Projet AMIBE

(**A**pplications **M**ultimodales pour **I**nterfaces et **B**ornes **E**voluées)

Responsable du projet
C. Montacié (LAFORIA-IBP)

Responsables dans chaque équipe
R. André-Obrecht (IRIT)
L.-J. Boë (ICP)
M.-J. Caraty (LAFORIA-IBP)
P. Deléglise (LIUM)
I. Herlin (INRIA)
H. Méloni (LIUAPV)

Membres

C. Abry (ICP)	P. Gilles (LIUAPV)	T. Spriet (LIUAPV)
A. Adjouani (ICP)	P. Jourlin (LIUAPV)	J.-L. Schwarzt (ICP)
C. Barras (LAFORIA-IBP)	T. Lallouache (ICP)	C. Senac (IRIT)
M. Baudry (LIUM)	B. Leroy (INRIA)	H. Yahia (INRIA)
I. Cohen (INRIA)	J.-L. Lefloch (LAFORIA-IBP)	
M. El-Beze (LIUAPV)	F. Pellegrino (IRIT)	

1. Objectifs scientifiques du projet AMIBE

L'étude d'une communication Homme-Machine performante a pour objectif un mode de fonctionnement ergonomique, a priori naturel, et grand public. La phase d'apprentissage de ce mode de communication doit être minimale. Le but du projet AMIBE est d'expérimenter la notion d'interfaces Homme-Machine multimodale dans un cadre multi-utilisateurs, d'une part pour fiabiliser la communication parlée qui est le mode de communication le plus naturel et spontané, d'autre part pour introduire la notion de confidentialité des informations échangées au cours de la communication Homme-Machine.

La reconnaissance de la parole est à elle seule un objectif difficile. AMIBE propose de résoudre en parti les problèmes de robustesse de la reconnaissance automatique multi-locuteurs de la parole par la prise en compte d'un mode supplémentaire : celui de l'image. Les modules introduits pour l'amélioration de la reconnaissance de la parole traitent du suivi du contour des lèvres et de la localisation de la source sonore.

La confidentialité de la communication est assurée par un module de vérification permanente de l'identité de l'utilisateur là aussi en associant le mode parole (vérification du locuteur) et le mode image (vérification des contours du visage).

Ce type d'interface multimodale est très adapté à l'accès sécurisé d'un nombre limité d'utilisateurs à des informations personnelles ou confidentielles. Les modes utilisés, Parole et Image, présentent l'avantage d'être transparents à l'utilisateur et ne nécessitent aucun ou peu d'apprentissage de celui-ci. Une application possible de ce projet est la conception d'interfaces bancaires évoluées pour lesquelles le vocabulaire et la syntaxe sont limités.

2. Approche proposée

Le système AMIBE (cf. synoptique) est constitué de deux modules principaux fondés sur une interaction forte des modes Parole et Image : la reconnaissance du message oral transmis par l'utilisateur et la vérification de son identité. En effet, à chaque instant (e.g., chaque centiseconde), les deux modes concourent à éliminer les ambiguïtés et à renforcer la fiabilité de l'information transmise (contenu du message oral et identité de l'utilisateur). Il faut également noter une seconde interaction entre les modules principaux :

— l'analyse permanente des caractéristiques du locuteur permet une adaptation au locuteur du système de reconnaissance,

— la connaissance du contenu phonétique du message oral peut améliorer la vérification du locuteur en se focalisant sur des sons caractéristiques du locuteur.

On distingue trois modules auxiliaires associés aux modules principaux: le suivi du contour des lèvres, la détection du mouvement de la source sonore et la détection des contours du visage.

2.1. Reconnaissance du message oral

Cette reconnaissance opère en synchronisant une “lecture labiale” avec un module de reconnaissance des formes acoustiques : la reconnaissance acoustico-labiale. La “lecture labiale” consiste en un suivi des lèvres et une mesure temps réel des paramètres labiaux. Cette opération permet la localisation d'événements labiaux caractéristiques et une détection silence/parole – chaînon assez faible dans les systèmes de reconnaissance automatique de la parole – et par conséquent un recalage temporel du module de reconnaissance. Le module de détection du mouvement de la source sonore a deux fonctions pour le décodage du message oral : le débruitage du signal de parole par technique de filtrage d'antenne et le recalage spatial nécessaire au module du suivi des lèvres obtenu par positionnement des caméras.

2.2. Vérification de l'identité de l'utilisateur

Au début de la transaction, l'utilisateur donne classiquement via un clavier ou un lecteur de cartes son code personnel. Le module de vérification va contrôler cette identité en permanence en comparant à chaque instant le modèle de la voix du locuteur et les contours du visage de l'utilisateur aux références respectives correspondant au code. Le module de détection du mouvement de la source sonore est utilisé ici pour le recalage spatial du module de détection des contours du visage.

3. Organisation

Chaque module principal est traité par plusieurs équipes : la reconnaissance du message oral est étudiée par l'IRIT, le LIUAPV et le LIUM. La vérification de l'identité de l'utilisateur est étudiée par le LAFORIA-IBP et l'INRIA. Les modules auxiliaires sont traités par un seul laboratoire, le suivi du contour des lèvres et la base de données acoustico-labiale par l'ICP, la détection du mouvement de source sonore par le LAFORIA-IBP et la détection des contours du visage par l'INRIA.

Suivi du contour des lèvres et connaissances acoustico-labiale de la parole

INRIA Rocquencourt- Institut National de Recherche en Informatique et Automatique

Suivi et vérification des contours du visage

IRIT - URA 1399 - Institut de Recherche en Informatique de Toulouse Université Paul Sabatier

modèles probabilistes segmentaux maître-esclaves

LAFORIA-IBP - URA 1095 - Laboratoire des Formes et d'Intelligence Artificielle Université Pierre et Marie Curie

Vérification du locuteur et localisation de source

LIUAPV - Laboratoire d'Informatique d'Avignon Faculté des Sciences

modèles probabilistes centisecondes

LIUM - Laboratoire d'Informatique de l'Université du Mans

Architecture logicielle intégrant modèles probabilistes et connaissances

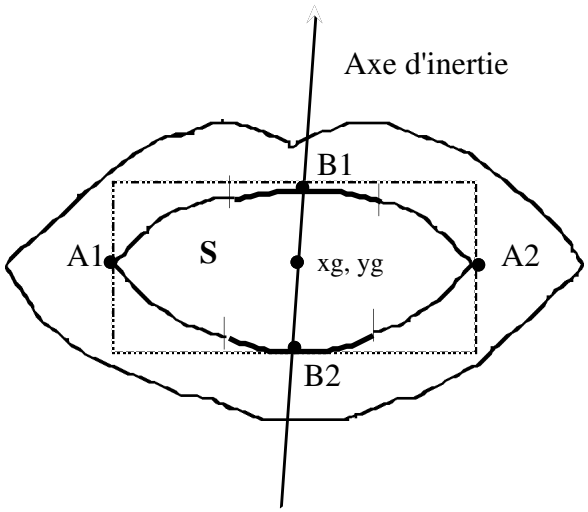
4. Tâches avancées du système AMIBE

Parmi les modules du système AMIBE, nous développerons dans ce rapport les tâches actuellement les plus avancées. Les bases de données, le système de reconnaissance acoustico-labiale de la parole. La fusion des données parole et image n'étant pas encore effectuée pour la vérification de l'utilisateur nous présenterons des résultats intermédiaires en vérification du locuteur et en détection des contours du visage.

4.1. Bases de données

Les diverses expériences de reconnaissance acoustico-labiale de la parole, de vérification des contours du visage et de vérification du locuteur, nécessitent des bases de données adéquates. Deux bases concernant les données acoustico-labiales et les données de visages sont ainsi conçues et constituées pour les besoins du projet.

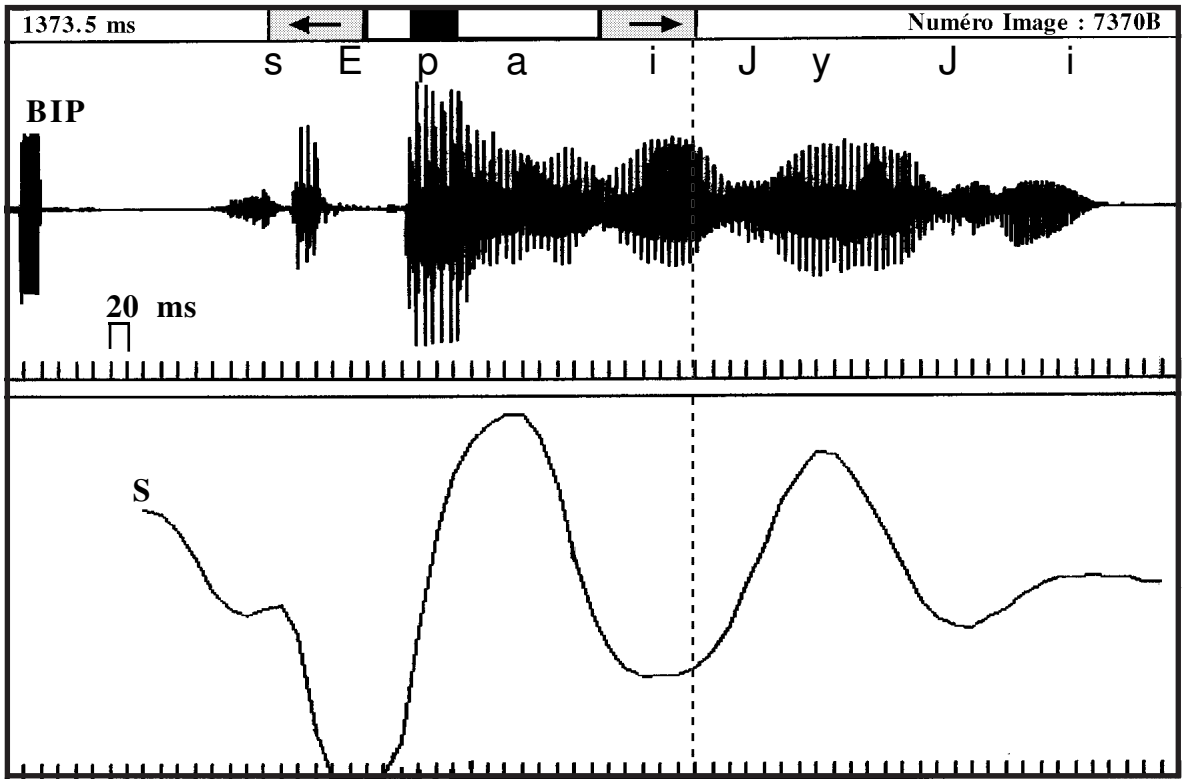
4.1.1. Données acoustico-labiales



Paramètres mesurés sur la bouche vue de face
S : aire intéro-labiale, A1A2 : étirement labial et B1B2 : séparation labiale

Cette base de données, constituée à l'ICP, est utilisée pour tester les systèmes de reconnaissance

acoustico-labiale de la parole. Elle consiste en l'enregistrement synchrone des informations acoustiques et des informations sur le suivi des lèvres. Un locuteur dont les lèvres sont maquillées en bleu est filmé de face par une caméra tri-CCD présentant une résolution de 500 lignes. L'image est traitée en temps réel via le chroma-key pour renforcer les lèvres (par incrustation d'un signal de luminance nulle dans tout l'espace du vermillon), ce qui facilite l'extraction automatique de leurs contours interne et externe. Le locuteur, à l'aide d'un bouton poussoir, commande manuellement l'insertion d'un bip sonore sur le canal audio en même temps que l'allumage d'un pavé de diodes électroluminescentes visibles dans le champ de la caméra. En même temps l'image est « time-codée » à l'aide du générateur de code image. Le son et l'image sont stockés sur bande magnétique. Le signal vidéo est numérisé et désormais prêt pour le traitement numérique (détection de contours, extraction des lèvres, mesure de paramètres, ...). On traite la suite des images dans une plage donnée et on extrait les paramètres intéressants ; notamment : l'étirement (distance entre A1 et A2), la séparation (distance entre B1 et B2) et l'aire intérolabiale (aire délimitée par le contour interne des lèvres) pour la vue de face.



Exemple de données acoustico-labiales d'une séquence "C'est pas IJUJIZE ? "
En haut, le signal audio échantillonné à 16 kHz est présenté sur une durée de 1,4 s. Les graduations de l'axe horizontal représentent les tops de trame vidéo synchronisés sur le début du bip sonore.
En bas, le signal articulatoire (courbe interpolant linéairement les points de mesure du paramètre S) est visualisé dans la limite de son analyse préalable

Le corpus des données acoustico-labiales est constitué de 38 groupes de quatre chiffres (corpus des chiffres) et cent groupes de quatre lettres (corpus des lettres) prononcés en continu. L'ensemble d'apprentissage est constitué des 14 premiers groupes des chiffres et des 71 premiers groupes des lettres, soit plus de 10.000 images. Le corpus a été enregistré par un seul locuteur masculin JLS. Neuf autres enregistrements de ce corpus sont prévus. Deux seront enregistrés par le même locuteur JLS, le premier pour tester l'effet de la taille de l'ensemble d'apprentissage sur les résultats de reconnaissance, le deuxième sans bleu sur les lèvres pour tester la reconnaissance acoustico-labiale dans des conditions plus réalistes. Les 7 autres enregistrements, prononcés par d'autres locuteurs, serviront à étudier la variabilité inter-locuteurs des données acoustico-labiales.

4.1.2. Données de visages

La base de données actuellement utilisée pour tester les techniques en vérification du visage a été enregistrée à l'INRIA. Cette base est relativement petite, elle est constituée d'acquisitions temporelles de visages de neuf personnes de sexe masculin. Au cours de la séquence d'acquisition, chaque personne prononce la phrase "cinq, six, sept, huit, neuf, dix, oui, non, la bise et le soleil". De cette façon, nous disposons d'un échantillon relativement large des expressions de la bouche et du visage. L'acquisition a été faite en studio, la personne de face sur fond uniforme, avec des conditions d'éclairage identiques sur une bande vidéo. La numérisation a ensuite été effectuée pour obtenir une image couleur rgb sous format vidéo (768 x 576), et un codage complémentaire en noir et blanc a été ensuite réalisé pour étudier le problème de la vérification en continu des visages. L'échantillonnage temporel au cours de la séquence est variable (entre 18 et 22 images par seconde). Seules les dix premières images de chaque personne, correspondant au début du "cinq", sont étudiées.

4.1.3. Données de vérification du locuteur

La vérification du locuteur indépendamment du texte est caractérisée par la possibilité de reconnaître un locuteur prononçant n'importe quelle phrase de test. De plus au cours de l'apprentissage, aucune phrase spécifique ne doit être utilisée. Le corpus choisi doit permettre la simulation d'une telle identification.

La base de données utilisée pour nos expériences est la base de données DARPA-TIMIT. Elle comprend l'enregistrement de 630 locuteurs (i.e., 192 femmes et 438 hommes). Huit "dialectes" régionaux représentent les différents accents de l'anglais américain.

Chaque locuteur prononce 10 phrases. 2 de ces phrases sont prononcées par chacun des locuteurs. Les 8 autres phrases sont différentes pour chaque locuteur. 5 sont des phrases "MIT", les 3 restantes sont des phrases "TI". Les phrases "MIT" (i.e., 450 phrases) sont construites pour avoir une grande variété des contextes phonétiques. Les phrases “TI” (i.e., 1890 phrases) sont extraites d'un corpus de texte écrit. Il n'y a eu qu'une seule session d'enregistrement par locuteur. Par conséquent, aucune étude de la dérive temporelle d'un locuteur ne pourra être effectuée.

4.2. Reconnaissance acoustico-labiale de la parole

La communication parlée est intrinsèquement bi-modale : dans la grande majorité des situations d’échange, les interlocuteurs utilisent les informations auditives et visuelles – signal de parole et visage de leur interlocuteur. Mais la nature multimodale de la parole est particulière : il ne s'agit pas d'une simple juxtaposition d'entrées parallèles, voire complémentaires, mais d'une interaction de deux modalités visant à identifier un processus sous-jacent commun. En 1978, Risberg & Lubker ont présenté à des sujets le visage seul d'un locuteur prononçant 25 phrases, puis le son correspondant dégradé (filtrage passe-bas) et enfin le son dégradé avec le visage. Dans le premier cas (perception visuelle) 1 % des mots ont été reconnus, 6 % dans le deuxième (perception auditive) et 45 % dans le troisième (perception audio-visuelle). Il faut relier à cette expérience les résultats de Erber (1969, 75) et ceux de Summerfield (1979). Ils confirment bien le gain en intelligibilité de la perception audio-visuelle sur l'audition seule, en fonction du bruit ajouté au signal acoustique. L'apport de la vision à l'intelligibilité de la parole est en effet fonction de la dégradation du signal acoustique : il est constant à partir d'un rapport signal sur bruit de -30 dB, ce qui correspond compte tenu du bruit à une seule lecture labiale, mais à -12 dB le taux de reconnaissance passe de 20 % pour la perception auditive seule à 80 % pour la perception auditive et visuelle. Le choix de l'intégration de la modalité visuelle se justifie pleinement dans les conditions naturelles d'une interface de communication mettant en jeu la parole, dans la mesure où le signal acoustique peut être bruité sans que le signal vidéo ne le soit. Dans une première phase d'évaluation de la reconnaissance acoustico-labiale, on s'intéressera à l'apport du mouvement des lèvres dans la compensation de la perte d'information due à un signal acoustique bruité.

4.2.1. Systèmes de reconnaissance acoustico-labiale

Le système de reconnaissance acoustico-labiale doit faire coopérer plusieurs processus participant à l'identification d'un même message. Comme dans toute application multi-canaux, ces processus travaillent à partir d'informations différentes pour obtenir un message complet ; toutefois dans le cadre de l'application envisagée, tous les processus sont de nature numérique - symbolique. En effet il s'agit d'obtenir, à partir des caractérisations numériques de phénomènes physiques (signal de parole, suite des images des lèvres), des descriptions symboliques (suites de mots prononcés, schémas des formes des lèvres).

Cette coopération est dissymétrique en raison de la nature distincte des paramètres labiaux et des paramètres acoustiques :

— Le traitement effectué sur les images des lèvres permet de caractériser de manière fiable la forme de ces dernières par trois paramètres (étirement labial, séparation labiale et aire intéro-labiale) ; par conséquent aucun retour à partir des résultats de décodage n'est nécessaire pour corriger la procédure d'extraction des paramètres. Ces paramètres ne sont pas toujours significatifs tant au niveau de leurs valeurs que de leurs variations, ceci peut poser des problèmes au niveau de l'apprentissage dans un processus purement probabiliste. Ce dernier point provient de la variation de l'influence des phonèmes sur la forme des lèvres. Les phonèmes peuvent se classer grossièrement en trois catégories ; ceux d'une première catégorie imposent une valeur assez précise de l'aire, ceux d'une deuxième n'imposent aucune contrainte de valeur mais seulement une contrainte de variation et pour ceux d'une troisième catégorie l'aire n'est fonction que du contexte phonétique.

— Au contraire, il est difficile d'interpréter exactement les paramètres acoustiques, mais l'expérience a montré que le niveau de leur valeur est significatif.

Cette coopération peut se réaliser de différentes manières selon le niveau d'intervention :

— Au niveau symbolique, celle-ci peut se réduire à une vérification de cohérence entre les résultats des divers processus : si les systèmes de reconnaissance acoustique et labiale ne fournissent pas de solutions alternatives, on ne peut que comparer les meilleures solutions obtenues par chacun des processus. La dissymétrie entre les processus peut être aussi utilisée dans une méthode intégrant les connaissances dont dispose l'ICP sur le mouvement des lèvres dans la production. Si cette méthode nécessite moins de données labiales il faut, en revanche, incorporer dans le processus de reconnaissance acoustique des informations symboliques obtenues par l'interprétation des paramètres labiaux. Ces informations sont essentiellement de deux types : – des impossibilités ou des restrictions sur les unités de reconnaissance pour une certaine durée ; – des zones permettant d'adapter les paramètres à l'environnement de reconnaissance.

— Au niveau numérique, plusieurs méthodes sont possibles. Le système de reconnaissance automatique de parole envisagé dans le cadre de ce projet est construit à partir de modèles probabilistes (modèles de Markov cachés), bases des systèmes actuellement les plus performants. Dans la mesure où les paramètres d'entrée sont hétérogènes, il est nécessaire de repenser l'utilisation des modèles de Markov cachés. Deux systèmes sont étudiés : l'un consiste en un modèle de Markov admettant comme paramètres d'entrée un vecteur d'observation dont les coordonnées décrivent les paramètres acoustiques et labiaux ; l'autre consiste en deux modèles de Markov cachés parallèles liés par une relation maître-esclave admettant comme paramètres d'entrée, pour l'un les paramètres labiaux et pour l'autre les paramètres acoustiques. Nous présenterons principalement dans ce rapport le premier système, développé au LIUAPV à partir du système de référence HTK, pour lequel nous disposons des premiers résultats.

4.2.2. Modèles de Markov cachés

C'est actuellement la méthode reconnue comme la plus performante et dont l'utilisation s'est largement répandue en Reconnaissance Automatique de la Parole. Un modèle de Markov caché

peut être représenté par un ensemble de NQ états discrets $\{q_i\}$, (i = 1 ,.., NQ) et de transitions $\{t_{ij}\}$, (i,j = 1 ,.., NQ) entre ces états. A chaque transition t_{ij} sont associées deux probabilités : la probabilité de transition a_{ij}, (i,j = 1 ,.., NQ) et la probabilité d'émission B_{ij}, (i,j = 1 ,.., NQ). La probabilité de transition a_{ij} représente la probabilité de passer dans l'état q_j alors que l'on est dans l'état q_i. La probabilité d'émission $B_{ij}(y_n)$ est une fonction de densité de probabilité correspondant à la probabilité d'émettre sur une transition, un symbole acoustique en fonction du vecteur spectral y_n présenté. Cette fonction de densité de probabilité peut être choisie discrète ou continue. L'estimation du modèle revient à la recherche, pour chaque transition, des deux ensembles de probabilité $\{a_{ij}\}$ et $\{B_{ij}\}$. Les algorithmes de Baum-Welch et de Viterbi sont utilisés pour cette estimation. La structure de la machine de Markov (i.e., type de fonction de probabilités, nombre d'états et de transitions) est définie a priori. L'ensemble des travaux sur cette méthode a permis de définir les topologies les plus performantes pour la reconnaissance de la parole.

Nous présenterons brièvement les modèles de Markov cachée parallèles maître-esclave, étudiés à l'IRIT, qui nous semblent bien adaptés à la reconnaissance acoustico-labiale. Il s'agit de deux modèles de Markov construits en parallèle, l'un est appelé modèle labial (M_{Lab}), dans la mesure où les vecteurs d'observation ne sont composés que des seuls paramètres labiaux, l'autre est appelé modèle acoustique (M_{Ac}) puisqu'il n'évalue que les paramètres acoustiques. Les modèles M_{Lab} et M_{Ac} peuvent être considérés liés par une relation maître-esclave. Le modèle M_{Lab} est le modèle maître, il est défini de manière classique. Le modèle M_{Ac} est l'esclave dans le sens où ses lois d'observation et ses lois de transition dépendent à chaque instant de l'état dans lequel se trouve le modèle M_{Lab}.

La construction du modèle M_{Lab} dépend de l'information que l'on veut extraire des signaux labiaux; sa topologie et les paramètres d'observation sont différents selon que l'on recherche une simple détection silence/parole ou que l'on recherche une détermination plus précise du type ouvert/fermé, arrondi/non arrondi.

4.2.3. Le modèle acoustico-labial

Considérant des modèles simples de Markov cachés, il s'agit de calculer la probabilité d'avoir émis une suite de vecteurs acoustiques avec un modèle X et la suite de vecteurs labiaux correspondants avec un modèle Y. A priori, si ces deux événements sont supposés statistiquement indépendants, il suffit de calculer le produit des probabilités et de choisir le couple de modèles qui a la plus forte probabilité d'avoir émis conjointement les deux suites de vecteurs. Ceci est faisable sur des mots isolés mais difficilement sur de la parole continue. En effet, l'algorithme de reconnaissance de mots connectés, permettant de trouver la phrase la plus probable, devient d'une trop grande complexité. Plusieurs solutions sous-optimales sont cependant possibles :

— Nous pouvons mettre en concurrence les deux types de modèles après un apprentissage séparé, et appliquer l'algorithme de décodage classique. L'inconvénient de ce système est que les probabilités ne sont pas comparables sur les deux types de données et il en résulte que, lors du décodage, le type de données ayant les plus fortes probabilités sera systématiquement choisi, sans que cela soit justifié.

— Une autre possibilité, nécessitant cette fois une modification de l'algorithme de reconnaissance consiste à mettre en parallèle les modèles labiaux et acoustiques, mais au lieu de propager les jetons sortant d'un modèle de l'unité de reconnaissance ayant la plus forte probabilité, on propage ceux qui ont la plus forte probabilité conjointe. Elle permet aux deux types de modèles d'être indépendants, c'est-à-dire d'avoir des probabilités de boucler ou de transiter différentes au niveau labial et au niveau acoustique. Cela permet de gérer l'anticipation et la rétention labiale à l'intérieur des unités à reconnaître, mais le problème persiste aux frontières de ces unités. Cette technique nous permet d'obtenir un système qui fonctionne mais qui nécessite néanmoins une pondération acoustico-labiale, cette dernière n'étant pas effectuée lors de l'apprentissage.

— Pour remédier au problème de l'anticipation et de la rétention labiale nous pouvons

associer au vecteur acoustique correspondant au temps t les vecteurs labiaux correspondants aux temps compris entre t - d_a et t - d_r (d_a étant l'anticipation et d_r la rétention maximale). Lors de l'apprentissage, les variances calculées pour les vecteurs labiaux non discriminants seront plus élevées que pour ceux qui sont réellement liés avec les vecteurs acoustiques. Par conséquent, lors du décodage, ce sont ces derniers qui auront le plus de poids.

Actuellement, le système de reconnaissance utilise un modèle unique pour lequel, sur chaque état, la probabilité d'émission d'un vecteur acoustico-labial est le produit des probabilités acoustiques et des probabilités labiales. L'avantage de cette solution est que l'apprentissage est fait conjointement, ce qui permet au système, pour un modèle et un état donné, de donner plus d'importance au type de vecteur le plus discriminant. En revanche, ceci contraint le modèle acoustique et le modèle labial à avoir le même nombre d'états et les mêmes probabilités de transition.

4.2.4. Expériences de reconnaissance

Les données originelles fournies au système de reconnaissance sont le signal acoustique échantillonné à 16 kHz et les trois signaux labiaux issus du module du suivi des lèvres (étirement labial, séparation labiale et aire intéro-labiale) échantillonnés à 50 Hz. Les paramètres acoustiques choisis sont les 12 coefficients cepstraux en échelle Mel, l'énergie du signal et la dérivée de chacun des paramètres précédents. L'analyse du signal vocal est centiseconde, un vecteur d'observation acoustique de 26 composantes est fourni toutes les 10 ms. Les paramètres labiaux retenus sont l'étirement, la séparation, l'aire, leurs dérivées et leurs accélérations réparties sur trois flots. Compte tenu de l'échantillonnage des signaux labiaux, un vecteur d'observation labiale de 9 composantes est fourni toutes les 20 ms. Les vecteurs d'observation acoustico-labiaux sont la concaténation des paramètres acoustiques et des paramètres labiaux. Compte tenu du sous-échantillonnage des paramètres labiaux une interpolation linéaire de ces paramètres est utilisée.

Un modèle est considéré par mot. Dans un premier temps, tous les modèles considérés ont la même topologie. Les modèles utilisés sont composés de huit états, chacun admettant une transition vers l'état suivant et une boucle sur lui-même. La probabilité d'émission d'un vecteur composé de plusieurs flots est le produit des probabilités données par les distributions associées à chacun des flots. Cette topologie est très pauvre et devra être adaptée en fonction du contenu phonétique des mots à reconnaître. Elle nous a permis cependant d'avoir des résultats intéressants sur de la parole bruitée.

Les expériences ont pour but d'estimer l'apport du suivi des lèvres dans la reconnaissance de la parole. Trois modèles sont évalués sur leur performance de reconnaissance, dans des conditions de bruit croissant, de la base de données des chiffres et des lettres (cf. §4.1.1.). Les modèles considérés sont : le modèle labial, le modèle acoustique et le modèle acoustico-labial qui admettent en entrée des vecteurs d'observation respectivement acoustique, labiale et acoustico-labial. Les performances de reconnaissance du modèle labial sont de 62 % sur les chiffres et de 34 % sur les lettres, ceux du modèle acoustique sans bruit de 96 % sur les chiffres et de 90 % sur les lettres, Ceux du modèle acoustico-labial de 96 % sur les chiffres et de 78 % sur les lettres. Nous n'avons donc pas pu pour l'instant mettre en évidence l'apport de la labialité sur les performances de reconnaissance de la parole propre, cela est sans doute dû à la simplicité du modèle de Markov caché utilisé.

Les mêmes expériences ont été conduites sur de la parole bruitée. Pour la simuler un bruit blanc uniforme a été additionné aux signaux acoustiques. Les résultats sont les suivants :

Rapport S/B	Corpus de chiffres	Corpus de chiffres	Corpus de lettres	Corpus de lettres
	Modèle acoustique	Modèle acoustico-labial	Modèle acoustique	Modèle acoustico-labial
-6 dB	89%	96%	53%	62%

-12 dB	79%	90%	20%	55%
-18 dB	40%	85%	4%	40%
- 24 dB	10%	70%	—	—

Taux de reconnaissance sur les différents corpus en fonction du rapport Signal/Bruit pour les modèles acoustiques et acoustico-labiale

Nous avons donc confirmé par un système automatique ce que nous avions appris par les études de perception. L'apport de la labialité permet de compenser en partie les faiblesses des systèmes de reconnaissance acoustique de la parole bruitée. Nous allons maintenant développer les systèmes de reconnaissance acoustico-labiaux plus sophistiqués que nous avons décrits précédemment.

4.3. Vérification du locuteur

Le module de vérification du locuteur développé au LAFORIA-IBP a pour but de vérifier en permanence l'identité prétendue de l'utilisateur pour éviter la possibilité d'un changement de locuteur au cours de la communication. L'approche que nous développons est fondée sur l'utilisation d'un modèle de l'évolution spectrale, la modélisation AR-Vectorielle (ARV), et d'une mesure de distorsion inter-locuteurs appelée Distance d'Itakura Vectorielle (DIV). Les modèles ARV permettent de décrire les trajectoires suivies par les paramètres d'analyse d'un segment de parole (i.e., phrase ou série de phrases). Ils sont interprétés comme une représentation des capacités articulatoires du locuteur (i.e., vitesse et accélération instantanées des paramètres spectraux). Un module de vérification du locuteur fondé sur cette approche présente trois caractéristiques indispensables à notre projet : – des performances excellentes – une utilisation indépendante du texte prononcé – un apprentissage rapide. En effet, la phase d'apprentissage d'un nouvel usager ne nécessite que l'apprentissage de son modèle ARV (i.e., 30 secondes de parole).

La modélisation AR-Vectorielle (ARV) est un outil classique en traitement du signal multi-dimensionnel. Soit $\{y_n\}$, (n= 1,..., N) une succession de N vecteurs cepstraux d'ordre p. Leur évolution est décrite par un modèle auto-régressif vectoriel d'ordre q :

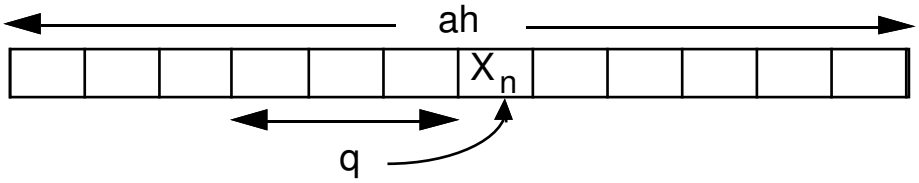
$$y_n = \sum_{i=1}^{q} A_i \, y_{n-i} + e_n$$

où $\{A_i\}$, (i= 1,..., q) sont des matrices (p, p) et $\{e_n\}$, (n = 1,..., N) est un bruit vectoriel blanc qui a pour matrice de covariance D. Les coefficients des matrices A_i sont estimés par l'algorithme de Levinson-Whittle-Robinson. Le critère à minimiser est la trace de la matrice D. Pour appliquer la modélisation ARV dans le cadre du projet AMIBE, deux études ont été nécessaires : La première pour obtenir une décision de vérification en continu et non à la fin d'une phrase comme dans la version originale, la seconde pour déterminer la fiabilité de cette vérification en fonction du type des segments de parole.

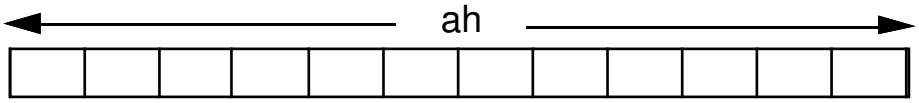
4.3.1. Vérification en continu du locuteur

Deux méthodes de vérification en continu (i.e., analytiques et segmentales) ont été développées et testées sur la base de données TIMIT : la vérification analytique et la vérification segmentale. Dans une vérification "analytique", tout ensemble de q+1 vecteurs cepstraux consécutifs (nécessaires à la prédiction vectorielle) permet une vérification du locuteur. La décision est prise en comparant le $(q+1)^{ème}$ vecteur avec la prédiction réalisée par les modèles ARV d'apprentissage sur les q premiers vecteurs. Dans une vérification "segmentale", un segment phonétique permet une vérification du locuteur à partir de la distance DIV entre le modèle ARV du segment et les modèles ARV appris. Deux variantes de la vérification ont été testées, l'une consiste à prendre un modèle ARV d'apprentissage unique pour chaque locuteur, l'autre à calculer

un modèle ARV d'apprentissage par type de segment phonétique.



Vérification analytique sur un vecteur
Comparaison entre le vecteur Xn et sa prédiction, par le modèle appris ARV,
à partir des q derniers vecteurs



Vérification segmentale sur un phonème
Comparaison entre le modèle ARV du segment phonétique et les modèles ARV appris

4.3.2. Fiabilité de la vérification en fonction des segments phonétiques

A priori, les phonèmes contenant le moins d'information caractéristique du locuteur, donc les moins fiables pour une vérification en continu, sont les phonèmes présentant un faible taux de vérification analytique et segmentale. Les résultats de vérification analytique du locuteur montrent que les phonèmes les plus discriminants, par ordre décroissant, sont : les diphtongues, les voyelles, les nasales/liquides, les glissées, les plosives/fricatives/affriquées. Le taux de vérification analytique du locuteur, pour les sons voisés, est généralement plus élevé lorsqu'il est calculé à partir du tiers central des phonèmes. Ceci est vraisemblablement dû à la stabilité du modèle ARV au centre de ces phonèmes. Les résultats de vérification segmentale sont faibles pour les phonèmes d'une durée de réalisation inférieure à une dizaine de centisecondes. Par contre, les diphtongues correspondent à de bons résultats explicables par leur importante durée de réalisation. Pour que les résultats de la vérification segmentale ne soient pas biaisés par la longueur du segment phonétique, nous avons développé une méthode pour concaténer les segments phonétiques de même nature jusqu'à l'obtention d'une seconde de signal. La vérification segmentale a lieu sur cette concaténation de segments en utilisant un modèle ARV spécialisé sur le type de segment phonétique considéré. Les résultats sont donnés ci-dessous :

Voyelles extrêmes	Voyelles	Nasales	Liquides/Glissées	Occlusives
82.1 %	94.3 %	73.7 %	47.6 %	12.5 %
Fricatives/affriquées	Silence	Diphtongues	Voisées	Non voisées
36.8 %	3.5 %	55.2 %	92.1 %	31 %

Taux de vérification du locuteur sur une seconde de segments concaténés
en fonction du type de segment phonétique

Cette étude montre qu'une vérification en continu n'est pas possible si on utilise seulement le signal acoustique. Si l'utilisateur prononce de nombreux sons peu caractéristiques de son identité comme les glissées, les fricatives ou les affriquées, ou encore s'il ne parle pas, un moyen supplémentaire de vérification est indispensable.

4.4. Vérification des contours du visage

Deux approches sont utilisées pour la reconnaissance du visage -l'approche globale qui n'utilise

aucune reconnaissance préalable sur la structure du visage -l'approche géométrique qui incorpore un modèle sur les caractéristiques spécifiques du visage . La première approche donne de bons résultats (96% à 64% de reconnaissance pour 16 personnes suivant les conditions d'éclairage et de position du visage) mais son coût de calcul (calcul des vecteurs propres de la matrice de covariance des images) la rend inutilisable pour une vérification en continu. L'approche géométrique que nous avons introduite à l'INRIA se base sur une extraction des contours caractéristiques du visage (yeux, bouche, nez) par des méthodes de modèles paramétriques déformables (snakes et représentation elliptique de Fourier). Le processus de vérification est basé sur des points caractéristiques de ces contours dans un espace modifié par un modèle déformable du visage.

4.4.1. Extraction des contours caractéristiques du visage

Deux approches ont été évaluées pour résoudre ce problème : l'utilisation des modèles déformables "snakes" et la définition de modèles paramétriques déformables spécifiques. Dans les deux cas l'extraction des contours caractéristiques est obtenue au moyen d'une initialisation préalable. Nous devons être capable pour cela de localiser un point peu éloigné du centre de l'objet recherché : pour l'oeil : ce point est fourni par la tâche de réflexion spéculaire obtenue au moment de l'acquisition, pour la bouche: il est possible d'utiliser le module de détection de la source sonore pour localiser ce point. Comme nous ne disposons pas encore de ce module nous utilisons pour l'instant une initialisation manuelle de la bouche. Pour initialiser le contour de l'objet recherché on utilise alors une courbe de forme et de taille constantes (en général, ce sera une ellipse) qui est translatée suivant l'image traitée. De plus, les paramètres du modèle utilisé pour la détection des contours sont constants sur toutes les images étudiées.

4.4.1.1. Utilisation des modèles déformables (snakes)

Cette méthode consiste à rechercher une courbe paramètrée minimisant une fonctionnelle d'énergie dépendant des informations recherchées dans l'image. Cette approche a été évaluée et abandonnée pour l'instant avec le constat que les résultats obtenus sont corrects dans le cas de la bouche, mais sont médiocres pour les yeux. Il n'a en effet pas été possible de trouver des valeurs de paramètres permettant de détecter les différentes formes d'oeil. De plus, l'utilisation des contours dans cette région du visage pose problème, car l'image de niveaux de gris contient de fortes irrégularités.

4.4.1.2. Modèle paramétrique déformable

Les contours des yeux et de la bouche ayant une forme proche de celle d'une ellipse, nous avons ensuite fait le choix de modéliser les contours caractéristiques recherchés par une représentation elliptique de Fourier d'ordre N. Pour N = 1, on obtient effectivement une courbe elliptique. Augmenter la valeur de N permet de modéliser des irrégularités de la courbe. En fait, la détection de l'oeil peut être effectuée en choisissant la valeur N = 6. (cf. Figure)

La détection des contours s'effectue par minimisation de l'énergie liée à la norme des gradients détectés dans l'image convoluée par un filtre de lissage. Cette minimisation est effectuée de manière itérative. On peut considérer le nombre d'harmoniques, N comme étant un paramètre de régularisation, plus N est grand et plus la courbe est à même d'épouser les détails du contour. Il faut toutefois noter qu'une valeur trop élevée de N peut donner lieu à des résultats fortement oscillants et irréguliers.

4.4.2. Description du schéma de vérification

Le but de notre approche est de ne pas effectuer d'identification à partir de l'information pixel mais d'utiliser le résultat des caractéristiques extraites de façon à ce que le processus soit moins coûteux en calculs et qu'il permette une vérification en continu. Il faut également que la méthode utilisée ne soit pas trop sensible à des déformations du visage non informantes sur

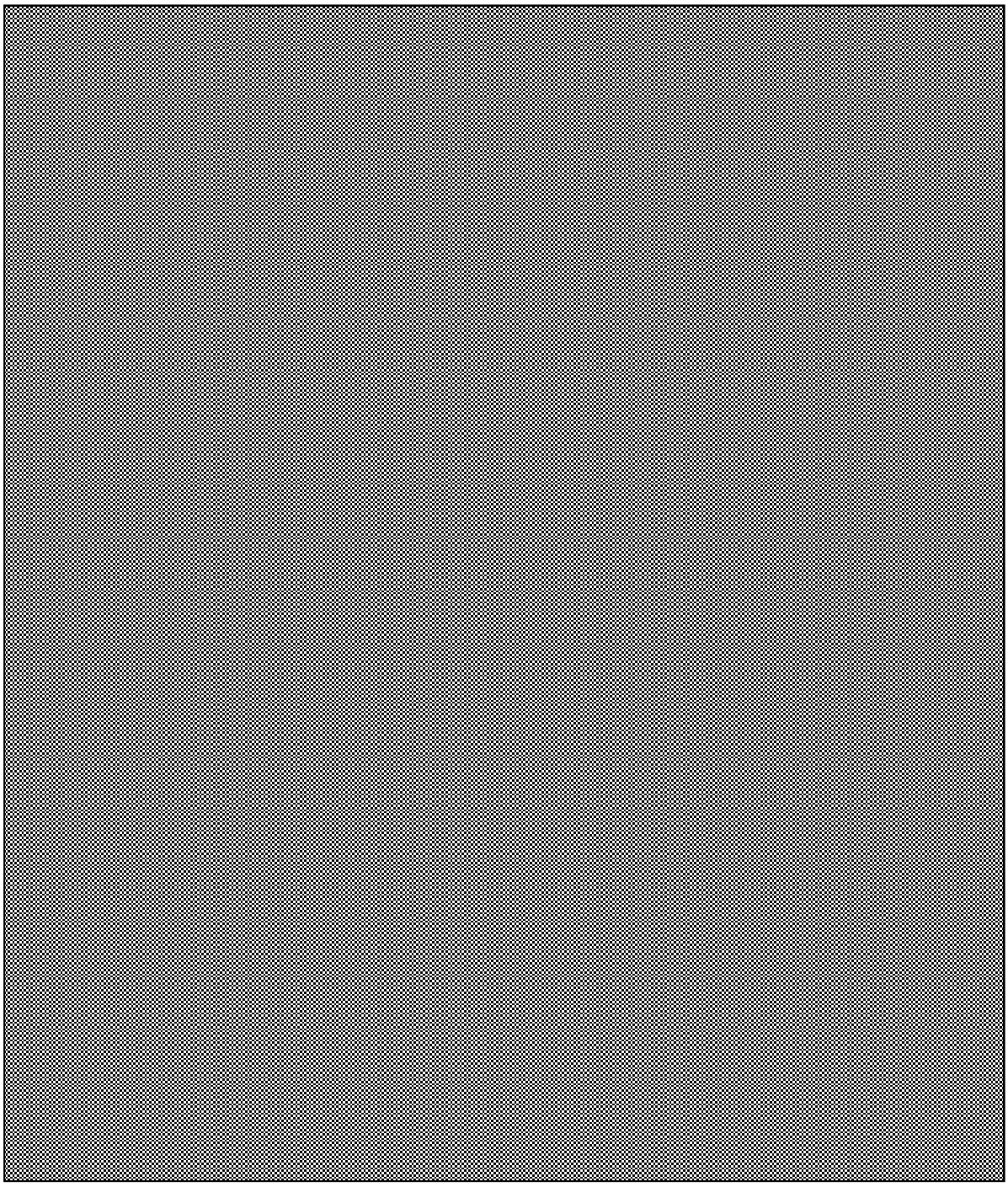
l'identité de l'utilisateur comme par exemple les contours de la bouche quand l'utilisateur parle.

Nous avons donc introduit un module de normalisation des points caractéristiques des contours par l'intermédiaire d'une grille de déformation ayant un comportement élastique et visqueux à la fois. On calcule d'abord le champ de déformation nécessaire pour transformer les points caractéristiques d'un contour de référence en ceux du visage à vérifier. On utilise ensuite ce champ de déformations à la grille entière pour obtenir les points caractéristiques normalisés. La sensibilité de la méthode de normalisation peut se régler pour chaque point de la grille en agissant sur les coefficients de ressort et de viscosité associés.

La vérification des contours du visage est effectuée sur les vecteurs propres de la matrice de covariance des points caractéristiques normalisés.

5. Conclusions

La communauté Parole a mis plus d'une vingtaine d'années à se structurer et l'informatique a été un facteur important de l'organisation de ce nouveau champ scientifique autrefois dispersé entre la phonétique, l'électronique et l'acoustique. Le PRC informatique Communication Homme-Machine a fort opportunément et judicieusement, favorisé une véritable rencontre des recherches sur la parole audible et la parole visible (le son et l'image) modalités essentielles et incontournables de la communication. Le moment semble d'autant plus favorable que les méthodes développées débouchent de part et d'autres sur des résultats véritablement exploitables et donc sur la perspective de plate-formes opérationnelles.



Exemples d'extraction du contour de l'oeil
La première image de chaque ligne représente l'initialisation ainsi que les données utilisées
De haut en bas : les contours, l'image de niveau de gris, l'image pic ou l'image vallée

Les différentes colonnes représentent les contours pour différentes valeurs de N : 2, 6 ou 10.