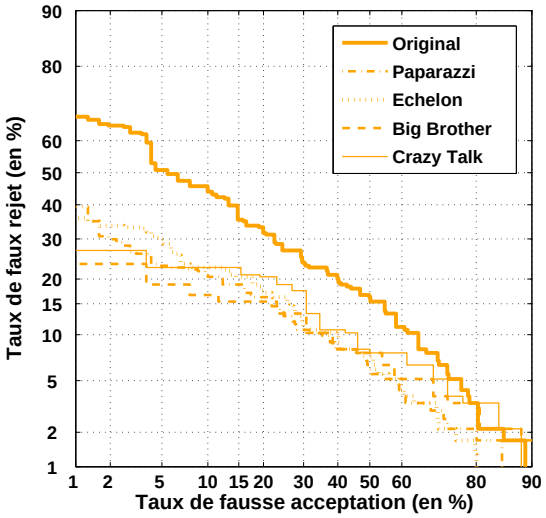
Groupe 1	Groupe 2
	
Protocole P original	DCF $= 8.21 \pm 0.73\%$ FAR $= 0.64 \pm 0.63\%$ FRR $= 75.75 \pm 3.89\%$
Attaque *Paparazzi*	DCF $= 7.58 \pm 0.39\%$ FAR $= 0\%$
Attaque *Echelon*	DCF $= 7.58 \pm 0.39\%$ FAR $= 0\%$
Attaque *Big Brother*	DCF $= 7.58\% \in [7\%, 15\%]$ FAR $= 0\% \in [0\%, 7\%]$
Attaque *Crazy Talk*	DCF $= 7.58\% \in [7\%, 15\%]$ FAR $= 0\% \in [0\%, 7\%]$

FIG. 8.3 – Performances du système de fusion par pénalisation

significative).

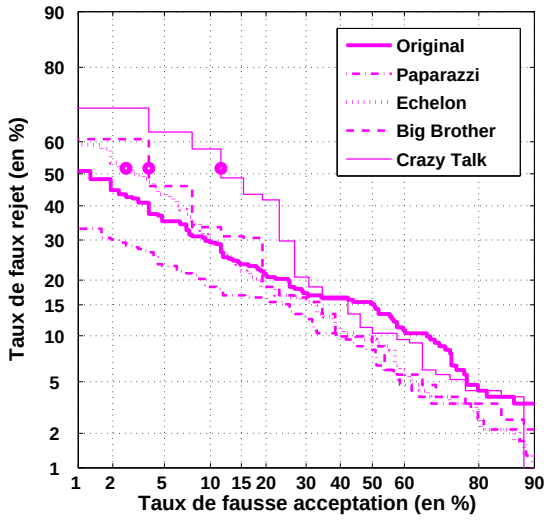
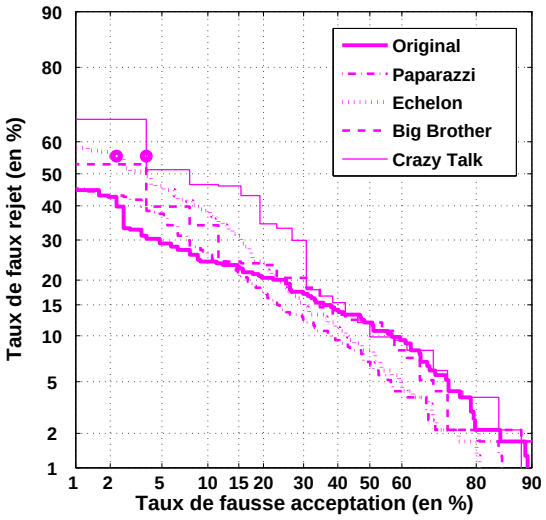
Groupe 1	Groupe 2
	
Protocole P original	DCF $= 6.00 \pm 0.77\%$ FAR $= 0.64 \pm 0.62\%$ FRR $= 53.65 \pm 4.53\%$
Attaque *Paparazzi*	DCF $= 5.68 \pm 0.63\%$ FAR $= 0.32 \pm 0.44\%$
Attaque *Echelon*	DCF $= 7.75 \pm 1.27\%$ FAR $= 2.41 \pm 1.20\%$
Attaque *Big Brother*	DCF $= 7.27\% \in [5\%, 16\%]$ FAR $= 1.92\% \in [0\%, 11\%]$
Attaque *Crazy Talk*	DCF $= 12.98\% \in [7\%, 25\%]$ FAR $= 7.69\% \in [2\%, 19\%]$

FIG. 8.4 – Performances du système de fusion par somme pondérée adaptative

8.3 Discussion

La figure 8.5 résume toutes ces expériences et inclut les performances du système de référence, de la modalité *synchronie* et des deux stratégies de fusion par pénalisation et par somme pondérée adaptative. Le compromis entre performance brute et robustesse aux attaques est mis en évidence en reportant en abscisse la valeur de DCF sur le protocole P original et en ordonnée la valeur de DCF face aux deux attaques les plus difficiles (*Big Brother* et son animation *Crazy Talk*). Alors qu'il possède les meilleures performances

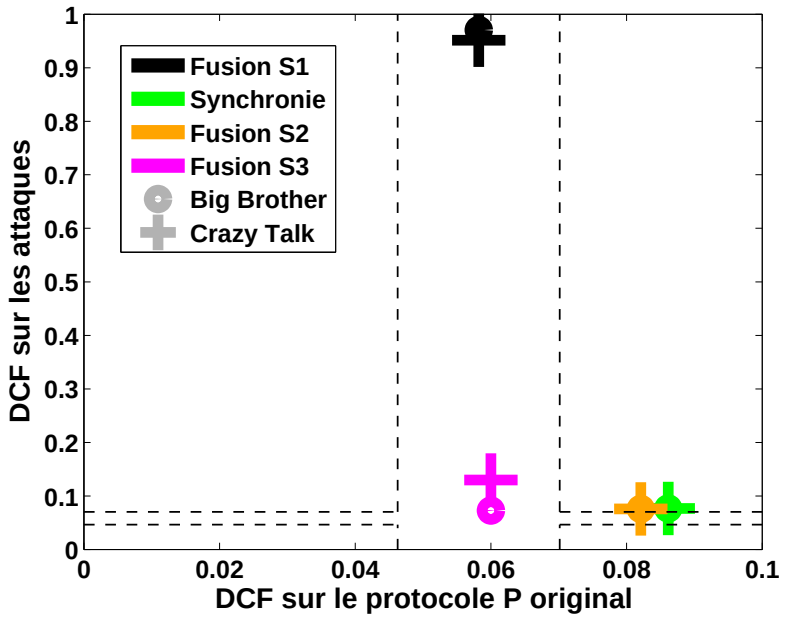


FIG. 8.5 – Compromis entre performance brute et robustesse aux attaques

brutes (dont l'intervalle de confiance est reporté en trait pointillé à la fois en abscisse et en ordonnée, pour référence), il apparaît clairement que le système de référence (équivalent au système de fusion S_1) est complètement inefficace face aux attaques. À l'inverse, la modalité *synchronie* est très robuste aux attaques (la valeur de DCF référence est atteinte pour les attaques) mais possède des performances brutes éloignées de la valeur de référence sur le protocole P original. Le système de fusion par somme pondérée adaptative bénéficie à la fois des performances brutes du système de référence *locuteur+visage* et de la robustesse de la modalité *synchronie*.

Ce meilleur compromis nous a ainsi permis d'atteindre l'objectif fixé dans la conclusion de la première partie : rendre le système de référence robuste aux attaques sans dégrader ses performances brutes.

Conclusions et perspectives

Conclusions

Les systèmes de vérification d'identité basée sur les visages parlants reposent pour la plupart sur la fusion multimodale de deux modules de vérification du locuteur et de reconnaissance du visage. Cette architecture parallèle constitue leur principal défaut : il suffit en effet de montrer une photographie d'une personne en utilisant un magnétophone pour rejouer un enregistrement de sa voix pour tromper ce type de système.

Première contribution Nous avons défini et simulé quatre scénarios d'attaques délibérées d'imposture (nommés *Paparazzi*, *Echelon*, *Big Brother* et *Crazy Talk*) et y avons confronté un système référence basé sur l'architecture classique de fusion des scores de vérification du locuteur et du visage. Nous avons ainsi mis en évidence la totale inefficacité de ce type de système face à des attaques pourtant facilement accessibles et réalisables par une personne malintentionnée.

Quelques parades simples ont déjà été proposées dans la littérature. La majorité d'entre elles se base sur l'exploitation des données vidéos uniquement : il s'agit d'analyser le visage et ses mouvements et d'en déduire une décision sur l'éventualité d'une attaque. Cependant, les logiciels d'animation de visage tels que *Crazy Talk* de la société *Reallusion* sont de plus en plus performants et proposent des animations très réalistes susceptibles de tromper ce type de parade.

Deuxième contribution Nous avons proposé quatre nouvelles mesures de synchronie audiovisuelle permettant de mesurer le degré de correspondance entre la voix acquise par le microphone et le mouvement des lèvres vues par la caméra. Elles sont basées sur l'analyse de corrélation canonique et l'analyse de coinertie des deux flux acoustique et visuel et apportent un gain en performance conséquent par rapport à la méthode dont elles sont inspirées. Les mesures Γ et Γ par morceaux permettent d'extraire une mesure de la synchronie intrinsèque d'une séquence audiovisuelle, sans apprentissage préalable de modèle de synchronie. La mesure Ω est basée sur un modèle de synchronie comparable à un modèle du monde en vérification du locuteur. Enfin, la mesure λ fait appel à un modèle de synchronie dépendant du client.

La comparaison de ces différentes mesures de synchronie pour la tâche de détection d'asynchronie a mis en avant la mesure de synchronie λ dépendante du client : ses bons résultats nous ont ensuite amenés à réfléchir à son application pour la vérification d'identité.

Troisième contribution À partir du postulat selon lequel chaque personne possède une façon qui lui est propre de synchroniser sa voix et ses lèvres, nous avons proposé une nouvelle modalité biométrique

basée sur la synchronie audiovisuelle. Lors de l'enrôlement, un modèle de synchronie dépendant du client constitué de deux matrices de projection est calculé par analyse de co-inertie entre les flux de parole acoustique et visuel. Au moment du test, les deux flux de parole acoustique et visuel sont transformés par les matrices du modèle de l'identité clamée. La mesure de synchronie est finalement utilisée comme score de vérification. Bien que ses performances brutes soient moins bonnes que le système de référence, la modalité *synchronie* est intrinsèquement robuste aux attaques délibérées d'imposture.

Ainsi, le système de référence et cette nouvelle modalité sont tout à fait complémentaires : quand l'un possède de bonnes performances brutes mais est inefficace face aux attaques, l'autre y est robuste mais possède des performances brutes moyennes.

Quatrième contribution Nous avons donc proposé deux nouvelles stratégies de fusion visant à tirer profit de cette complémentarité. Elles font toutes deux appel à une nouvelle mesure de confiance (basée sur la mesure de synchronie) en le système initial. La première stratégie de fusion (dite de *pénalisation*) vise à pénaliser les accès dont la mesure de confiance est faible. La seconde stratégie est une somme pondérée adaptative (en fonction de la mesure de confiance) des scores issus du système initial *locuteur+visage* et de la modalité *synchronie*. Elle donne plus de poids au système initial lorsque la mesure de confiance est élevée. Inversement, elle privilégie la modalité *synchronie* lorsque la mesure de confiance est faible.

Au final, la stratégie de fusion par somme pondérée adaptative des scores du système de référence et de la modalité *synchronie* apporte le meilleur compromis possible : elle permet de concilier les performances brutes du système initial et la robustesse aux attaques de la modalité *synchronie*. Nous avons ainsi apporté une solution originale et efficace au problème de robustesse aux attaques délibérées d'imposture rencontré par les systèmes de vérification biométrique d'identité basés sur les visages parlants. Même les attaques de type *Crazy Talk* (que l'on considère comme les plus difficiles à contrer) ne parviennent pas à tromper le système final, là où les méthodes proposées dans la littérature et basées sur la seule analyse de la partie visuelle du signal auraient échoué.

Perspectives à court terme

Après analyse des résultats obtenus par la modalité *synchronie*, il apparaît que la grande majorité des erreurs qu'elle commet est issue d'une mauvaise segmentation de la zone des lèvres. Cette étape cruciale au traitement de la parole audiovisuelle mériterait donc à l'avenir de recevoir toute notre attention.

Parmi les attaques délibérées d'imposture que nous avons proposées, l'attaque de type *Crazy Talk* constitue la menace la plus difficilement détectable. Cependant, dans le cas où une phrase aléatoire différente est demandée à chaque nouvel accès, elle serait inopérante puisqu'il est très peu probable que l'enregistrement audio préalable contiennent cette même phrase. Aussi, il conviendra, à court terme, de se pencher sur la question de l'élaboration d'attaques plus élaborées : une solution serait de faire appel conjointement à des techniques de conversion et/ou synthèse de voix et d'animation du visage. La voix de l'imposteur prononçant la phrase demandée serait transformée de façon à ressembler à celle de la cible et une photographie du visage serait animée à l'aide du logiciel *Crazy Talk*.

Des expériences préliminaires ont déjà été menées et montrent qu'une simple transformation dans le domaine cepstral suffit à augmenter de façon drastique le taux de fausse acceptation d'un système de vérification du locuteur. Ceci a fait l'objet de la publication [Perrot *et al.*, 2007] reportée en annexe (page 182).

Perspectives à long terme

La vérification de l'identité d'un visage parlant est loin d'être la seule application de ces nouvelles mesures de synchronie audiovisuelle.

Par exemple, elles pourraient être utilisées pour noter les synthétiseurs audiovisuels de parole et ainsi fournir une mesure objective à mettre en relation avec les mesures subjectives généralement utilisées dans ce domaine. Un tel outil d'évaluation objective pourrait, par exemple, être utilisé dans une campagne d'évaluation de synthétiseurs audiovisuels.

Dans l'industrie du cinéma, la qualité du doublage de longs métrages en langue étrangère pourrait aussi être évaluée à l'aide de telles mesures. Il suffirait par exemple d'acquérir le mouvement des lèvres du doubleur et de le comparer à la voix de l'acteur original. La meilleure prise pourrait être alors automatiquement choisie en comparant leurs mesures de synchronie. Enfin, en la couplant à un système de vérification du locuteur fournissant un score de ressemblance avec l'acteur doublé, un score global de doublage pourrait être obtenu, notant à la fois la qualité du doublage et la ressemblance de la voix.

Une dernière application originale consiste à utiliser ces nouvelles mesures de synchronie pour localiser, parmi plusieurs personnes apparaissant à l'écran, celle qui est effectivement en train de parler. Il suffit pour cela de mesurer la synchronie entre la voix entendue et le mouvement des lèvres de chaque personne : la personne dont la mesure de synchronie est la plus élevée est celle qui parle. Ceci ouvre la voie à de

nouvelles applications dans le domaine de l'indexation de séquences audiovisuelles. En la couplant à des techniques de segmentation en locuteurs et de suivi et reconnaissance du visage, il devient possible de constituer automatiquement un modèle d'identité audiovisuelle de chacun des acteurs d'un long métrage et à terme de faciliter l'archivage voire de proposer un mode de navigation intelligent dans les bases de films, axé sur les acteurs présents à l'image.

Annexe A

Technovision IV2

Le projet *Technovision IV2* a pour but de créer des ressources et les conditions d'une évaluation à l'échelle nationale et internationale de différents systèmes liés à l'information du visage, de l'iris et de la voix, dans des milieux semi-contraints. Une base de données biométriques a été constituée dans le but d'évaluer les performances de systèmes d'identification par l'iris, par le visage 2D et 3D et par l'analyse de visages parlants, de systèmes de détection de la position des yeux dans les images 2D et de systèmes de reconstruction 3D du visage.

A.1 Base *Technovision IV2*

Les séquences acquises lors de la campagne d'acquisition des données *visage parlant* ont été enregistrées simultanément à l'aide d'un caméscope DV et d'une webcam. Les personnes devaient lire, face à la caméra, une quinzaine de phrases constituant un corpus phonétiquement équilibré, correspondant à environ une minute de parole en français par session d'enregistrement :

1. `0 - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9`
2. `9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1 - 0`
3. `Oui - Non`
4. `Il se garantira du froid avec un bon capuchon.`
5. `Annie s'ennuie loin de mes parents.`
6. `Les deux camions se sont heurtés de face.`
7. `Un loup s'est jeté immédiatement sur la petite chèvre.`

8. Dès que le tambour bat, les gens accourent.
9. Mon père m'a donné l'autorisation.
10. Vous poussez des cris de colère.
11. Ce petit canard apprend à nager.
12. La voiture s'est arrêtée au feu rouge.
13. La vaisselle propre est mise sur l'évier.
14. Alors que monsieur Gorbatchev regagnait Moscou au terme d'un difficile voyage en Lituanie, une partie du Caucase s'est embrasée.
15. Chaque jour ils reçoivent dans la bonne humeur la visite du commissaire des renseignements généraux qui suit de loin l'opération.

Parmi toutes les personnes ayant participé à la campagne d'acquisition de données *Technovision IV2* et dont nous avons obtenu les données, seules 54 personnes ont enregistré deux sessions ou plus : 51 d'entre elles ont participé à deux sessions exactement et les 3 autres à trois sessions. Les 111 séquences webcam correspondantes ont été extraites et constituent la base de test *Technovision IV2 - Visage parlant*.

A.2 Protocole d'évaluation *Technovision IV2*

Ce petit nombre de personnes multi-sessions et le faible nombre de sessions ne nous permettent pas de définir deux groupes de test disjoints comme c'est le cas pour la base de données BANCA. Un seul ensemble de test composé de la totalité des 111 séquences est ainsi constitué et le protocole d'évaluation *Technovision IV2* est défini comme suit :

Enrôlement Afin de maximiser le nombre de tests *client*, chacune des 111 séquences est utilisée pour constituer un modèle d'identité.

Tests *client* Pour chaque modèle λ, toutes les séquences de la même personne (autres que la séquence utilisée pour constituer le modèle) sont comparées au modèle λ. Au final, le protocole *Technovision IV2* prévoit 2 tests *client* pour les 51 personnes ayant participé à deux sessions et 6 tests *client* pour les 3 personnes ayant participé à trois sessions, soit seulement 120 tests *client* au total.

Tests *imposteur* Pour chaque modèle λ, toutes les séquences des autres personnes (différentes de celle correspondant au modèle) sont comparées au modèle. Au final, le protocole *Technovision IV2* prévoit 12090 tests *imposteur*.

A.3 Évaluation

La figure A.2 récapitule les performances obtenues sur le protocole *Technovision IV2* par le système optimisé sur le protocole P de la base BANCA.

Les performances relativement mauvaises obtenues par le système basé sur la modalité *voix* ($\text{DCF}_{\text{IV2}} = 8.5\%$ vs. $\text{DCF}_{\text{BANCA}} = 5.8\%$) peuvent paraître d'autant plus surprenantes que les séquences *Technovision IV2* sont environ trois fois plus longues que les séquences BANCA. Néanmoins, elles peuvent s'expliquer par le fait que le modèle du monde a été constitué à partir d'enregistrements en langue anglaise alors même que la base *Technovision IV2* est en français. L'adaptation d'un modèle du monde anglais à l'aide de séquences en français entraîne sans doute la création de modèles peu robustes. De même, la modalité *synchronie* est paradoxalement beaucoup moins performante sur la base de données *Technovision IV2* ($\text{DCF}_{\text{IV2}} = 10\%$ vs. $\text{DCF}_{\text{BANCA}} = 8.6\%$). La figure A.1 présente la distribution des scores de synchronie sur les bases BANCA et *Technovision IV2*. Il apparaît clairement que les mauvaises performances obtenues sur le protocole *Technovision IV2* sont dues au fait que les distributions diffèrent largement entre les deux protocoles : un seuil optimisé sur la base BANCA entraîne un rejet systématique de tout accès (*client* ou *imposteur*) du protocole *Technovision IV2*. Les performances obtenues par le système basé sur

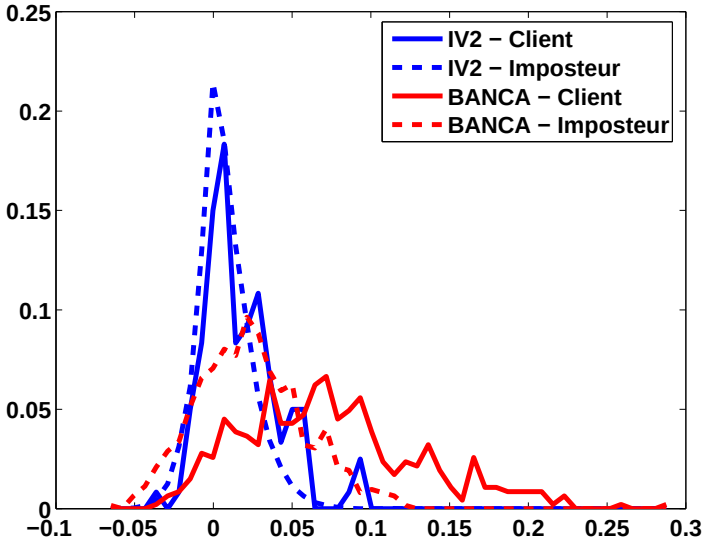
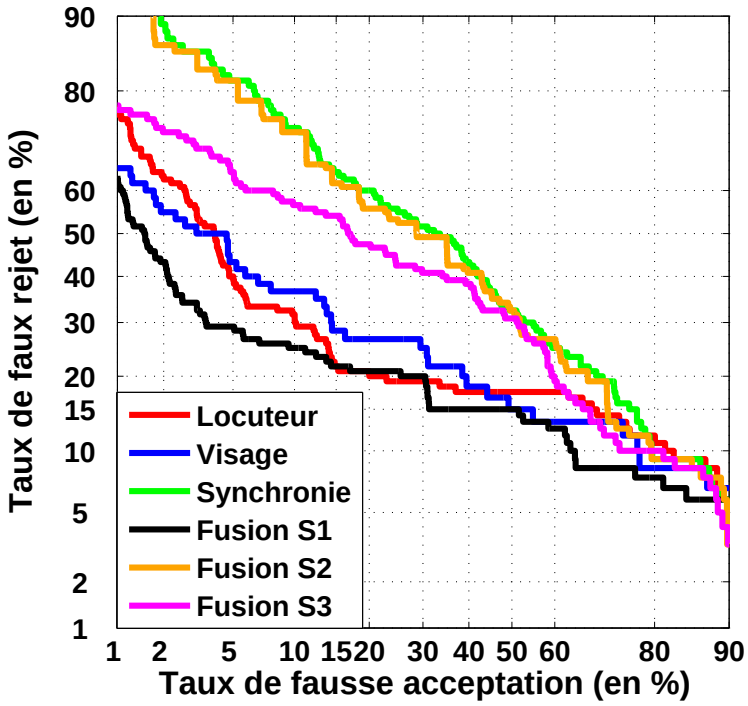


FIG. A.1 – Distribution des scores de synchronie pour les bases BANCA et *Technovision IV2*

la modalité *visage* sont équivalentes sur les deux bases BANCA et *Technovision IV2* ($\text{DCF}_{\text{IV2}} = 7.6\%$ vs.

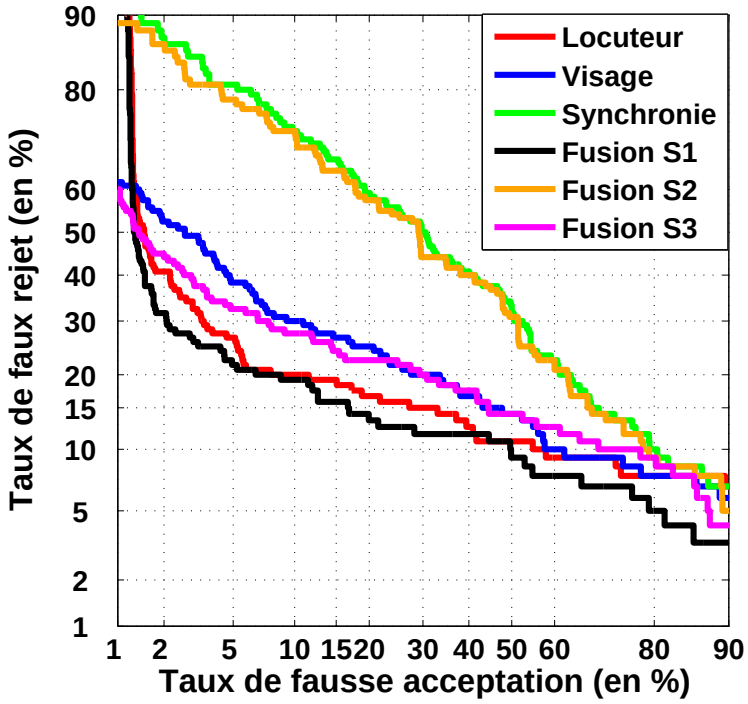
$\text{DCF}_{\text{BANCA}} = 8.0\%$). La modalité *visage* apparaît ici comme la plus stable des trois modalités *voix*, *visage* et *synchronie*. Enfin, les systèmes basés sur les trois stratégies de fusion (somme pondérée, pénalisation et somme pondérée adaptative) pâtissent inévitablement des mauvaises performances des deux modalités *voix* et *synchronie*.

Le faible nombre de clients multi-sessions n'a pas permis de définir d'ensemble de développement sur la base de données *Technovision IV2*. Néanmoins, en utilisant l'ensemble de test comme ensemble de développement, il est possible d'obtenir une mesure des performances optimales que l'on peut espérer atteindre : ceci est résumé dans la figure A.3. Notons en outre que, dans ce cadre, nous avons aussi entraîné un modèle du monde en langue française à partir des données *Technovision IV2* des clients mono-session (qui ne font pas partie de l'ensemble de test). La stratégie de somme pondérée adaptative (la meilleure en termes de compromis entre la performance brute et la robustesse aux attaques) donne des performances équivalentes sur les deux bases BANCA et *Technovision IV2* ($\text{DCF}_{\text{IV2}} = 6.2\%$ vs. $\text{DCF}_{\text{BANCA}} = 6.0\%$).



Locuteur	DCF $= 8.55 \pm 0.94\%$
	FAR $= 3.17 \pm 0.31\%$
	FRR $= 54.17 \pm 8.92\%$
Visage	DCF $= 7.64 \pm 0.90\%$
	FAR $= 1.65 \pm 0.23\%$
	FRR $= 60.00 \pm 8.77\%$
Synchronie	DCF $= 10.00 \pm 0.00\%$
	FAR $= 0.00 \pm 0.00\%$
	FRR $= 100.00 \pm 0.00\%$
Somme pondérée	DCF $= 9.23 \pm 0.90\%$
	FAR $= 6.63 \pm 0.44\%$
	FRR $= 26.67 \pm 7.91\%$
Pénalisation	DCF $= 10.00 \pm 0.00\%$
	FAR $= 0.00 \pm 0.00\%$
	FRR $= 100.00 \pm 0.00\%$
Somme pondérée adaptative	DCF $= 9.36 \pm 0.48\%$
	FAR $= 0.12 \pm 0.06\%$
	FRR $= 92.50 \pm 4.71\%$

FIG. A.2 – Performances sur le protocole *Technovision IV2*, avec apprentissage sur la base BANCA.



Locuteur	DCF $= 5.83 \pm 0.91\%$
	FAR $= 1.76 \pm 0.23\%$
	FRR $= 40.83 \pm 8.79\%$
Visage	DCF $= 7.09 \pm 0.89\%$
	FAR $= 0.93 \pm 0.17\%$
	FRR $= 61.67 \pm 8.70\%$
Synchronie	DCF $= 9.68 \pm 0.32\%$
	FAR $= 0.02 \pm 0.02\%$
	FRR $= 96.67 \pm 3.21\%$
Somme pondérée	DCF $= 4.93 \pm 0.86\%$
	FAR $= 1.78 \pm 0.24\%$
	FRR $= 31.67 \pm 8.32\%$
Pénalisation	DCF $= 9.31 \pm 0.50\%$
	FAR $= 0.15 \pm 0.07\%$
	FRR $= 91.67 \pm 4.95\%$
Somme pondérée adaptative	DCF $= 6.20 \pm 0.92\%$
	FAR $= 1.71 \pm 0.23\%$
	FRR $= 45.00 \pm 8.90\%$

FIG. A.3 – Performances optimales sur le protocole *Technovision IV2*.

Annexe B

Publications

Articles de journaux

- **H. Bredin** et G. Chollet, *Audiovisual Speech Synchrony Measure : Application to Biometrics*, EURASIP Journal on Advances in Signal Processing, 2007 (2007), pp. Article ID 70186, 11 pages. doi :10.1155/2007/70186.
- E. Argones-Rúa, **H. Bredin**, G. Chollet et D. G. Jiménez, *Audio-Visual Speech Asynchrony Detection using Co-Inertia Analysis and Coupled Hidden Markov Models*, submitted to Pattern Analysis and Applications Journal, (2007).

Chapitres d'ouvrages

- B. Abboud, **H. Bredin**, G. Aversano et G. Chollet, ch. *Audio-Visual Identity Verification : an Introductory Overview*, Progress in Nonlinear Speech Processing, no. 4391 in Lecture Notes in Computer Science, Springer, 2007, pp. 118–134.

Conférences internationales

- K. McTait, **H. Bredin**, S. Colon, T. Fillon et G. Chollet, *Adapting a High Quality Audiovisual Database to PDA Quality*, 4th International Symposium on Image and Signal Processing and Analysis (ISPA'05), Zagreb, Croatia, Septembre 2005, pages 262-267.
- **H. Bredin**, A. Miguel, I. H. Witten et G. Chollet, *Detecting Replay Attacks in Audiovisual Identity Verification*, in 31st IEEE International Conference on Acoustics, Speech, and Signal Processing

(ICASSP'06), vol. 1, Toulouse, France, Mai 2006, pp. 621–624 | voir page 156

- **H. Bredin**, N. Dehak et G. Chollet, *GMM-based SVM for Face Recognition*, in 18th International Conference on Pattern Recognition (ICPR'06), Hong-Kong, Août 2006, pp. 1111–1114 | voir page 161
- **H. Bredin** et G. Chollet, *Measuring Audio and Visual Speech Synchrony : Methods and Applications*, in IET International Conference on Visual Information Engineering (VIE'06), Bangalore, Inde, Septembre 2006, pp. 255–260.
- **H. Bredin** et G. Chollet, *Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification*, in 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07), Honolulu, USA, Avril 2007
- R. Landais, **H. Bredin**, L. Zouari et G. Chollet, *Vérification Audiovisuelle de l'Identité*, in Traitement et Analyse de l'Information : Méthodes et Applications (TAIMA'07), Hammammet, Tunisie, Juin 2007, pp. 27–32 | voir page 166
- P. Perrot, **H. Bredin** et G. Chollet, *Biometrics and Forensic Sciences : the Same Quest for Identification ?*, in International Crime Science Conference, London, UK, Juillet 2007 | voir page 182

Conférences nationales

- F. Brugger, L. Zouari, **H. Bredin**, A. Amehraye, G. Chollet, D. Pastor et Y. Ni, *Reconnaissance Audio-Visuelle de la Parole par VMike*, in XXVIèmes Journées d'Étude sur la Parole (JEP'06), Dinard, France, June 2006, pp. 417–420
- E. Argones-Rúa, C. García-Mateo, **H. Bredin** et G. Chollet, *Aliveness Detection using Coupled Hidden Markov Models*, in First Spanish Workshop on Biometrics (SWB'07), Girona, Espagne, Juin 2007 | voir page 173

Workshops

- **H. Bredin**, G. Aversano, C. Mokbel et G. Chollet, *The Biosecure Talking-Face Reference System*, in 2nd Workshop on Multimodal User Authentication (MMUA'06), Toulouse, France, Mai 2006 | voir page 147
- J. Koreman, A. C. Morris, D. Wu, S. Jassim, H. Sellahewa, J.-H. Ehlers, G. Chollet, G. Aversano, **H. Bredin**, S. Garcia-Salicetti, L. Allano, B. L. Van et B. Dorizzi, *Multimodal Biometric Authentication on the SecurePhone PDA*, in 2nd Workshop on Multimodal User Authentication (MMUA'06), Toulouse, France, 2006.

– **H. Bredin** et G. Chollet, *Synchronisation Voix/Lèvres pour la Vérification d'Identité*, Journée Jeunes Chercheurs - Visage/Geste/Mouvement, Paris, France, October 27, 2006.

The BioSecure Talking-Face Reference System

Hervé Bredin [1], Guido Aversano [1], Chafic Mokbel [2] and Gérard Chollet [1]
[1] CNRS-LTCI, GET-ENST (TSI Department), 46 rue Barrault, 75013 Paris, France
[2] University of Balamand, El Koura, BP 100, Tripoli, Lebanon
{bredin, aversano, chollet}@tsi.enst.fr, chafic.mokbel@balamand.edu.lb

Abstract

In the framework of the BioSecure Network of Excellence, a talking-face identity verification reference system was developed: it is open-source and made of replaceable modules. This is an extension of the BECARS speaker verification toolkit, implementing the GMM approach. In this paper, the audio and visual features extraction front-ends are presented. The performance of the system on the Pooled protocol of the BANCA database are described.

1. Introduction

In the framework of identity verification, it has been noticed that it is very difficult (if not impossible) to compare two different methods from two different articles in the literature, even though they deal with the very same task. It poses a real problem when one wants to know if a new original method performs better than the current state of the art, for example. This can be explained by the fact that a lot of research laboratories own their own test database and are the only one performing experiments on it, which are subsequently impossible to reproduce. Reference systems bring an easy yet efficient answer to this problem. Since they are open-source and freely available for everybody, when publishing results on a specific database, experiments using the reference system can be added as a way of calibrating the difficulty of this particular database.

Developing a reference system made of replaceable modules is also of great interest. Researchers often work on a specific part of the system and do not have the time nor the interest in building a complete system from A to Z. A researcher could show the improvement of his new features extraction algorithm simply by replacing the corresponding module in the reference system and without having to bother about the pattern recognition algorithm.

On a pragmatic side, using a reference system as a basis for researching a specific area is also a good way to save time, human resources and money and therefore to facilitate advances of the state of the art.

This reference system addresses the relatively new area of identity verification based on talking-faces. This biometric modality is intrinsically multimodal. Indeed, not only does it contain both voice and face modalities, but it also integrates the combined dynamics of voice and lips motion. Identity verification based on talking-faces is a growing subject of research in the recent literature. In [8], fusion of speech, face and visual speech information for text-dependent identification is presented. In this purpose, the authors use HTK Speech Recognition Toolkit[1] for speech features extraction and Hidden Markov Model (HMM) modelling.

Since our system is designed to perform text-independent identity verification, it uses the Gaussian Mixture Model (GMM) approach for each of the three modalities. GMM for speaker verification is well-known as being very efficient [12]. However, GMM for video-based face recognition is relatively new. It aims at improving robustness of face recognition against light changes, pose variations, etc. In [8], the GMM approach for mouth-based identity verification was concluded to be sufficient (compared to HMM) but not tested.

Therefore, our system mainly consists in an audiovisual front-end extension of the existing open-source BECARS speaker verification GMM toolkit [2]. It also includes a module allowing the detection of basic replay attacks using the synchronisation between voice dynamics and lip motion [3].

Section 2 quickly overviews the BECARS toolkit. The new open-source GET-ENST Online Speech Processing Evaluation Library (GOSPEL) is introduced in section 3. It was developped in the aim of being portable to embedded devices such as PDA and SmartPhones. The face and lips visual front-end is described in section 4. It is made of modules allowing face and mouth tracking, eigenfaces features and lips features extraction. A simple yet efficient algorithm tackling replay attacks is quickly described in section 5. A more detailled description of the algorithm and its per-

[1] http://htk.eng.cam.ac.uk/

formance in simple replay attacks scenarios is available in [3]. In section 6, the question of the fusion of these different modalities is discussed. Sections 7 and 8 report about the experiments and corresponding results performed on the widely available audiovisual BANCA database. Section 9 draws conclusions and presents our plan and perspective to improve the Biosecure Talking-Face Reference System.

2. Speaker Verification Algorithm

Speech is a biometric modality that may be used to verify the identity of a speaker. The speech signal represents the amplitude of an audio waveform as captured by a microphone. To process this signal a feature extraction module calculates relevant feature vectors on a signal window that is shifted at a regular rate. In order to verify the identity of the claimed speaker a stochastic model for the speech generated by the speaker is generally constructed. New utterance feature vectors are generally matched against the claimed speaker model and against a general model of speech that may be uttered by any speaker called the world model. The most likely model identifies if the claimed speaker has uttered the signal or not. In text independent speaker recognition, the model should not reflect a specific speech structure, i.e. a specific sequence of words. Therefore in state-of-the-art systems Gaussian Mixture Models (GMM) are used as stochastic models.

Given a feature vector $\mathbf{x}$, the GMM defines its probability distribution function as follows:

$$\sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^d \|\Gamma_i\|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Gamma_i^{-1}(\mathbf{x}-\mu_i)\right) \quad (1)$$

This distribution can be seen as the realizations of two successive processes. In the first process, the mixture component is selected and based on the selected component the corresponding Gaussian distribution defines the realization of the feature vector. The GMM model is defined by the set of parameters $\lambda = (\{w_i\}, \{\mu_i\}, \{\Gamma_i\})$. To estimate the GMM parameters speech signals are generally collected. The unique observation of the feature vectors provides incomplete data insufficient to allow analytic estimation, following the maximum likelihood criterion, of the model parameters, i.e. the Gaussian distributions weights, mean vectors and covariance matrices. The Estimation Maximization (EM) algorithm offers a solution to the problem of incomplete data [7]. The EM algorithm is an iterative algorithm, an iteration being formed of two phases: the Estimation (E) phase and the Maximization (M) phase. In the E phase the likelihood function of the complete data given the previous iteration model parameters is estimated. In the M phase new values of the model parameters are determined by maximizing the estimated likelihood. The EM algorithm ensures that the likelihood on the training data does not decrease with the iterations and therefore converges towards a local optimum. This local optimum depends on the initial values given to the model parameters before training. Thus, the initialization of the model parameters is a crucial step. The LBG algorithm is used to initialise the model parameters.

The direct estimation of the GMM parameters using the EM algorithm requires a large amount of speech feature vectors. This causes no problem for the world model where several minutes from several speakers may be collected for this purpose. For the speaker model, this would constrain the speaker to talk for large duration and may not be acceptable. To overcome this, speaker adaptation techniques may be used [2]. In the current work, BECARS [2] software has been used for speaker recognition. BECARS implements GMM and includes several adaptation techniques, i.e. Bayesian adaptation, maximum likelihood linear regression (MLLR), and the unified adaptation technique defined in [10]. Using the adaptation techniques few minutes of speech become sufficient to determine the speaker model parameters.

At recognition, feature vectors are extracted from a speech utterance. The log likelihood ratio between the speaker and world models is computed and compared to a threshold. This allows to verify the identity of the claimed speaker.

3. Audio Front-End

One of the goals of our research is the realisation of a real-time talking-face verification interface that can also run on limited resource platforms such as PDAs or SmartPhones. To this purpose, we have developed and validated a new software module for speech parameterization, named "GOSPEL" (GET-ENST Online Speech Processing Embedded Library for prototyping and evaluation).

This library, written in ANSI C, is compatible with the UNIX, Windows and Windows CE platforms. The GOSPEL programming interface has been specially designed for being used within a reference system for scientific evaluation and for being integrated, without extra efforts, into demonstrators and commercial prototypes.

The current version of GOSPEL allows for MFCC feature extraction (including standard options such as delta calculation, cepstral mean subtraction, or liftering) from online or offline audio. It also provides buffering mechanisms suitable for multithreaded applications. A diagram representing typical operations performed by the GOSPEL audio front-end is shown in figure 1.

The "running-CMS" option of GOSPEL makes possible to perform cepstral mean subtraction (CMS) without requiring that a whole utterance is recorded. The online estimation of cepstral mean is obtained by running-average over

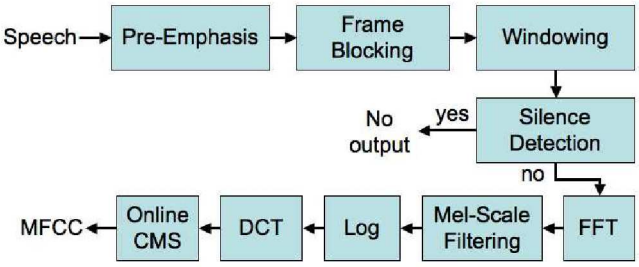


Figure 1. Audio front-end: GOSPEL

the cepstral vector sequence.

An online silence/voice detection function is also provided. Discrimination between speech and silence is based on energy thresholding, with either fixed or exponential adaptive thresholds.

Moreover, GOSPEL supports fixed-point arithmetics: fixed-point optimization, that can be chosen at compile time, is exploited to achieve faster processing throughput on those platforms which do not provide a floating-point unit (like nowadays SmartPhones and PDAs).

The GOSPEL library has been intensively tested and evaluated in speaker verification experiments on the BANCA database. Verification accuracy results, for all the features described above (running-CMS, online silence detection, fixed-point optimisation), are given in section 8.

4. Visual Front-End

The visual front-end is divided into two parts, performing the features extraction for face and lips respectively, as shown in figure 2. First, face is tracked in the video and face features are extracted. Then, for each frame of the video, within the detected face, lips are located and lips features extraction is performed. It was developped using the *Open CV* open-source C/C++ library, freely available over the Internet [2].

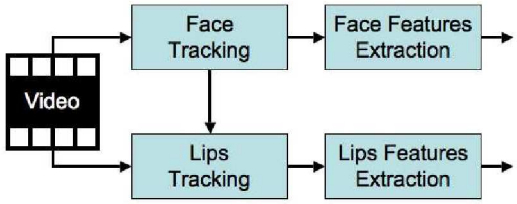


Figure 2. Visual front-end

[2] http://www.intel.com/research/mrl/research/opencv/

4.1 Face module

4.1.1 Face detection

The *OpenCV* library face detector algorithm is first used to get a rough idea of the location of the face: it is an implementation of Viola's algorithm [14], well-known for being very efficient and very fast. The bounding box of the resulting face candidates is then used as a region of interest where to look for a face in the second step of the algorithm. Figure 3 shows the face candidates and the corresponding bounding box on a sample from the BANCA database [1]. Within

Figure 3. Face candidates bounding-box

this region of interest, a moving window scans every possible rectangle at every position with many sizes. For each candidate, the Distance From Face Space (DFFS) [11] is computed and the candidate with the lowest DFFS is chosen as the location of the face. It is defined as the distance between the face candidate and its projection in the eigenface space [13]. Figure 4 summarizes how this distance is computed. A temporal median filter is then applied on the

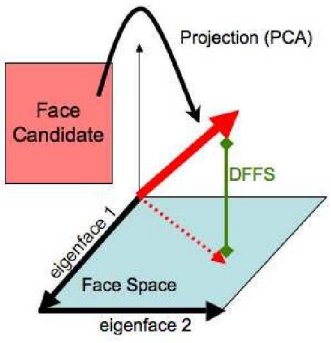


Figure 4. Distance From Face Space

location and size of the detected faces in the video in order to avoid local detection problems. Figure 5 shows the final result of face detection on the same example as before.

These two steps need a preliminary training phase. We used the frontal-face Haar cascade available in *OpenCV* to

find the first face candidates and its corresponding bounding box. The principal component analysis (PCA) needed for the computation of the DFFS is learned based on a training set extracted from the BANCA database (see section 7 for more details) and using the PCA-related functions available in *OpenCV*.

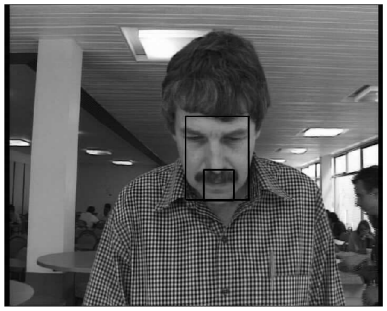


Figure 5. Face and lips tracking

4.1.2 Face tracking

Since this exhaustive search for the best face candidate is very CPU-consuming, a simple tracking algorithm is used: given the location and size of the face in frame $n-1$, the face in frame n is searched in its neighborhood, allowing only a small difference in size. To avoid any divergence in tracking, the algorithm is reinitialized every 20 frames.

4.1.3 Features extraction

Once the face is detected, it is size-normalized to the size of the previously learned eigenfaces. The decomposition of the detected face on the eigenfaces is computed and used as features for face recognition.

4.2 Lips module

4.2.1 Mouth Detection

The very same algorithm as in the first step of face tracking is applied for mouth detection. In each detected face, its lower part is searched for a mouth candidate using the Viola's algorithm. Thus, no real tracking of the lips is performed in this module: it would rather be considered as a *mouth area detector*. Hence, a Haar cascade is learned based on rectangle mouth images extracted from the BANCA database: the effective lips contour tracking is still being investigated since, in our knowledge, no open-source libraries or software for this task is available yet. As one can notice for face detection in figure 3, a lot of false mouth candidates may be detected. Then, a simple algorithm is applied: the biggest detected mouth candidate in the lower part of the face is chosen as the right one. A temporal median filter is then applied in order to avoid local detection problems. Figure 5 shows an example of the output of the mouth detection module.

4.2.2 Features extraction

Once the mouth area is detected, it is size-normalized to 64x64 and a Discrete Cosine Transform (DCT) is applied. Among these 4096 DCT features available, only 50 are kept as lips features. Their selection is performed based on a training set: the ones with the highest energy are chosen. DCT is performed using the *OpenCV* library.

5. Replay Attacks Detection

The talking-face modality is one of the biometrics the most likely to be defeated by replay attacks. As a matter of fact, it is based on the identification of a person using his/her voice and his/her face: two pieces of information which can easily be recorded (which is not that easy for iris or fingerprint, for example). Thus, an imposter could show to the camera a picture of the face of his/her target while playing a recording of the latter's voice previously acquired without any consent nor knowledge of the impersonated person.

This particular scenario (called *Paparazzi*) was proposed in [3], along with the *Big Brother* scenario where the imposter not only owns a picture of the face but a whole video of his/her target. In this paper, a replay attacks detection algorithm is also developed. It is based on a measure of correlation between two streams: one representing the voice dynamics and the other one representing the lip motion. The initial observation that led to this algorithm is presented in figure 6. The upper signal is the energy of speech and the

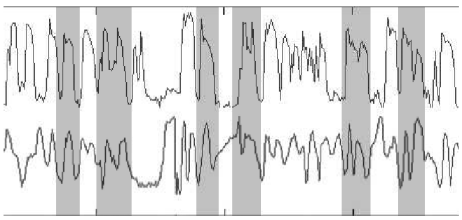


Figure 6. Speech energy vs. Mouth openness

bottom one is the openness of the mouth, both extracted from the same audiovisual sequence. The shadowed parts of the curves emphasize how similar and correlated these two signals can be.

Preliminary results with features as simple as the log-energy of the audio signal and the average value of gray level pixel of the mouth area for the visual signal give en-

couraging results for future improvements (see [3] for more details).

6. Fusion

6.1. Score fusion

Score fusion consists in the combination of the scores of two or more monomodal identity verification algorithms. In [6], this kind of fusion has already been studied using multiple face recognition algorithms on the BANCA database. We used the open-source Support Vector Machine (SVM) library *libSVM* [4] to perform fusion of speaker verification and face recognition scores. More precisely, a Support Vector Classifier with a linear kernel is learned and applied in the 2-dimensional bimodal score space, after a preliminary normalisation step.

6.2. Feature fusion

Feature fusion consists in the combination of two or more monomodal feature vectors into one multimodal features vector to be used as the input of a common multimodal identity verification algorithm.

Audio and visual frame rates are different. Typically, 100 audio feature vectors are extracted per second whereas only 25 video frames are available during the same period. Therefore, one solution is to perform linear interpolation of the visual feature vectors. Another one is to downsample the audio features to reach the video frame rate.

Only simple concatenation of audio and visual feature vectors has been investigated so far. As expected (yet, it still had to be tested), the concatenation-based system is worse than the best monomodal system (see results in section 8). However, more elaborated combination methods still need to be investigated. For example, a transformation such as Principal Component Analysis, Independant Component Analysis or Linear Discriminant Analysis might intrinsically model the correlation between voice dynamics and lips motion. The open-source pattern classification libraries Torch [3] or LNKnet [4] will be used for this purpose.

7. Experiments

7.1 The BANCA database

The BANCA audiovisual database [1] contains 52 speakers divided into 2 groups G1 and G2 of 26 speakers each (13 females and 13 males). 12 sessions were recorded in 3 different conditions (controlled, adverse and degraded). In each session and for each speaker, 2 recordings were performed: one client access where the speaker pronounces digits and his/her (fake) own address and one impostor access where he/she pronounces digits and the address of another person.

[3] http://www.torch.ch/

[4] http://www.ll.mit.edu/SST/lnknet/

7.2 The Pooled BANCA Protocol

The experiments are performed following the Pooled BANCA protocol. In each modality (voice, face, lips and concatenation of voice and lips), for each subject, a GMM is adapted from a world GMM using only the features extracted from the client access of the first controlled session. 312 imposter (one per client per session) and 234 (one per client per session, except the first one used for training) client accesses are performed for each group. The world GMM is learned on the features extracted from the 20 controlled videos of the world model (more than 11000 visual samples).

The face space (used for the face tracking algorithm and the eigenface projection) is built using the manually located face from the world model of the English still images BANCA dataset (300 faces from 30 different subjects).

7.3 Extracted features

In our experiments the BANCA audio (from the first, high-quality, microphone) has been resampled to 16kHz. Speech preprocessing is performed on 20 ms Hamming-windowed frames, with 10 ms overlap. For each frame, 15 MFCC coefficients and their first-order deltas are extracted in the full frequency range, with 20 MEL-scaled triangular filters.

Automatic face tracking is performed on every video of the BANCA database and 80 eigenface coefficients are extracted from each frame (about 400 frames or more per video). Similarly, 50 DCT coefficients are extracted from the mouth area (size-normalized to 64x64), for each frame of each video. Among the 4096 DCT coefficients, the 50 with highest energy (in the world model) are kept as the most significant.

7.4 Score normalisation and fusion

In order to achieve good results during the score fusion process, scores have to be normalised so that scores from different modalities vary in the same predefined range of values.

For that purpose, we used the fact that groups G1 and G2 are two completely distincts sets of subjects: no cross access is performed between them. A linear transformation is learned on scores from G2 to constrain them between -1 and 1 and the SVM classifier is trained on G2. Then, the

same linear transformation is applied on scores from G1, on which the SVM classifier is applied.

8. Results

8.1 Voice

Several "online" speech preprocessing techniques (provided by our audio front-end and described in section 3) have been evaluated. All the speaker verification experiments have been performed following the BANCA P protocol and using our BECARS-based GMM classifier, with 128 gaussians. Speaker models are obtained by MAP adaptation (adapting just the mean of the distributions) from a gender-independent world model (trained on the "controlled" part of BANCA world model data). The GMM training/testing is done only on frames detected as speech, that is frames whose total energy exceeds a given threshold. Unless otherwise stated this threshold is fixed to a very low value (corresponding to an average signal power of about -70 dB compared to full saturation).

Validation of the GOSPEL library Some tests have been conducted to validate our new audio front-end GOSPEL against the previously adopted HTK speech parameterisation module, using the same configuration for both of them. The GOSPEL module produced a small improvement for the equal error rate (+0.6% averaged on both BANCA speaker groups).

Thus, we concluded that the two software modules are equivalent for standard MFCC parameterisation, within statistical errors.

Online CMS The "running-CMS" algorithm, implemented in GOSPEL, as been tested and compared to standard offline CMS. Results show that online cepstral mean estimation does not deteriorate verification performance. On the contrary an increased accuracy is obtained on both BANCA groups (5.8% EER against 6.4% EER on group G1, 7.4% EER against 7.7% EER on group G2). Figure 7 shows DET curves for the G1 case. The grey regions in the plot represent 95% confidence intervals for our tests. The darkest region correspond to a large-sample approximation of the confidence interval (which is optimistic, considering the size of the BANCA database). The ligther grey shading corresponds to the most pessimistic limit (the so-called Chernoff limit [9]), for the confidence interval.

Considering confidence intervals, we can conclude that running-CMS performs as well as standard CMS on BANCA protocol P.

Fixed-point processing Fixed-point voice feature extraction (using an approximated representation of fractional numbers on 16-bit integers) has also been tested. As figure 7 shows, we observe a degradation in terms of verification performance (about -3.5%, averaged on both BANCA groups), compared to the floating-point case. This difference has the same order of magnitude as the confidence intervals. This loss in accuracy corresponds to a considerable gain in terms of processing speed on limited resource devices. We have tested our library on a SmartPhone equipped with an Intel PXA263 processor that does not provide a floating-point unit. On this platform, the optimized part of the algorithm runs about 3.5 times faster than its floating-point correspondent.

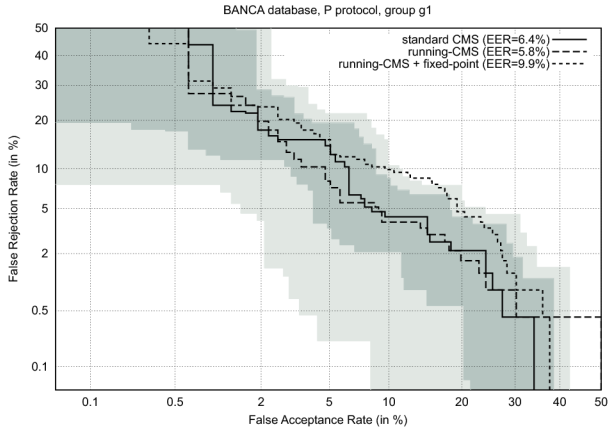


Figure 7. Running-CMS and fixed-point processing

Adaptive silence thresholding The experiments presented in this section compare verification accuracy for both fixed and adaptive silence thresholding. Firstly, a baseline threshold as been estimated on the world model data, by fitting the distribution of frame energy with two gaussians. Then, the energy threshold for silence deletion has been fixed to $\mu_s - 2\sigma_s$, where μ_s and σ_s are the mean and the standard deviation of the rightmost gaussian. This threshold value has been either kept fixed or used as an initialisation for the adaptive thresholding (thresholding is reinitialised for each sentence). Figure 8 shows that, for the BANCA P protocol, adaptive thresholding performs significantly better than a fixed threshold approach, giving 5.4% EER on G1 and 4.8% EER on G2.

8.2 Face and lips

Figures 9 and 10 present the performance of GMM modelling for identity verification based on face and lips re-

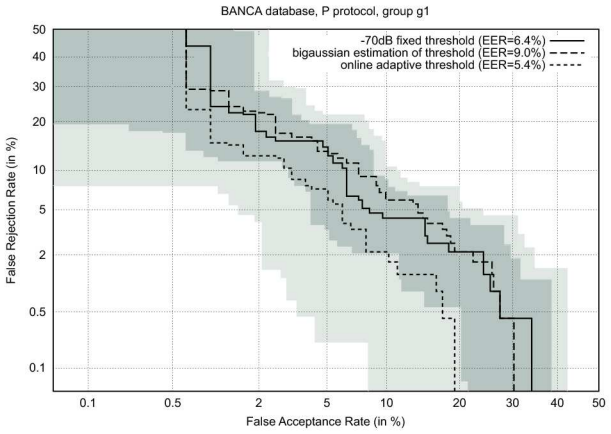


Figure 8. Different silence detection methods

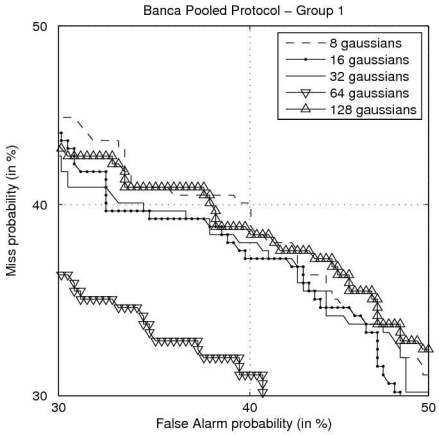


Figure 10. GMM on lips features

spectively. For face recognition, using 32 or 64 gaussians gives the best result: around 28% Equal Error Rate (EER). These relatively poor results can be explained by the simplistic features used to model face: eigenfaces with no normalisation of any kind (rotation of the head, light changes, etc.). Lips-based recognition reaches at best 34% EER for

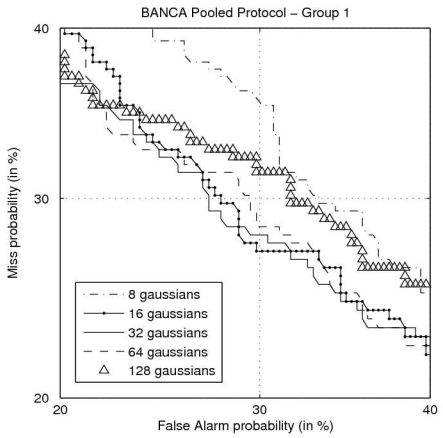


Figure 9. GMM on face features

64 gaussians, whereas all the other systems with less or more gaussians stand around $38 - 39\%$. The same kind of performance was achieved on G2 (not plotted).

8.3 Feature fusion

Figure 11 presents the result of the experiments we performed about feature fusion. Lips feature vectors were linearly interpolated to reach the audio frame rate. Then, a simple concatenation of lips feature vectors and voice feature vectors was performed. Voice only (with 64 gaussians)

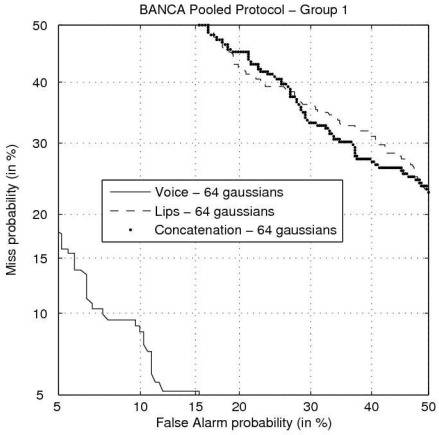


Figure 11. Fusion of voice and lips features

gives an EER of 8.5%, lips only (still with 64 gaussians) gives an EER of 34%. Combining them strongly degrades the performance (compared to voice only) and gives an EER of 32%.

8.4 Score fusion

Following the process described in section 7, we performed score fusion using an SVM classifier with linear kernel: results are presented in figure 12. Voice only (with 256 gaussians) gives an EER of 8.1%, face only (with 256 gaussians) gives an EER of 31%. Performing score fusion brings a non-significant improvement over the voice only systems: 7.7% EER.

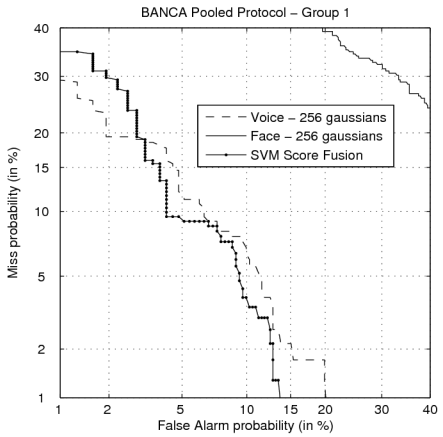


Figure 12. Score fusion with SVM

9. Conclusion and future work

The BioSecure Talking-Face reference system has been introduced in this paper. It is based on the open-source software BECARS initially developed for speaker verification. Our new audio front-end performs "live" feature extraction, including online CMS and silence deletion. Moreover it is implemented both with floating and fixed point operations, which makes it usable on portable devices such as PDA or SmartPhones. The usability of the GMM approach for face- and lips-based recognition was also demonstrated. The reference system makes extensive use of open-source libraries and is freely available on request to the authors. An original way of using the intrinsic bimodal nature of talking-faces has been reported: the detection of a lack of correspondence between the voice and the lips motion is of great help when dealing with simple replay attacks.

In the future, using this reference system as a basis, we plan to improve some of the modules. More precisely, much more efficient face tracking algorithms based on Active Appearance Modelling (AAM) [5] can be investigated. For that purpose, we plan to use the open-source AAM library available on the internet[5]. This might as well help to lead to an improved lips tracking algorithm and consequently the replay attacks detection module. Finally, though eigenface coefficients have been used as a reference in the field of face recognition, better features can be extracted: promising KCFA [15] shall be investigated, for example.

10. Acknowledgments

This work was supported by the EC BioSecure IST-2002-507634 project, the EC SecurePhone IST-2002-506883 project and by the Paris Institute of Technology (ParisTech). The authors would like to thank Santa Rossi for her contribution to lip tracking.

[5] `http://www.imm.dtu.dk/~aam/`

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.

[2] R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez, and G. Chollet. BECARS: a Free Software for Speaker Verification. In *ODYSSEY 2004*, pages 145 – 148, 2004.

[3] H. Bredin, A. Miguel, I. H. Witten, and G. Chollet. Detecting Replay Attacks in Audiovisual Identity Verification. Accepted for ICASSP 2006, May 2006.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 681 – 685. June 2001.

[6] J. Czyk, M. Sadeghi, J. Kittler, and L. Vandendorpe. *Decision Fusion for Face Authentication*, volume 3072/2004, chapter Biometric Authentication: First International Conference, ICBA 2004, pages 686 – 693. Springer-Verlag GmbH, July 2004.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of Royal Statistical Society*, 39(1):1 – 22, 1977.

[8] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly. Robust Automatic Human Identification using Face, Mouth, and Acoustic Information. In *AMFG 2005*, pages 264 – 278, 2005.

[9] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik. What Size Test Set Gives Good Error Rate Estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):52 – 64, January 1998.

[10] C. Mokbel. Online Adaptation of HMMs to Real-Life Conditions: A Unified Framework. In *IEEE Transactions on Speech and Audio Processing*, volume 9, pages 342 – 357. 2001.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent Advances in the Automatic Recognition of Audiovisual Speech. In *IEEE*, volume 91, September 2003.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19 – 41, 2000.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71 – 86, 1991.

[14] P. Viola and M. Jones. Robust Real-Time Object Detection. *Int. Journal of Computer Vision*, 2002.

[15] C. Xie, M. Savvides, and B. V. Kumar. Kernel Correlation Filter Based Redundant Class-Dependence Feature Analysis (KFCA) on FRGC2.0 Data. In *AMFG 2005*, pages 32 – 43, 2005.

DETECTING REPLAY ATTACKS IN AUDIOVISUAL IDENTITY VERIFICATION

Hervé BREDIN [1]*, Antonio MIGUEL* [2]*, Ian H. WITTEN* [3] *and Gérard CHOLLET* [1]

[1] Ecole Nationale Supérieure des Télécommunications, Dept. TSI, Paris, France
[2] Communication Technologies Group (GTC), I3A, University of Zaragoza, Spain
[3] University of Waikato, Dept. Computer Science, Hamilton, New Zealand

ABSTRACT

We describe an algorithm that detects a lack of correspondence between speech and lip motion by detecting and monitoring the degree of synchrony between live audio and visual signals. It is simple, effective, and computationally inexpensive; providing a useful degree of robustness against basic replay attacks and against speech or image forgeries. The method is based on a cross-correlation analysis between two streams of features, one from the audio signal and the other from the image sequence.

We argue that such an algorithm forms an effective first barrier against several kinds of replay attack that would defeat existing verification systems based on standard multimodal fusion techniques. In order to provide an evaluation mechanism for the new technique we have augmented the protocols that accompany the BANCA multimedia corpus by defining new scenarios. We obtain 0% equal-error rate (EER) on the simplest scenario and 35% on a more challenging one [1].

1. INTRODUCTION

Numerous studies have exposed the limits of biometric identity verification based on a single modality (such as fingerprint, iris, hand-written signature, voice, face). Consequently many researchers are exploring whether the coordinated use of two or more modalities can improve performance. The "talking-face" modality, which includes both face recognition and speaker verification, is a natural choice for multimodal biometrics in many practical applications—including face-to-face scenarios, remote video cameras, and even future personal digital assistants.

Talking faces provide richer opportunities for verification than does ordinary multimodal fusion. The signal contains not only voice and image but also a third source of information: the simultaneous dynamics of these features. Natural lip motion and the corresponding speech signal are synchronized. However, most work on audiovisual speech-based biometrics ignores this third information source: it uses the audio and video streams separately and performs fusion at the score level [1] [2]. Nevertheless, some research in speech recognition has shown that it is helpful to take into account the synchronized lip motion, particularly in noisy environments [3] [4].

The aim of this paper is to exploit this novel characteristic of the talking-face modality within the specific framework of identity verification. Section 4 presents a simple method for detecting and quantifying the synchronization between speech and lip motion, based on the correlation between primitive measures of audiovisual activity. The technique can be used to augment an existing audio-visual verification system without excessive computational cost. Doing so thwarts a number of deliberate (so-called "high-effort") attacks that would defeat a standard system.

Many databases are available to the research community to help evaluate multimodal biometric verification algorithms, such as BANCA [5], XM2VTS and BIOMET [6]. Different protocols have been defined for evaluating biometric systems on each of these databases, but they share the assumption that impostor attacks are zero-effort attacks. For example, in the particular framework of the BANCA database, each subject records one client access and one impostor access per session. However, the only difference between the two is the particular message that the client utters—their name and address in the first case; the target's name and address in the second. Thus the "impersonation" takes place without any knowledge of the target's face, age, and voice. These zero-effort impostor attacks are unrealistic—only a fool would attempt to imitate a person without knowing anything about them. In this work we adopt more realistic scenarios in which the impostor has more information about the target.

This article is organized as follows. The next section presents the deliberate (as opposed to "zero-effort") impostor attacks that we have defined. The following section describes the features that our new algorithm uses, while the one after that describes the algorithm itself. Section 5 describes the evaluation methodology, followed by a presentation of performance results for the algorithm. The final section summarizes the results and draws some conclusions.

2. DELIBERATE IMPOSTOR ATTACKS

A major drawback of using the talking-face modality for identity verification is that an impostor can easily obtain a sample of any client's audiovisual identity. Contrast this with iris

[1] This work was initiated in the framework of the First Biosecure Residential Workshop - http://www.biosecure.info

recognition: it is quite difficult to acquire a sample of another person's iris. But numerous small devices allow an impostor to take a picture of the target's face without being noticed, and some mobile phones are even able to record movies. Of course, it is even easier to acquire a recording of the target's voice. Therefore, protocols to evaluate audiovisual identity verification systems should recognize this fact, for example by adding replay attacks to their repertoire of envisaged impostor accesses.

2.1. *Paparazzi* scenario

In this scenario, prior to the attack the impostor takes a still picture of the target's face and acquires an audio recording of their voice. Then, when trying to spoof the system, the impostor simply places the picture in front of the camera and plays the audio recording. The purpose of this scenario is to illustrate the limits of a system that does not take into account the dynamics of lips motion. It has already been tackled in [7].

2.2. *Big Brother* scenario

In this scenario, prior to the attack the impostor records a movie of the target's face, instead of a still picture, and acquires a voice recording as before. However, the audio and video do not come from the same utterance, so they are not synchronised. This is a realistic assumption in situations where the identity verification protocol chooses an utterance for the client to speak. Using the same process as in the *Paparazzi* scenario, the impostor tries to spoof the system by a simple replay attack. In this paper, we address this kind of impostor attack by detecting lack of synchronisation between the audio and video streams.

2.3. Forgery scenarios

More elaborate impostor attacks can include voice and face forgery. Perrot *et al.* [8] use a recording of the target's voice, and automatically transform the impostor's voice so that it resembles the recorded utterance. Abboud and Chollet [9] track the impostor's lip motion throughout a video sequence, and then animate the target's face in a way that moves their lips to match the impostor's. A combination of these two forgeries would be a real threat for a talking-face-based identity verification system.

3. AUDIOVISUAL FEATURES

3.1. Audio features

Let y be the audio signal from a BANCA sequence. Every 10 ms, a 20 ms window is extracted on which the log-energy is computed:

$$e = \log \sum_{n=1}^{N} y(n)^2 \qquad (1)$$

Therefore, 100 samples are extracted per second. Then, a simple voice activity detector based on a bi-gaussian modeling of signal energy distribution is applied: this gives the time stamps allowing to distinguish between silence and voice activity.

3.2. Visual features

For each frame, the lip area is manually located with a rectangle r of size proportional to 20x30 and centered on the mouth (as shown in figure 1) and converted to gray-level. Finally,

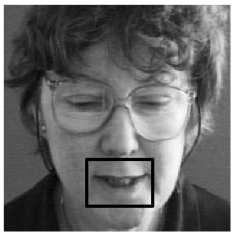


Fig. 1. Manual location of the lips

the mean of the values of the pixels of the lip area (of width W and height H) is computed:

$$m = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} r(i,j) \qquad (2)$$

Audiovisual sequences of the BANCA database are recorded at 25 frames per second. Therefore, 25 samples are extracted per second.

3.3. Different sample rates

As a result of these separate processes of features extraction, audio and visual features are sampled at two different rates. The proposed algorithm deals with audio and video features that must have the same sample rate. Three techniques are proposed to balance the sample rates:

Downsampling the audio signal Every 4 audio samples, only their average value is kept;

Duplicating samples of the visual signal After every sample, 3 identical samples are added;

Linearly interpolating the visual signal Between two samples, 3 linearly interpolated samples are added.

4. AUDIOVISUAL SYNCHRONY MODELLING

4.1. State of the art

Very few previous works on the particular subject of liveness detection based on speech/lips synchronisation were found in the literature. In [7], a Gaussian Mixture Model is learnt on the concatenated audio (MFCC coefficients) and visual (eigenlips projection) features. An Equal Error Rate (EER) of 2% is reached on the equivalent of the *Paparazzi* scenario.

4.2. Preliminary observation

The initial observation that led to a simple model based on correlation between audio and video features is presented in Figure 2. The upper signal is the energy of speech and the bottom one is the openness of the mouth, both extracted from the same audiovisual sequence. The shadowed parts of the curves emphasize how similar and correlated these two signals can be. In our particular case, we chose the mean of

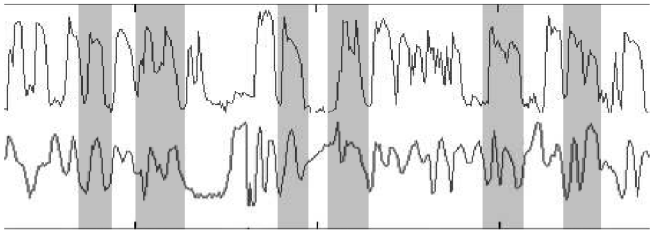


Fig. 2. Speech energy vs. Mouth openness

pixels value instead of the openness because it is easier and faster to compute, supposing that when the mouth is open, pixels are darker and *vice versa*.

4.3. Cross-correlation

Let $A(t)$ and $V(t)$ be two one-dimensional random variables representing respectively the audio and the visual samples. The cross-correlation $X(d)$ ($d \in [-L, L]$) between A and V is defined as follows:

$$X(d) = \mathbf{E}(\tilde{A}(t) \cdot \tilde{V}(t-d)) \tag{3}$$

where $\tilde{S}$ is the centered and variance-normalized version of $S \in \{A, V\}$. In our case, where $A(t)$ and $V(t)$ are only defined for $t \in [1, T]$, we can approximate X by:

$$\hat{X}(d) = \frac{1}{T-d} \sum_{t=1}^{T} \tilde{A}(t) \cdot \tilde{V}(t-d) \tag{4}$$

assuming that $\tilde{V}(t) = 0$ for $t < 1$ and $t > T$.

4.4. Training

$$L_{max}(X) = \text{argmax}_{d \in [-L,L]} |X(d)| \tag{5}$$

is the delay for which the correlation between A and V is maximum. Figures 3 and 4 show how it is computed and what is its distribution on two training sets: synchronised and artificially desynchronised (audio from one sequence, video from another one). Then, L_{sync} is defined as the delay corresponding to the peak in the *synchronised* training set distribution.

4.5. Testing

When testing the synchronisation of a new sequence $AV = \{A, V\}$, the score s of AV is computed as follows:

$$s(AV) = 1 - \frac{|L_{max}(X) - L_{sync}|}{L} \tag{6}$$

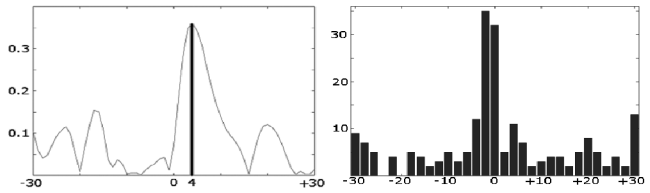


Fig. 3. Example of $L_{max}(X)$ and its distribution on 208 synchronised sequences

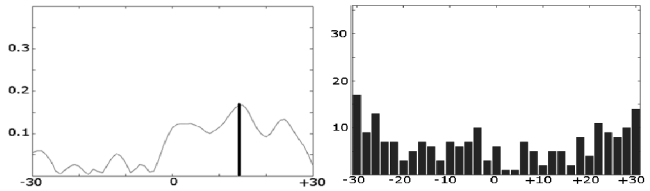


Fig. 4. Example of $L_{max}(X)$ and its distribution on 208 not-synchronised sequences

According to a given threshold $\theta \in [0, 1]$, the sequence AV is decided to be synchronised if $s(AV) \geq \theta$ and not synchronised if $s(AV) < \theta$.

5. EXPERIMENTS

The protocols we used are inspired by the original BANCA Mc protocol [5]. Thus the 52 speakers are divided into two groups (G1 and G2) with 13 females and 13 males in each one. Each speaker recorded four sessions (S1 to S4) during which two accesses were performed (client and impostor). These two groups are completely independent: when G1 is used for training tests are performed on G2, and *vice versa*.
For reasons stated in the introduction, we adapted them to simulate more realistic scenarios. Two new protocols were designed in which training and client access sequences are identical to the original BANCA Mc protocol, but with modified impostor access sequences:

***Paparazzi* protocol** The video is made of only one repeating frame, while the audio is kept unchanged;

***Big Brother* protocol** The video is taken from a different sequence, while the audio is kept unchanged.

6. RESULTS

The system obtained 0% equal-error rate (EER) on the *Paparazzi* scenario, because the visual signal for the impostor was constant and thus completely uncorrelated with the audio signal. In the more challenging *Big Brother* scenario, the system with the best tuned parameters obtained 35% EER.
Figure 5 shows the influence of parameter L (which was introduced in section 4.3). It appears that the best value lies between 20 and 50, which corresponds to a delay of between 1 and 2 seconds.
Using time-stamps of voice activity, silence frames were deleted in the audio and visual signals. Indeed, it has been noticed that when people are taking breath between two utterances,

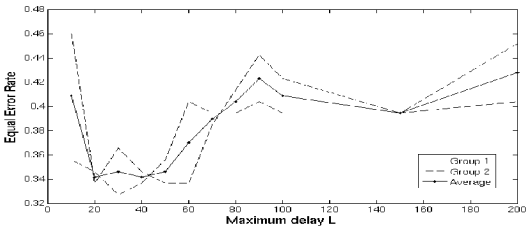


Fig. 5. Influence of maximum delay L on Equal Error Rate

they sometimes open their mouths: this fact is an obvious potential source of error for our system. Figure 6 shows that deleting silence frames gives better performance.

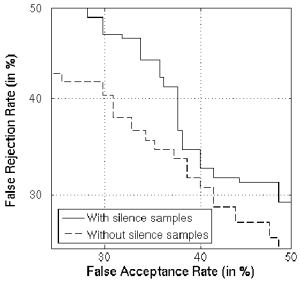


Fig. 6. Influence of silence frames deletion

Figure 7 shows the influence on performance of the method used to balance sample rates. The left curve compares linear interpolation with the duplication of visual samples. It appears that the latter is slightly better, probably because no artificial data is produced. The right curve shows that upsampling the visual samples or downsampling the audio samples does not cause any significant difference.

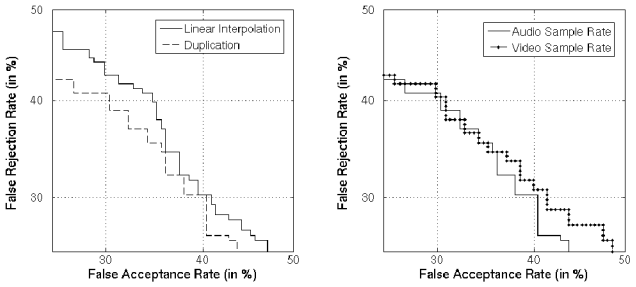


Fig. 7. Influence of sample rate balance

7. CONCLUSION

This paper has argued that account should be taken of the synchronization between the audio and video signals in audiovisual identify verification, in order to defeat sophisticated attacks and forgeries. Since the problem of skilled attacks is not treated by standard evaluation techniques, we have defined new protocols for the BANCA database in order to augment the existing evaluation methodology.

A simple algorithm has been developed to detect and measure synchrony, and tested against two realistic attack scenarios using the BANCA database. An error rate of 0% was reached on the simplest scenario, *Paparazzi*, where a still picture is placed before the camera. Preliminary work using features related to a more accurate shape of the mouth (such as its openness), instead of the simple features we have described, suggest encouraging results. However, robust automatic lip tracking is still needed to further improve the method, and we plan work in this area in order to further improve defences against higher-effort impostor attacks.

8. REFERENCES

[1] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1065 – 1074, September 1999.

[2] A. Jain, L. Hong, and Y. Kulkarni, "A Multimodal Biometric System Using Fingerprint, Face, and Speech," in *Audio- and Video-based Biometric Person Authentication*, 1999.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proc. IEEE*, vol. 91, no. 9, September 2003.

[4] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recogntion," in *EURASIP J. Appl. Signal Processing*, November 2002, vol. 2002, pp. 1260 – 1273.

[5] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Lecture Notes in Computer Science*, January 2003, vol. 2688, pp. 625 – 638.

[6] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J.-L. Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities," *Audio- and Video-Based Biometric Person Authentication*, pp. 845 – 853, June 2003.

[7] G. Chetty and M. Wagner, ""Liveness" Verification in Audio-Video Authentication," in *8th International Conference on Spoken Language Processing*, October 2004.

[8] P. Perrot, G. Aversano, G. Chollet, and M. Charbit, "Voice Forgery Using ALISP: Indexation in a Client Memory," in *ICASSP 2005*, 2005.

[9] B. Abboud and G. Chollet, "Appearance based Lip Tracking and Cloning on Speaking Faces," in *ISPA 2005*, September 2005.

GMM-based SVM for face recognition

Hervé BREDIN, Najim DEHAK and Gérard CHOLLET
CNRS-LTCI, GET-ENST (TSI Department)
46 rue Barrault, 75013 Paris, France
{bredin, dehak, chollet}@tsi.enst.fr

Abstract

A new face recognition algorithm is presented. It supposes that a video sequence of a person is available both at enrollment and test time. During enrollment, a client Gaussian Mixture Model (GMM) is adapted from a world GMM using eigenface features extracted from each frame of the video. Then, a Support Vector Machine (SVM) is used to find a decision border between the client GMM and pseudo-impostors GMMs. At test time, a GMM is adapted from the test video and a decision is taken using the previously learned client SVM. This algorithm brings a 3.5% *Equal Error Rate (EER) improvement over the BioSecure reference system on the Pooled protocol of the BANCA database.*

1. Introduction

The wide majority of face recognition algorithms shares a common framework. At enrollment time, one or a few training pictures of the subject are taken, discriminative features are extracted and saved as the model of the subject. At test time, another small set of pictures (sometimes only one) of the subject is taken and features are extracted following the same process. These features can be geometrical (such as distance between the eyes, length of the nose, etc.) or pixel-based features (such as eigenface coefficients [13], wavelet transform, etc.), or a combination of them (Active Appearance Model [8]). Some transformations are often performed on these features to reduce dimensionality and increase their discriminative power: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (IDA), etc. Then, a comparison is performed between the model and these features in order to decide if the subject is the person he/she pretends to be. This comparison is often performed by computing the distance (euclidean, L1-norm, L2-norm, correlation) between the training and the test pictures. But it can also use classical classification algorithms such as K Nearest Neighbors (KNN) or One-Class Support Vector Machines (OC-SVM). These 1-to-1 or few-to-few comparisons were mostly induced by the protocols defined on the available evaluation databases. Thus, the FERET database only includes still images of face. The BANCA and XM2VTS databases do contain video sequences of talking-faces but evaluation protocols have not used them until now (for example, only 5 pictures per video are used in the BANCA protocols [1]). In [6], Gaussian Mixture Models (GMM) are used to take into account the intra-subject variability (such as face rotation, lips motion or changing light conditions), though the authors only had a few pictures manually annotated and did not use every frame of the videos. Our work starts from the same idea, but using every frame of the video in which the face is automatically located. The face tracking algorithm is quickly described in section 2 along with the features extraction process. Sections 3 and 4 present the core of our algorithm: how it is possible to apply SVM in the GMM space. The experiments we performed and the corresponding results are described in sections 5 and 6.

2. Visual front-end

Any automatic face recognition algorithm needs a preliminary step of face detection. Thus, the face has to be located before it is even possible to recognize it. To allow a fair comparison with previously published results, we used the visual front-end of the Biosecure talking-face reference system which is quickly described in the following paragraphs (a full description is available in [5]).

2.1. Face detection and tracking

The *OpenCV* implementation of [15] is first used to get a rough idea of the location of the face. Then, a moving window scans every possible rectangle in this region of interest at every position with many sizes. For each candidate, the Distance From Face Space (DFFS) [11] is computed and the candidate with the lowest DFFS is chosen as the location of the face. A temporal median filter is then applied on the

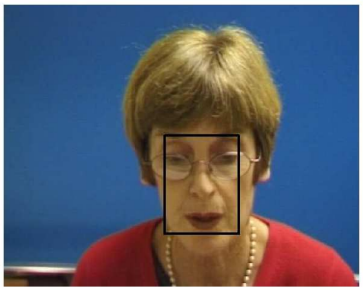


Figure 1. Face tracking

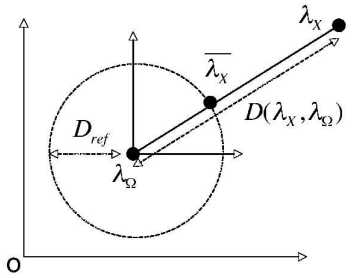


Figure 2. Normalization

location and size of the detected faces throughout the video in order to avoid local detection problems (see figure 1).

2.2. Features extraction

Once the face is detected, it is size-normalized to the size of the eigenfaces. The decomposition of the detected face on the eigenfaces is computed and used as features for face recognition [13].

3. Gaussian Mixture Model

3.1. Principle

The use of GMMs for speaker verification has been studied in depth in the literature [12]. Given a subject X and a corresponding training set $\mathbf{x} = \{x_t, t = 1...N\}$ of D-dimensional features extracted from a video of subject X, a gaussian mixture model $\lambda_X = \{w_i, \mu_i, \Sigma_i, i = 1...M\}$ is learned maximizing the following log-likelihood:

$$\log p(\mathbf{x}|\lambda_X) = \frac{1}{N}\sum_{t=1}^{N} \log p(x_t|\lambda_X) \quad (1)$$

where

$$p(x_t|\lambda_X) = \sum_{i=1}^{M} w_i p_i(x_t|\lambda_X) \quad (2)$$

and

$$p_i(\bullet|\lambda_X) \sim \mathcal{N}(\mu_i, \Sigma_i) \quad (3)$$

Then, at test time, the likelihood $p(\mathbf{y}|\lambda_X)$ that the samples $\mathbf{y}$ come from subject X is computed and compared to a threshold θ: if it is higher than θ then the subject Y is decided to be the subject X.

Training sets for each subject usually contain a relatively small number of samples which may lead to unreliable models. Therefore, using this small training set, subject GMMs are adapted from a world model λ_Ω that was previously trained on a much larger set of samples: we used MAP adaptation in our experiments (see [3] for more details). At test time, the log-ratio of $p(\mathbf{y}|\lambda_X)$ and $p(\mathbf{y}|\lambda_\Omega)$ is compared to a threshold θ. In the following, all models λ are adapted from a common world model λ_Ω.

3.2. Distance between GMMs

In [2] Ben introduces a distance between two GMMs, based on the Kullback-Leibler (KL) divergence. At test time, given a set of samples $\mathbf{y}$, a test model λ_Y is adapted from the world model λ_Ω. In the particular case where neither the weights w_i nor the covariances Σ_i are adapted (gaussians means adaptation only) and using diagonal covariance matrices, the distance is defined as follows:

$$D(\lambda_Y, \lambda_X) = \sqrt{\sum_{i=1}^{M}\sum_{d=1}^{D} w_i^\Omega \frac{\left(\mu_{i,d}^X - \mu_{i,d}^Y\right)^2}{\Sigma_{i,d}^\Omega}} \quad (4)$$

Following the same idea as in the classical GMM case, this distance can be normalized by the distance to the world model and then compared to a threshold:

$$D(\lambda_Y, \lambda_\Omega) - D(\lambda_Y, \lambda_X) > \theta \quad (5)$$

3.3. Normalization

In [3] an additional normalization step is proposed which happens in the adapted GMMs space. Considering the world model λ_Ω as the origin of the space, every adapted GMM λ_X is normalized so that the distance between λ_X and λ_Ω equals a reference distance D_{ref} (see figure 2). In our particular case (MAP with means adaptation only), this normalization can be summarized by the equation 6 (see [3] [2] for more details).

$$\overline{\mu_{i,d}^X} = \frac{D_{ref}}{D(\lambda_X, \lambda_\Omega)}\mu_{i,d}^X + \left(1 - \frac{D_{ref}}{D(\lambda_X, \lambda_\Omega)}\right)\mu_{i,d}^\Omega \quad (6)$$

4. Support Vector Machines

4.1. Principle

SVMs [14] are classifiers used to find the *best* separator between two classes. They are very efficient in solving clustering problems which are not linearly separable. The fundamental idea is to project (using a mapping function ϕ) the input vectors into a new feature space of greater dimension in which it is possible to find a hyperplane producing a linear separation between classes.
In practice, SVMs use kernel functions to perform the computation of scalar products in the feature space without using the definition of ϕ. The *best* hyperplane is chosen in order to maximize the distance between the separating hyperplane and the training vectors the closest to the border: they are called support vectors. The classification of a sample x is given by equation 7

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \tag{7}$$

where k is the kernel function, x_i are the training samples and $y_i \in \{-1, +1\}$ their respective class labels, α_i and b are the parameters of the model obtained after training. At test time, $f(x)$ is compared to a threshold θ.

4.2. SVM in the GMM space

In this paper, we propose to apply SVM directly in the GMM space. The idea is to train a SVM for each client X in order to separate the model λ_X from the rest of the world. In this purpose, pseudo-impostors models λ_{PI_i} are introduced. Typically, they are models adapted from the world model λ_Ω using subsets of the original world model dataset. GMMs are normalized following the equation 6 and the best hyperplane separating the client model λ_X (class +1) from the pseudo-impostors models λ_{PI_i} (class -1) gives the classification function f_X of client X. We used the probabilistic distance kernel given by equation 8, which is a particular case of kernels introduced in [10]:

$$k(\lambda_X, \lambda_Y) = \exp\left(-D^2(\lambda_X, \lambda_Y)\right) \tag{8}$$

where D was previously defined in equation 4. This method was first introduced and successfully applied for speaker verification by Dehak et al. in [9].

5. Experiments

5.1. The BANCA database

The BANCA audiovisual database contains 52 speakers divided into 2 groups G1 and G2 of 26 speakers each (13 females and 13 males). This division into 2 disjoint groups allows to use G2 as the world model when testing on G1 (and reciprocally). In the following, *world model* always refers to the group which is not currently tested. 12 sessions were recorded in 3 different conditions (controlled, adverse and degraded). In each session and for each speaker, 2 recordings were performed: one client access where the speaker pronounces digits and his/her (fake) own address and one impostor access where he/she pronounces digits and the address of another person.

5.2. Protocols

The experiments are performed following the Pooled BANCA protocol. The face space (needed for both the face tracking algorithm and the eigenface projection [5]) is built using all faces from the world model. Automatic face tracking is then performed on every video of the BANCA database and 80 eigenface coefficients are extracted from each frame (about 450 frames or more per video). The world GMM is learned on the features extracted from the videos of the world model. One pseudo-impostor GMM per video of the world model is adapted from the world GMM: this makes about 600 pseudo-impostor GMMs. For each subject, a GMM is adapted from the world GMM using only the features extracted from the client access of one controlled session (4-fold cross validation is achieved by using successively the 4 controlled session for client GMM training). Therefore, 1248 imposter and 936 client accesses are performed for each group: confidence at 95% is less than 3%.

A simple face recognition algorithm based on only 5 pictures randomly extracted from the recordings was also tested as a way of calibrating the difficulty of the BANCA Pooled protocol. It uses the same eigenface features and computes, at test time, the minimum euclidean distance between the five feature vectors of the client model and the five feature vectors of test. Note that, as in the SVM-GMM case, faces are located automatically (the annotation given in BANCA are not used): this might explain the poor performance of this simple algorithm (in comparison to the results obtained in the literature [6]).

5.3. Tools

GMM training, adaptation and scoring are performed using the open-source software *BECARS* [4]. SVM training and scoring are performed using the library *libSVM* [7].

6. Results

Figure 3 shows that the GMM algorithm outperforms the simple *minimum euclidean distance* algorithm. Hence,

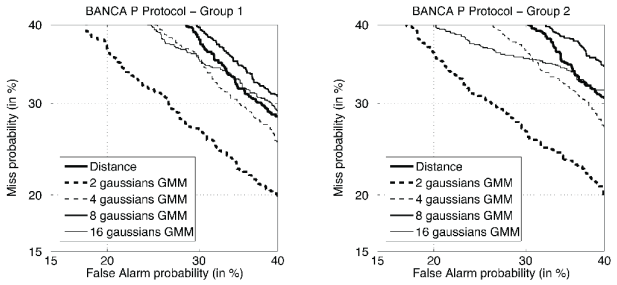


Figure 3. Euclidean distance vs. GMM

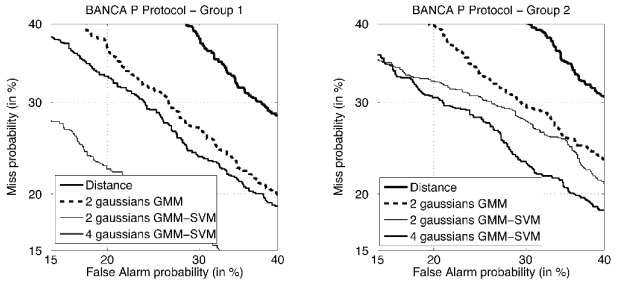


Figure 4. GMM vs. GMM-based SVM

the *distance* algorithm gets a 34% Equal Error Rate (EER) when averaged on groups G1 and G2. The best GMM algorithm leads to a 29% EER, which is obtained when only 2 gaussians per model are used. This might be explained by the fact that the training dataset available for each client is very small and therefore does not allow a good estimation of additional gaussian parameters.

Figure 4 shows the improvements given by the use of GMM-based SVM for face recognition. In average (on groups G1 and G2), it brings a significant 3.5% EER improvement. Hence, the GMM-SVM algorithms leads to a 25.5% EER with 2 gaussians, and 26.8% with 4 gaussians.

7. Conclusions and future work

A fully automatic face recognition system has been presented in this paper. We showed that using every frame of the video in a GMM framework outperforms a simple system based on a distance between features extracted from only a few frames. The main contribution of this paper stays in the use of SVM in the GMM space. It brings another 3.5% equal error rate improvement.

Many points have yet to be improved. For instance, no normalization pre-processing was applied on the automatically tracked face: head rotation normalization and histogram equalization might drastically improve performances. Moreover, we expect to improve the GMM-SVM system by using more than one client model: training a SVM with only one example in one class can lead to inaccurate training.

Finally, the use of face and voice combined in a talking-face modality for identity verification is still a great challenge and we plan to improve the system by adding a speaker verification step, by fusing score and/or audiovisual features.

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.

[2] M. Ben. *Approches Robustes pour la Vérification Automatique du Locuteur par Normalisation et Adaptation Hiérarchique*. PhD thesis, University of Rennes I, 2004.

[3] M. Ben and F. Bimbot. D-MAP: a Distance-Normalized MAP Estimation of Speaker Models for Automatic Speaker Verification. In *IEEE-ICASSP*, volume 2, 2003.

[4] R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez, and G. Chollet. BECARS: a Free Software for Speaker Verification. In *ODYSSEY 2004*, pages 145 – 148, 2004.

[5] H. Bredin, G. Aversano, C. Mokbel, and G. Chollet. The Biosecure Talking-Face Reference System. In *2nd Workshop on Multimodal User Authentication*, May 2006.

[6] F. Cardinaux, C. Sanderson, and S. Bengio. Face Verification Using Adaptative Generative Models. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[8] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 681 – 685. IEEE, June 2001.

[9] N. Dehak and G. Chollet. Support Vector GMMs for Speaker Verification. In *IEEE Odyssey 2006*, 2006.

[10] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *NIPS*, 2003.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent Advances in the Automatic Recognition of Audiovisual Speech. In *IEEE*, volume 91, pages 1306 – 1326, September 2003.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19 – 41, 2000.

[13] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71 – 86, 1991.

[14] V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer Verlag, Berlin, 2000.

[15] P. Viola and M. Jones. Robust Real-Time Object Detection. *Int. Journal of Computer Vision*, 2002.

Vérification audiovisuelle de l'identité

Rémi Landais, Hervé Bredin, Leila Zouari, et Gérard Chollet

École Nationale Supérieure des Télécommunications,
Département Traitement du Signal et des Images, Laboratoire CNRS LTCI
46 rue Barrault, 75634 PARIS Cedex 13 - FRANCE.
Tél : int+ 33 1 45 81 72 63, Fax : int+ 33 1 45 81 37 94
remi.landais@enst.fr

Résumé Un nouveau système de vérification d'identité, tirant parti de la fusion des modalités *visage*, *voix* et *synchronie* est présenté. Chacun des 3 systèmes est décrit : vérification des visages basée sur l'utilisation conjointe d'une représentation globale (par *eigenfaces*) et locale (par des descripteurs SIFT) ; vérification du locuteur par modèles de mélange de gaussiennes (GMM) à l'aide de la boîte à outil BECARS et authentification par analyse de la synchronie entre le mouvement des lèvres et la voix. L'ensemble est intégré par fusion *a posteriori* des scores de chaque système. Le système est évalué sur la base de données biométrique multimédia BANCA.

Mots clés Biométrie, multimodalité, vérification d'identité, fusion, vidéo, BANCA.

1 Introduction

La méthode de vérification d'identité multimodale proposée dans cet article s'appuie sur trois modalités différentes : *visage*, *voix* et *synchronie*. Cette dernière modalité utilise la corrélation entre le signal de parole acoustique et le signal visuel (mouvement des lèvres). L'utilisation de la synchronie permet de faire face à de nouvelles menaces, notamment lorsqu'un imposteur dispose d'un enregistrement vocal de la personne dont il cherche à usurper l'identité.
Nous reviendrons successivement sur chacune de ces modalités. La vérification du *visage* utilise conjointement une représentation globale et une représentation locale des visages. La vérification du locuteur est basée sur les modèles de mélange de gaussiennes (GMM) avec la boîte à outil BECARS [2]. La méthode d'analyse de la synchronie est, quant à elle, basée sur l'estimation de la corrélation *lèvres/voix*. La dernière partie portera sur l'exposé des résultats obtenus sur la base vidéo BANCA [1].

2 Modalité *visage*

La base de données BANCA est constituée de séquences vidéo de personnes énonçant un texte devant une caméra équipée d'un microphone. De cette façon, toutes les trames de la vidéo sont disponibles pour mener à bien la vérification d'identité.

2.1 Méthode de détection de visages

La première étape obligatoire consiste à détecter le visage dans chacune des trames de la vidéo. Elle est ici réalisée à l'aide de la boîte à outil *Machine Perception Toolbox* dont l'algorithme de détection repose sur les modèles génératifs [6]. Sachant que chaque vidéo fait apparaître un unique visage, la majorité des fausses alarmes est supprimée en conservant la plus grande zone détectée. Un filtre temporel médian évite finalement d'obtenir des positions aberrantes et permet de produire une position du visage dans les images où celui-ci n'a pas été initialement détecté.
Le système détecte ensuite la position des yeux en s'appuyant sur une méthode similaire en prenant l'*a priori* que deux yeux exactement doivent être détectés. La position des yeux ainsi obtenue permet de normaliser les visages par alignement spatial et égalisation d'histogramme.
Une phase de filtrage des visages détectés permet de réduire le nombre de résultats erronés pouvant nuire au processus de vérification. L'inverse de la distance entre chaque zone détectée et sa projection dans l'espace de visage défini par les *eigenfaces* (voir la section suivante) est utilisé comme indice (noté Rel) de la qualité de la détection. En effet, une zone détectée représente effectivement un visage si celle-ci est proche (au sens euclidien) de sa projection dans l'espace des visages. L'ensemble des résultats est alors classé selon l'indice Rel et seules les zones dont l'indice est compris entre l'indice maximal Rel_{max} (estimé sur la séquence vidéo) et $\alpha \times Rel_{max}$ (α fixé expérimentalement à 2/3) sont conservées. Par la suite, seuls les 100 meilleurs visages pour l'analyse par *eigenfaces* et les 5 meilleurs pour l'analyse par descripteurs SIFT (un plus grand nombre dégradant les résultats) sont conservés.

2.2 Représentation globale des visages

La méthode des *eigenfaces* [14] constitue un ajustement de l'analyse en composantes principales (PCA) à la reconnaissance de visages. Etant donné un ensemble d'apprentissage de visages de face, la PCA permet d'obtenir un espace dans lequel la dispersion des données est maximisée. Les directions définissant cet espace constituent les *eigenfaces*. Une fois la dimension de l'espace de visage (le nombre d'*eigenfaces*) choisie, chaque nouvelle image peut y être projetée, les coefficients de projection constituant alors sa représentation globale. L'ensemble d'apprentissage utilisé contient 300 visages de la base BANCA (30 personnes) ainsi que 500 environ de la base BIOMET [7] (130 personnes).

2.3 Représentation locale des visages

Les descripteurs SIFT comptent parmi les descripteurs locaux les plus efficaces [11]. Dans un premier temps, les points robustes aux changements d'échelle sont extraits des images en s'appuyant sur leur représentation *scale-space* [9]. L'étape suivante consiste à préciser la position de ces points et à déterminer l'échelle leur étant associée. L'ensemble de points obtenu est filtré selon des contraintes liées au contraste et à la géométrie locale.

Finalement, un vecteur de description (de dimension 128 dans cette étude) est associé à chaque point retenu en analysant l'orientation et la magnitude du gradient dans son voisinage. Par ailleurs, un vecteur à 4 composantes, incluant la position, l'échelle ainsi que l'orientation, est associé à chaque point clé extrait.

2.4 Comparaison des représentations par une méthode de *matching* basée sur la décomposition SVD

La méthode de *matching* basée sur la décomposition en valeurs singulières (SVD) [13] permet d'associer les points d'intérêt extraits de deux images différentes. Elle s'appuie sur l'analyse d'une matrice de proximité : $G_{ij} = g(R_{ij}) = \exp^{-R_{ij}^2/2\sigma^2}$ où R_{ij} définit la distance euclidienne entre les points i et j. Les associations recherchées sont *exclusives* : un point d'une image est associé à un unique point de l'autre image et inversement. Pour faciliter la recherche de ces paires de points, l'idée est de rechercher une projection permettant de *rapprocher* la matrice G de la matrice identité. Cette recherche est facilitée par une procédure dérivée de la solution au *problème orthogonal de Procrustes* [8] :

1. Calculer la SVD de la matrice R : $G = UDV'$
2. Calculer la matrice Q définie par $Q = UV'$
3. Rechercher les paires (i, j) telles que Q_{ij} soit le maximum de la ligne i et de la colonne j.

Une première amélioration est proposée dans [12] : la matrice G prend la forme $G_{ij} = f(C_{ij}) * g(R_{ij})$, où C_{ij} définit la corrélation entre les niveaux de gris des voisinages des points d'intérêt i et j ; et où f peut prendre une forme exponentielle ($f(C_{ij}) = \exp^{-(C_{ij}-1)^2/2\gamma 2}$) ou linéaire $f(C_{ij}) = (C_{ij} + 1)/2$. Un seuil sur la valeur de la corrélation C_{ij} permet de conserver uniquement les meilleurs *matchings*.
Une seconde amélioration consiste à prendre en compte la corrélation entre les descripteurs SIFT associés aux points d'intérêt [4]. Le *matching* de notre méthode basée sur les descripteurs SIFT relève de cette dernière amélioration, tout en considérant pour le calcul des distances R_{ij}, les vecteurs à 4 composantes précédemment cités.
Concernant les représentations *eigenfaces*, la même méthode de *matching* est mise en oeuvre à ceci prêt que la composante *euclidienne* de la matrice G n'est plus prise en compte. À la différence des *matchings* SIFT établis entre différents points d'intérêts, les associations produites par l'application de la méthode de *matching* SVD dans le cas de représentations *eigenfaces* sont établies entre les images des vidéos considérées lors de la vérification (la vidéo *modèle* et la vidéo de *test*).
Cette méthode met en jeu de nombreux paramètres. Les paramètres σ et γ sont fixés selon les recommandations de Pilu [12], validées expérimentalement. Les autres paramètres sont estimés par validation croisée entre les deux groupes composant la base BANCA [10]. Le seuil sur la corrélation C_{ij} est ainsi fixé à 0.4 ; la forme de la fonction f optimale est exponentielle pour la méthode SIFT et linéaire pour la méthode *eigenfaces*. Enfin, la dimension de l'espace de projection pour la méthode *eigenfaces* est fixée à 97.
Quelle que soit la méthode *visage* appliquée, le nombre de *matchings* calculés constitue le

score de vérification. Dans le cas des descripteurs SIFT, les représentations des 5 images retenues dans chaque vidéo sont comparées deux à deux. 25 scores de vérification sont ainsi obtenus pour chaque test. Chacun de ces 25 scores de matching est normalisé relativement au nombre de descripteurs contenus dans chacune des deux images concernées. Le score moyen constitue par la suite le score de vérification final. En ce qui concerne le *matching* des représentations *eigenfaces*, un unique score de *matching* (entre images des deux vidéos) est mesuré et utilisé comme score final de vérification.

3 Modalité *voix*

La vérification du locuteur repose sur l'utilisation des modèles de mélange de gaussiennes (GMM) et est réalisée à l'aide de la boîte à outils *open-source* BECARS [2]. Dans un premier temps, un modèle du monde Ω est construit à partir de paroles prononcées par un grand éventail de locuteurs. Ensuite, un modèle de locuteur λ est estimé par adaptation MAP (Maximum A Posteriori) du modèle du monde à l'aide de données propres à ce locuteur. Cette technique permet de surmonter le manque de données d'apprentissage disponibles pour chaque locuteur. De façon classique, 13 MFCC (*Mel Frequency Cepstral Coefficients*) sont calculés toutes les 10ms sur une fenêtre glissante de 20ms, auxquels sont concaténés les dérivées premières et secondes. Au moment du test, étant donnée l'observation x des MFCC sur un segment de test, le rapport de vraisemblance $P(x|\lambda)/P(x|\Omega)$ fournit le score de la modalité *voix*.

4 Modalité *synchronie*

L'objectif est ici de reconnaître une personne par sa façon de synchroniser sa voix et ses lèvres. La méthode est décrite en détails dans [3]. Étant donnée la séquence d'enrôlement de la personne λ, les signaux acoustique X_λ et visuel Y_λ sont extraits. Il s'agit des coefficients MFCC extraits toutes les 10ms et des coefficients DCT (*Discrete Cosine Transform*) de la zone de la bouche (localisée à l'aide de la méthode détaillée dans [3]). L'analyse de coinertie [5] (CoIA pour *CoInertia Analysis*) permet de calculer les matrices de projection A_λ et B_λ maximisant la covariance des projections des vecteurs X_λ et Y_λ, ces matrices constituant alors le modèle de la personne λ. Au moment du test, une personne ϵ prétend être la personne λ. Les caractéristiques X_ϵ et Y_ϵ associées sont calculées et transformées à l'aide des matrices de projection A_λ et B_λ de la personne λ. Le score de la modalité *synchronie* est alors obtenu par la formule suivante :

$$S_{X_\epsilon,Y_\epsilon} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{cov}(a_{\lambda,k}^T X_\epsilon, b_{\lambda,k}^T Y_\epsilon) \tag{1}$$

où K désigne la dimension de l'espace de projection retenue (c'est à dire le nombre de colonnes $a_{\lambda,k}$ et $b_{\lambda,k}$ conservés dans les matrices A_λ et B_λ).

5 Expérimentations

Les expérimentations ont été menées sur la base BANCA [1]. Elle regroupe 52 personnes divisées en 2 groupes disjoints (G1 et G2). Chacune des personnes a été enregistrée à 8 reprises (4 accès client et 4 accès imposteur) dans 3 conditions d'enregistrement différentes (*controlled*, *degraded* et *adverse*). Le protocole d'évaluation utilisé (le protocole P pour *Pooled*) met en oeuvre des enregistrements produits selon ces trois conditions et constitue ainsi le protocole le plus strict. Pour chaque groupe, 234 accès client et 312 accès imposteurs sont réalisés à l'aide de chacun des 4 systèmes (visage PCA, visage SIFT, voix et synchronie). Les 4 scores sont fusionnés à l'aide d'un SVM (pour *Support Vector Machine*) à noyau RBF : la séparation entre les classes *client* et *imposteur* est apprise sur G1 et testée sur les scores obtenus sur G2 et inversement. L'ensemble des résultats obtenus est résumé dans la figure 1.

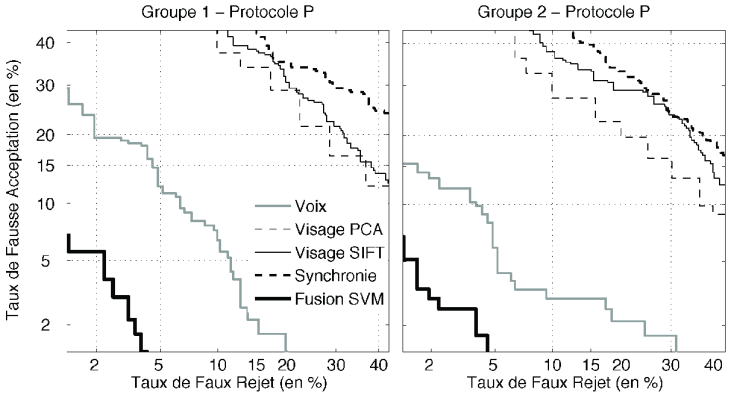


Fig. 1. Résultats des systèmes de vérification sur les données BANCA

Les courbes de la figure 1 montrent l'impact de la fusion sur les performances globales du système : si la modalité *voix* s'avère produire les meilleurs résultats, il n'en demeure pas moins que les performances sont accrues par l'ajout des modalités *visage* et *synchronie*. Conjointement aux courbes DET, nous avons calculé les Taux d'Erreur Pondérée (WER) : $WER(R) = (P_{FR} + RP_{FA})/(1 + R)$ pour plusieurs valeurs de R (0.1, 1 et 10), où P_{FR} désigne le taux de faux rejet et P_{FA} le taux de fausse acceptation. Ces résultats sont résumés dans le tableau 1 dans lequel les intervalles de confiance des mesures obtenues sont précisés. Le non-recouvrement de ces derniers permet de conclure sur le caractère significatif de l'amélioration apportée par la fusion multimodale.

6 Conclusion

Le système de vérification présenté dans cet article s'appuie sur trois modalités : *voix*, *visage* et *synchronie*. Une originalité de ce travail relève de l'utilisation conjointe

Tab. 1. WER obtenus selon les différentes modalités et selon leur fusion.

	R=0.1		R=1		R=10		
Groupe	G1	G2	G1	G2	G1	G2	Moyenne
Voix	2.79	3.65	8.29	5.14	3.79	2.37	4.34 [3.69-5.09]
Eigenfaces	8.25	10.19	23.44	19.80	7.10	6.33	12.52 [11.43-13.70]
SIFT	8.98	8.51	25.76	24.66	7.73	7.69	13.89 [12.75-15.12]
Synchronie	9.73	9.03	27.24	26.10	7.17	7.41	14.45 [12.49-16.66]
Fusion	**0.90**	**2.59**	**3.74**	**2.51**	**2.43**	**0.78**	**2.16 [1.45-3.21]**

de représentations globale et locale pour la vérification des visages. L'utilisation de la modalité *synchronie* constitue la seconde contribution. Les résultats obtenus sur la base BANCA sont satisfaisants puisque ils montrent clairement l'apport de la fusion multimodale relativement à l'utilisation de systèmes mono-modaux. Le travail futur consistera à améliorer chacune des briques constituant le système actuel. Il sera notamment envisagé de prendre en compte un ensemble d'apprentissage plus grand pour la détermination des *eigenfaces* et de définir une méthode de *matching* qui puisse prendre en compte des associations multiples.

Références

1. E. Bailly-Baillière and S. Bengio *et al.* The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.
2. R. Blouet, C. Mokbel, H. Mokbel, E. Sanchez, and G. Chollet. BECARS : a Free Software for Speaker Verification. In *ODYSSEY 2004*, pages 145 – 148, 2004.
3. H. Bredin and G. Chollet. Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification. In *Proc. of the IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 2007.
4. E. Delponte, F. Isgrò, F. Odone, and A. Verri. SVD-Matching using SIFT Features. In *Proc. of the Int. Conf. on Vision, Video and Graphics*, pages 125–132, 2005.
5. S. Dolédec and D. Chessel. Co-Inertia Analysis : an Alternative Method for Studying Species-Environment Relationships. *Freshwater Biology*, 31 :277–294, 1994.
6. I. Fasel, B. Fortenberry, and J.R. Movellan. A Generative Framework for Real-Time Object Detection and Classification. *Computer Vision and Image Understanding*, pages 182–210, 2004.
7. S. Garcia-Salicetti and C. Beumier *et al.* BIOMET : a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities. *Audio- and Video-Based Biometric Person Authentication*, pages 845 – 853, June 2003.
8. G.H. Golub and C.F. Van Loan. *Matrix Computations 3rd Edition.*
9. J. Koenderink. The Structure of Images. *Biological Cybernetics*, 50 :363–370, 1984.
10. R. Landais, H. Bredin, and G. Chollet. Multilevel Face Verification involving SIFT Descriptors and Eigenfaces, 2007. (soumis à IAPR/IEEE International Conference on Biometrics).
11. D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal of Computer Vision*, 60(2) :91–110, 2004.
12. M. Pilu. A Direct Method for Stereo Correspondence based on Singular Value Decomposition. In *Proceedings of CVPR*, pages 261–266, 1997.
13. G.L. Scott and H.C. Longuet-Higgins. An Algorithm for Associating the Features of Two Images. *Proc. of the Royal Society of London. Series B. Biological Sciences.*
14. M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1) :71 – 86, 1991.

Aliveness Detection Using Coupled Hidden Markov Models

Enrique Argones Rúa[1], Carmen García Mateo[1], Hervé Bredin[2], and Gérard Chollet[2] *

[1] UVigo, ST Group, STC Dept., Vigo (Spain)
[2] GET-ENST, Dépt. TSI, Paris (France)

Abstract. A biometric system must verify the identity of a person. Furthermore, it should ensure that the biometric cues have actually been acquired from that person at the moment of the identity verification. The aliveness check ensures that the acquired biometric cue is actually acquired from a live person actually present at the time of capture. This paper compares different techniques to check the aliveness by measuring the synchrony between speech and lip movement in an audio-visual framework. This statistical relationship between speech and lip movement is checked with four different statistical tests based on coupled hidden Markov models.

1 Introduction

It is well known that oral communication between people is intrinsically multimodal. Speech, lip movements and even gestures can help to understand the message. Since gestures usually depend a lot on the person who is talking, the useful information about the message is usually concentrated on the speech itself and on the lip movements. Blind people can obviously understand speech since they can listen to it, and deaf people can understand speech since they can lip-read.

Multimodal biometric systems based on face verification and speaker verification usually make a score level fusion of the face expert and speaker expert outputs. Nonetheless, some of them try to use visual speech information to improve the overall verification performance [1]. One of the major weakness in multimodal biometric systems based on face verification and speaker verification is that they do not take into account realistic impostor attacks scenarios. If a previously recorded segment of speech uttered by the user is used jointly with a photograph of the user's face, even a perfect speaker verifier and a perfect face verifiers fused at the score level would be easily cheated.

Liveness detection based on the synchrony detection of lip movements and speech has been recently proposed in the literature [2]. On the other hand, Coupled Hidden Markov Models (CHMM) have been used for audio-visual speech

* This project has been partially supported by Spanish MEC under the project PRESA TEC2005-07212 and the European Union through the NoE BioSecure and K-Space

recognition [3–5], since they are well suited to model dynamic relationships between several signals. This paper is organized as follows. In section 2 the audio-visual features and further processing necessary to adapt the features to the CHMM framework are presented. In section 3 the CHMM audio-visual modelling is shown. In section 4 four different hypothesis tests to perform the asynchrony detection based on CHMM are presented. The experimental framework is explained in section 5. Results are shown in section 6, and the paper is drawn to conclusion in section 7.

2 Audio-visual Feature Extraction

Any aliveness check based on the link between the lip movement and the speech produced needs at least two information streams. One of them must encode the acoustic information whilst the other must encode the lip movement information.

2.1 Acoustic Features

Mel-Frequency Cepstral Coefficients (MFCC) are classical acoustic speech features in automatic speech processing. They are state-of-the-art features in many applications, including automatic speech recognition and speaker verification. Every 10 ms, a 20 ms long window is extracted from the acoustic signal and 12 MFCCs and the signal energy are computed, in order to get 13-dimensional acoustic speech features. First and second order time-derivatives are then appended. Finally, a 39-dimensional feature vector is extracted every 10 ms.

2.2 Lip Features

The mouth detection algorithm described in [6] was used to locate the lip area, as shown in figure 1. A Discrete Cosine Transform (DCT) is then applied on the gray level size-normalized ROI, and the first 30 DCT coefficients (in a zig-zag manner, corresponding to the low spatial frequency) are kept as the visual speech features. In the same way as acoustic features, first and second order derivatives are appended to the static visual features, and finally 90-dimensional visual features are produced every video frame. Visual features have been linearly interpolated in order to equilibrate visual and acoustic sample rates. After the interpolation, both acoustic and visual features have a sample rate of 100 Hz.

2.3 Coinertia Analysis Transform

The CoInertia Analysis (CoIA) was first introduced by Doledec et al. [7] to solve statistical problems in ecology. CoIA aims at providing a two sets of axes, one for each data stream, on which the projections of the data maximize the covariance of the projections. Given two multivariate random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ of covariance matrix $C_{XY} = \mathrm{E}\{(X - \mu_X)(Y - \mu_Y)^t\}$, where the operator $\mathrm{E}\{\cdot\}$ is the expectation operator, CoIA finds the orthogonal vectors $\{\mathbf{a_1}, \ldots, \mathbf{a_d}\}$

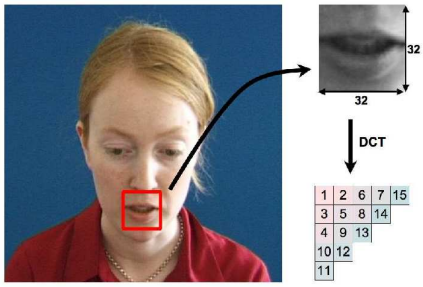


Fig. 1. Appearance-based features extraction

and $\{\mathbf{b_1}, \ldots, \mathbf{b_d}\}$, where d is the rank of C_{XY}, which maximize the covariance between projections of X and Y. These vectors are learnt from a training subset and then applied to all the features. These projected variables with maximum covariance are $\mathcal{X} = \{\mathbf{a_1}^t X, \ldots, \mathbf{a_d}^t X\}$ and $\mathcal{Y} = \{\mathbf{b_1}^t Y, \ldots, \mathbf{b_d}^t Y\}$, sorted by the covariance. Covariance is a compromise between correlation, maximized by the CANonical CORrelation (CANCOR) and variance [2], maximized by the Principal Component Analysis (PCA), and hence between inter-set and intra-set modelization. CoIA, as a compromise between PCA and CANCOR, provides us with a mechanism to reduce the dimension of both visual and acoustic streams while keeping the most covariance as possible just keeping the K first components of the transformed features $\mathcal{X}$ and $\mathcal{Y}$. This is necessary to alleviate the curse of dimensionality in the CHMM training.

3 Dynamic Modelling

A CHMM can be seen as a collection of HMM where the state at time t for every HMM in the collection is conditioned by the states at time $t-1$ of all the HMM in the collection. This is illustrated in figure 2.A CHMM can be completely described by the parameters $\boldsymbol{\lambda} = \{\lambda^i\} = \{\pi^i_{s^i}, a^i_{s^i|\mathbf{r}}, b^i_{s^i}\}$, for every stream $i \in \{1, \ldots, N_h\}$, where N_h is the number of streams; $\mathbf{q_t} = \{q_t^1, \ldots, q_t^{N_h}\}$ is the composite state at time t, where $q_t^i \in \{1, \ldots, NS_i\}$ is the state of stream i and NS_i is the number of possible states for that stream; $\pi^i_{s^i}$ is the initial probability of the state s^i for the stream i; $a^i_{s^i|\mathbf{r}}$ is the state transition probability for the stream i and state s_i from the composite state $\mathbf{r} = \{r^1, \ldots, r^{N_h}\}$; and $b^i_{s^i}$ is the output distribution for stream i and state s^i. The transition probabilities for the stream i are defined as:

$$a^i_{s^i|\mathbf{r}} = \mathrm{P}\left(q_t^i = s^i | q_{t-1}^1 = r^1, \ldots, q_{t-1}^{N_h} = r^{N_h}\right) \tag{1}$$

The output distribution function for every state s^i and stream i is a gaussian mixture model (GMM) with $M^i_{s^i}$ mixtures. Let o_t^i be the observation of the stream i at time t. The output distribution can be written as:

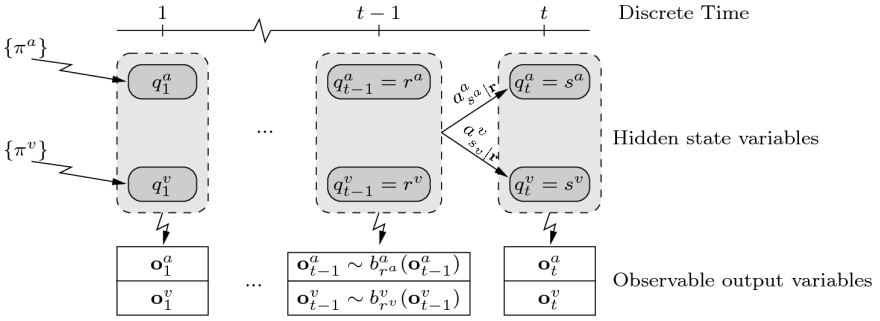


Fig. 2. CHMM state sequence depends for every HMM on the state of all the HMM in the CHMM. Only 2 streams, denoted as a and v, are used.

$$b^i_{s^i}(o^i_t) = \mathrm{P}\left(o^i_t | q^i_t = s^i\right) = \sum_{m=1}^{M^i_{s^i}} w^i_{s^i,m} \mathcal{N}(o^i_t; \mu^i_{s^i,m}, \sigma^i_{s^i,m}) \tag{2}$$

The initial states for the training sequences are obtained using the 5 internal states of an energy-based voice activity detector (VAD), applied to the most informative acoustic and visual features $\mathcal{X}_1$ and $\mathcal{Y}_1$ as described in subsection 2.3. The state transition probabilities $a^i_{s^i|\mathbf{r}}$ are initially estimated from the state transitions obtained by the VAD evolution for all the training sequences: $a^i_{s^i|\mathbf{r}} = n^i_{s^i|\mathbf{r}}/n^i_{\mathbf{r}}$, where $n^i_{s^i|\mathbf{r}}$ is the number of transitions to the state s_i of the stream i from the composite state $\mathbf{r} = \{r^1, \ldots, r^{N_h}\}$, and $n^i_{\mathbf{r}}$ is the total number of times that the CHMM visits the composite state $\mathbf{r}$ before the last sample for every training sequence. The initial state probabilities $\pi^i_{s^i}$ can be estimated as $\pi^i_{s^i} = n^i_{s^i}/ns$, where $n^i_{s^i}$ are the number of training sequences of which first state of the stream i is the state s^i, and ns is the total number of training sequences.

The Baum-Welch algorithm adapted to the CHMM framework is iterated 20 times to train the CHMM. The Viterbi algorithm is used to calculate the sequence of states for every stream and the frame loglikelihoods. This framework has been derived in previous works such as [3].

4 Asynchrony Detection

In order to detect the asynchrony between the acoustic and visual streams X and Y, a hypothesis test can be performed with the following hypothesis:

- $\mathcal{H}_\mathbf{0}$: Both streams are likely produced synchronously, and thereby there is a dependence of the state evolution of one stream with the other one. This hypothesis is represented by the CHMM $\boldsymbol{\lambda}$.
- $\mathcal{H}_\mathbf{1}$: Both streams are produced by independent sources, and hence there is not any dependence between both streams' state sequences. This hypothesis is represented by a two stream HMM as described in [8], namely $\boldsymbol{\lambda'}$.

Four different hypothesis have been derived within this framework:

1. The *first approach* is a slight modification of the classical Bayesian test:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \Longleftrightarrow \frac{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q'|\boldsymbol{\lambda'}\right)} > \theta, \tag{3}$$

   where Q and Q' are the most likely state sequence. These likelihoods are provided by the Viterbi algorithm. This test approximates the classical Bayesian test whether there is a state sequence much more likely than the others.
2. The *second approach* is derived from the previous one:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \Longleftrightarrow \frac{\mathrm{P}\left(Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(Q'|\boldsymbol{\lambda'}\right)} > \theta, \tag{4}$$

   since $\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right) = \mathrm{P}\left(\mathcal{X},\mathcal{Y}|Q,\boldsymbol{\lambda}\right)\mathrm{P}\left(Q,\boldsymbol{\lambda}\right)$. This test eliminates the mismatch due to the differences in the trained output distributions of $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda'}$.
3. The *third approach* performs the test:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \Longleftrightarrow \frac{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(\mathcal{X},\mathcal{Y},Q'|\boldsymbol{\lambda}'_{\boldsymbol{u}}\right)} > \theta, \tag{5}$$

   where $\boldsymbol{\lambda}'_{\boldsymbol{u}}$ is an uncoupled version of $\boldsymbol{\lambda}$.
4. The *fourth approach* is a combination of the second and the third one:

$$\mathcal{H}_{\mathbf{0}} \text{ is accepted } \Longleftrightarrow \frac{\mathrm{P}\left(Q|\boldsymbol{\lambda}\right)}{\mathrm{P}\left(Q'|\boldsymbol{\lambda}'_{\boldsymbol{u}}\right)} > \theta, \tag{6}$$

   where $\boldsymbol{\lambda}'_{\boldsymbol{u}}$ is an uncoupled version of $\boldsymbol{\lambda}$.

The two stream HMM $\boldsymbol{\lambda}'_{\boldsymbol{u}}$ used in the third and fourth approaches shares the parameters $\{\pi^i_{s^i}\}$ and $\{b^i_{s^i}(o^i_t)\}$ with $\boldsymbol{\lambda}$. The state transition vectors $\{a^i_{s^i|r_i}\}$ are generated from the CHMM $\boldsymbol{\lambda}$ parameters abiding the following relation:

$$\begin{aligned} a^i_{s^i|r^i} &= \mathrm{P}\left(q^i_t = s^i | q^i_{t-1} = r^i\right) \\ &= \sum_{\mathbf{q_{t-1}}|q^i_{t-1}=r^i} \mathrm{P}\left(q^i_t = s^i|\mathbf{q_{t-1}} = \mathbf{r}\right) \prod_{j=1,j\neq i}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right) \\ &= \sum_{r_1=1}^{NS_1} \cdots \sum_{r^{i-1}=1}^{NS_{i-1}} \sum_{r^{i+1}=1}^{NS_{i+1}} \cdots \sum_{r^{N_h}=1}^{NS_{N_h}} a^i_{s^i|\mathbf{r}} \prod_{j=1,j\neq i}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right) \end{aligned} \tag{7}$$

The probability $\mathrm{P}\left(q^i_t = r^i\right)$ can be calculated. It depends on the time, but it is not desirable to work with time-dependent state transition probabilities. Therefore, since the quantity $\lim_{t\to\inf}\mathrm{P}\left(q^i_t = r^i\right)$ converges fastly for ergodic models, it is computed following this iterative procedure:

1) Initialization: for $t = 1$,	$\mathrm{P}\left(q^i_1 = s_i\right) = \pi^i_{s^i}$
2) Induction:	$\mathrm{P}\left(q^i_t = s^i\right) = \sum_{\mathbf{r}} a^i_{s^i\|\mathbf{r}} \prod_{j=1}^{N_h} \mathrm{P}\left(q^j_{t-1} = r^j\right)$
3) Stop condition:	$\left\|\frac{\mathrm{P}\left(q^i_t=s^i\right)-\mathrm{P}\left(q^i_{t-1}=s^i\right)}{\mathrm{P}\left(q^i_t=s^i\right)}\right\| < 10^{-6}$

5 Experimental Framework

The experiments conducted in this paper have been performed on the English part of the BANCA Database [9]. A new protocol focused on detecting audio-visual asynchrony has been designed. Asynchronous recordings are artificially built using audio and video from two different recordings from the same subject. Only client accesses recordings, in which true identity is claimed, were used. All the sessions are used in this protocol, including controlled, degraded and adverse condition recordings. Finally, 622 synchronized videos and 6820 desynchronized videos are used for testing purposes. The *World Model* part of the database, a total of 60 video sequences, 20 from each environment in the database, is used to train the models.

The BANCA database is divided into two disjoint groups, namely group 1 and group 2. Performance in one group is calculated using the thresholds that fix the working point in the other group to the equal error rate (EER), the so-called a priori EER threshold. Half Total Error Rate (HTER) and the Detection Error Tradeoff (DET) curves [10] are provided for performance comparison. HTER will be calculated taking into account both group 1 and group 2 using the thresholding approach described previously:

$$HTER = \frac{1}{2}\left(\frac{FA_1 + FA_2}{NI_1 + NI_2} + \frac{FR_1 + FR_2}{NC_1 + NC_2}\right) \tag{8}$$

where FA_i is the number of not synchronized videos classified as synchronized in group i, NI_i is the number of not synchronized videos in group i, FR_i is the number of synchronized videos classified as not synchronized in group i, and NC_i is the number of synchronyzed videos in group i.

6 Experimental Results

Several CHMM configurations have been tried, and finally all the experiments used CHMMs with 8 Gaussians in every GMM output distribution and output dimension 8 for both acoustic and visual streams (parameter K in the CoIA formulation). A number of Gaussians too big leads to poor performance due to the lack of training examples, whilst a too small number of Gaussians in the GMMs leads to a poor modeling, and therefore poor performance. Performance remains similar and very close to the optimal for a number of Gaussians around 8, and hence this value has been used as a reference for performance comparison of the different methods. The curse of dimensionality makes dimension 8 a suitable value: much bigger dimensions make the training fail, whilst too small dimensions do not provide enough information.

Table 1 shows the HTER related to every asynchrony detection method. Besides, the asynchrony detection performances of the methods described in this paper can be visually compared in figure 3. Mismatch between output distributions drives the first approach to the worst performance. Second approach gets better performances, although it is far from the performances of both third

Table 1. HTER for the different hypothesis tests

Hypothesis test	HTER (%)
First approach	37.58
Second approach	27.30
Third approach	13.06
Fourth approach	17.65

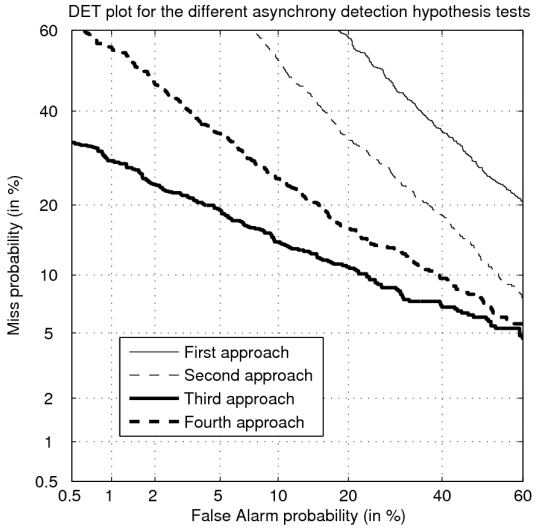


Fig. 3. DET curves for the different hypothesis tests

and fourth approaches. The uncoupling procedure used in third and fourth approaches enhances the synchrony information and gets the best results. Fourth approach is eliminating information that should not be removed, since in that approach the output distributions are the same for both $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'_{\boldsymbol{u}}$. The third and fourth approaches have an additional advantage: they avoid the training of a two stream HMM, since model $\boldsymbol{\lambda}'_{\boldsymbol{u}}$ is built from the already trained CHMM by means of a simple and inexpensive uncoupling procedure.

7 Conclusions

Different asynchrony detection methods have been derived from the CHMM theory and checked in an audio-visual liveness detection task. Results show that the synchrony detection can become an effective anti-spoofing technique. However these asynchrony detection tasks have application wherever the CHMM can be

used in order to determine the degree of coupling between two or more different streams.

The relative structural simplicity of the CHMM proposed and used here is a contrast to the structural complexity of the CHMM used for audio-visual speech recognition, where many different models are trained, one for each phonetic or visual unit. The lack of training sequences did not allow us to train such a family of CHMM, which could result in a much more accurate synchrony detection performance. Future works come up to use these hypothesis contrasts using more complex CHMM structures, where more accurate states are defined more strongly related to the analyzed signal information.

Possible applications of the principles shown here can emerge in different fields not directly related to biometrics, such as automatic video and soundtrack alignment in a movie postproduction or dubbing evaluation.

References

1. Claude C. Chibelushi, Farzin Deravi, and John S.D. Mason. A Review of Speech-Based Bimodal Recognition. *IEEE Trans. Multimedia*, 4(1):23–37, 2002.
2. Hervé Bredin and Gérard Chollet. Measuring Audio and Visual Speech Synchrony: Methods and Applications. In *IET International Conference on Visual Information Engineering 2006 (VIE 2006)*, pages 255 – 260, Bangalore, India, September 2006.
3. X. Liu, L. Liang, Y. Zhaa, X. Pi, and A. V. Nefian. Audio-visual Continuous Speech Recognition using a Coupled Hidden Markov Model. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.
4. Xiaozheng Zhang, Russell M. Mersereau, and Mark Clements. Bimodal Fusion in Audio-Visual Speech Recognition. In *IEEE 2002 International Conference on Image Processing*, volume 1, pages 964–967, September 2002.
5. Ara V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao, and Kevin Murphy. A Coupled HMM for Audio-Visual Speech Recognition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP02)*, May 2002.
6. Hervé Bredin, Guido Aversano, Chafic Mokbel, and Gérard Chollet. The Biosecure Talking-Face Reference System. In *2nd Workshop on Multimodal User Authentication*, May 2006.
7. Sylvain Dolédec and Daniel Chessel. Co-Inertia Analysis: an Alternative Method for Studying Species-Environment Relationships. *Freshwater Biology*, 31:277–294, 1994.
8. Sabri Gurbuz, Zekeriya Tufekci, Tufekci Patterson, and John N. Gowdy. Multi-Stream Product Modal Audio-Visual Integration Strategy for Robust Adaptive Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, 2002.
9. Enrique Bailly-Bailliére et al. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.
10. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *European Conference on Speech Communication and Technology*, pages 1895 – 1898, 1997.

BIOMETRICS AND FORENSIC SCIENCES: THE SAME QUEST FOR IDENTIFICATION?

P. Perrot [(1)(2)], H. Bredin[(2)], G. Chollet [(2)]

[(1)]Institut de Recherche Criminelle de la Gendarmerie Nationale
[(2)]CNRS - LTCI - Ecole Nationale Supérieure des Télécommunications
perrot/bredin/chollet@tsi.enst.fr

Keywords: identification, verification, forensic sciences, biometrics

Abstract

Identification is one of the main challenges of the forensic sciences in order to provide some evidences to investigators or magistrates. The aim of forensic specialists is to determine the identification or not of a suspect and to evaluate the power of the identification evidence. In this field, biometric techniques are particular relevant because they provide a measurement of the identification power of an individual. The link between biometry and forensic sciences is really significant, even if the aim is not completely the same. The Forensic Research Institute of French Gendarmerie is involved in a program of research in the field of speaker recognition based on a free automatic system "Becars", in order to use the voice modality for identification of suspects. The technique and the results obtained in the field of speaker recognition open interesting ways of applications. Nevertheless, the use of modalities in the case of biometry and forensic sciences are answering to different constraints: The case of speaker recognition reveals the common interest of biometry and forensic sciences for identification, but it also reveals the limits in the applications between these both domains. So, trough the example of an automatic speaker recognition approach, this paper presents the interest of using biometric modalities in forensic sciences and the different limits of this use.

1 Introduction

Anthropometry constitutes the same origin of biometric and forensic sciences and reveals the interest of both matters on the question of identification. By identification, we mean identification in a closed set and verification that is to say identification in an open set. Most of biometric systems regard identification and most of forensic applications regard verification. The difference between identification and verification depends on the application. Many biometric systems are linked to security and commercial applications. In the case of security, the use of ID card or password is today insufficient in term of efficiency because of the ease of forgery or more simply forgetting. This is the reason why some modalities like iris, retina, or fingerprints could be a solution. The development of commercial applications causes a necessity of new modalities easier to use like voice, face and so on. The challenge of forensic sciences is to identify a suspect in criminal offences. Terrorism attacks in the USA or UK reveal the need to increase the identification of suspect person from their face for instance. Different modalities are today used in this perspective: DNA, fingerprints, writing…. Some other modalities are under investigation like voice, face, ear shape. Biometric techniques constitute both a solution to identify suspects and also a potential future source of criminality. Speaker recognition is a good example to understand this problematic. So, this paper presents in a first section an overview of biometric systems in order to understand the interest but also the limits of their applications in forensic sciences, then in a second part the example of speaker recognition will be detailed and at last the difficulties and the limits will be presented.

2 Principle

The principle of biometry is to extract some measurable characteristics from an individual and to compare them to a database to establish a correlation between the features and the models. This principle is based on the same scheme for the different modalities. It is divided in two main parts: an enrolment phase, where the system is trained on a reference database and a test phase that consists in deciding if the test sample is a client or an impostor of the system.

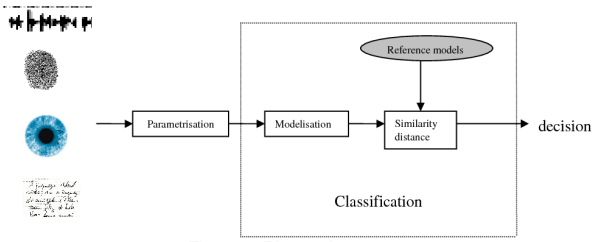


Figure1: Biometric system

Such a system illustrated in Figure 1, can generate two kinds of errors: false rejection or false acceptance. The false rejection consists in considering as impostor, a person who has been enrolled in the training phase. The false acceptance consists in considering as authentic an impostor who has not been enrolled. These both errors are evaluated for each kind of system and presented in a DET curve as illustrated below.

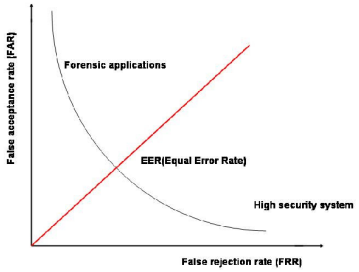


Figure 2: DET curve

According to the application, the chosen threshold decides the level of security of the system. A system will be very secure if it does not accept any impostor (high False Rejection Rate) but the risk is to reject clients. On contrary if the system is very cautious, it accepts all clients but also take the risk to accept impostors (high False Acceptance Rate). This last case is more adapted to forensic applications. International evaluations on multiple biometric modalities will take place in 2007 during the Biosecure workshop. (http://www.biosecure.info/)

3 Voice modality

Biometric modalities could be divided in two main parts: biological and behavioural characteristics. The first one is linked to physiological and anatomical features like retina, fingerprints… and the second one is linked to individual behaviour like the gait for instance. The voice is especially interesting because this is a modality at the bounder between physiological and behavioural characteristics. Automatic speaker recognition is used in many commercial applications but also in the forensic field. Most of the best systems are annually evaluated at the NIST speaker recognition evaluation. From the origin in 1970 till now the classification algorithm has moved from the DTW (Dynamic Time Warping) to VQ (Vector Quantization) and statistical methods (GMM: Gaussian mixture models – SVM: support vector machines). The performance of these systems are very linked to the databases and to the algorithm used. Different parameters influences on the quality of the voice: emotional or health state, environmental noise or channel distortion. Nevertheless in the field of forensic sciences voice appears as a very important modality because of abusive call or terrorist claim. Sometimes the voice is the only element to recognize a person. This is the reason why it is very important to use biometric system but being conscious of the limits and knowing the performance of the system. One of the limits is the capacity of spoofing automatic systems. Different automatic methods exist to convert the voice of a speaker (source voice) in order to imitate another speaker (target voice). The aim is to find the function that minimizes the following expression:

$$\varepsilon = E\left[\left\| y - F(x) \right\|^2\right]$$

where $x=[x_1,x_2, \ldots x_N]$ the source spectral vector
and $y = [y_1,y_2, \ldots y_N]$ the target spectral vector

This figure 3 presents the moving of a free speaker recognition system (BECARS) [1] performance on the DET curve after a conversion based on a spectral transformation [4]. A training corpus of 10 digits pronounced by the source and the target speaker is first aligned by DTW (Dynamic Time Warping). Then 20 MFCC coefficients are extracted from the speech and clustered in 64 classes in two codebooks. The mapping between both codebooks is realized and a conversion matrix for each class is applied on a test sentence pronounced by the source. The conversion matrix M minimizes the square error between two sets of normalised vectors.

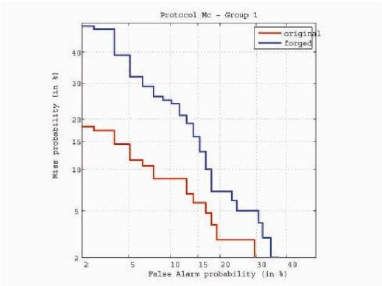


Figure 3: DET for verification test on original and forged voice

The result of the automatic system is largely degraded. This kind of attack is a real threat for automatic speaker recognition system. So, the use of voice in the case of forensic sciences encounters the following difficulties:

- voice quality of the questioned speech
- database quality
- voice disguise
- voice forgery [3]

Nevertheless, voice constitutes an element that participates to the identification of an individual, if the system used is very adapted and the performance known.

3 Limits and difficulties

Fingerprints, voice, DNA, writing, signature are the current modalities used in forensic sciences and ear shape, odor, DNA, face, footprint, gait, keystroke, hand shape will be the next. As presented above in the case of speaker recognition, the use of biometric modalities in forensic sciences must be very cautious. All kind of modalities can be spoofed. This is not very difficult to realize a fingerprint based on a silicon model, to imitate a vein modality by using a vein scanner and so on. A solution to increase the robustness of biometric system applicable in forensic sciences is to fuse different modalities like for instance face and voice, or voice and lip movement [2] to identify a hooded individual. It will be the next challenge for biometric systems but also a constraint in forensic sciences because of the difficulties to collect several modalities. Do not forget that in general in biometry people are volunteers for the collect of modality contrary to forensic sciences.

Conclusion

As a conclusion the aim of this paper is to prepare forensic sciences to the unavoidable use of biometric techniques. The problematic of identification will not be absolutely solved by biometry but this matter constitutes a progress in this way to the condition of knowing the performance of the system used. This performance will have to weight the power of an identification decision.

References

[1] Raphaël Blouet and al. Becars: a free software for speaker verification" Odyssey pages 145-148 - 2004
[2] Hervé Bredin, Gérard Chollet "Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification" IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP Honolulu, USA, 2007
[3] Patrick Perrot, Guido Aversano, Gérard Chollet "Voice forgery using ALISP" – ICASSP 2005
[4] H. Valbret, E. Moulines, J.P. Tubach « Voice conversion using PSOLA technique » - ICASSP 1992

Bibliographie

[Aleksic et Katsaggelos, 2006] Petar S. ALEKSIC et Aggelos K. KATSAGGELOS. Audio-Visual Biometrics. Dans *Proceedings of the IEEE*, volume 94, pages 2025–2044, November 2006.

[André-Obrecht *et al.*, 1997] Régine ANDRÉ-OBRECHT, Bruno JACOB, et Nathalie PARLANGEAU. Audio-Visual Speech Recognition and Segmental Master Slave HMM. Dans *Workshop on Audio-Visual Speech Processing (AVSP'97)*, pages 49–52, September 1997.

[Argones-Rúa *et al.*, 2007a] Enrique ARGONES-RÚA, Hervé BREDIN, Gérard CHOLLET, et Daniel González JIMÉNEZ. Audio-Visual Speech Asynchrony Detection using Co-Inertia Analysis and Coupled Hidden Markov Models. *submitted to Pattern Analysis and Applications Journal*, 2007.

[Argones-Rúa *et al.*, 2007b] Enrique ARGONES-RÚA, Carmen GARCÍA-MATEO, Hervé BREDIN, et Gérard CHOLLET. Aliveness Detection using Coupled Hidden Markov Models. Dans *First Spanish Workshop on Biometrics (SWB'07)*, Girona, Spain, June 2007.

[Arsic *et al.*, 2006] Ivana ARSIC, Roger VILAGUT, et Jean-Philippe THIRAN. Automatic Extraction of Geometric Lip Features with Application to Multi-Modal Speaker Identification. Dans *IEEE International Conference on Multimedia and Expo (ICME'06)*, pages 161–164, 2006.

[AT&T Laboratories Cambridge, 1994] AT&T LABORATORIES CAMBRIDGE. AT&T Database of Faces. 1994.

[Bailly-Baillière *et al.*, 2003] Enrique BAILLY-BAILLIÈRE, Samy BENGIO, Frédéric BIMBOT, Miroslav HAMOUZ, Josef KITTLER, Johnny MARIÉTHOZ, Jiri MATAS, Kieron MESSER, Vlad POPOVICI, Fabienne PORÉE, Belen RUIZ, et Jean-Philippe THIRAN. The BANCA Database and Evaluation Protocol. Dans *4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA'03)*, volume 2688 de *Lecture Notes in Computer Science*, pages 625 – 638, Guildford, UK, January 2003. Springer.

[Barker *et al.*, 1998] Jon BARKER, François BERTHOMMIER, et Jean-Luc SCHWARTZ. Is Primitive AV Coherence an Aid to Segment the Scene ? Dans Denis BURNHAM, Jordi ROBERT-RIBES, et Eric

VATIKIOTIS-BATESON, éditeurs, *Auditory-Visual Speech Processing Workshop (AVSP'98)*, pages 103–108, Sydney, Australia, December 1998.

[Barker et Berthommier, 1999a] Jon P. BARKER et François BERTHOMMIER. Estimation of Speech Acoustics from Visual Speech Features : a Comparison of Linear and Non-Linear Models. Dans *Audio-Visual Speech Processing (AVSP'99)*, pages 112–117, Santa Cruz, USA, August 1999.

[Barker et Berthommier, 1999b] Jon P. BARKER et François BERTHOMMIER. Evidence of Correlation between Acoustic and Visual Features of Speech. Dans *14th International Congress of Phonetic Sciences (ICPhS'99)*, pages 199–202, San Francisco, USA, August 1999.

[Ben, 2004] Mathieu BEN. *Approches Robustes pour la Vérification Automatique du Locuteur par Normalisation et Adaptation Hiérarchique*. PhD thesis, University of Rennes I, 2004.

[Ben et Bimbot, 2003] Mathieu BEN et Frédéric BIMBOT. D-MAP : a Distance-Normalized MAP Estimation of Speaker Models for Automatic Speaker Verification. Dans *28th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 2, pages 69–72, Hong-Kong, April 2003.

[Bengio, 2003] Samy BENGIO. An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition. Dans S. BECKER, S. THRUN, et K. OBERMAYER, éditeurs, *Advances in Neural Information Processing Systems 15*, pages 1213–1220. MIT Press, 2003.

[Bicego *et al.*, 2005] Manuele BICEGO, Enrico GROSSO, et Massimo TISTARELLI. Face Authentication using One-Class Support Vector Machines. Dans Stan Z. LI, Tieniu TAN, Sharath PANKANTI, Gérard CHOLLET, et David ZHANG, éditeurs, *International Workshop on Biometric Recognition Systems*, volume 3781 de *Lecture Notes in Computer Science*, page 15, 2005.

[Bicego *et al.*, 2006] Manuele BICEGO, Enrico GROSSO, et Massimo TISTARELLI. Person Authentication from Video of Faces : a Behavioral and Physiological Approach using Pseudo Hierarchical Hidden Markov Models. Dans *International Conference on Biometrics*, volume 3832 de *Lecture Notes in Computer Science*, pages 113–120, Hong-Kong, January 2006.

[Bimbot *et al.*, 2004] Frédéric BIMBOT, Jean-François BONASTRE, Corinne FREDOUILLE, Guillaume GRAVIER, Ivan MAGRIN-CHAGNOLLEAU, Sylvain MEIGNIER, Teva MERLIN, Javier ORTEGA-GARCIA, Dijana PETROVSKA-DELACRÉTAZ, et Douglas A. REYNOLDS. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 4 :430–451, 2004.

[Blouet *et al.*, 2004] Raphael BLOUET, Chafic MOKBEL, Hoda MOKBEL, Eduardo SANCHEZ, et Gérard CHOLLET. BECARS : a Free Software for Speaker Verification. Dans Javier ORTEGA-GARCIA, Joaquin GONZÀLEZ-RODRIGUEZ, Frédéric BIMBOT, Jean-François BONASTRE, Joseph CAMPBELL, Ivan

MAGRIN-CHAGNOLLEAU, John S.D. MASON, Renana PERES, et Douglas A. REYNOLDS, éditeurs, *ODYSSEY 2004 - The Speaker and Language Recognition Workshop*, pages 145–148, Toledo, Spain, May 2004.

[Bolle et Pankanti, 1998] Ruud BOLLE et Sharath PANKANTI. *Biometrics - Personal Identification in Networked Society*. Kluwer Academic Publishers, 1998.

[Bradski, 1998] Gary R. BRADSKI. Real-Time Face and Object Tracking as a Component of a Perceptual User Interface. Dans *4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, pages 214–219, Princeton, NJ, USA, October 1998.

[Bredin *et al.*, 2006a] Hervé BREDIN, Guido AVERSANO, Chafic MOKBEL, et Gérard CHOLLET. The Biosecure Talking-Face Reference System. Dans *2nd Workshop on Multimodal User Authentication (MMUA'06)*, Toulouse, France, May 2006.

[Bredin et Chollet, 2007] Hervé BREDIN et Gérard CHOLLET. Audio-Visual Speech Synchrony Measure for Talking-Face Identity Verification. Dans *32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Honolulu, USA, April 2007.

[Bredin *et al.*, 2006b] Hervé BREDIN, Najim DEHAK, et Gérard CHOLLET. GMM-based SVM for Face Recognition. Dans *18th International Conference on Pattern Recognition (ICPR'06)*, pages 1111–1114, Hong-Kong, August 2006.

[Bredin *et al.*, 2006c] Hervé BREDIN, Antonio MIGUEL, Ian H. WITTEN, et Gérard CHOLLET. Detecting Replay Attacks in Audiovisual Identity Verification. Dans *31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, volume 1, pages 621–624, Toulouse, France, May 2006.

[Bregler et Konig, 1994] Christoph BREGLER et Yochai KONIG. "Eigenlips" for Robust Speech Recognition. Dans *19th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, volume 2, pages 19–22, Adelaide, Australia, April 1994.

[BT-DAVID, 1996] BT-DAVID. http ://eegalilee.swan.ac.uk/. 1996.

[Castrillón Santana *et al.*, 2005] M. CASTRILLÓN SANTANA, J. LORENZO NAVARRO, O. DÉNIZ SUÁREZ, et A. FALCÓN MARTEL. Multiple Face Detection at Different Resolutions for Perceptual User Interfaces. Dans *2nd Iberian Conference on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005.

[Chetty et Wagner, 2004] Girija CHETTY et Michael WAGNER. "Liveness" Verification in Audio-Video Authentication. Dans *10th Australian International Conference on Speech Science and Technology (SST'04)*, pages 358–363, Sydney, Australia, December 2004.

[Chibelushi *et al.*, 2002] Claude C. CHIBELUSHI, Farzin DERAVI, et John S.D. MASON. A Review of Speech-Based Bimodal Recognition. *IEEE Transactions on Multimedia*, 4(1) :23–37, 2002.

[Chibelushi *et al.*, 1997a] Claude C. CHIBELUSHI, John S.D. MASON, et Farzin DERAVI. Feature-Level Data Fusion for Bimodal Person Recognition. Dans *Sixth International Conference on Image Processing and its Applications*, volume 1, pages 399–403, 1997.

[Chibelushi *et al.*, 1997b] Claude C. CHIBELUSHI, John S.D. MASON, et Farzin DERAVI. Integrated Person Identification Using Voice and Facial Features. Dans *IEE Colloquium on Image Processing for Security Applications*, numéro 4, pages 1–5, London, UK, March 1997.

[Choudhury *et al.*, 1999] Tanzeem CHOUDHURY, Brian CLARKSON, Tony JEBARA, et Alex PENTLAND. Multimodal Person Recognition using Unconstrained Audio and Video. Dans *2nd International Conference on Audio-Video Based Person Authentication*, pages 176–180, Washington, USA, March 1999.

[Chowdhury *et al.*, 2002] A.R. CHOWDHURY, Rama CHELLAPPA, S. KRISHNAMURTHY, et T. VO. 3D Face Reconstruction from Video using a Generic Model. Dans *IEEE International Conference on Multimedia and Expo (ICME'02)*, volume 1, pages 449–452, Lausanne, Switzerland, August 2002.

[Cutler et Davis, 2000] Ross CUTLER et Larry DAVIS. Look Who's Talking : Speaker Detection using Video and Audio Correlation. Dans *IEEE International Conference on Multimedia and Expo (ICME'00)*, volume 3, pages 1589–1592, New-York, USA, July 2000.

[Dean *et al.*, 2005] David DEAN, Patrick LUCEY, Sridha SRIDHARAN, et Tim WARK. Comparing Audio and Visual Information for Speech Processing. Dans *Eighth International Symposium on Signal Processing and its Applications*, volume 1, pages 58–61, August 2005.

[Dehak et Chollet, 2006] Najim DEHAK et Gérard CHOLLET. Support Vector GMMs for Speaker Verification. Dans *IEEE ODYSSEY 2006 - The Speaker and Language Recognition Workshop*, pages 1–4, San Juan, Puerto Rico, June 2006.

[Dempster *et al.*, 1977] Arthur P. DEMPSTER, Nan LAIRD, et Donald B. RUBIN. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[Deng et Huang, 2004] Li DENG et Xuedong HUANG. Challenges in Adopting Speech Recognition. *Communication of the ACM - Special Issue : Multimodal Interfaces that Flex, Adapt, and Persist*, 47 :69–75, 2004.

[Dolédec et Chessel, 1994] Sylvain DOLÉDEC et Daniel CHESSEL. Co-Inertia Analysis : an Alternative Method for Studying Species-Environment Relationships. *Freshwater Biology*, 31 :277–294, 1994.

[Dumas *et al.*, 2005] B. DUMAS, C. PUGIN, J. HENNEBERT, D. PETROVSKA-DELACRÉTAZ, A. HUMM, F. EVÉQUOZ, R. INGOLD, et D. Von ROTZ. MyIdea - Multimodal Biometrics Database, Description of Acquisition Protocols. Dans *Third COST 275 Workshop (COST 275)*, pages 59–62, Hatfield, UK, October 2005.

[Eveno et Besacier, 2005a] Nicolas EVENO et Laurent BESACIER. A Speaker Independent Liveness Test for Audio-Video Biometrics. Dans *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 3081–3084, Lisboa, Portugal, September 2005.

[Eveno et Besacier, 2005b] Nicolas EVENO et Laurent BESACIER. Co-Inertia Analysis for "Liveness" Test in Audio-Visual Biometrics. Dans *4th International Symposium on Image and Signal Processing and Analysis (ISISPA'05)*, pages 257–261, Zagreb, Croatia, September 2005.

[Fairhurst *et al.*, 2004] Michael C. FAIRHURST, Farzin DERAVI, et J. GEORGE. Towards Optimised Implementations of Multimodal Biometric Configurations. Dans *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety (CIHSPS'04)*, pages 113–116, July 2004.

[Fasel *et al.*, 2004] Ian FASEL, Bret FORTENBERRY, et J. R. MOVELLAN. A Generative Framework for Real-Time Object Detection and Classification. *Computer Vision and Image Understanding - Special Issue on Eye Detection and Tracking*, 98(1) :182–210, 2004.

[Fisher *et al.*, 2001] John W. FISHER, Trevor DARRELL, William T. FREEMAN, et Paul VIOLA. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. Dans T. K. LEEN, T. G. DIETTERICH, et V. TRESP, éditeurs, *Advances in Neural Information Processing Systems 13*, pages 772–778. MIT Press, 2001.

[Fox et Reilly, 2003] Niall FOX et Richard B. REILLY. Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features. Dans *4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA'03)*, volume 2688 de *Lecture Notes in Computer Science*, pages 743–751, Guildford, UK, January 2003. Springer.

[Fox *et al.*, 2007] Niall A. FOX, Ralph GROSS, Jeffrey F. COHN, et Richard B. REILLY. Robust Biometric Person Identification using Automatic Classifier Fusion of Speech, Mouth and Face Experts. *IEEE Transactions on Multimedia*, 9(4) :701–714, June 2007.

[Furui, 1997] Sadaoki FURUI. Recent Advances in Speaker Recognition. Dans *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'97)*, pages 237–251, Crans-Montana, Switzerland, March 1997.

[Garcia-Salicetti *et al.*, 2003] S. GARCIA-SALICETTI, C. BEUMIER, G. CHOLLET, B. DORIZZI, J.-L. JARDINS, J. LUNTER, Y. NI, et D. PETROVSKA-DELACRETAZ. BIOMET : a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities. Dans *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, pages 845 – 853, Guildford, UK, June 2003.

[Gauvain et Lamel, 2000] Jean-Luc GAUVAIN et Lori LAMEL. Large-Vocabulary Continuous Speech Recognition : Advances and Applications. Dans *Proceedings of the IEEE*, volume 88, pages 1181–1200, 2000.

[Georgia Institute of Technology, 1999] GEORGIA INSTITUTE OF TECHNOLOGY. Georgia Tech Face Database - http ://www.anefian.com/face_reco.htm. 1999.

[Goecke et Millar, 2003] Roland GOECKE et Bruce MILLAR. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. Dans *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP'03)*, pages 133–138, Saint-Jorioz, France, September 2003.

[Guyon *et al.*, 1998] Isabelle GUYON, John MAKHOUL, Richard SCHWARTZ, et Vladimir VAPNIK. What Size Test Set Gives Good Error Rate Estimates ? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1) :52–64, January 1998.

[Hershey et Movellan, 1999] John HERSHEY et Javier MOVELLAN. Audio-Vision : Using Audio-Visual Synchrony to Locate Sounds. Dans Michael S. KEARNS, Sara A. SOLLA, et David A. COHN, éditeurs, *Advances in Neural Information Processing Systems 11*, pages 813–819. MIT Press, 1999.

[Hyvärinen, 1999] Aapo HYVÄRINEN. Survey on Independent Component Analysis. *Neural Computing Surveys*, 2 :94–128, 1999.

[Iyengar *et al.*, 2003] G. IYENGAR, H.J. NOCK, et Chalapathy NETI. Audio-Visual Synchrony for Detection of Monologues in Video Archives. Dans *IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 1, pages 329–332, Baltimore, USA, July 2003.

[Jain *et al.*, 1999] Anil JAIN, Lin HONG, et Yatin KULKARNI. A Multimodal Biometric System Using Fingerprint, Face, and Speech. Dans *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington, USA, March 1999.

[Jain *et al.*, 2005] Anil JAIN, Karthik NANDAKUMAR, et Arun A. ROSS. Score Normalization in Multimodal Biometric Systems. *Pattern Recognition*, 38(12) :2270–2285, 2005.

[Jain et Ross, 2002] Anil K. JAIN et Arun A. ROSS. Learning User-Specific Parameters in a Multibiometric System. Dans *9th IEEE International Conference on Image Processing (ICIP'02)*, volume 1, pages 57–60, New-York, USA, September 2002.

[Jee *et al.*, 2006] Hyung-Keun JEE, Sung-Uk JUNG, et Jang-Hee YOO. Liveness Detection for Embedded Face Recognition System. *International Journal of Biomedical Sciences*, 1(4) :235–238, 2006.

[Jourlin *et al.*, 1997] Pierre JOURLIN, Juergen LUETTIN, Dominique GENOUD, et Hubert WASSNER. Acoustic-Labial Speaker Verification. Dans *First International Conference on Audio- and Video-based Biometric Person Authentication*, volume 18, pages 853–858, Crans-Montana, Switzerland, 1997.

[Kollreider *et al.*, 2005] K. KOLLREIDER, H. FRONTHALER, et Josef BIGUN. Evaluating Liveness by Face Images and the Structure Tensor. Dans *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, pages 75–80, 2005.

[Krueger et Zhou, 2002] Volker KRUEGER et Shaohua ZHOU. Exemplar-based Face Recognition from Video. Dans *7th European Conference on Computer Vision*, volume 4, page 732, Copenhagen, Denmark, May 2002.

[Landais *et al.*, 2007] Rémi LANDAIS, Hervé BREDIN, Leila ZOUARI, et Gérard CHOLLET. Vérification Audiovisuelle de l'Identité. Dans *Proceedings of Traitement et Analyse de l'Information : Méthodes et Applications*, pages 27–32, Hammamet, Tunisia, June 2007.

[Li et Jain, 2005] Stan Z. LI et Anil K. JAIN. *Handbook of Face Recognition*. Springer, 2005.

[Lucey *et al.*, 2005] Simon LUCEY, Tsuhan CHEN, Sridha SRIDHARAN, et Vinod CHANDRAN. Integration Strategies for Audio-Visual Speech Processing : Applied to Text-Dependent Speaker Recognition. *IEEE Transactions on Multimedia*, 7(3) :495–506, June 2005.

[Mahalanobis, 1936] Prasanta Chandra MAHALANOBIS. On the Generalised Distance in Statistics. Dans *Proceedings of the National Institute of Science of India 12*, pages 49–55, 1936.

[Martin *et al.*, 1997] Alvin F. MARTIN, George R. DODDINGTON, T. KAMM, M. ORDOWSKI, et M. PRZYBOCKI. The DET Curve in Assessment of Detection Task Performance. Dans *European Conference on Speech Communication and Technology (Interspeech'1997 - Eurospeech)*, volume 4, pages 1895–1898, Rhodes, Greece, 1997.

[Martin et Przybocki, 2000] Alvin F. MARTIN et Mark A. PRZYBOCKI. The NIST Speaker Recognition Evaluation - an Overview. *Digital Signal Processing*, 10 :1–18, 2000.

[Matthews et Baker, 2004] Iain MATTHEWS et S. BAKER. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2) :135–164, 2004.

[Messer *et al.*, 2004] Kieron MESSER, Josef KITTLER, Mohammad SADEGHI, Miroslav HAMOUZ, Alexey KOSTIN, Fabien CARDINAUX, Sébastien MARCEL, Samy BENGIO, Conrad SANDERSON, Norman POH, Yann RODRIGUEZ, Jacek CZYK, Luc VANDENDORPE, Chris MCCOOL, Scott LOWTHER, Sridha SRIDHARAN, Vinod CHANDRAN, Roberto Parades PALACIOS, Enrique VIDAL, Li BAI, LinLin SHEN, Yan WANG, Chiang YUEH-HSUAN, Liu HSIEN-CHANG, Hung YI-PING, Alexander HEINRICHS, Marco MUELLER, Andreas TEWES, Christoph von der MALSBURG, Rolf WURTZ, Zhenger WANG, Feng XUE, Yong MA, Qiong YANG, Chi FANG, Xiaoqing DING, Simon LUCEY, Ralph GOSS, et Henry SCHNEIDERMAN. Face Authentication Test on the BANCA Database. Dans *17th International Conference on Pattern Recognition (ICPR'04)*, volume 4, pages 523–532, Cambridge, UK, August 2004.

[Messer *et al.*, 1999] Kieron MESSER, Jiri MATAS, Josef KITTLER, Juergen LUETTIN, et G. MAITRE. XM2VTSDB : The Extended M2VTS Database. Dans *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, pages 72–77, Washington, USA, March 1999.

[Morris *et al.*, 2006] Andrew C. MORRIS, Jacques KOREMAN, Harin SELLAHEWA, Johan-Hendrik EHLERS, Sabah JASSIM, Lorene ALLANO, et Sonia GARCIA-SALICETTI. The SecurePhone PDA Database, Experimental Protocol and Automatic Test Procedure for Multi-Modal User Authentication. Rapport Technique, Saarland University, Institute of Phonetics, 2006.

[Nefian et Liang, 2003] Ara V. NEFIAN et Lu Hong LIANG. Bayesian Networks in Multimodal Speech Recognition and Speaker Identification. Dans *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 2004–2008, 2003.

[Nock *et al.*, 2002] H. J. NOCK, G. IYENGAR, et Chalapathy NETI. Assessing Face and Speech Consistency for Monologue Detection in Video. Dans *10th ACM International Conference on Multimedia*, pages 303–306, Juan-les-Pins, France, 2002.

[Patterson *et al.*, 2002] E. PATTERSON, S. GURBUZ, Z. TUFEKCI, et J.N. GOWDY. CUAVE : a new Audio-Visual Database for Multimodal Human-Computer Interface Research. Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, volume 2, pages 2017–2020, Orlando, Florida, May 2002.

[Perrot *et al.*, 2007] Patrick PERROT, Hervé BREDIN, et Gérard CHOLLET. Biometrics and Forensic Sciences : the Same Quest for Identification ? Dans *International Crime Science Conference*, London, UK, July 2007.

[Potamianos *et al.*, 2003] Gerasimos POTAMIANOS, Chalapathy NETI, Guillaume GRAVIER, Ashutosh GARG, et Andrew W. SENIOR. Recent Advances in the Automatic Recognition of Audiovisual Speech. Dans *Proceedings of the IEEE*, volume 91, pages 1306–1326, September 2003.

[Potamianos *et al.*, 2004] Gerasimos POTAMIANOS, Chalapathy NETI, Juergen LUETTIN, et Iain MATTHEWS. Audio-Visual Automatic Speech Recognition : An Overview. Dans G. BAILLY, Eric VATIKIOTIS-BATESON, et P. PERRIER, éditeurs, *Issues in Visual and Audio-Visual Speech Processing*, Chapitre 10. MIT Press, 2004.

[Reynolds, 2002] Douglas A. REYNOLDS. An Overview of Automatic Speaker Recognition Technology. Dans *27th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, volume 4, pages 4072–4075, Orlando, Florida, May 2002.

[Reynolds *et al.*, 2000a] Douglas A. REYNOLDS, George R. DODDINGTON, Mark A. PRZYBOCKI, et Alvin F. MARTIN. The NIST Speaker Recognition Evaluation - Overview Methodology, Systems, Results, Perspective. Dans *Speaker Recognition and its Commercial and Forensic Applications*, volume 31 de *Speech Communication*, pages 225–254, 2000.

[Reynolds *et al.*, 2000b] Douglas A. REYNOLDS, Thomas F. QUATIERI, et Robert B. DUNN. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10 :19 – 41, 2000.

[Ross *et al.*, 2006] Arun A. ROSS, Karthik NANDAKUMAR, et Anil K. JAIN. *Handbook of Multibiometrics*. Springer, 2006.

[Saeed *et al.*, 2006] Usman SAEED, Federico MATTA, et Jean-Luc DUGELAY. Person Recognition based on Head and Mouth Dynamics. Dans *IEEE International Workshop on Multimedia Signal Processing (MMSP'06)*, Victoria, Canada, October 2006.

[Saporta, 1978] Gilbert SAPORTA. *Théories et Méthodes de la Statistique*. Technip, Paris, 1978.

[Sargin *et al.*, 2006] Mehmet Emre SARGIN, Engin ERZIN, Yucel YEMEZ, et A. Murat TEKALP. Multimodal Speaker Identification using Canonical Correlation Analysis. Dans *31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, volume 1, pages 613–616, Toulouse, France, May 2006.

[Slaney et Covell, 2000] Malcolm SLANEY et Michele COVELL. FaceSync : A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. Dans *Advances in Neural Information Processing Systems 13*. MIT Press, 2000.

[Smaragdis et Casey, 2003] Paris SMARAGDIS et Michael CASEY. Audio/Visual Independent Components. Dans *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA'03)*, pages 709–714, Nara, Japan, April 2003.

[Sodoyer *et al.*, 2003] David SODOYER, Laurent GIRIN, Christian JUTTEN, et Jean-Luc SCHWARTZ. Speech Extraction based on ICA and Audio-Visual Coherence. Dans *7th International Symposium on Signal Processing and its Applications (ISSPA'03)*, volume 2, pages 65–68, Paris, France, July 2003.

[Sodoyer *et al.*, 2002] David SODOYER, Jean-Luc SCHWARTZ, Laurent GIRIN, Jacob KLINKISCH, et Christian JUTTEN. Separation of Audio-Visual Speech Sources : A New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli. *EURASIP Journal on Applied Signal Processing*, 11 :1165–1173, 2002.

[Sugamura et Itakura, 1986] Noboru SUGAMURA et Fumitada ITAKURA. Speech Analysis and Synthesis Methods developed at ECL in NTT–From LPC to LSP. *Speech Communications*, 5(2) :199–215, June 1986.

[Turk et Pentland, 1991a] Matthew TURK et Alex PENTLAND. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1) :71–86, 1991.

[Turk et Pentland, 1991b] Matthew TURK et Alex PENTLAND. Face Recognition using Eigenfaces. Dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'91)*, pages 586–591, Maui, USA, June 1991.

[Vatikiotis-Bateson *et al.*, 2006] Eric VATIKIOTIS-BATESON, Gérard BAILLY, et Pascal PERRIER. *Audio-Visual Speech Processing*. The MIT Press, 2006.

[Viola et Jones, 2002] Paul A. VIOLA et Michael J. JONES. Robust Real-Time Object Detection. *International Journal of Computer Vision*, 57(2) :137–154, 2002.

[Weber, 1999] Markus WEBER. CALTECH Face Database - http ://www.vision.caltech.edu/html-files/archive.html. 1999.

[Weenink, 2003] David WEENINK. Canonical Correlation Analysis. Dans University of AMSTERDAM, éditeur, *Institute of Phonetic Sciences*, volume 25, pages 81–99, 2003.

[Yang *et al.*, 2002] M.H. YANG, D. KRIEGMAN, et N. AHUJA. Detecting Faces in Images : a Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 :34–58, 2002.

[Yehia *et al.*, 1998] Hani YEHIA, Philip RUBIN, et Eric VATIKIOTIS-BATESON. Quantitative Association of Vocal-Tract and Facial Behavior. *Speech Communication*, (28) :23–43, 1998.

[Yoshimi et Pingali, 2002] Billibon H. YOSHIMI et Gopal S. PINGALI. A multimodal speaker detection and tracking system for teleconferencing. Dans *Tenth ACM international conference on Multimedia (MULTIMEDIA'02)*, pages 427–428, New York, NY, USA, 2002. ACM Press.

[Young, 2001] Steve YOUNG. Statistical Modelling in Continuous Speech Recognition (CSR). Dans *17th International Conference on Uncertainty in Artificial Intelligence*, pages 562–571, Seattle, USA, August 2001.

[Zhao *et al.*, 2003] Wen-Yi ZHAO, Rama CHELLAPPA, P.J. PHILLIPS, et Azriel ROSENFELD. Face Recognition : a Literature Survey. *ACM Computing Surveys*, 35(4) :399–458, 2003.

[Zhou *et al.*, 2004] Shaohua ZHOU, Rama CHELLAPPA, et Baback MOGHADDAM. Visual Tracking and Recognition using Appearance-Adaptive Models in Particle Filters. *IEEE Transactions on Image Processing*, 13(11) :1491–1506, 2004.