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

THÈSE

Présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

APPROCHE BIMODALE DU TRAITEMENT AUTOMATIQUE DE LA PAROLE : APPLICATION À LA RECONNAISSANCE DU MESSAGE ET DU LOCUTEUR

par

Pierre Jourlin

Soutenance le 24 Avril 1998 devant le Jury composé de :

MM	Régine	André-Obrecht	Rapporteur
	Gérard	Chollet	Rapporteur
	Paul	Deléglise	Examinateur
	Renato	De Mori	Examinateur
	Marc	El-Bèze	Directeur de thèse
	Henri	Méloni	Examinateur

Leurs yeux sont morts et leurs lèvres sont molles,
Et l'on entend à peine leurs paroles.
Paul Verlaine

RÉSUMÉ

Ces travaux constituent une étude sur la possibilité d'intégrer les informations visuelles constituées par le mouvement et la forme des lèvres dans les systèmes de traitement automatique de la parole.

Les différentes approches et méthodes relatives à cette problématique sont abordées d'une façon théorique et expérimentale. Une description technique des phénomènes d'asynchronie (ou d'indépendance temporelle) présents dans cette source d'information bimodale est tout d'abord établie. Différentes manières de gérer ces phénom'enes dans les systèmes de reconnaissance de la parole sont étudiées et comparées. Nous définissons alors une nouvelle approche fondée sur un produit d'automates à transitions valuées. En outre, la combinaison de deux modalités, qui peuvent avoir des niveaux de fiabilité totalement différents, pose un certain nombre de problèmes relatifs à la pondération. Nous étudions donc les divers critères et méthodes permettant de trouver une pondération optimale.

Il est également fait état de différentes expérimentations effectuées dans le domaine du traitement de la parole acoustico-labiale. Nos propres expérimentations dans le domaine de la reconnaissance de la parole bimodale et dans le projet AMIBE (financé par le CNRS) sont décrites. Les résultats des nouvelles méthodes développées dans ces travaux sont également présentés.

Enfin, nous abordons également le domaine de la vérification d'identité acoustico-labiale. Nous présentons les résultats obtenus par le système que avons réalisé en collaboration avec l'IDIAP dans le cadre du projet européen M2VTS (programme ACTS). Ces expérimentations furent parmi les toutes premières réalisées dans le domaine de la reconnaissance et vérification acoustico-labiale du locuteur.

ABSTRACT

This work addresses the problem of integrating visual information about the lip movements into acoustic-based speech processing systems.

In a first part, the different problems and methods related to this particular approach are discussed. The first chapter is dedicated to the asynchrony phenomenon (or temporal independence) of the two sources of information and to the different ways of handling it in HMM-based speech processing systems. The second chapter concerns the problem of combining two modalities which may have very different behaviours from a reliability point of view : Different methods for finding an optimal weighting are discussed.

In the second part are reported the different experiments I have done in the acoustic-labial speech processing field. In the chapter 3 are reported my own experiments in the acoustic-labial speech recognition field and in the framework of AMIBE (French CNRS project). I also present in this chapter the results of the novel methods described in chapter 1 and 2. Chapter 4 is dedicated to acoustic-labial person authentication systems and experiments, in the framework of the European project M2VTS (program ACTS). These experiments are the very first conduced in the field of acoustic-labial speaker recognition and verification.

Pour l'intérêt qu'ils ont spontanément témoigné à mes activités de recherche, je tiens à exprimer mes remerciements et ma sincère reconnaissance aux membres du jury :

Marc El-Bèze, Professeur à l'Université d'Avignon et des Pays de Vaucluse et actuellement directeur du Laboratoire d'Informatique de la dite Université, qui a dirigé mes recherches et m'a constamment encouragé de ses précieux commentaires, conseils et critiques ;

Régine André-Obrecht, Chargée de Recherche CNRS, qui s'est consciencieusement impliquée dans mon sujet de recherche et qui m'a fait l'honneur d'examiner ce travail ;

Gérard Chollet, Directeur de Recherche CNRS, qui a pris le temps d'étudier ce mémoire et m'a fait profiter de ses réflexions ;

Paul Deléglise, Professeur à l'Université du Maine, qui m'a conseillé et encouragé lors du projet AMIBE et qui m'a fait l'honneur de faire parti de mon jury de thèse ;

Renato De Mori, Professeur à l'Université d'Avignon et des Pays de Vaucluse, qui m'a fait profiter de son expérience pendant ma dernière année thèse au LIA ;

Henri Méloni, Professeur à l'Université d'Avignon et des Pays de Vaucluse et actuellement président de la dite Université, pour m'avoir acceuilli si chaleureusement au LIA, lorsqu'il en était le directeur.

Table des matières

Table des figures

Liste des tableaux

AVANT-PROPOS

Dans le domaine du traitement automatique de l'information, un problème récurrent est celui de la classification. Suivant la nature des données, on peut envisager un nombre considérable d'approches. Quel que soit le contexte, cette information peut-être considérée soit dans sa globalité, soit comme étant un ensemble d'informations plus ou moins indépendantes entre elles. Cette possible indépendance peut d'ailleurs s'exprimer sur plusieurs niveaux : le contexte, le temps, la fiabilité, etc. Dans ce travail, deux questions sous-jacentes mais d'une importance capitale se posent : considérer l'information non plus de façon globale mais comme un ensemble d'informations de différente nature peut-il enrichir la problématique initiale ? Dans l'affirmative, peut-on mettre à profit cet enrichissement pour améliorer le processus de classification ?

Ce premier paragraphe généraliste justifie, sinon explique les développements que j'ai pu faire autour du thème initial de ma thèse qui était le traitement automatique de la parole audiovisuelle basé sur les modèles de Markov cachés. Il convient également de noter que les problèmes identifiés et traités dans ce mémoire dans un contexte de recherche précis, peuvent de surcroît être transposés pour d'autres contextes de classification.

Les études effectuées en sciences cognitives sur l'intégration audiovisuelle peuvent apporter nombre d'idées pour la recherche en traitement automatique de la parole. Cependant, le lien qui existe entre ces deux domaines est loin d'être simple : le fonctionnement de notre cerveau, autant qu'on puisse le comprendre, n'est pas forcément le plus adapté pour réaliser des systèmes automatiques. C'est la raison pour laquelle je ne tenterai aucun parallèle entre ces deux domaines, que ce soit au niveau théorique ou expérimental.

INTRODUCTION

Un message oral contient à la fois des informations d'ordre linguistique et des informations liées au locuteur. Il existe de très nombreux sous-domaines de traitement automatique de la parole. Nous nous limiterons aux deux principaux types d'application : La reconnaissance automatique de la parole et celle du locuteur.

Les premières recherches et expérimentations en traitement automatique de la parole, ont fait apparaître l'extrême complexité de ce problème.

C'est pourquoi, plutôt que d'essayer de le résoudre dans sa globalité, on a tout d'abord posé des contraintes le simplifiant : vocabulaire réduit, système mono-locuteur, élocution en mode "mots isolés", et environnement acoustique peu bruité. Deux principaux buts de recherche apparaissent alors : la réduction des contraintes sur les conditions de fonctionnement et l'amélioration des taux de reconnaissance de ces systèmes.

Dans cette optique, on porte un intérêt grandissant à l'intégration d'autres sources d'informations pour compléter la composante acoustique. Ces dernières peuvent être des connaissances situées à des niveaux différents : articulatoire, auditif, syntaxique, morphologique, sémantique, prosodique, etc. Mais, on peut aussi envisager de tenir compte des informations visuelles qui accompagnent, voire conditionnent, l'émission de sons. Les lèvres participent à ce processus de production de la parole et transmettent par conséquent un message sur les deux modalités.

Ceci nous offre deux perspectives d'amélioration des systèmes de traitement automatique de la parole. Tout d'abord, les distances entre les différentes unités à classifier ne sont pas les mêmes sur le plan acoustique et sur le plan visuel. Il nous faudra donc vérifier dans quelle mesure ceci nous permettra de lever les ambiguïtés propres à une source ou à une autre. Par exemple, les phonèmes [p] et [t] sont assez proches au niveau acoustique mais très différents au niveau labial : la prononciation du [p] exige une closion labiale, ce qui n'est pas le cas du [t] (closion dentale). Il en va de même en ce qui concerne le locuteur : deux personnes peuvent avoir des voix proches, mais une articulation labiale différente. D'autre part, dans des conditions naturelles, la source acoustique est très souvent bruitée, ce qui peut être corrigé par l'utilisation de la source visuelle.

Ce document s'articulera en deux parties, la première étant consacrée aux méthodologies et la seconde aux expérimentations. L'objectif avoué d'un tel découpage étant d'éviter toute comparaison hâtive des méthodes et principes d'intégration sur la base des expérimentations. En effet, dans un domaine aussi récent que celui-ci, l'absence de bases de données et de systèmes de référence rend extrêmement hasardeux le processus d'évaluation.

L'intégration de ces deux types d'informations pose un certain nombre de problèmes. Leur indépendance temporelle est source d'une variabilité nouvelle qui s'ajoute à celle liée à chacune d'entre elles ; ce point sera l'objet du chapitre 1.

Nous verrons que l'information labiale seule ne permet que des performances très inférieures à celles que l'on peut obtenir en utilisant la partie acoustique du signal. La pondération de ces deux sources est donc un élément crucial dans le processus de fusion et sera l'objet du chapitre 2.

Le chapitre 3 comporte le compte-rendu des différentes expérimentations effectuées en reconnaissance automatique de la parole acoustico-labiale. Mes propres conditions expérimentales ainsi que les résultats obtenus y seront présentés.

Le chapitre 4 sera consacré à mes travaux en matière d'identification et de vérification d'identité. La prise en compte des deux modalités acoustiques et visuelles de la parole constitue une première dans ce domaine.

Enfin, nous établirons un bilan du travail accompli et au-delà des conclusions que nous pouvons en tirer, des perspectives d'amélioration et d'extension seront également évoquées.

Première partie

Méthodologies

Chapitre 1

Asynchronie

RÉSUMÉ

Dans ce chapitre, sont abordés les problèmes liés à l'aspect asynchrone des informations relatives à deux modalités différentes.

En nous plaçant à un niveau purement théorique, nous étudions les différentes façons d'aborder les phénomènes d'asynchronie dans les systèmes statistiques de classification et de segmentation intégrant deux modalités.

Dans une première partie (section 1.3) sont passées en revue les différentes méthodes et approches décrites dans la littérature qui sont susceptibles de gérer ces problèmes. Nous verrons qu'il est parfois difficile de quantifier la part d'asynchronie qu'ils prennent en compte. **Dans une deuxième partie** (section 1.4) est donc définie notre propre approche, construite dans le but de dépasser cette difficulté. Elle nous permettra d'évaluer, dans le chapitre 3 la part d'erreurs de reconnaissance qui est due à l'absence de prise en compte de l'indépendance temporelle dans les systèmes de type classique.

Ces différentes méthodologies sont décrites, analysées et autant que possible, comparées sur le plan théorique.

1.1 Introduction

Lorsque qu'un locuteur prononce la même suite d'unités phonétiques, les différents événements produits par chacun des organes phonatoires ne sont pas réalisés de façon synchrone. Par exemple, le temps qui s'écoule entre le début du mouvement des lèvres et l'émission effective d'un son peut être très variable. Cette durée est en effet liée à un nombre considérable de paramètres, tels que l'état psychologique du locuteur, les phonèmes qu'il doit prononcer, etc.

Ces phénomènes, appelés également *anticipation* et *rétention* ont été étudiés en détail dans d'autres domaines, par exemple en production par Abry et Lallouache (1995) et en perception par Cathiard et al. (1995). Cette variabilité est amplifiée par la prise en considération de contextes linguistiques ou de locuteurs différents.

D'un point de vue statistique, entre deux réalisations différentes de la même séquence de sons, la mise en correspondance d'événements acoustiques et labiaux est temporellement différente. Il ne s'agit pas simplement de décalages temporels, mais aussi d'une variation possible du rapport de durée d'un événement acoustique et de l'événement labial correspondant.

Dans un système de traitement automatique de la parole bimodale fondé sur une méthode d'apprentissage donnée, il faudra donc modéliser ces différentes possibilités de synchronisation audiovisuelle pour chaque unité à classifier.

C'est ce type de variabilité que nous appelons par la suite asynchronie. Dans ce chapitre, nous étudions dans cette optique le fonctionnement des divers systèmes décrits dans la littérature. Nous présentons également notre propre approche, ses intérêts, ses faiblesses et ses perspectives d'améliorations.

1.2 Cas des mots isolés

Nous entendons par *mots isolés*, les cas où l'on dispose d'une segmentation en classe. Autrement dit, lors de la phase de reconnaissance, nous n'avons pas à rechercher une segmentation, mais uniquement à classifier une portion de signal de parole. C'est par exemple le cas lorsque le locuteur prononce les mots isolément et qu'un détecteur de silence les extrait de l'ensemble du signal.

Les suites d'observations de la première modalité O_1 et de la deuxième O_2 correspondent à la prononciation d'un mot (ou, plus généralement, unité de reconnaissance) dont on cherche l'identité.

Le mot recherché fait partie d'un vocabulaire $\{W_1, \ldots, W_n\}$. À ces n mots sont associés les modèles $\{M_{1,1}, \ldots, M_{1,n}\}$ et $\{M_{2,1}, \ldots, M_{2,n}\}$ respectivement attachés à la modalité 1 et 2.

Nous recherchons par conséquent le modèle $\mathcal{M}$ qui maximise la probabilité conjointe :

$$\begin{aligned} \mathcal{M} &= \arg\max_i P(M_{1,i}, M_{2,i}|O_1, O_2) \\ \text{avec} \quad & i \in [1, n]. \end{aligned}$$

En appliquant la règle de Bayes, nous obtenons :

$$P(M_{1,i}, M_{2,i}|O_1, O_2) = \frac{P(O_1, O_2|M_{1,i}, M_{2,i}) \cdot P(M_{1,i}, M_{2,i})}{P(O_1, O_2)}$$

$P(O_1, O_2)$ est une constante au regard de i. $P(M_{1,i}, M_{2,i})$ peut être calculée par un modèle de langage. Cependant, pour certaines applications, prononciation d'un code par exemple, elles sont égales pour tout i et peuvent donc être ignorées. La valeur à maximiser est donc : $P(O_1, O_2|M_{1,i}, M_{2,i})$.

Si l'on suppose l'indépendance statistique de $(O_1, M_{1,i})$ et de $(O_2, M_{2,i})$ il ne reste plus qu'à calculer $P(O_1|M_{1,i}) \cdot P(O_2|M_{2,i}) \cdot P(M_{1,i}) \cdot P(M_{2,i})$, soit de rechercher :

$$\mathcal{M} = \arg\max_i P(O_1|M_{1,i}) \cdot P(O_2|M_{2,i}) \tag{1.1}$$

avec $i \in [1, n]$.

L'asynchronie n'apparaît pas ici comme un problème, puisque la probabilité associée à chacune des sources est calculée séparément.

En revanche, dans le cas de la parole continue et en conservant cette méthode, il s'agirait de rechercher non plus le couple de modèles mais le couple de suites de modèles maximisant cette probabilité. La complexité résultant de ce problème étant ingérable, nous devons donc avoir recours à des solutions sous-optimales.

La première possibilité qui vient à l'esprit consiste à utiliser des MMC classiques émettant des vecteurs composés en partie d'observations acoustiques et en partie d'observations labiales. Moyennant un grand nombre de paramètres à estimer, provenant d'un grand nombre d'allophones et de mixtures de gaussiennes, l'asynchronie sera modélisée. Néanmoins, il faudrait un corpus de taille beaucoup plus importante pour assurer l'apprentissage correct de l'asynchronie. Or, à l'heure actuelle, personne ne dispose de tels corpus (voir chapitre 3, section 3.1, pages 50-52).

Nous allons donc explorer dans ce chapitre les différentes méthodologies mises au point pour résoudre ce problème. Les avantages et limites de chacune d'entre elles seront étudiés d'un point de vue théorique.

1.3 Étude bibliographique

1.3.1 Asynchronie et modèles d'intégration

Robert-Ribes et al. (1995b) ont introduit une classification et une taxonomie des modèles d'intégration suivant une approche cognitive du problème. Elle semble avoir été assez bien adoptée par la communauté des chercheurs sur la parole audiovisuelle, c'est pourquoi il peut être intéressant de placer les problèmes d'asynchronie dans ce contexte.

La première catégorie contient les modèles *sans représentation commune*, aussi appelés *modèles à identification directe*. Dans ces derniers, un vecteur bimodal est soumis à un unique classifieur. Certains modèles de cette catégorie peuvent intégrer une gestion de l'asynchronie (modèle produit, maître-esclave, etc.), d'autres non (MMC de type classique, RNMDT de type classique ou multi-état).

La deuxième classe comprend les modèles à représentation commune et *intégration tardive*, aussi appelés *modèles à identification séparée*. Dans ces derniers, la fusion s'opère après classification sur chacune des modalités. Nous trouvons donc dans cette classe des modèles gérant l'asynchronie (Méthodes N-MEILLEURES mais l'on peut difficilement imaginer que des méthodes synchrones en fassent partie : La classification, comme la segmentation est réalisée séparément pour chacune des sources d'informations.

Les troisième et quatrième catégories sont constituées par les modèles à intégration précoce et représentation commune. Dans la troisième, l'espace de représentation commun est la modalité dominante. Dans la quatrième, c'est un espace *amodal*, l'espace des paramètres articulatoires en l'occurence. Du point de vue de l'asynchronie, ces deux catégories n'ont pas lieu d'être puisque les modèles en faisant partie peuvent être rangés dans les deux premières classes. En effet, le recodage dans un espace commun ne change pas cette problématique.

En conclusion, la gestion de l'asynchronie peut s'appréhender dans les 4 catégories. Il faut cependant noter qu'elle est implicite pour les modèles à identification séparée.

1.3.2 Fusion sur les scores

Le Décodage N-MEILLEURES

Une possibilité, proposée par Alissali et al. (1996) consiste a réaliser un décodage acoustique proposant les N meilleures suites de mots possibles auxquelles sont associées leurs probabilités. Pour chacune de ces suites de mots, il est alors possible de calculer un score visuel et, par conséquent, trouver celle qui a la plus forte probabilité conjointe.

Dans une première approche, un système acoustico-labial à intégration *directe* propose N meilleures suites de phonèmes possibles. Ces séquences sont évaluées, sans remise en cause de frontières inter-unités, par un modèle purement labial.

Une fonction de décision linéaire appliquée aux sorties de ces deux sous-systèmes fournit ainsi la séquence de phonèmes reconnue. Cette méthode est enrichie par l'utilisation de visèmes au niveau du modèle visuel. Le nombre de visèmes étant largement inférieur au nombre de phonèmes, on permet ainsi une meilleure estimation des paramètres du modèle visuel.

Les résultats expérimentaux montrent pourtant une dégradation non significative des résultats, justifiée par les auteurs par un choix inapproprié des classes de visèmes.

L'apport d'une telle approche, relatif à la gestion dynamique de l'asynchronie acoustico-labiale est indéniable. Cependant, si l'asynchronie est ici gérée au niveau de la phrase, elle l'est de façon incomplète.

Plaçons-nous à un niveau théorique et considérons un système qui choisirait pour solution parmi l'ensemble E_1 des N_1 meilleures solutions de la modalité 1 et l'ensemble E_2 des N_2 meilleures solutions de la modalité 2, celle qui a la plus forte probabilité conjointe. Pour pouvoir affirmer que la solution proposée est la meilleure, il faut qu'elle appartienne à $E_1 \cap E_2$.

En effet, si cette solution S_i (suite de mots) de probabilité $p_{1,i}$ et $p_{2,i}$, optimale sur $E_1 \cup E_2$ appartient à $E_1 \cap E_2$, alors toute solution S_j de probabilités $p_{1,j}$ et $p_{2,j}$ extérieure à $E_1 \cup E_2$, vérifiera : $p_{1,j} \times p_{2,j} < p_{1,i} \times p_{2,i}$, la définition du système imposant : $p_{1,j} < p_{1,i}$ et $p_{2,j} < p_{2,i}$. Dans le cas contraire, il peut exister une $N + x$ meilleure solution de la modalité 1 qui soit supérieure en probabilité conjointe à cette solution S_i. Dans ce cas, nous n'obtenons pas réellement la solution de l'équation 1.1. Nous ne pouvons donc pas affirmer que l'asynchronie a été gérée de façon optimale.

Nous pouvons prendre un exemple pour nous en persuader :

- Soit un vocabulaire : $A, B, C, ..., Y, Z$
- Soit $N = 3$ le nombre des meilleures solutions S_{1_i} de la modalité 1.
- Soit $N' = 2$ le nombre des meilleures solutions S_{2_j} de la modalité 2.
- Soit $p(m, S_{m',n})$ le score pour la modalité m de la solution $S_{m',n}$.

Les différentes valeurs de $p(m, S_{1,i})$ sont :

$S_{1,i}$	$p(1, S_{1,i})$	$p(2, S_{1,i})$	$p(1, S_{1,i}) \cdot p(2, S_{1,i})$
ABC	0.3	0.01	0.003
AABD	0.2	0.005	0.001
ABECE	0.1	0.004	0.0004

Les différentes valeurs de $p(m, S_{2,i})$ sont :

$S_{2,i}$	$p(2, S_{2,i})$	$p(1, S_{2,i})$	$p(1, S_{2,i}) \cdot p(2, S_{2,i})$
APC	0.1	0.003	0.0003
AAPD	0.09	0.002	0.00018

La meilleure solution bimodale est ABC avec une score conjoint de 0.003. Cependant, il peut exister une solution $S' =$ ABED telle que $p_{1,S'} = 0.09$ est inférieur à $p(1, S_{1,N})$ et $p_{2,S'} = 0.08$ est inférieur à $p(2, S_{1,N'})$ mais dont le score conjoint $p_{1,S'} \cdot p_{2,S'} = 0.0072$ est supérieur à 0.003.

On peut également noter que vouloir conserver le principe d'optimalité dans la recherche des N *meilleures solutions* (qu'elles soient différentes à un niveau phonétique ou à un niveau segmental) amène à des problèmes de complexité liés à la valeur de N.

La figure 1.1 représente le système N-MEILLEURES développé au LIUM par Alissali et al. (1996).

Les réseaux de neurones à délai temporel multi-états

Bregler et al. (1993) proposent une architecture basée sur les RNMDT (voir figure 1.2).

Deux RNMDT permettent de classifier d'une part les données acoustiques et d'autre part les données visuelles et par conséquent de les associer à deux états, l'un acoustique, l'autre visuel. À un couple d'états acoustique et visuel est associé un état dans la couche des états bimodaux. Une couche d'alignement (DTW) permet de trouver le chemin d'états bimodaux optimal, et par conséquent de générer les hypothèses d'unités de reconnaissance (voir figure 1.2).

En fait, cette architecture est assez proche des MMC de type classique, l'alignement se faisant après combinaison de la classification des deux sources d'information. La dépendance temporelle des deux sources est une hypothèse très forte dans un tel système.

1.3.3 Fusion sur les lois : Les modèles maître-esclave

Une seconde possibilité est l'utilisation d'un modèle acoustique *piloté* par un modèle labial (Jacob et Senac, 1996), voir Figure 1.3.

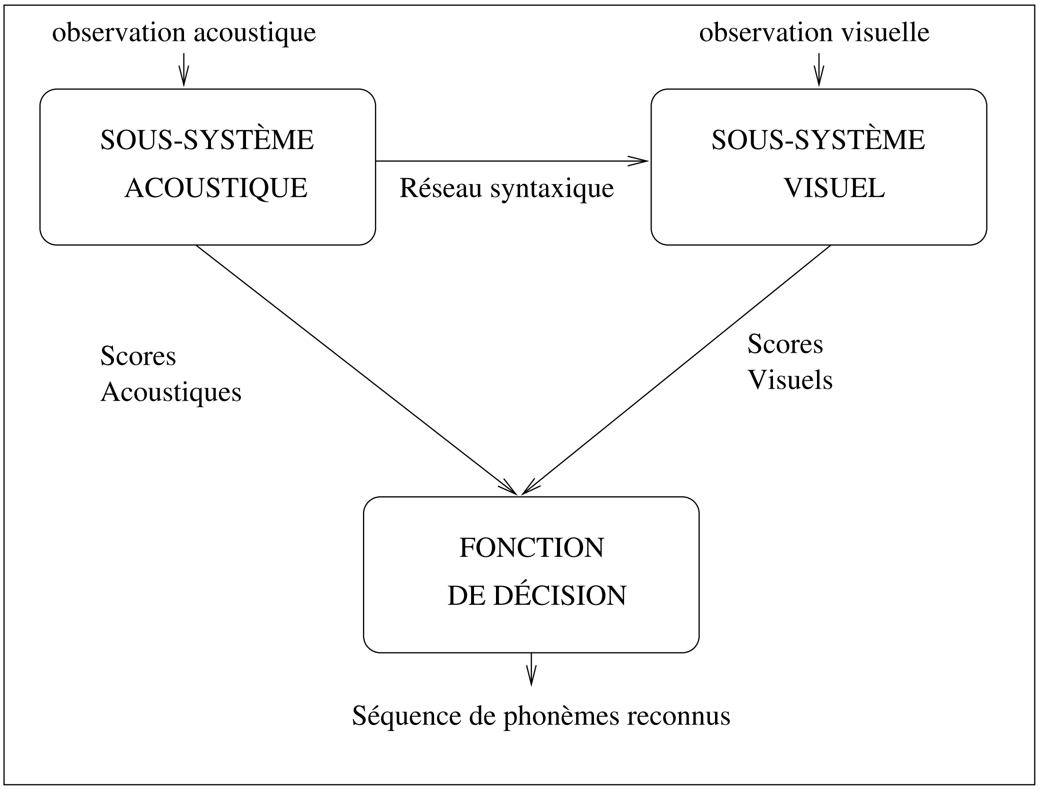

FIG. 1.1: Schéma du système N-MEILLEURES de Alissali et al. (1996)

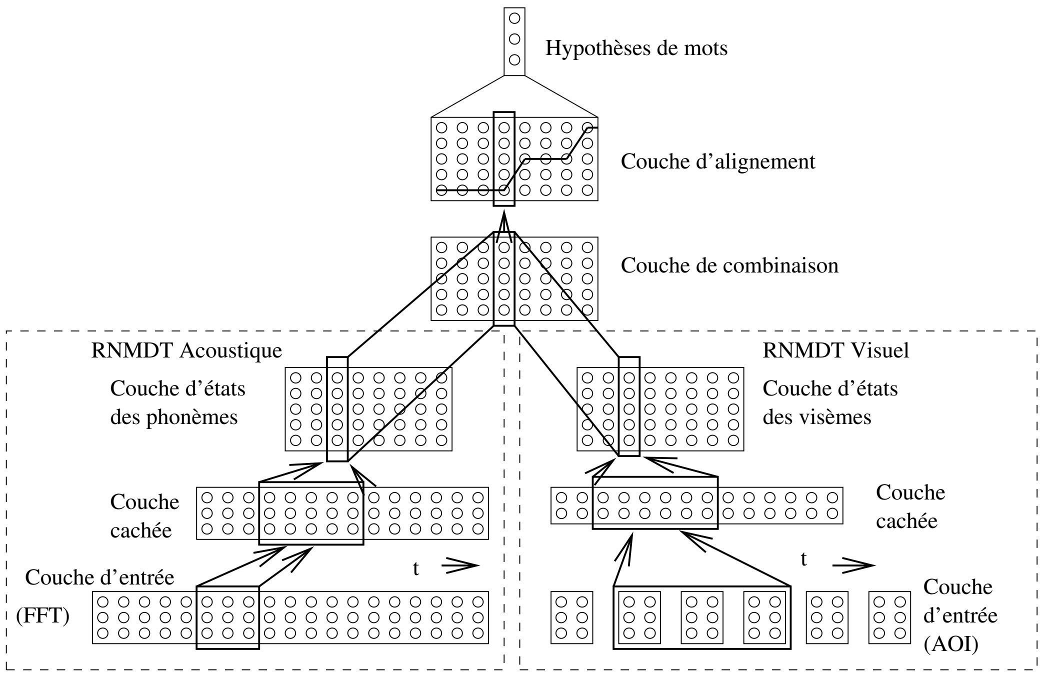

FIG. 1.2: Architecture du système de Bregler et al. (1993)

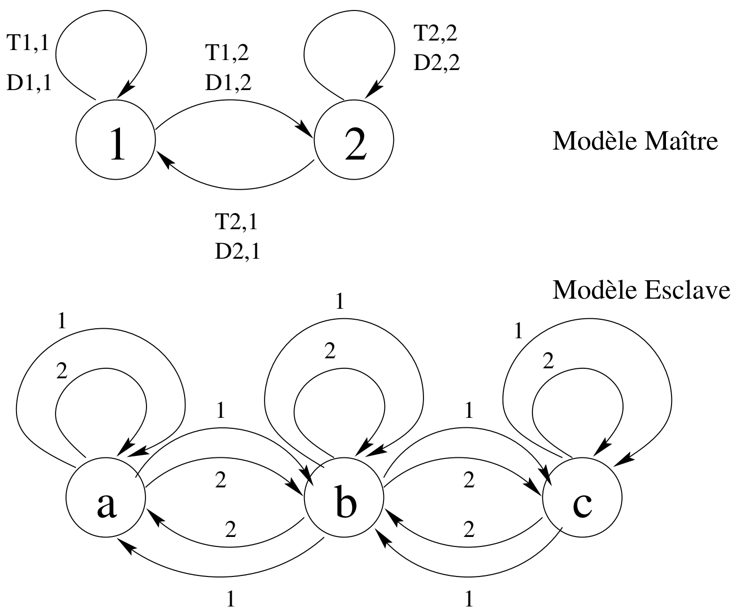

FIG. 1.3: Exemple de modèle maître-esclave

Les bases théoriques d'une telle approche proviennent des travaux de Brugnara et al. (1992). Un MMC *parallèle* λ est défini comme un quadruple processus stochastique (X', Y', X'', Y''), Y' et Y'' étant les séquences d'observations de deux processus stochastiques et X' et X'' les processus cachés associés.

Dans le cas d'un modèle maître-esclave, les paramètres de $\lambda'' = (X'', Y'')$ sont une *fonction probabiliste* de l'état de $\lambda' = (X', Y')$. Autrement dit, pour passer d'un état de λ'' vers un autre, il y a autant de valeurs de probabilités de transition et d'émission qu'il y a d'états dans λ'.

L'indépendance temporelle est donc traitée ici comme une dépendance entre deux processus stochastiques. Pratiquement, un modèle maître à N états représentant l'ensemble des configurations labiales possibles conditionne les distributions et transitions des modèles acoustiques correspondants à chacune des unités de reconnaissance.

Sont donc étudiées N possibles mises en correspondance d'observations acoustiques et labiales pour chaque trame t. L'asynchronie sera donc gérée sur une fenêtre temporelle de N trames.

Pour passer d'un état du modèle acoustique à un autre, il existe autant de transitions qu'il y a d'états dans le modèle maître (labial). Le modèle ainsi créé doit être soumis à une phase d'apprentissage qui, étant donné le grand nombre de paramètres à estimer, demande un très grand nombre de données.

En pratique, avec un corpus de faible taille, il faudra simplifier énormément le modèle labial. Au niveau expérimental, il s'agira d'un unique modèle à 3 états (lèvres ouvertes, semi-ouvertes ou fermées) qui *pilotera* tous les modèles acoustiques.

1.3.4 Décomposition et recombinaison

Cette section fait état de travaux effectués dans d'autres domaines que celui de la parole bimodale mais dans lesquels l'information est considérée comme étant dissociable. Les problèmes abordés sont donc très proches, à un niveau théorique, de ceux traités dans ce chapitre.

Décomposition en parole et bruit

Les travaux de Varga et Moore (1990) (repris par Gales et Young (1992)) se situent dans le cadre de la reconnaissance de parole en environnement sonore bruité. Ils s'écartent du schéma classique de pré-débruitage en proposant d'intégrer une reconnaissance du bruit à celle de la parole.

Leur approche est caractérisée par la combinaison de deux modèles, le premier étant entraîné sur le signal de la parole *propre*, l'autre uniquement avec les différents bruits.

Pour ce modèle *produit*, la probabilité d'être au temps t dans l'état i du modèle de parole et dans l'état j du modèle de bruit est :

$$P_t(i,j) \quad = \quad max_{u,v} P_{t-1}(u,v) \cdot a1_{u,i} \cdot a2_{v,j} \cdot b1_i \otimes b2_j(O_t)$$

où $a1_{u,i}$ est la probabilité de transiter de l'état u vers l'état i dans le modèle de parole, $a2_{v,j}$ la probabilité de transiter de l'état v vers l'état j dans le modèle de bruit et $b1_i \otimes b2_j(O_t) = \int P(O1_t, O2_t|i,j)$ la probabilité conjointe d'émission de O_t.

Nous ne détaillerons pas le processus permettant de séparer l'observation liée au signal de parole de celle liée au bruit et qui permet de calculer cette probabilité d'émission. Le principe de ce processus consiste à étudier les corrélations temporelles des différents bancs de filtres.

Notons simplement que le processus de décodage mettra en relation tous les états du premier modèle avec tous ceux du second. Excepté le calcul des probabilités d'émission, cette combinaison est ce que nous appellerons par la suite *produit de* MMC.

Je montrerai dans la section 1.4 que ce type de modèle est très approprié à la gestion des phénomènes d'asynchronie.

Décomposition en bandes de fréquences

Bourlard et Dupont (1996), dans le cadre de la parole unimodale, proposent de traiter indépendamment les différentes bandes de fréquence présentes dans le spectre acoustique. Les motivations d'une telle approche sont très semblables à celles que l'on retrouve dans le cadre de la parole audiovisuelle :

- Le bruit peut n'affecter que certaines bandes de fréquences.
- L'information pertinente se situe à des endroits différents pour des unités de reconnaissance différentes.
- Gestion des phénomènes d'asynchronie.
- Modélisation adaptée à chaque sous-bande de fréquence.

La topologie utilisée ici (voir figure 1.4) est une mise en parallèle des modèles correspondant à chaque modalité. Les états initiaux et finaux ont une fonction de recombinaison, synchronisation et pondération (voir section 2.3.1).

Des travaux similaires ont été réalisés par Tibrewala et Hermansky (1997). D'autre part, récents travaux sur la séparation des différentes bandes de fréquence ont été récemment réalisés au LIA par Besacier et Bonastre (1997).

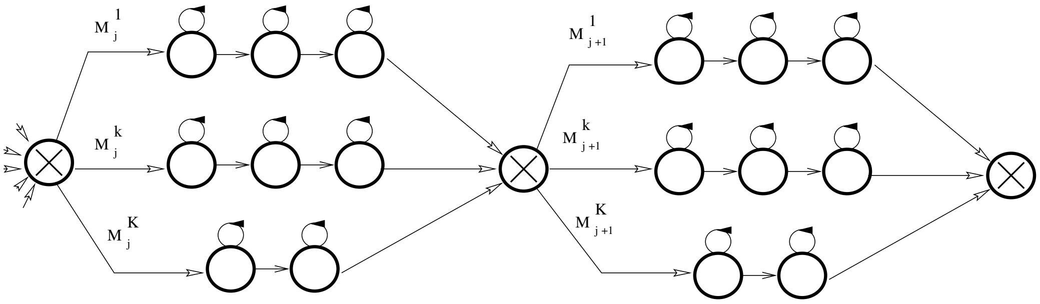

FIG. 1.4: Modèles multi-bandes de Bourlard et Dupont (1996)

Ils diffèrent cependant de ceux de Bourlard et Dupont (1996) dans leur optique (division optimale du domaine fréquentiel), la méthode de traitement (statistiques du second ordre) et par leur domaine d'application (reconnaissance du locuteur).

1.4 Le produit de modèles

1.4.1 Introduction

Nous proposons un autre type de combinaison dont l'idée forte consiste à réaliser au décodage le simple produit d'un modèle acoustique et d'un modèle labial (voir figure 1.5).

Empiriquement nous voyons que ce modèle nous permettra d'obtenir des séquences d'états différentes au niveau acoustique et au niveau labial. En effet, le modèle acoustique peut boucler sur un état pendant que le modèle labial transite d'un état vers le suivant. La figure 1.6 représente l'association états-observations du chemin marqué en gras sur la figure 1.5.

Avoir une vision empirique du comportement d'une telle topologie est important mais en aucun cas suffisant. Nous définissons donc tout d'abord ce qu'est le produit de deux automates à transitions valuées. Nous montrons alors quelles propriétés y sont attachées. Nous appliquons ensuite ces définitions aux modèles de Markov cachés continus et ce, dans le but d'utiliser ce produit dans des systèmes de reconnaissance de la parole.

1.4.2 Définition d'un automate à transitions valuées

Un automate à transitions valuées (ATV) est défini par $(S, \mathcal{O}, (I, \cdot), p, d)$, où :

- $\mathcal{O}$ est un ensemble quelconque,

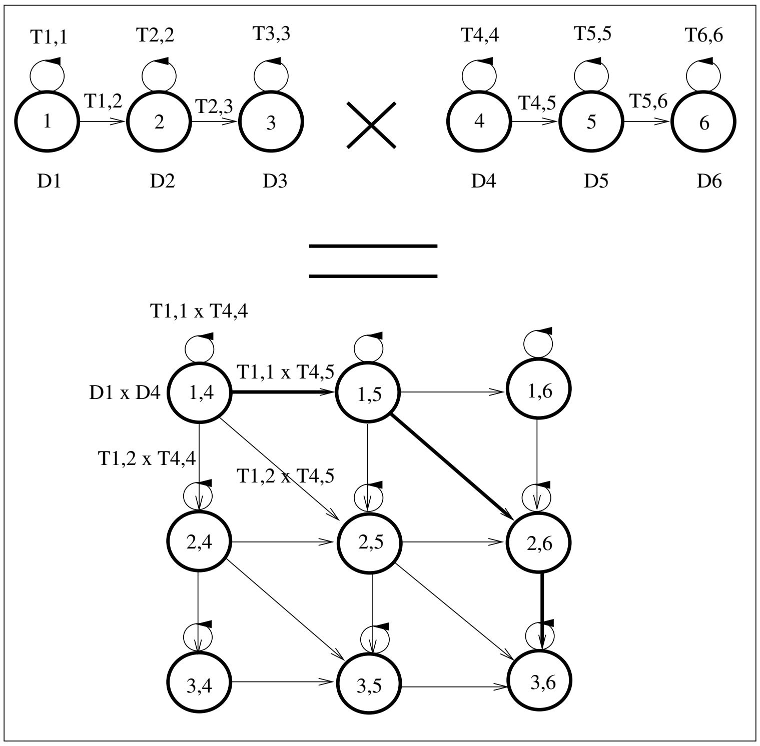

FIG. 1.5: Exemple de produit

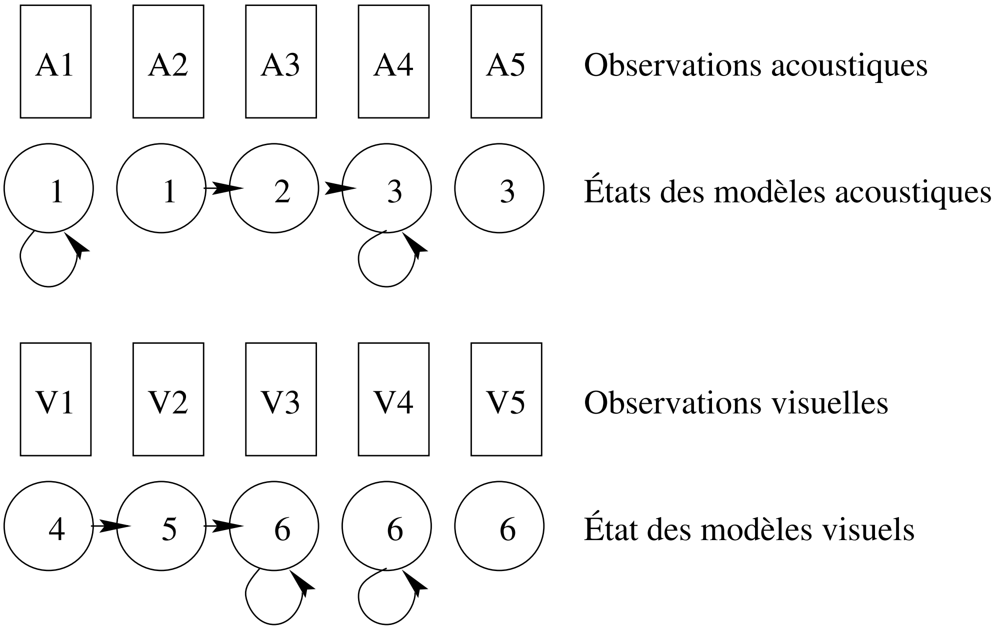

FIG. 1.6: Exemple d'associations états-observations

- S est un ensemble fini quelconque,
- $(I, \cdot)$ est un demi-groupe commutatif,
- L est l'ensemble des fonctions de $\mathcal{O}$ dans I,
- p est une fonction de S^2 dans I,
- d est une fonction de S dans L.

1.4.3 Définition du produit de deux ATV

Considérons les deux ATV A et A', respectivement définis par $(S, \mathcal{O}, (I, \cdot), p, d)$ et $(S', \mathcal{O}', (I, \cdot), p', d')$. Nous appellerons produit de A et A' l'opération " $*$ " telle que $A'' = A * A'$ défini par $(S'', \mathcal{O}'', (I, \cdot), p'', d'')$, où :

- $S'' = S \times S'$
- $\mathcal{O}'' = \mathcal{O} \times \mathcal{O}'$
- $p'' : (S \times S')^2 \to I$
- $d'' : S \times S' \to L''$
- L'' est l'ensemble des fonctions de $\mathcal{O} \times \mathcal{O}'$ dans I
- $p''(((s_w, s'_x), (s_y, s'_z))) = p((s_w, s_y)) \cdot p'((s'_x, s'_z)), \forall (s_w, s_y) \in S^2,\ \forall (s'_x, s'_z) \in S'^2$
- $d''((s, s')) = l'' \in L''$ telle que $l''((o, o')) = l(o) \cdot l'(o')$, où $l = d(s)$ et $l' = d'(s')$, $\ \forall (s, s') \in S \times S',\ \ \forall (o, o') \in \mathcal{O} \times \mathcal{O}'$

Il est évident que A'' est un ATV.

1.4.4 Évaluation des séquences de symboles produites en empruntant un chemin dans un ATV

Soit (C) un ensemble de suites d'éléments de $S \times \mathcal{O}$. Appelons évaluation d'une suite d'éléments de $S \times \mathcal{O}$ dans un ATV $A : (S, \mathcal{O}, (I, \cdot), p, d)$ la fonction f telle que :

$$f : A \times (C) \to I$$

$$f(S, \mathcal{O}, (I, \cdot), p, d, (c)) = d_{s_n}(o_n) \cdot \prod_{i=1}^{n-1} (d_{s_i}(o_i) \cdot p(s_i, s_{i+1}))$$

Où $n = |(c)|$, $(s_i, o_i) \in (c)$, $\ s_i \in S$, $\ o_i \in \mathcal{O}$, $\forall i$ tel que $1 \leq i \leq n$.

Propriété

Soit $A(S, \mathcal{O}, (I, \cdot), p, d)$, $A'(S', \mathcal{O}', (I, \cdot), p', d')$ et $A''(S'', \mathcal{O}'', (I, \cdot), p'', d'')$ trois ATV tels que $A'' = A * A'$. Considérons maintenant l'ensemble (C'') des couples de chemins $((c), (c'))$ de même longueur appartenant à $(C) \times (C')$.

La propriété suivante est vérifiée :

$$\forall((c), (c')) \in (C'')\ f(A, (c)) \cdot f(A', (c')) = f(A'', (c''))$$

où $\forall i$ tel que $1 \leq i \leq n\ :\ c''_i = ((s_i, s'_i), (o_i, o'_i)),\ (s_i, o_i) \in (c)$ et $(s'_i, o'_i) \in (c')$

Démonstration

$f(A, (c)) \cdot f(A', (c'))$
$= d_{s_n}(o_n) \cdot \prod_{i=1}^{n-1}(d_{s_i}(o_i) \cdot p(s_i, s_{i+1})) \cdot d'_{s'_n}(o'_n) \cdot \prod_{i=1}^{n-1}(d'_{s'_i}(o'_i) \cdot p'(s'_i, s'_{i+1}))$
où $(s_i, o_i) \in (c)$ et $(s'_i, o'_i) \in (c')\ \forall i\ 1 \leq i \leq n$
Or, $\cdot$ est une loi associative et commutative dans l'ensemble I :

$$= d_{s_n}(o_n) \cdot d'_{s'_n}(o'_n) \cdot \prod_{i=1}^{n-1} d_{s_i}(o_i) \cdot d'_{s'_i}(o'_i) \cdot p(s_i, s_{i+1}) \cdot p'(s'_i, s'_{i+1})$$

Par définition de d'' et de p'' :

$$= d''_{(s_n, s'_n)}((o_n, o'_n)) \cdot \prod_{i=1}^{n-1}(d''_{(s_i, s'_i)}((o_i, o'_i)) \cdot p''((s_i, s'_i), (s_{i+1}, s'_{i+1})))$$

Et par définition de (c'') :
$= f(A'', (c''))$

1.4.5 Évaluation d'une suite de symboles produite par un ATV

Soit (O) l'ensemble des suites d'éléments de $\mathcal{O}$ et (C) l'ensemble des suites d'éléments de $S \times \mathcal{O}$. Soit $(C)_{(o)}$ le sous-ensemble des suites de (C) tel que si (s_j, o_j) est le $j^{\text{ème}}$ élément de $(C)_{(o)}$ alors o_j est le $j^{\text{ème}}$ élément de (o).

Appelons évaluation d'une suite d'éléments de (O) pour un ATV donné, la fonction g telle que :

$$g : (S, \mathcal{O}, (I, \cdot), p, d) \times (O) \rightarrow I$$

$$g(S, \mathcal{O}, (I, \cdot), p, d, (o)) = \sum_{(c) \in (C)_{(o)}} f(S, \mathcal{O}, (I, \cdot), p, d, (c))$$

Propriété

Soit A, A' et A'' trois ATV tels que $A'' = A * A'$:

$$g(A,(o)) \cdot g(A',(o')) \ = \ g(A'',(o''))$$

$$\forall((o),(o')) \in (O) \times (O') \text{ tel que } |(o)| = |(o')|$$

Démonstration

$g(A,(o)) \cdot g(A',(o'))$

$= (\sum_{(c)\in(C)_{(o)}} f(S,\mathcal{O},(I,\cdot),p,d,(c)))\cdot(\sum_{(c')\in(C')_{(o')}} f(S',\mathcal{O}',(I,\cdot),p',d',(c')))$

$= \sum_{(c)\in(C)_{(o)},(c')\in(C')_{(o')}} f(S,\mathcal{O},(I,\cdot),p,d,(c))) \cdot f(S',\mathcal{O}',(I,\cdot),p',d',(c'))$

$= \sum_{(c'')\in(C'')_{(o'')}} f(S'',O'',(I,\cdot),p'',d'',(c'')))$

$= g(A'',(o''))$

1.4.6 Application aux modèles de Markov cachés

Un modèle de Markov étant un automate d'état finis, nous pouvons remplacer :

- $\mathcal{O}$ par $\mathcal{R}^n$
- S par l'ensemble des états du modèle.
- $(I,\cdot)$ par le demi-groupe commutatif $([0,1],\cdot)$ issu de R.
- L par l'ensemble des distributions de probabilités.
- p par la fonction qui associe une probabilité à chaque transition.
- d par la fonction qui associe une distribution de probabilités à chaque état émetteur.
- f par la fonction qui calcule la probabilité d'émettre une suite d'observations donnée en suivant un chemin donné dans un modèle donné.
- g par la fonction qui estime la probabilité *a posteriori* d'émettre une suite d'observations donnée avec un modèle donné.

Un modèle de Markov caché (MMC) est entièrement défini par (S,p,d).

Soient :

- $A(S,p,d)$, $A'(S',p',d')$, $A''(S'',p'',d'')$, trois MMC tels que $A'' = A * A'$ (on montre que $\forall s''_1, s''_2 \in S'', p''(s''_1, s''_2) > 0, \forall s''_1 \in S'', \sum_{s''_2 \in S''} p''(s''_1, s''_2) = 1$ et que $\forall s'' \in S'', \int d''_{s''}(O'')dO'' = 1$.
- (o) une suite de vecteurs de $\mathcal{O}$ et (o') une suite de vecteurs de $\mathcal{O}'$ telles que $|(o)| = |(o')|$.
- (s) un chemin d'états de A et (s') un chemin d'états de A'.
- (c), (c'), (c'') tels que $c_i = (s_i, o_i)$, $c'_i = (s'_i, o'_i)$ et $c''_i = ((s_i, s'_i), (o_i, o'_i))$ avec $c_i \in (c)$, $c'_i \in (c')$ et $c'_i \in (c')$ pour tout i compris entre 1 et $|(o)|$.

Par définition du produit de deux ATV, nous avons les propriétés suivantes :

$$\begin{aligned} f(A,(c)) \cdot f(A',(c')) &= f(A'',(c'')) \\ g(A,(o)) \cdot g(A',(o')) &= g(A'',(o'')) \end{aligned}$$

Pour une suite d'observations donnée, la probabilité calculée avec le modèle produit est égale au produit des probabilités données par le modèle acoustique et le modèle labial. L'originalité de cette approche est donc qu'elle gère de façon optimale et complète l'asynchronie à l'intérieur des unités de reconnaissance. Ceci nous permet d'évaluer les perturbations introduites par cette forme de variabilité dans des MMC de topologie classique.

De plus, le produit est réalisé après un apprentissage séparé des deux modèles. Cet aspect de la méthode, n'impose pas le partage d'une même topologie pour les deux modalités. D'autre part, contrairement aux modèles maître-esclave, il n'impose pas l'utilisation de modèles labiaux par trop simplifiés.

Les modèles ainsi obtenus sont utilisés dans l'algorithme classique du *token passing model* (Young et al., 1993). Dans cet algorithme de décodage, un jeton porte la probabilité d'être dans un état donné d'un modèle donné à un instant donné. Le passage à la tranche de temps suivante se caractérise par la propagation des jetons aux états successeurs, accompagné d'une augmentation de leur valeur. La figure 1.7 montre un exemple d'utilisation des modèles produit dans un réseau de type DAP (Décodage acoustico-phonétique). Nous pouvons remarquer que retenir uniquement les *diagonales* des modèles produit, revient à utiliser des modèles synchrones classiques. Les modèles synchrones classiques ne sont donc qu'un cas particulier des modèles produit.

D'autre part, réaliser le produit des probabilités de deux modèles unimodaux sur leur état final est équivalent lorsque la segmentation est fixée. En revanche, procéder de la sorte en parole continue serait une erreur : Sur la figure 1.7 on voit qu'un jeton sortant d'un état final d'un modèle produit serait ré-injecté dans le premier état d'un modèle unimodal. À ce niveau, il faudrait donc comparer une probabilité conjointe avec une probabilité unimodale (boucle sur le premier état émetteur).

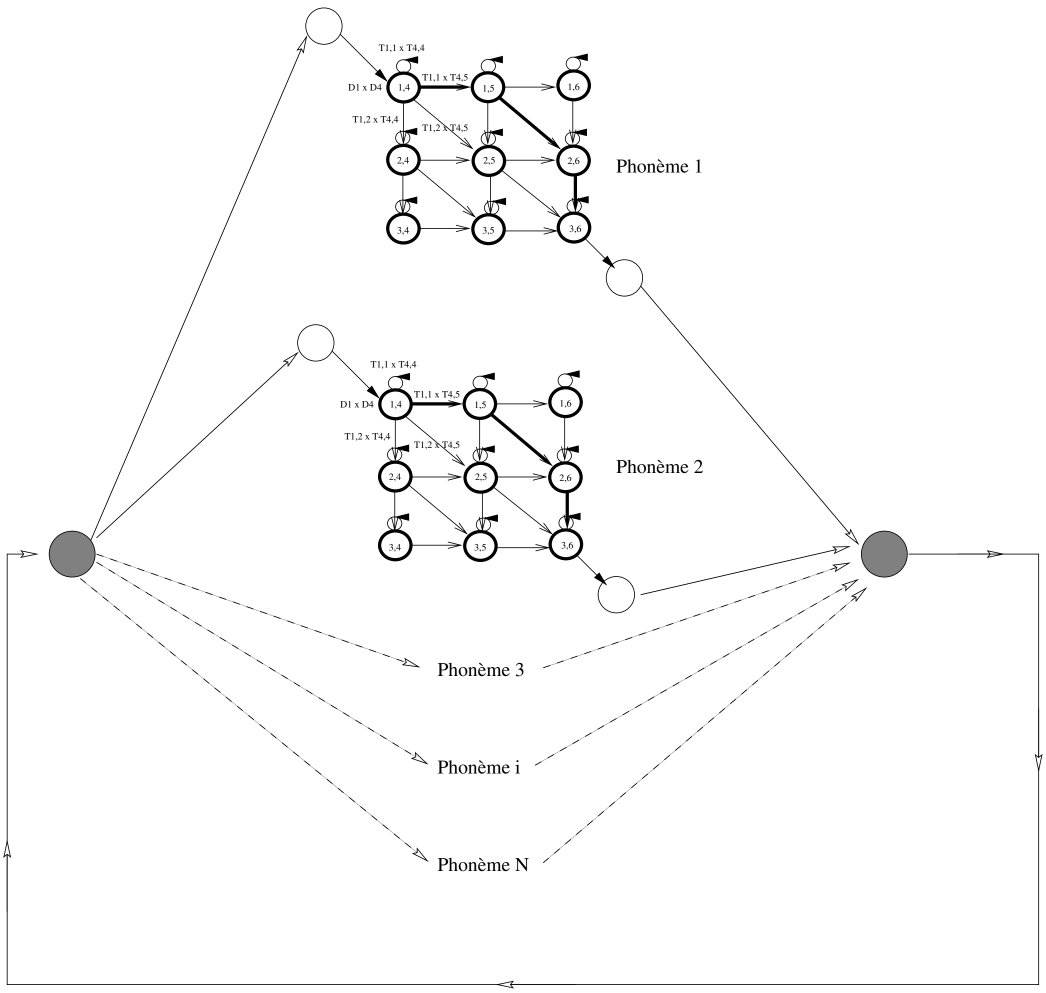

FIG. 1.7: Exemple d'utilisation du produit de MMC pour le décodage acoustico-phonétique

Dans son fonctionnement actuel, cet algorithme ne permet de gérer l'asynchronie qu'à l'intérieur des unités de reconnaissance. Autrement dit, la contrainte imposée à un tel système est le partage d'une unique segmentation pour les 2 types de sources d'information. Nous avons en revanche montré qu'il le fait de façon complète et optimale à l'intérieur d'un même modèle.

1.4.7 Parallèle entre modèles Maître-esclave et modèles produit

L'approche que nous proposons est d'une certaine façon assez proche des modèles maître-esclaves (MME). En effet, Brugnara et al. (1992) montrent une relation forte entre leur modèle et le nôtre. En fait, on peut transformer un MME entraîné en un MMC de *topologie produit*. Les paramètres d'un MME peuvent être définis pour une topologie *produit*, de la façon suivante :
Modèle maître :

- transitions : $a'_{x',z'} = Pr[X'_t = z'|X'_{t-1} = x']$
- émissions : $b'_{x',z'}(y') = Pr[Y'_t = y'|X'_{t-1} = x', X'_t = z']$
- probabilités initiales : $\pi'_{x'} = Pr[X'_0 = x']$

Modèle esclave :

- transitions : $a''_{x'',z'',z'} = Pr[X''_t = z''|X'_t = z', X''_{t-1} = x'']$
- émissions : $b''_{x'',x'',z'}(y'') = Pr[Y''_t = y''|X''_{t-1} = x'', X'_t = z']$
- probabilités initiales : $\pi''_{x'',x'} = Pr[X''_0 = x''|X'_0 = x']$

Le modèle MME correspondant est défini ainsi :

- transitions : $a_{x,z} = Pr[X_t = z|X_{t-1} = x]$
  $= a'_{x',z'} \times a''_{x'',z'',z'}$
- émissions : $b_{x,z}(y) = Pr[Y_t = y|X_{t-1} = x, X_t = z]$
  $= b'_{x',z'}(y') \times b''_{x'',x'',z'}(y'')$
- probabilités initiales : $\pi_x = Pr[X_0 = x]$
  $= \pi'_{x'} \times \pi''_{x'',x'}$

Inversement, entraîner un modèle de *topologie produit* en respectant les contraintes *Maître-esclave* sur les transitions ($t_{(x',y'),(x'',y'')} = t_{x',x''} \times t_{y',y'',x'}$, $\sum_{x''} t_{x',x''} = 1$ et $\sum_{y'''} t_{y',y'',x'} = 1$) et sur les probabilités d'émission revient à créer un MME (voir figure 1.8).

En revanche, ne pas respecter les contraintes de transitions lors de l'entraînement d'un modèle de topologie produit revient à créer un modèle standard. Bien entendu, toute augmentation du nombre de paramètres dans un modèle entraîne une plus grande faculté de modélisation, même en ce qui concerne l'asynchronie. C'est le cas si les contraintes produit, modélisées par des liens sur les transitions et les distributions ne sont pas respectées. Notons toutefois que la suppression des contraintes entraîne une très grande augmentation du nombre de paramètres et par voie de conséquence une augmentation de la taille du corpus nécessaire pour estimer ces paramètres. D'autre part, l'asynchronie serait *apprise* et non plus gérée dynamiquement.

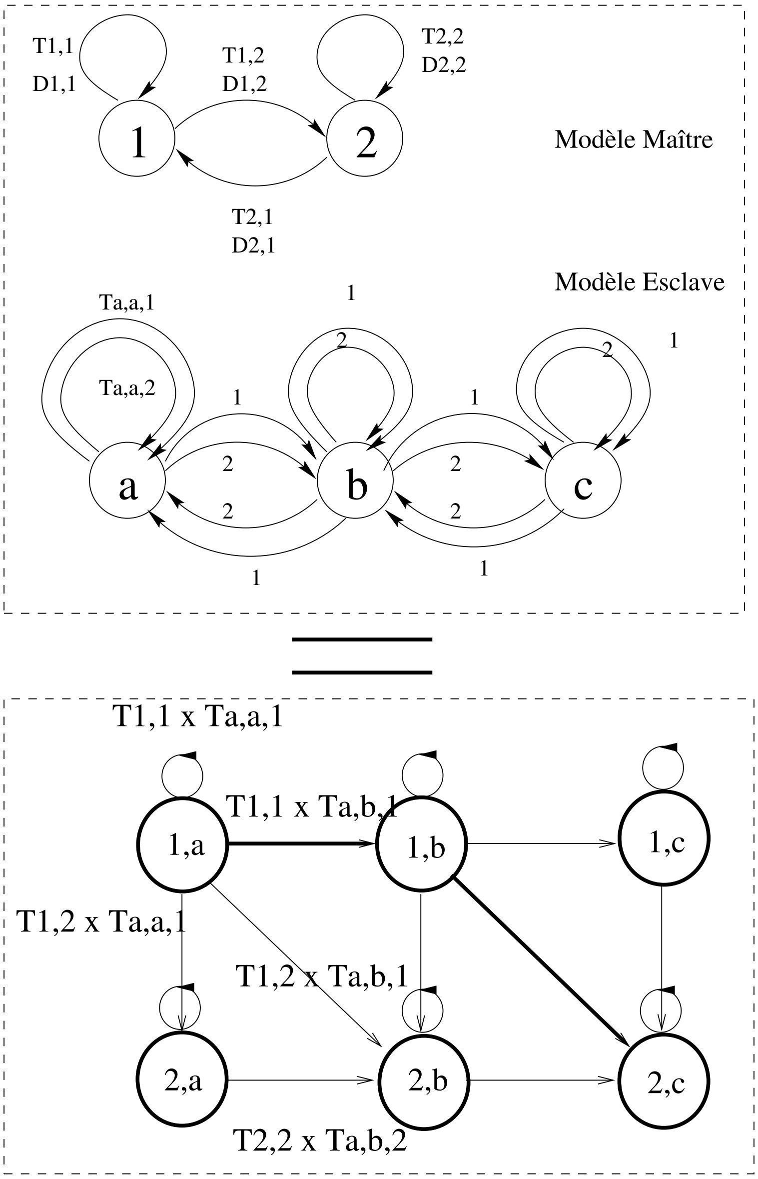

FIG. 1.8: Topologie produit des modèles maître-esclave

Respecter les contraintes de transitions et d'émissions *produit* ($t_{(x',y'),(x'',y'')} = t_{x',x''} \times t_{y',y''}$, $\sum_{x''} t_{x',x''} = 1$ et $\sum_{y'''} t_{y',y''} = 1$) lors de l'apprentissage revient à créer un modèle produit.

Enfin, en créant un modèle produit, que ce soit à partir de deux modèles entraînés séparément ou par apprentissage, nous posons la forte hypothèse de l'indépendance temporelle des deux sources d'informations. En revanche, les MME sont caractérisés par la dépendance temporelle d'un modèle sur l'autre. La prise en compte de cette dépendance temporelle provoque une augmentation conséquente du nombre de paramètres à estimer et par conséquent, du nombre de données d'apprentissage.

1.4.8 Conclusion

Nous avons vu, au cours de ce chapitre, les différentes approches proposées dans la littérature, qui intègrent une gestion des phénomènes d'asynchronie dans le processus conjoint de classification et segmentation. Nous avons alors introduit notre propre méthode et étudié ses relations avec les précédentes.

L'originalité de notre approche consiste en la démonstration de la gestion complète de l'asynchronie à l'intérieur des unités de reconnaissance et l'apprentissage séparé des deux modèles initiaux. Ce dernier point est crucial étant donné la faible taille des bases audiovisuelles disponibles : Les modèles labiaux peuvent être ainsi d'une complexité comparable à celle des modèles acoustiques. On peut notamment utiliser un modèle labial par unité de reconnaissance, et non pour l'ensemble du signal.

Un intérêt évident de cette méthode est la possibilité de fusionner deux MMC de topologies différentes. Cet aspect n'a pas été exploité au niveau expérimental, mais il va de soi que la topologie qui est optimale pour une source d'information ne l'est pas forcément pour l'autre.

Une telle approche présente l'avantage de ne pas nécessiter de modification des procédures d'apprentissage ou de décodage : Le modèle produit est un MMC de type classique. Il faut cependant noter que la complexité de la phase de décodage est augmentée. Néanmoins, nombre d'optimisations, optimales ou non, sont envisageables (liens sur les distributions, élagage par seuil, réduction du nombre d'états par classification). Ces avantages semblent particulièrement bien adaptés au problème de la reconnaissance de la parole audiovisuelle. On peut d'ailleurs noter que Tomlinson et al. (1996) adoptent exactement la même approche.

Enfin, sur le plan des applications, cette méthode est également valide dans le contexte de la parole unimodale. On peut considérer que le signal est composé de N flux de nature différente, par exemple les paramètres statiques et dynamiques ou les différentes bandes de fréquences, etc. Il s'agit de construire un ensemble de modèles pour chaque flux et de les fusionner par produit. C'est une idée que j'ai proposée dans (Jourlin, 1996a). On peut également la retrouver dans (Tomlinson et al., 1997).

Chapitre 2

Pondération

RÉSUMÉ

Nous abordons dans ce chapitre un des problèmes essentiels dans l'intégration de plusieurs sources d'information dans un même processus de classification.

En effet, fusionner plusieurs modalités n'a d'intérêt que si ces dernières sont complémentaires. Suivant le contexte, il faudra donc attacher plus ou moins d'importance à une source plutôt qu'à une autre.

Nous passons en revue différents systèmes élaborés dans divers domaines : fusion de bandes de fréquence, de paramètres statiques et dynamiques et, bien entendu, fusion d'informations acoustiques et visuelles.

Dans cette étude, nous nous focalisons sur le problème de la pondération, en mettant un éclairage particulier sur le niveau où elle est effectuée (données, modèles ou classification). Le problème de l'estimation des valeurs de cette pondération est étudié dans cette même optique.

Enfin, en prenant en compte ces différentes analyses, nous proposons une nouvelle méthode, applicable aux méthodes fondées sur les modèles de Markov cachés.

2.1 Introduction

Seule une faible partie de l'ensemble des paramètres articulatoires contrôlant la production acoustique de la parole est *visible*. Dans un environnement sonore protégé, la source acoustique contient par conséquent beaucoup plus d'informations sur le message oral ou l'identité du locuteur que l'information visuelle portée par le mouvement des lèvres.

Cela peut être le cas dans des conditions de laboratoire où l'on peut disposer de chambres sourdes et anéchoïques afin de réaliser des bases de données de parole de très bonne qualité acoustique.

En revanche, dans de nombreuses situations plus proches de la réalité, le canal acoustique peut perdre une grande part de sa fiabilité (bruits, conversations multiples, etc.) sans que le signal visuel ne soit perturbé.

De plus, malgré la supériorité de la source acoustique, les deux types d'informations ne sont pas redondantes : Prendre en compte les caractéristiques labiales peut mener à une amélioration des systèmes de traitement automatique de la parole. C'est en tous cas ce qu'indiquent les résultats de l'ensemble des recherches menées dans ce domaine.

C'est la raison pour laquelle, que l'environnement sonore soit protégé ou non, il est très important de pondérer la contribution relative de chacune des modalités au processus de classification.

Les valeurs de cette pondération peuvent être liées à divers critères tels que l'identité du locuteur, le phonème prononcé ou la qualité de chaque source d'information.

C'est pourquoi un problème majeur dans la réalisation d'un système de TAP audiovisuel est d'estimer une pondération optimale entre information labiale et acoustique et surtout de définir des critères d'optimalité adaptés à ce type de systèmes.

2.2 Fusion de données

Une première approche pour classifier des données multimodales est de considérer que l'ensemble des informations d'origines différentes peut se ramener à un unique vecteur d'observation. Un seul classifieur permettra donc de modéliser l'information multimodale.

Dans notre cas, l'avantage d'une telle approche est que le processus d'estimation des pondérations ne souffre pas de la complexité induite par les différents processus mis en œuvre pour la classification et la segmentation. Cette façon de procéder est valide mais l'estimation ne peut-être réalisée ici que de manière subjective. En effet, si une partie des critères de fiabilité ou de qualité des données peut-être estimée uniquement sur celles-ci (*ex.* rapport signal/bruit), une autre partie est très dépendante de la modélisation et du processus de classification situé en amont.

En ce qui concerne non plus l'estimation, mais la pondération effective, il faut savoir que dans le cadre d'une modélisation statistique, la correspondance entre les valeurs des pondérations fixées pour l'obtention du vecteur multimodal et celles implicitement utilisées dans les processus de modélisation, de segmentation et de décision finale est très difficile à définir.

À moins d'utiliser un processus de classification différent, ce sera le cas pour les 3 types d'architecture correspondant à une fusion de donnée dans la classification introduite par Robert-Ribes (1995a) et reprise par Adjoudani (1997) (voir figure 2.1). En effet, la pondération est effectuée après modélisation uniquement dans les modèles *à identification séparée*.

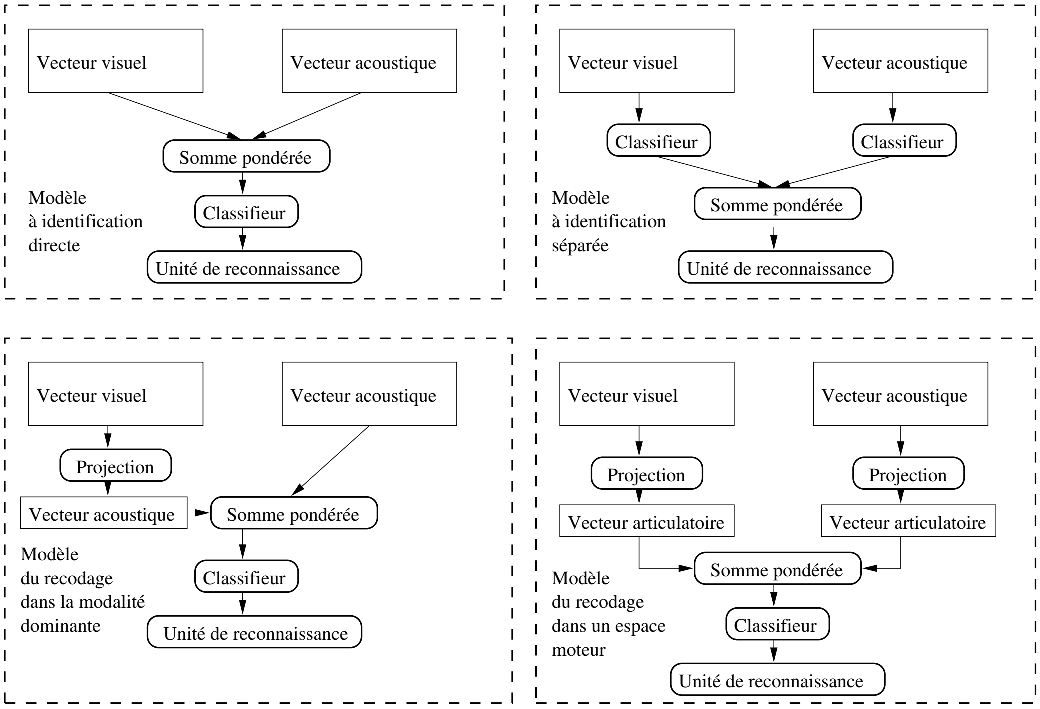

FIG. 2.1: Classification des systèmes d'intégration suivant Robert-Ribes (1995a)

Nous voyons donc l'intérêt d'effectuer une estimation et une pondération effective, au minimum en amont du premier processus de modélisation.

2.3 Fusion de scores

Dans un premier temps, il semble préférable de se placer dans l'hypothèse où les différents processus de classification des informations multimodales n'incluent pas de processus de segmentation. En effet, si la prise en compte de ce dernier peut augmenter considérablement la complexité d'une technique de pondération, elle ne change pas les différents types de critères qui peuvent être utilisés lors de la phase d'estimation.

De nombreuses règles peuvent être appliquées pour fusionner soit les scores provenant de N modèles, soit les décisions prises par N classifieurs, dans le cas où l'on dispose de N sources d'information. Quelles que soient l'approche ou les règles de fusion choisies, le problème se ramène de façon plus générale à celui d'une classification à N dimensions et à N' classes. Un cas simple et graphiquement représentable est la vérification bimodale du locuteur (classe *client* ou classe *imposteur*, scores acoustiques et visuels). La figure 2.2 en montre un exemple.

	Modalité 1	Modalité 2	
Classe 1	score 11	score 12	S11 + S12 / 2
Classe 2	score 21	score 22	S21 + S22 / 2

Règle de décision : Maximum du score moyen

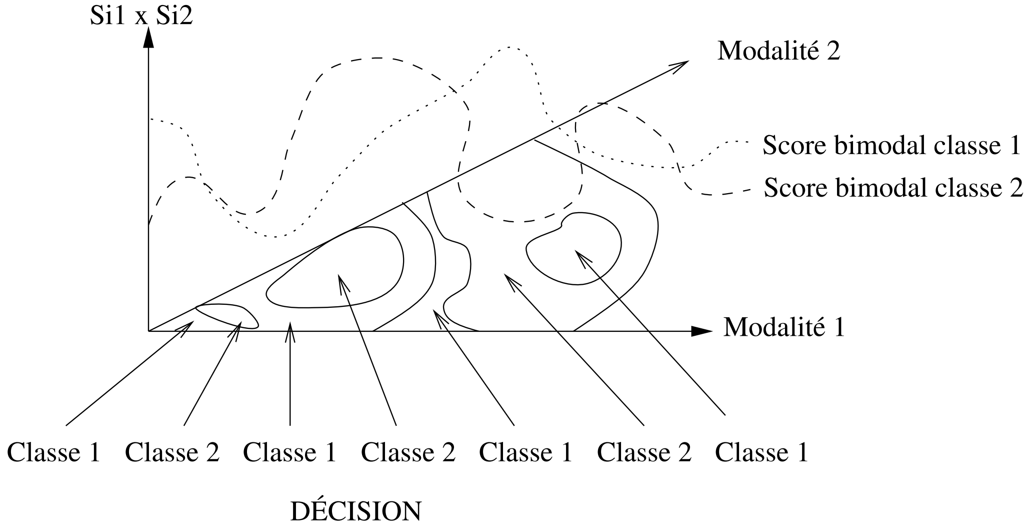

FIG. 2.2: Exemple de classification à deux classes et deux modalités

On peut se poser le problème de savoir quelles formes doivent prendre ces classes. Suivant la règle de fusion utilisée (somme ou moyenne pondérée, vote majoritaire, mixtures de gaussiennes ou réseaux de neurones), les séparations inter-classes peuvent prendre des formes différentes[1].

Quoi qu'il en soit, il ne s'agit que d'une étape de plus dans le processus de classification : La première étape permet pour chaque modalité m de transformer le vecteur (ou matrice) d'observation O_m en un vecteur V_m dont chaque composante $V_{m,c}$ représente la vraisemblance pour que l'observation appartienne à une classe donnée c.

On peut, dans une deuxième étape, facilement obtenir une classification pour chaque modalité (par exemple $\arg\max_c V_{m,c}$) et appliquer une règle de décision pour concilier les différents *experts* (Bigün et al., 1997).

Mais d'une manière plus générale, cette deuxième étape consiste globalement en un classement définitif de l'observation V. La figure 2.3 schématise ce processus général de fusion de scores.

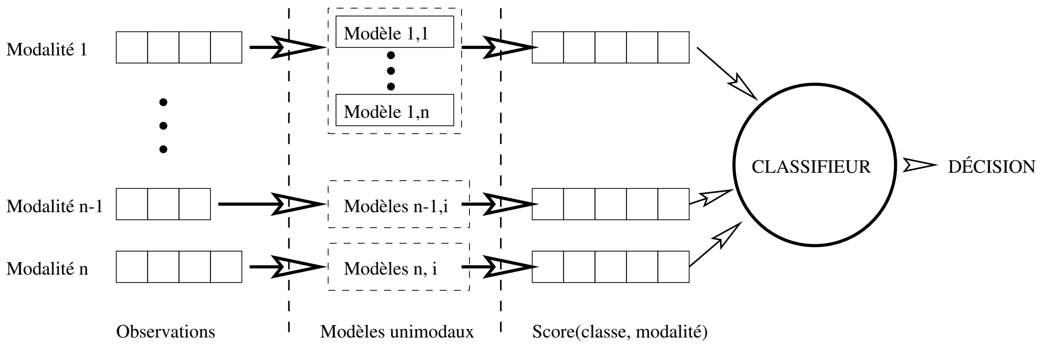

FIG. 2.3: Schéma général d'un système de fusion de scores

Par conséquent, quel que soit le type de classification choisi pour la première ou la deuxième étape du processus, il semblerait que l'élément commun et déterminant les performances du système multimodal soit celui de la pondération.

La question de savoir quel est le nombre de poids différents à utiliser (un poids pour chaque modalité, pour chaque classe, pour chaque sous-classe, etc.), comme celle de savoir quelle forme de classe utiliser (séparation linéaire, loi de décision, etc.) est liée à notre capacité à estimer correctement ces paramètres. Cette dernière est d'ailleurs très fortement liée au nombre et à la représentativité des données disponibles pour l'estimation. Autrement dit, bien que revêtant un aspect théorique, ces questions dépendent du corpus d'apprentissage et n'ont pas de réponse en dehors d'un cadre expérimental.

[1]voir (Kittler et al., 1997) pour une comparaison de différentes règles et stratégies d'intégration non pondérées.

En revanche, la question du critère d'estimation de cette pondération nous paraît revêtir une importance capitale. Un autre point reste de savoir à quel niveau se situer, celui de la modélisation ou celui de la classification.

Nous allons donc passer en revue les différentes approches décrites dans la littérature par niveau croissant, de l'estimation indépendante de la modélisation à l'estimation dépendante de la classification.

2.3.1 Estimation indépendante de la modélisation

Bourlard et Dupont (1996) proposent de pondérer la contribution de chaque bande de fréquence du signal acoustique par une estimation de son rapport signal sur bruit. Ces paramètres peuvent donc être estimés de façon dynamique, et cumulés avec d'autres critères de pondération.

C'est également ce que font Adjoudani et Benoît (1996), puis Rogozan et al. (1997) dans le cadre de la reconnaissance de la parole acoustico-labiale. Rogozan et al. (1997) déterminent dynamiquement un unique poids acoustico-labial qui est fonction du rapport signal sur bruit acoustique. Ils donnent leur préférence à ce type d'approche, plutôt qu'à une estimation sur un corpus d'apprentissage (Silsbee et Su, 1996) en raison de la trop faible taille du corpus dans leurs propres expérimentations.

Ils se basent par conséquent sur les travaux de Meier et al. (1996) [2]dans lesquels est proposée, entre autres, une fonction linéaire par morceaux, projetant le rapport signal sur bruit en pondération acoustico-labiale (voir figure 2.4).

Ces différentes mesures de qualité de l'information ont un avantage indéniable : Elles sont absolument indépendantes du système de classification utilisé et des données d'apprentissage. Néanmoins, ces critères sont subjectifs et peuvent se révéler assez éloignés du critère optimal : la réduction du nombre d'erreurs de reconnaissance.

2.3.2 Estimation dépendante de la modélisation

Critère d'entropie

(Meier et al., 1996) proposent d'utiliser un calcul d'entropie sur les scores (appelés dans ce cadre *activations*) acoustiques et visuels pour estimer les poids. L'entropie S_x est, dans ce cadre, un indicateur de la distance qui sépare les scores obtenus pour chacune des classes et pour une modalité donnée X,

[2]Méthode de classification : RNMDT-ME

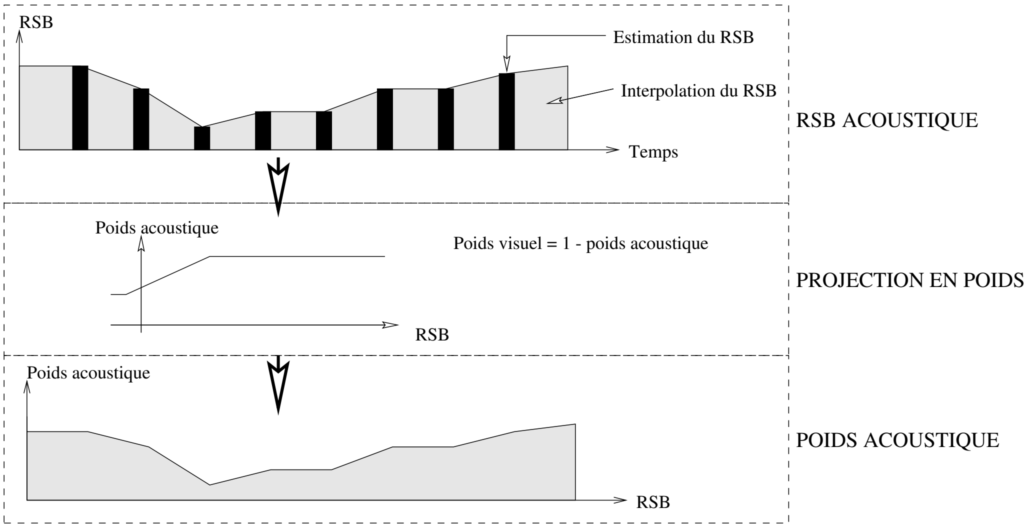

FIG. 2.4: Estimation d'un poids acoustico-labial, en fonction du rapport signal sur bruit et suivant Meier et al. (1996)

Les poids acoustique λ_A et visuel λ_V sont calculés de la manière suivante :

$$\lambda_A = b + \frac{S_V - S_A}{\Delta S_{\text{max-sur-corpus}}}, \text{ avec } \lambda_V = 1 - \lambda_A$$

où b est un biais permettant de corriger cette estimation en fonction de la qualité des données acoustiques.

Utilisation des rangs

Brunelli et Falavigna (1995), dans le cadre d'un système multimodal d'identification du locuteur, proposent une approche basée sur la dispersion des scores. $S'_{ij} \in [0,1]$ étant le score normalisé du classifieur j pour la personne i, le score final pour un modèle donné sera :

$$S_i = \left(\prod_j (S'_{ij})^{w_j} \right)^{1/\sum_j w_j}$$

avec :

$$w_j = \frac{S'_{i_1 j} - 1/2}{S'_{i_2 j} - 1/2} - 1$$

Le poids w_j affecté à un classifieur j est donc fonction de l'écart entre le meilleur candidat i_1 et le second i_2. Une autre technique de performances similaires bien que plus complexe est proposée, elle repose quant à elle sur l'ensemble des scores et des rangs pour chaque modalité.

Un point crucial dans cette approche est qu'elle repose sur l'hypothèse que les scores normalisés et les rangs affectés aux différents candidats pour une modalité donnée sont représentatifs de la fiabilité de cette dernière. La différence qu'il peut y avoir entre une fonction des rangs et la fiabilité ou la qualité réelle d'une modalité ou d'un modèle reste cependant à évaluer.

2.3.3 Estimation dépendante de la classification

Fonction d'un taux d'erreur

Une autre proposition de Bourlard et Dupont (1996) consiste à pondérer les différentes bandes de fréquences en fonction de leur fiabilité relative pour le système de reconnaissance. La méthode consiste à calculer un taux d'erreur sur les phonèmes en n'utilisant qu'une seule bande de fréquence comme source d'information. Ce taux, une fois normalisé sur l'ensemble des bandes servira de pondération pour la bande ainsi isolée. En utilisant un taux d'erreur par phonème, issu d'une matrice de confusion, on pourra donc obtenir un poids pour chaque unité de reconnaissance et pour chaque modalité.

Classification des scores

Bigün et al. (1997) à un niveau théorique et Duc et al. (1997) à un niveau expérimental, proposent l'utilisation d'un modèle bayésien de *conciliation d'experts* dans le cadre d'un système de vérification d'identité basé sur les diverses caractéristiques du visage et sur le signal de parole.

Le modèle théorique est exposé dans cet article dans un contexte très fort de vérification d'identité et il serait fastidieux d'en exposer les détails. Le principe général consiste à disposer d'un *superviseur*, modélisant la fiabilité de chaque *expert* pour chacune des classes. Les modèles *experts* ainsi que le modèle *superviseur* sont entraînés sur un corpus d'apprentissage.

Dans une proposition de Bourlard et Dupont (1996), la pondération est effectuée en utilisant un classifieur supplémentaire (perceptron multicouche), ayant pour entrée les scores de chaque modalité et en sortie, la décision finale.

Dans le cadre de la reconnaissance acoustico-labiale de la parole, un système de même architecture a été mis au point par Kabré (1995) : Deux ensembles de filtres flous sont entraînés pour extraire les traits acoustiques et labiaux. L'intégration des deux types de traits est réalisée par un module basé sur les réseaux de neurones.

C'est une approche similaire qu'utilisent Silsbee et Su (1996) et qui leur permet d'estimer des pondérations au niveau du phonème.

Nous voyons donc apparaître un certain engouement pour ce type d'approche de l'intégration. Il faut cependant noter que dans les systèmes décrits ci-dessus, le problème de la segmentation n'est pas abordé.

2.3.4 Synthèse des approches exposées

Nous avons vu ici se dégager une taxonomie du problème de la pondération : D'une part, la pondération doit-elle s'effectuer au niveau des données, de la modélisation, de la classification ? La même question se pose quant au critère sur lequel doit être fondée l'estimation des poids. Notons qu'il n'est en rien nécessaire d'apporter la même réponse aux deux questions.

Nous avons vu qu'il était difficile de pondérer les données elles-mêmes dans notre cadre précis (voir pages 30-31). En revanche, nous avons vu que la pondération devait se situer au niveau le plus proche possible de la modélisation, ceci afin d'avoir une influence sur la segmentation.

D'autre part, il est évident que notre but final est trouver une pondération qui optimise les performances du système. Or, ces performances sont évaluées après le processus de classification. C'est donc à ce niveau que devra être estimée la fiabilité relative de chacune des modalités.

Ces deux choix ont été à l'origine de nos propres travaux. Voyons maintenant comment ils peuvent se traduire dans des systèmes de classification et de segmentation d'informations multimodales fondés sur les MMC.

2.4 Pondération des émissions dans les MMC

2.4.1 Critères d'estimation des paramètres d'un MMC

Nous avons précédemment décrit les divers critères que l'on peut utiliser en vue d'estimer la fiabilité relative de plusieurs sources d'information ou de plusieurs types de modèles.

Il est maintenant nécessaire de décrire les critères les plus utilisés pour estimer les autres paramètres (probabilités de transition et d'émission) d'un système fondé sur les MMC.

Critère du maximum de vraisemblance

Ce critère est le premier à avoir été utilisé en reconnaissance de la parole avec des méthodes statistiques (voir Baker (1975), Jelinek (1976)). Il est cependant encore très utilisé de nos jours. Il s'agit de rechercher pour chaque modèle les valeurs des paramètres maximisant la vraisemblance de production des exemples d'apprentissage qui lui correspondent.

On peut s'apercevoir assez facilement qu'il ne sera pas adapté à la recherche de pondérations optimales : En effet, nous devons trouver la pondération $\alpha \in [0,1]$ telle que $\sum_e [\alpha \cdot V1(O1_e) + (1-\alpha) \cdot V2(O2_e)]$ soit maximum[3], e étant l'exemple considéré, $V1(O1_e)$ la log-vraisemblance du modèle pour l'observation $O1_e$ et la modalité 1, respectivement $V2(O2_e)$ pour la modalité 2.

Ce problème se ramène à trouver $\alpha \in [0,1]$ tel que $\alpha \cdot C_1 + (1-\alpha) \cdot C_2$, soit maximum. Or, $C_1 = \sum_e V1(O1_e)$ et $C_2 = \sum_e V2(O2_e)$ sont des constantes, on a donc :

$$\alpha = \begin{cases} 1 & \text{si } C_1 > C_2 \\ 0 & \text{si } C_1 < C_2 \\ \forall x \in [0,1] & \text{si } C_1 = C_2 \end{cases}$$

Ce critère est donc inutilisable en vue d'optimiser des pondérations, que ce soit par apprentissage ou de manière dynamique. C'est également la conclusion qu'a tirée Chow (1990) dans le domaine de la pondération des paramètres statiques vis-à-vis des paramètres dynamiques. Les problèmes sont très similaires, les paramètres statiques paraissant plus fiables que les dynamiques (Bocchieri et Wilpon, 1993).

De même, Normandin et al. (1994), del Álamo et al. (1995) et Hernando et al. (1995) utilisent des critères discriminants pour venir à bout de ce problème. Ce n'est que très récemment que (Hernando, 1997) utilise ce critère, en posant certaines contraintes sur les poids. Ces contraintes dépendent de deux constantes qui ne sont pas déterminées automatiquement.

Critères discriminants

En utilisant les critères basés sur les probabilités de classe, tels que *l'information mutuelle* et *l'entropie conditionnelle* on considère la probabilité qu'une observation ait été produite par un modèle particulier parmi un ensemble de modèles possibles (voir Viterbi et Omura (1979), Bahl et al. (1986), Bridle (1989), Niles et al. (1990)).

Il s'agit donc cette fois de maximiser, pour un exemple donné, le rapport de la log-vraisemblance du modèle lui correspondant et de la somme des log-vraisemblances des autres modèles. On peut aisément mesurer les avantages de tels critères. L'apprentissage devient *discriminant*, les paramètres d'un modèle étant optimisés pour maximiser sa différence avec les autres modèles.

Plaçons nous maintenant dans le cas de la recherche d'une pondération optimale, nous cherchons l'ensemble des poids $\alpha_m \in [0,1]$ correspondant aux modèles $m \in M$ maximisant :

[3] Il s'agit en fait de $\prod_e \left[V1(O1_e)^{\alpha} \cdot V2(O2_e)^{(1-\alpha)}\right]$, l'usage du logarithme nous permet cette facilité d'écriture.

$$\sum_{m \in M} \frac{\sum_e \left(\alpha_m \cdot V_{m,1}(O_{e,m,1}) + (1-\alpha_m) \cdot V_{m,2}(O_{e,m,2})\right)}{\sum_{i \neq m} \sum_e \left(\alpha_i \cdot V_{i,1}(O_{e,m,1}) + (1-\alpha_i) \cdot V_{i,2}(O_{e,m,2})\right)} \tag{2.1}$$

Où M est l'ensemble des modèles du système, $O_{e,i,s}$ l'observation associée à l'exemple e du modèle i pour la modalité s, $V_{i,s}$ la fonction de calcul de log-vraisemblance pour le modèle i et la modalité s.

La fonction à maximiser en fonction du vecteur de poids α est un rapport de polynômes de α. Cette fonction a donc une valeur maximale pour des valeurs de α_m différentes de 0 et de 1. Ce critère est donc théoriquement applicable aux cas de l'estimation des pondérations.

Il est de plus très proche du critère de choix utilisé lors du décodage. Mais une autre classe de critères peut se révéler encore plus adaptée au problème de l'estimation des pondérations.

Critères approchant le taux d'erreur

Il est évident qu'en utilisant un apprentissage *discriminant* nous nous rapprochons de la véritable valeur à optimiser. Mais nous pouvons aller encore plus loin : Ney (1995a) propose de se rapprocher le plus possible du critère idéal : minimiser le taux d'erreur du système.

Lorsqu'il s'agit d'estimer tous les paramètres du système, la méthode d'optimisation nécessitera de nombreuses itérations avec des calculs d'erreurs qui peuvent être très complexes. Il est alors indispensable d'utiliser une méthode à convergence *rapide*, par exemple basée sur une descente de gradient.

Pour cela, l'estimation de ce taux d'erreur devra être une fonction dérivable. Différentes méthodes permettent d'obtenir une telle fonction à partir d'un comptage des erreurs de classification.

Dans notre cas, nous n'avons que très peu de paramètres à estimer (un paramètre par unité ou sous-unité de reconnaissance), il devient donc envisageable d'utiliser une méthode d'optimisation à convergence *lente* et avec une fonction d'erreur *complexe*.

2.4.2 Description de la méthode utilisée

L'idée directrice de cette méthode est constituée par la recherche des valeurs de pondérations minimisant un taux d'erreur de classification sur les données d'apprentissage.

Calculer un taux d'erreur complet (suppressions, insertions et substitutions) impliquerait une recherche du chemin d'états et de modèles optimal pour chaque phrase de d'un corpus de développement. En termes de complexité algorithmique, il est quasiment impossible d'effectuer pour chaque itération de la méthode d'optimisation, ce décodage de *Viterbi* orienté état pour toutes ces phrases.

Par conséquent, nous avons dans un premier temps choisi d'ignorer, dans la phase d'estimation, les effets de la pondération sur la précision de la segmentation.

En revanche, il nous a semblé important que la nouvelle pondération puisse modifier l'alignement et la segmentation. La décision qui a été prise n'est donc pas de pondérer des scores calculés après un décodage mais de pondérer les probabilités d'émission avant la phase de décodage.

Le critère à minimiser étant une fonction de calcul du taux d'erreurs qui n'est pas dérivable, nous ne pourrons pas utiliser une méthode basée sur une descente de gradient. Néanmoins d'autres algorithmes sont applicables à ce problème, par exemple la méthode classique du *simplexe*.

2.4.3 Pondération des distributions

Pour chaque état de chaque modèle, la distribution est calculée de la façon suivante :

$$b_{m,j}(O_t) = b_{a,m,j}(O_{a,t})^{W_m} \times b_{l,m,j}(O_{l,t})^{1-W_m} \tag{2.2}$$

Où m est l'unité de reconnaissance (ou le modèle correspondant), j l'état considéré, O_t le vecteur d'observation au temps t, $b_{x,m,j}$ la fonction de distribution (produit de mixtures de gaussiennes) associée à la modalité x et aux observations $O_{x,t}$, et W_m le poids acoustico-labial correspondant au modèle m.

Il est important de préciser à ce niveau, que pondérer des probabilités n'a aucun sens en soi. La pondération effectuée ici est en fait une modification des variances, supposées mal estimées en raison d'un corpus d'apprentissage insuffisamment représentatif du corpus de test (ce qui peut être fait manuellement, voir (Wilpon et al., 1991)). L'utilisation d'un corpus de *développement* nous permet d'estimer cet écart, et de modifier en conséquence les valeurs des variances.

Malheureusement, les variances ne sont pas les seules à être modifiées et la fonction résultante de la pondération n'est pas une distribution de probabilité.

En effet : Soit M l'ensemble des modalités, E_m l'espace des observations pour la modalité $m \in M$, D l'ensemble des distributions $d_m : E_m \to [0,1]$ telles que $\int_{-\infty}^{+\infty} d_m = 1$.

On peut facilement vérifier que la distribution multimodale

$$\begin{aligned} d : (\prod_{m\in M} E_m) &\rightarrow [0,1], \\ d(x_1, x_2, \cdots, x_m, \cdots, x_n) &= \prod_{m\in M} d_m(x_m) \\ \text{est telle que : } \int_{-\infty}^{+\infty} d &= 1. \end{aligned}$$

Néanmoins, ce n'est pas le cas si $d(x_1, x_2, \cdots, x_m, \cdots, x_n) = \prod_{m\in M} d_m(x_m)^{w_m}$ où w_m est le poids affecté à la modalité m.

Signalons que dans la littérature, ce fait est très souvent [4] reconnu (Hernando et al. (1995), Normandin et al. (1994)), mais peu de travaux en tiennent compte, probablement pour des raisons de simplification du processus d'estimation.

Par exemple, Su et Silsbee (1996) proposent d'utiliser des distributions de probabilités classiques tant que la distance entre l'observation et la moyenne de la distribution considérée reste dans un certain intervalle. Dans le cas contraire, cette distribution est remplacée par une fonction pondérée, d'intégrale non égale à 1.

Dans un premier temps nous allons également ne pas tenir compte des contraintes probabilistes, ceci afin de simplifier la description de la méthode suivante. Nous verrons un peu plus tard comment rester dans un cadre probabiliste.

Considérons la suite d'observations (O_a, O_l) associée à un segment s et une suite d'états S du modèle m, de même longueur que (O_a, O_l). La probabilité pour que le modèle MMC classique m ait produit (O_a, O_l) en suivant un chemin d'états S est :

$$\begin{aligned} P((O_a, O_l), S|m) &= \prod_1^{|s|} a_{s_{i-1},s_i} \cdot b_{m,s_i}(O_{a,i}, O_{l,i}) \\ &= \prod_1^{|s|} a_{s_{i-1},s_i} \cdot b_{a,m,s_i}(O_{a,i})^{W_m} \cdot b_{l,m,s_i}(O_{l,i})^{1-W_m} \end{aligned}$$

Par conséquent : $\log P((O_a, O_l), S|m)$

$$= \sum_1^{|s|} \left[\log a_{s_{i-1},s_i} + Wm \cdot \log b_{a,m,s_i}(O_{a,i}) + (1 - W_m) \cdot \log b_{l,m,s_i}(O_{l,i})\right]$$

$$= \left(\sum_1^{|s|} \log a_{s_{i-1},s_i}\right) + Wm \cdot \left(\sum_1^{|s|} \log b_{a,m,s_i}(O_{a,i})\right)$$

[4] mis à part Rogina et Waibel (1990) qui affirment le contraire

$$+(1 - W_m) \cdot \left(\sum_{1}^{|s|} \log b_{l,m,s_i}(O_{l,i}) \right)$$

En posant :

$$\begin{aligned}
T_{s,m} &= \log P((O_a, O_l), S|m) \\
A_{s,m} &= \sum_{1}^{|s|} log b_{a,m,s_i}(O_{a,i}) \\
L_{s,m} &= \sum_{1}^{|s|} b_{l,m,s_i}(O_{l,i}) \\
C_{s,m} &= \sum_{1}^{|s|} log a_{s_{i-1},s_i}
\end{aligned}$$

nous obtenons l'équation suivante :

$$T_{s,m} = W_m \cdot A_{s,m} + (1 - W_m) \cdot L_{s,m} + C_{s,m} \tag{2.3}$$

2.4.4 Cadre probabiliste

Pour rester dans un cadre probabiliste, il ne s'agit donc plus de pondérer les distributions, mais les variances qui leur sont associées. En effet, augmenter ou diminuer artificiellement les variances d'une distribution permet de conserver les contraintes probabilistes.

D'autre part, en procédant de la sorte, nous mettons simplement en application l'hypothèse selon laquelle le corpus d'apprentissage est insuffisant pour décrire la variabilité des paramètres pour une distribution donnée.

Pratiquement, pour une observation donnée x, une distribution gaussienne $d_{\mu,\Sigma}(x)$ de moyenne μ et de matrice de variance-covariance Σ nous avons $w_1 > w_2 \Rightarrow d_{\mu,w_1 \cdot \Sigma}(x) > d_{\mu,w_2 \cdot \Sigma}(x)$. Les constantes w_1 et w_2 ont donc bien un effet de pondération sur les probabilités d'émission.

Cela n'implique pas de modification importante du système de pondération précédemment décrit. L'équation 2.2 devient :

$$b_{m,j}(O_t) = b_{a,m,j,W_m \cdot V_{a,m}}(O_{a,t}) \times b_{l,m,j,(1-W_m) \cdot V_{l,m}}(O_{l,t}) \tag{2.4}$$

où $b_{x,m,j,W \cdot V_{x,m}}$ est la distribution de probabilités associées à l'état j du modèle m et la modalité x mais dans laquelle les matrices de variance-covariance $V_{x,m}$ d'origine sont remplacées par $W \cdot V_{x,m}$.

L'équation 2.3 devient :

$$T_{s,m} = A_{s,m,W_m \cdot V_{a,m}} + \cdot L_{s,m,(1-W_m) \cdot V_{l,m}} + C_{s,m} \tag{2.5}$$

où A et L sont les log-vraisemblances calculées en utilisant respectivement $a_{m,j,W_m \cdot V_{a,m}}$ et $l_{m,j,(1-W_m) \cdot V_{l,m}}$. Pratiquement, il ne suffira plus de pré-calculer les valeurs de A et de L pour chaque modèle appliqué à chaque segment avant de chercher à optimiser les poids W_m. Le calcul de $T_{s,m}$ en fonction de W_m devient beaucoup plus long et le nombre de constantes à stocker avant la phase d'optimisation beaucoup plus important. C'est pourquoi les expérimentations menées au cours du chapitre 3 (pages 64 et 68) sont effectuées dans le cadre non-probabiliste. À un niveau théorique, rester dans un cadre probabiliste n'entraîne cependant aucun changement fondamental dans la méthode décrite dans la section suivante.

2.4.5 Estimation des poids

Le corpus d'apprentissage est séparé en deux parties distinctes. Nous appellerons A_1 la partie sur laquelle seront estimés les paramètres des modèles tels que les moyennes, variances et probabilités de transition. La partie A_2 sera réservée au calcul des pondérations.

Les données d'apprentissage étant segmentées et étiquetées, nous entraînons un ensemble de modèles avec des poids initiaux égaux. Pour chaque partie du corpus et respectivement chaque ensemble de modèles, nous utilisons alors la procédure suivante :

Pour chaque segment s des phrases d'apprentissage de A_2, nous calculons les constantes $A_{s,m}$ et labiales $L_{s,m}$ et $C_{s,m}$ relatives à chaque HMM m.

Considérons maintenant la fonction d'erreur suivante :

$$E(W) = \sum_s e(s, W)$$

Les composantes du vecteur W sont les poids acoustico-labiaux affectés à chaque modèle et :

$$e(s, W) = \begin{cases} 0 & \text{si le modèle qui donne la plus forte} \\ & \text{valeur de } T_{s,m} \text{ est l'étiquette pour } s \\ 1 & \text{sinon} \end{cases} \tag{2.6}$$

Bien que moins prometteurs car plus éloignés du critère idéal, les critères discriminants peuvent également être placés à ce niveau, par exemple :

$$e(s, W) = \frac{T_{s,n(s)}}{\sum_{m \neq n(s)} T_{s,m}} \tag{2.7}$$

Où $n(s)$ est la fonction qui associe au segment s le modèle étiqueté dans le corpus d'apprentissage. Nous utilisons maintenant la méthode du *simplexe*

(Nelder et Mead (1965), Daniels (1978)) afin de déterminer le vecteur W des poids optimaux, minimisant $E(W)$.

Les nouvelles valeurs des poids permettront au système non seulement de changer son comportement en classification, mais aussi, par voie de conséquence, en segmentation.

La phase d'apprentissage des modèles hors contexte consiste en l'estimation des transitions, moyennes et variances d'un modèle sur l'ensemble des segments lui correspondant. La phase d'apprentissage en contexte permet de remettre en cause ces frontières en réestimant les paramètres de la suite de modèles qui correspond à une phrase entière.

Afin d'ajouter à ce critère de minimisation du nombre d'erreurs de classification, la prise en compte des erreurs de segmentation, nous pouvons appliquer l'algorithme itératif suivant :

1. L'ensemble des modèles dont les pondérations ont été réestimées, est soumis à une nouvelle phase d'apprentissage en contexte, en utilisant le critère de maximum de vraisemblance.

   Nous pouvons remarquer que la segmentation implicitement utilisée durant cette étape prend en compte la nouvelle pondération.
2. Les nouveaux modèles fournissent de nouvelles vraisemblances (au sens de l'équation 2.3).
3. Les pondérations sont réestimées avec le nouveau jeu de vraisemblances.
4. Le processus entier est réitéré jusqu'à convergence des valeurs des pondérations.

La figure 2.5 résume ce processus. Ne disposant pas de preuve de la convergence de ce processus, il faudra fixer arbitrairement un nombre maximal d'itérations à effectuer. Notons que cette façon de procéder est très souvent utilisée dans les procédures classiques d'estimation par maximum de vraisemblance.

2.5 Conclusion

En étudiant différents travaux relatifs à des problèmes de fusion de sources d'informations, de connaissances, voire même de compétences, nous avons vu émerger deux notions : La fusion de données et la fusion de décisions. Sans vouloir rejeter complètement cette classification, on peut facilement concevoir que la limite entre les deux approches est floue.

Ainsi, dans la réalisation d'un système combinant différents types d'information, de connaissances ou de méthodes, la fusion peut apparaître à tous les niveaux du processus qui mène à la décision finale.

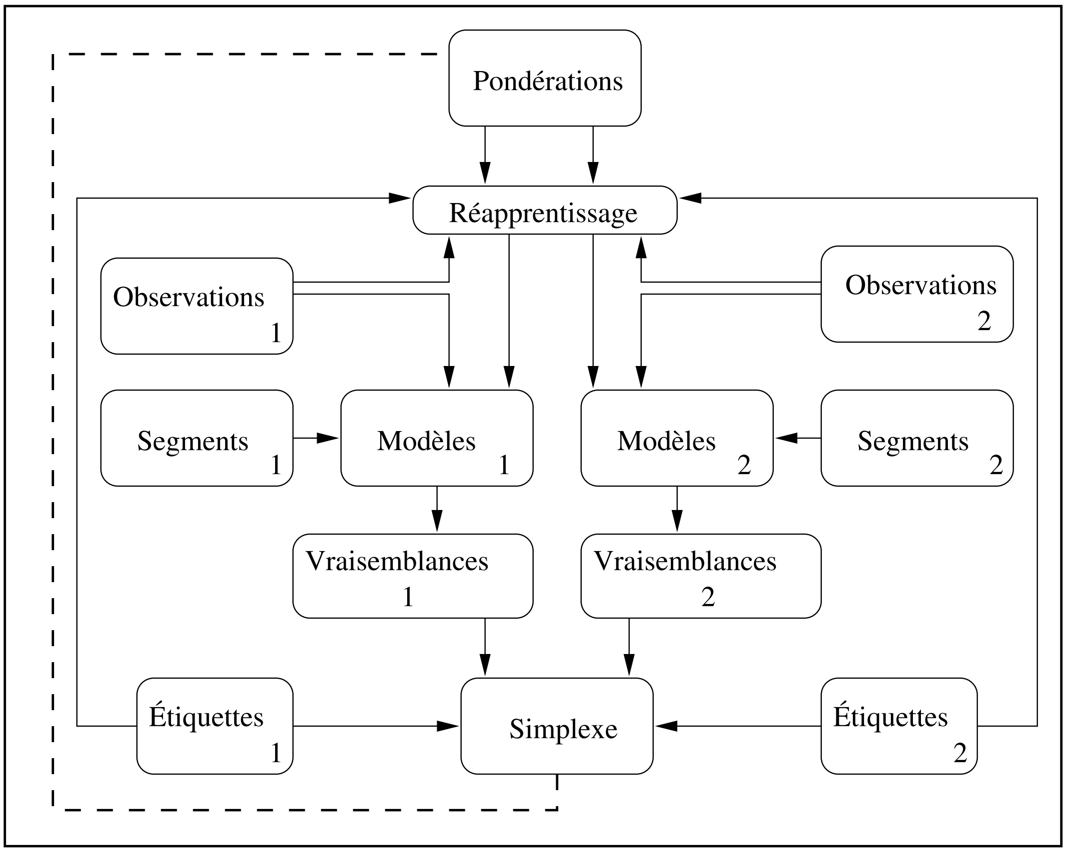

FIG. 2.5: Système d'estimation des pondérations

De même, la pondération telle qu'elle a été définie en introduction peut apparaître dans tous les niveaux d'intégration et donc être contenue dans des termes tels que normalisation ou projection dans un espace commun.

Après avoir étudié les différents critères pouvant mener à une estimation correcte des différentes pondérations affectées à chaque source d'information, nous avons fait le choix de nous rapprocher le plus possible d'un critère idéal.

Nous avons alors mis au point une méthode permettant de maximiser ces pondérations sur un critère théoriquement très proche du critère idéal : celui du nombre d'erreurs de classification. L'utilisation d'un processus itératif, nous a permis de modifier ce critère pour y inclure une prise en compte des erreurs de segmentation.

Les pondérations étant appliquées au niveau du calcul de la probabilité d'émission d'une trame d'information nous ne nous situons que très légèrement au-dessus du niveau de la fusion de données. En revanche, l'estimation de ces pondérations *optimales* est réalisé en minimisant une estimation du taux d'erreur de classification et de segmentation : Le niveau considéré dans ce processus est donc celui de la fusion de décision.

C'est cette méthode de fusion à plusieurs niveaux qui va nous permettre d'obtenir les pondérations de chaque modèle qui minimisent les taux d'insertion, de suppression et de substitution.

Deuxième partie

Expérimentations

Chapitre 3

Reconnaissance de la parole

AVANT-PROPOS ET RÉSUMÉ

Dans ce chapitre, sont décrites les différentes expérimentations en reconnaissance de la parole audiovisuelle que nous avons menées aux cours de ces travaux. Dans les deux premiers chapitres, nous avons décrit nos propres modèles et méthodes d'intégration. Il est question dans ce chapitre de mettre en pratique ces différentes réflexions théoriques afin de se faire une idée sur les améliorations que l'on peut en attendre.

Nous avons, en effet, pris le parti de ne pas tirer de conclusions expérimentales trop importantes tant que les bases audiovisuelles disponibles seront de taille insuffisante.

L'ensemble des tests ont été effectués sur la base AMIBE, constituée de deux parties. La première a été réalisée à l'ICP par Lallouache (1991), la seconde au LIA dans le cadre de ces travaux[1]. Les expérimentations sont conduites dans le bruit pour les modèles synchrones. Les modèles produits et synchrones sont également évalués sans ajout de bruit, mais avec les pondérations obtenues avec l'approche que nous avons décrite dans le chapitre 2.

[1]Les détails sur les données que j'ai utilisées lors des phases d'apprentissage et de test figurent pp. 77-79

3.1 Bases de données audiovisuelles existantes

Nous avons débuté ces travaux en reconnaissance de la parole audiovisuelle en 1994, c'est-à-dire 10 ans après les toutes premières expérimentations dans ce domaine. La partie visuelle des bases de données nécessaires posait et pose encore d'importants problèmes matériels, liés aux phases d'acquisition et de stockage des données visuelles.

En conséquence, un lecteur appartenant au domaine du traitement de la parole acoustique pourra s'étonner de la faible taille de ces bases et du nombre de tests effectués lors des expérimentations. Il me semble toutefois qu'une comparaison des bases de parole, acoustique d'une part et audiovisuelle d'autre part ne peut être menée dans l'ignorance de leur évolution respective.

C'est pourquoi il nous faut dresser ici une liste des bases audiovisuelles utilisées par d'autres chercheurs, de 1984 à 1997. Les bases que j'ai personnellement utilisées ou créées, seront décrites plus loin (pp. 53-59 pour le projet AMIBEet pp. 77-79 pour le projet M2VTS).

3.1.1 Mots isolés

Bases orientées reconnaissance de la parole

- Les premiers travaux en reconnaissance de la parole audiovisuelle ont été réalisés par Petajan (1984). Afin de tester le système mis au point, ce dernier a dû créer une base de données ayant les caractéristiques suivantes :
  - Un seul locuteur
  - Corpus en langue américaine
  - Mots isolés
  - Corpus des chiffres de 0 à 9 (10 répétitions)
  - Corpus des lettres de A à Z (4 répétitions)
  - Corpus de 64 mots (2 répétitions)
  - Fréquence d'acquisition vidéo : 60Hz.

  Trois années plus tard, le nombre de locuteurs passa à 4 (Petajan et al., 1987).
- Finn et Montgomery (1988) ont réalisé une base uniquement visuelle à l'aide de marqueurs réfléchissants. Il s'agit de 23 consonnes anglaises en contexte /aCa/, un locuteur masculin prononce 2 fois chacun des logatomes.

- Yuhas et al. (1990) ont créé une base mono-locuteur de 9 voyelles dans un contexte /CVC/. Les données visuelles ont une fréquence de 60Hz.
- Une base contenant 10 répétitions des 22 consonnes de l'anglais en contexte /aCa/, prononcées par un seul locuteur a été réalisée par Silsbee (1994), le film a une fréquence de 60Hz.
- La base AVIM (Graf et al., 1995) est composée de deux locuteurs prononçant les voyelles de l'anglais.
- Sonoda et al. (1990) proposent pour le japonais, une base de 2 locuteurs prononçant 3 répétitions de :
  - 5 voyelles isolées
  - 5 voyelles en contexte /eVe/
  - Mots symétriques de la forme /eCVCe/, $C \in p, b, m, n, t, k$ et $V \in a, i, u, e, o$

  Cette base est uniquement visuelle, la fréquence des données est de 200Hz.
- Watanabe et Kohda (1990) disposent d'une base visuelle de deux locuteurs masculins prononçant en japonais 40 répétitions de 5 voyelles.
- Otani et Hasegawa (1995) possèdent une base contenant la prononciation par un seul locuteur de 2 répétitions de 5 voyelles japonaises. Le film de la zone buccale a été échantillonné à la fréquence de 30 Hz.
- Pour l'italien, Cosi et al. (1994) proposent une base visuelle de 2 locuteurs masculins et 2 féminins prononçant 5 répétitions de logatomes /VCV/, $C \in p, t, k, b, d, g$ et $V \in a, I, u$.

Bases orientées reconnaissance du locuteur

- La base DAVID (Chibelushi et al., 1993a; Chibelushi et al., 1993b) est composée de chiffres et de lettres en anglais, de 1 à 5 sessions par personne et 5 répétitions par session. 100 personnes, hommes et femmes ont participé à l'enregistrement. Le film est échantillonné à une fréquence de 25 Hz.
- Wolff et al. (1994) proposent une base de 10 locuteurs anglais prononçant 5 fois 5 consonnes en contexte /Ca/, avec une fréquence visuelle de 60 Hz.
- La base TULIPS (Movellan et Chadderdon, 1995) contient deux répétitions des chiffres anglais de 1 à 4, à 30 Hz, mais prononcés par 12 locuteurs, hommes et femmes.

- En ce qui concerne l'italien, la seule base de ce type disponible à ce jour a été réalisée par Falavigna et Brunelli (1994). Elle est composée de 89 locuteurs prononçant les chiffres de 0 à 9, en 5 sessions séparées d'environ 1 semaine. En revanche, une seule image est disponible pour la prononciation d'un chiffre.
- La base audiovisuelle allemande de Wagner et Dieckmann (1994) est composée de mots isolés, répétés 10 fois par 101 locuteurs, les films de la zone buccale ont une fréquence de 128 Hz.
- Pour le japonais, Wu et al. (1991) disposent d'une base audiovisuelle de 11 locuteurs masculins et 3 féminins prononçant 10 répétitions de 5 voyelles.

3.1.2 Parole continue

- La base de Goldschen et al. (1994) est constituée par la prononciation de 450 phrases de la base acoustique TIMIT (anglais-américain) par un seul locuteur mâle. Les données visuelles ont une fréquence de 60 Hz.
- Brooke et al. (1994) proposent une base mono-locuteur composée de 2 répétitions de suites de 3 chiffres, en anglais et avec une fréquence visuelle de 25 Hz.
- La base SpellingLips (Vo et al., 1995) est composée de 200 séquences d'environ 6 lettres allemandes, prononcées par un seul locuteur mâle. Le film est échantillonné à une fréquence de 25 Hz.
- Bregler et Konig (1994), utilisent une base de mêmes caractéristiques, avec des séquences plus courtes (de 2 à 3 lettres) mais prononcées par 4 locuteurs masculins et 2 féminins.
- Très récemment, Potamianos et al. (1997) ont construit une base audiovisuelle de 50 locuteurs américains, chacun prononçant d'une part 25 mots de type /CVC/ et d'autre part 25 séries de 4 lettres enchaînées. Le film correspondant a été échantillonné à la fréquence de 60 Hz. La partie audio est constituée par 4 enregistrements provenant de 4 microphones ayant des caractéristiques différentes.

3.1.3 Conclusion sur les bases existantes

À la date de mise à disposition de notre propre base de données (1995), les bases de parole audiovisuelle avaient les caractéristiques suivantes : les langues disponibles étaient l'allemand, l'américain, l'anglais, l'italien et le japonais. Le nombre de locuteurs maximum était de 100 pour la parole en mode *mots isolés* et 4 pour la parole continue. Il faut attendre 1997 pour trouver une base de 50 locuteurs pour la parole continue.

Quoi qu'il en soit, il n'existait en 1995, aucune base de donnée de parole audiovisuelle continue en langue française.

3.2 Bases de données AMIBE

Les corpus utilisés dans les expérimentations que j'ai effectués sont issus du projet AMIBE
(Montacié et al., 1995). Deux bases de données en français ont été créées dans ce contexte. Elles sont issues de deux laboratoires, de deux systèmes d'acquisition et deux locuteurs différents. Dans chaque cas, un unique locuteur prononce en continu des séquences de 4 lettres[2] comprises entre A et Z. L'intérêt de ce corpus est principalement constitué par sa difficulté :

- Unités de reconnaissance relativement courtes.
- Multiples contextes phonétiques (diphones de type /CC/, /CV/, /VC/, /VV/, etc..)
- Problèmes de segmentation (enchaînements des lettres K-A, I-X, I-Y, etc.)
- Unités très proches d'un point de vue acoustique (ex : {M,N}, {B,P,D} etc.)
- Unités très proches au niveau visuel (ex : {B,P}, {G,J}, etc.)
- Absence de modèle de langage (probabilités équi-réparties).

3.2.1 Première base

Un premier corpus, réalisé à l'ICP nous a été fourni (locuteur : jls) dans le cadre du projet AMIBE. Prise de vue et enregistrement se déroulent dans une chambre sourde anéchoïque. Le locuteur est assis, sa tête étant maintenue stable dans le champ des caméras par un casque solidaire de son siège. Ses lèvres sont maquillées en bleu et des lunettes opaques le protègent d'un spot halogène de 800 Watts qui assure le contraste de l'image, ce dernier point rendant obligatoire l'intervention d'un répétiteur. Les images sont enregistrées sur magnétoscope, l'extraction des paramètres labiaux ne se faisant pas en temps réel.

[2] La longueur de la séquence de lettres n'est cependant pas supposée connue lors de nos test de reconnaissance.

Le rôle des lèvres, en tant qu'organe articulatoire, est de modifier la forme et l'aire de la sortie du conduit vocal. On peut donc raisonnablement penser que la largeur, hauteur et aire de cette forme *intero-labiale* par laquelle s'échappe le son provenant du conduit vocal sont des paramètres pour le moins intérressants. Or, les points de contacts entre lèvres supérieure et inférieure ayant une position variable, l'usage de marqueurs ponctuels ne permet donc pas une mesure précise de la la largeur intero-labiale.

Un système de *chroma-key* électronique nous permet ici d'isoler les pixels bleus qui, dans l'image de la zone buccale, ne peuvent appartenir qu'aux lèvres. L'avantage d'une telle technique est de permettre une mesure précise de la *largeur intero-labiale*.

La lourdeur de ces manipulations explique la faible quantité de données dont nous avons pu disposer. En revanche, on se doit de reconnaître la très bonne qualité de ces données, les paramètres labiaux ayant une précision inférieure au millimètre. Les données brutes qui nous ont été fournies sont la hauteur, la largeur et l'aire intero-labiale synchronisées avec le signal acoustique (Lallouache, 1991).

Il faut noter que la largeur, mesurée à l'intérieur des lèvres est systématiquement nulle lors d'une fermeture labiale (sauf problèmes de sous-échantillonnage[3]). Cela nous oblige à fixer un seuil sur la variance de ce paramètre, qui pourrait se retrouver nulle pour des phonèmes tels que [p], [b] et [m].

La figure 3.1 montre un exemple de l'évolution de ces paramètres sur une séquence de lettres épelées de façon continue.

3.2.2 Deuxième base

Nous avons réalisé un deuxième corpus au Laboratoire d'Informatique d'Avignon (locuteur : pj). Le précédent système d'acquisition a été mis au point dans l'optique d'obtenir des mesures d'une précision inférieure au millimètre.

Nous avons pour notre part préféré la facilité d'utilisation, ceci dans la perspective de créer des corpus de taille plus grande. La qualité des paramètres obtenue est donc probablement inférieure à celle du précédent système. Nous verrons dans la section 3.5.1 (pages 65-66) l'influence de la précision des paramètres labiaux sur un système de lecture labiale.

[3]Si la durée d'une fermeture labiale est inférieure à la période d'échantillonage, la fermeture complète des lèvres peut ne pas être observée.

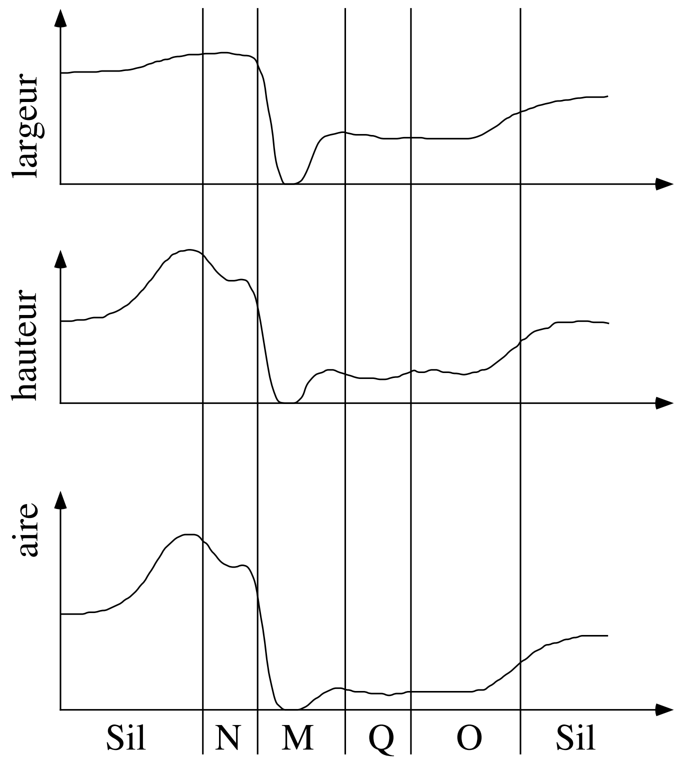

FIG. 3.1: Exemple de paramètres labiaux, base audiovisuelle de l'ICP, locuteur jls

Système d'acquisition

Le système d'acquisition est constitué d'un caméscope de bonne qualité relié à une carte de compression JPEG. Une lampe classique éclaire le visage, ceci afin d'obtenir un meilleur contraste. Le locuteur peut cependant regarder l'écran sur lequel apparaissent les suites de lettres à prononcer.

Un algorithme remplace le système électronique de chroma-key, permettant d'extraire les pixels bleus appartenant aux lèvres. Chaque pixel est représenté sous la forme d'un vecteur à trois composante (rouge, vert et bleu). On peut déterminer la moyenne, valeurs maximales et minimales de ce vecteur sur un ensemble d'images de lèvres et ainsi filtrer de nouvelles images. Les propriétés géométriques des lèvres rendent alors l'extraction des largeurs et hauteurs extero et intero-labiales assez aisée. Extraire les paramètres labiaux sur des lèvres non maquillées aurait entraîné l'utilisation d'un système de détection des contours labiaux plus complexe et de fiabilité inférieure.

D'autre part le temps nécessaire pour l'extraction ainsi que son manque de fiabilité aurait impliqué un stockage, non plus des paramètres, mais des séquences d'images. Il faut à ce propos signaler que la localisation, le suivi et l'extraction de contours labiaux dans des conditions réelles fait partie du domaine de la recherche.

Grâce au choix ainsi fait, le traitement complet d'une séquence de lettres ne prend que quelques secondes et ne nécessite qu'un opérateur qui peut-être le locuteur lui-même. La figure 3.2 montre un schéma de ce système d'acquisition.

Paramètres utilisés

Après désinterlaçage, le film a une fréquence de 50 Hz. La mise au point de ce système nous a permis de créer une base de taille et de contenu très proche de la première base.

Quelques différences doivent toutefois être notées : nous ne cherchons à extraire que les paramètres de largeur et hauteur, intero-labiale et extero-labiale.

Le paramètre constitué par l'aire intero-labiale étant très fortement corrélé au produit de la hauteur par la largeur (voir Abry et Boë (1986)). Nous remplaçons donc ce paramètre par le produit des deux premiers, ce qui nous permet d'accélérer le processus d'extraction.

L'utilisation du produit de la largeur et de la hauteur en tant que paramètre d'une modélisation statistique peut paraître absurde si on dispose également d'un vecteur hauteur-largeur. Il faut cependant remarquer que ce produit concentre en un seul paramètre de l'information provenant de deux

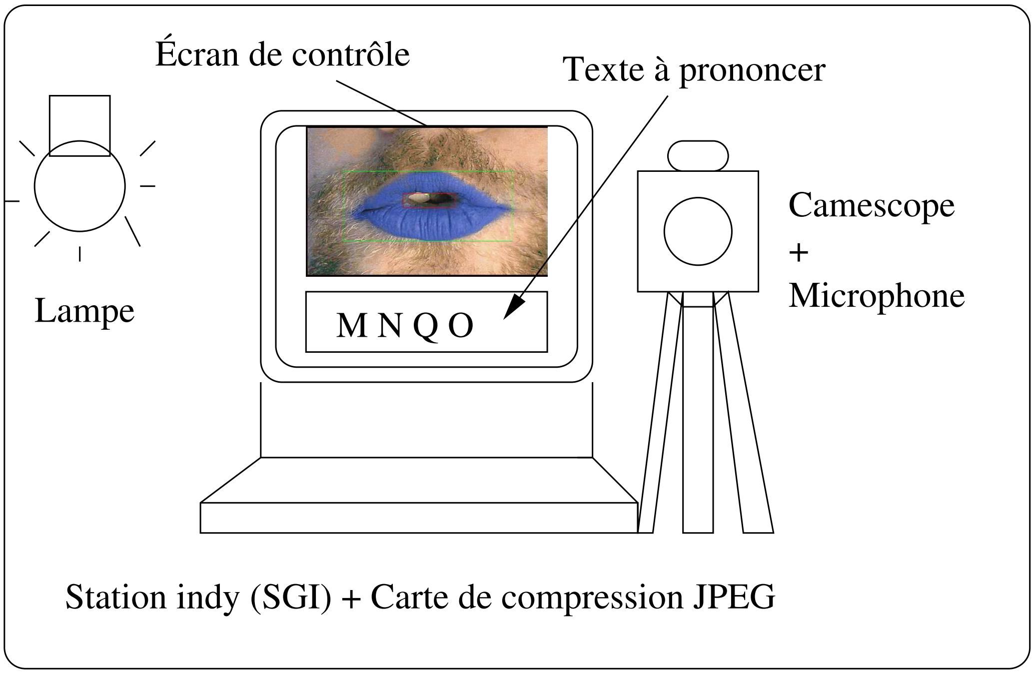

FIG. 3.2: Schéma du système d'acquisition audiovisuelle du LIA

autres mesures réelles. Sa variabilité sera donc moins importante que le couple hauteur-largeur et l'information qu'il véhicule plus discriminante, ce qui est un avantage indéniable sur un ensemble d'apprentisssage de faible taille.

Nous n'utiliserons pas les paramètres extero-labiaux dans nos expérimentations, ceci afin de ne pas biaiser les comparaisons entre les deux locuteurs Le signal acoustique est échantillonné à 16kHz via le microphone du camescope. Ce signal est transformé en vecteurs de 12 coefficients MFCC [4], énergie et dérivées premières. Ces vecteurs sont produits à une fréquence de 100Hz.

Pour le début et la fin de prononciation d'une séquence de lettres, nous imposons au locuteur d'avoir les lèvres fermées. Ceci nous permet en premier lieu de rendre les paramètres labiaux indépendants des variations d'échelle en les normalisant par la largeur extero-labiale des lèvres au repos. Cette contrainte imposée au locuteur permet également une vérification automatique de la qualité de l'enregistrement.

La figure 3.3 montre un mouvement d'ouverture labiale extrait de cette base (50 images par seconde). L'ensemble des données contient deux cents séquences de quatre lettres de l'alphabet. Cette base est actuellement accessible au public[5]. C'est sur ces deux bases (locuteurs jls et pj) que nous

[4]coefficients cepstraux sur une échelle de Mel

[5]`http ://www-laforia.ibp.fr/PAROLE/montacie/amibe/corpLIA.html`

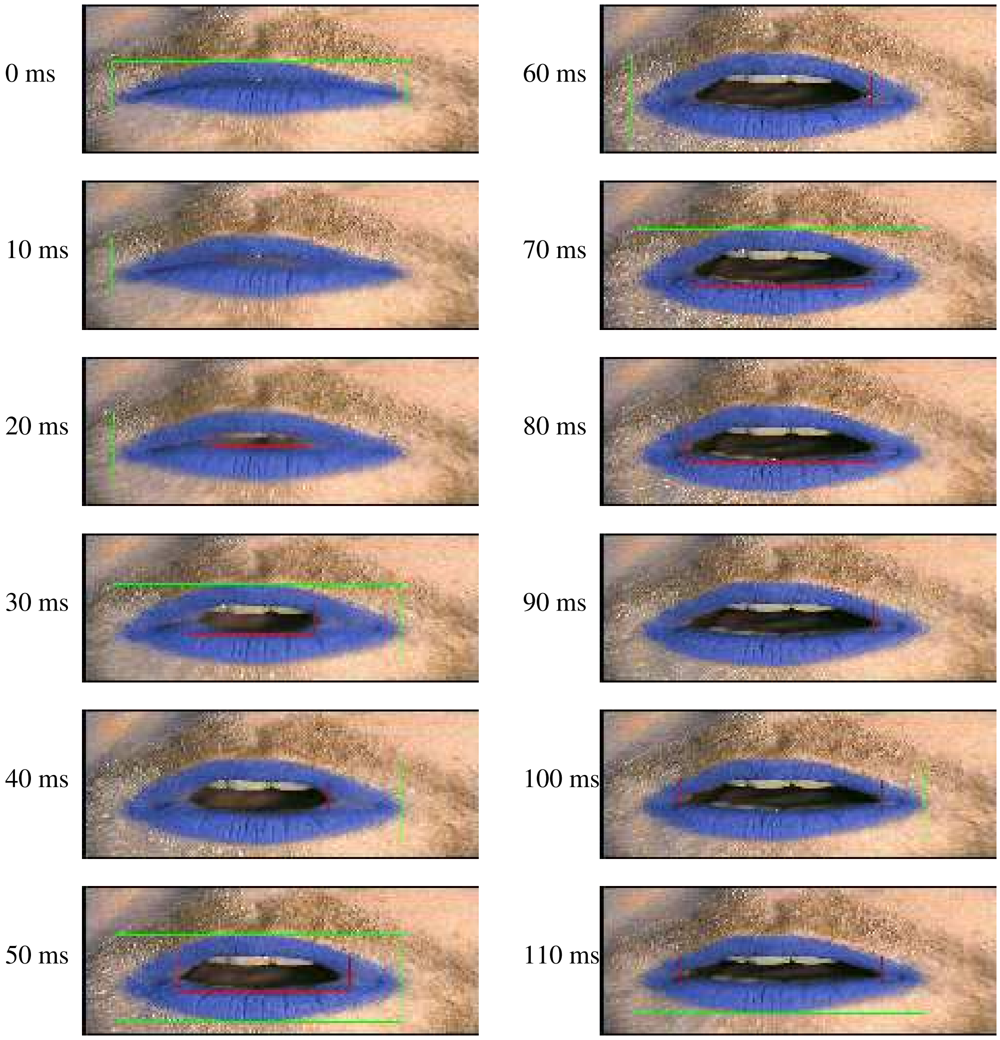

FIG. 3.3: Exemple de mouvements labiaux dans notre propre base audiovisuelle (locuteur pj)

avons effectué les expériences de reconnaissance bimodale. Le découpage de cet ensemble entre données d'apprentissage et de test va être décrit dans le chapitre suivant.

3.3 Protocole de test

Pour les deux locuteurs, les données ont été séparées en un ensemble d'apprentissage (éventuellement découpé en deux parties égales pour créer un sous-ensemble de développement) et un ensemble de test. Les séquences de 4 lettres sont réparties de façon aléatoire dans les 2 ensembles. Cette répartition a été fixée pour l'ensemble des participants du projet AMIBE. Les deux bases et leur répartition peuvent être obtenues en adressant une demande au responsable du projet.

Le premier ensemble contient 70% des données (soit 35% pour l'apprentissage et 35% pour le développement), l'ensemble de test contient quant à lui 30% des données. Pour les besoins d'une expérimentation rigoureuse, l'intersection entre ces deux ensembles est vide.

Les pourcentages de reconnaissance sont calculés sur les données de test, suivant la formule classique :

$$P = \frac{N - I - S - D}{N} \times 100$$

N étant le nombre de lettres à reconnaître, I le nombre d'insertions, S le nombre de substitutions, D le nombre de suppressions et P le pourcentage de reconnaissance. Le nombre de ces erreurs de différents types est calculé à l'aide d'une méthode d'alignement dynamique (Young et al., 1993). À titre indicatif, et en faisant l'hypothèse que les reconnaissances correctes suivent une distribution binomiale, les différents intervalles de confiance à 99% sont reportés le tableau 3.1. Ces intervalles ont été calculés avec $N = 240$, nombre de lettres à reconnaître lors d'une phase de test dans ces expérimentations.

P	Borne inférieure	Borne supérieure
70	62	77
75	67	81
80	73	86
85	78	90
90	84	94
95	90	98

TAB. 3.1: Intervalles de confiances sur les bases de test AMIBE

La largeur de ces intervalles de confiance nous impose la plus grande prudence quant à l'interprétation de nos résultats.

3.4 Description des modèles

3.4.1 Apprentissage des modèles

Initialisation

Nous utilisons pour toutes les phases d'apprentissage, la boîte à outils HTK[6] du CUED[7]. Chaque lettre est modélisée par un unique MMC. Pour chaque exemple d'une lettre dans le corpus d'apprentissage, un alignement de Viterbi sur les états du modèle correspondant est effectué. Nous pouvons ainsi associer un ensemble de vecteurs d'observation à chaque état du modèle et par conséquent estimer les paramètres des distributions.

Ce processus est répété jusqu'à la convergence de la valeur de la probabilité pour qu'un modèle donné produise les exemples lui correspondant.

Estimation hors-contexte

La différence essentielle avec la phase d'estimation est la réestimation par l'algorithme de Baum-Welsh. Cela implique une recherche de la probabilité d'être à un instant donné dans un état donné d'un modèle en utilisant l'algorithme du *Forward-Backward*. L'utilisation de cette méthode permet une estimation plus fine des paramètres : l'affectation des vecteurs d'observation à un état donné est cette fois réalisée conjointement à l'estimation des valeurs des paramètres. Notons que cette étape nécessite des frontières temporelles relatives aux états de début et de fin de chaque modèle.

Estimation en contexte

Cette phase est très proche de la précédente mais nous cherchons cette fois à estimer les paramètres de la concaténation des modèles correspondant à une phrase entière des données d'apprentissage. Les frontières de mots qui ont été posées manuellement suivant un critère subjectif, sont donc remises en cause en vue de maximiser la probabilité qu'une suite de modèles produise une suite d'observations.

---

[6] HMM toolkit

[7] Cambridge University Engineering Deparment

3.4.2 Modèles unimodaux

Une partie des paramètres de ces modèles est très difficile à estimer :

- La topologie : nombre d'états, nombre et structure des transitions.
- Nombre de gaussiennes par mixture d'un flux, d'un état, d'un modèle.
- Nombre de paramètres dans chaque vecteur d'observation.
- Nombre d'itérations d'apprentissage

Ces paramètres peuvent être estimés sur un corpus de développement, mais contrairement aux pondérations, ils nécessitent une phase complète d'apprentissage pour chaque calcul d'erreur. D'autre part, le nombre optimal d'itérations d'apprentissage dépend du rapport entre quantité de paramètres à estimer et nombre de données disponibles. Nous avons donc fixé arbitrairement tous ces paramètres.

Tous les modèles (un par lettre) ont six états émetteurs, un état initial et un état final non-émetteurs. Chaque état émetteur comporte une transition vers lui-même et une vers l'état suivant. Un système plus efficace peut pourrait être réalisé en cherchant une topologie optimale pour chaque locuteur, unité de reconnaissance et modalité. Cela n'est cependant pas l'objet de ces expérimentations.

La faible taille de la base de donnée d'apprentissage, ne nous permet pas d'utiliser comme distribution des mixtures de plus de deux gaussiennes. De même, et en accord avec l'hypothèse d'indépendance statistique[8], nous utilisons uniquement des matrices de variance-covariance diagonales. Les deux premières phases d'apprentissage sont menées jusqu'à la convergence des probabilités, la troisième jusqu'à la sixième itération.

Les modèles acoustiques émettent des vecteurs composés de douze coefficients MFCC et de l'énergie du signal auxquels sont ajoutées leurs dérivées (26 paramètres au total). L'ajout de l'accélération augmenterait le nombre de paramètres et rendrait par conséquent plus difficile leur estimation. La distribution de probabilité associée à chaque état est le produit de deux mixtures de gaussiennes, l'une pour les paramètres statiques, l'autre pour les paramètres dynamiques. Nous dirons par la suite qu'ils constituent deux *flux* de données.

Les modèles labiaux émettent des vecteurs composés de 9 paramètres : la hauteur, la largeur, l'aire intero-labiale ainsi que leurs vitesses et accélérations respectives. Paramètres statiques, dérivées et accélérations sont réparties sur trois flux.

---

[8]Cette hypothèse, bien que très grossière, permet une simplification non négligeable du modèle statistique utilisé. Ne pas faire cette hypothèse reviendrait approximativement à élever à la puissance deux le nombre de paramètres à estimer.

Dans notre cas, le nombre de paramètres à estimer peut être calculé ainsi : $M \times (E \times N \times G + T)$, où $M = 27$ est le nombre de modèles (un par lettre de l'alphabet, plus un pour le silence), $E = 6$ le nombre d'états par modèle, N le nombre total de composantes du vecteur d'observation multiplié par 2 (vecteur de moyennes et vecteur de variances), G le nombre de gaussiennes par mixture, et $T = 6$ le nombre de transitions à estimer par modèle[9].

Il y a donc 17010 paramètres à estimer dans les modèles acoustiques et 5944 pour les modèles labiaux.

3.4.3 Modèles bimodaux synchrones

Nous avons à notre disposition deux flux produits par deux sources d'origines différentes (acoustique et visuelle) mais correspondant à la même suite sous-jacente d'unités de reconnaissance (voir Figure 3.4). Les paramètres acoustiques et labiaux n'ont la même fréquence d'échantillonnage mais une simple interpolation permet d'apporter une solution parmi d'autres à ce problème.

Les modèles acoustico-labiaux synchrones ont la même topologie que leurs correspondants unimodaux mais émettent des vecteurs qui sont la concaténation des vecteurs acoustiques et labiaux. L'ensemble de leurs paramètres est par conséquent réparti sur cinq flux (22842 paramètres au total).

Dans ce cadre, le rapport entre nombre de données et nombre de paramètres est le même pour les trois types de modèles. Ceci est très important en ce qui concerne de futures comparaisons de résultats.

3.4.4 Modèles bimodaux produit

Il s'agit simplement de réaliser le produit (décrit dans le chapitre 1, section 1.4, pages 17-26) sur les deux modèles unimodaux décrits ci-dessus, ces derniers étant entraînés séparément.

Nous pouvons constater que le rapport entre le nombre de données et le nombre de paramètres à estimer est quasiment le même pour les modèles produits (22954 paramètres) et les modèles synchrones (22842 paramètres). Il en va de même pour le nombre d'itérations d'apprentissage. Cette caractéristique limite considérablement le biais expérimental.

---

[9] Si un état possède n transitions, seulement $n-1$ probabilités sont à estimer, la somme des probabilités devant être égale à 1

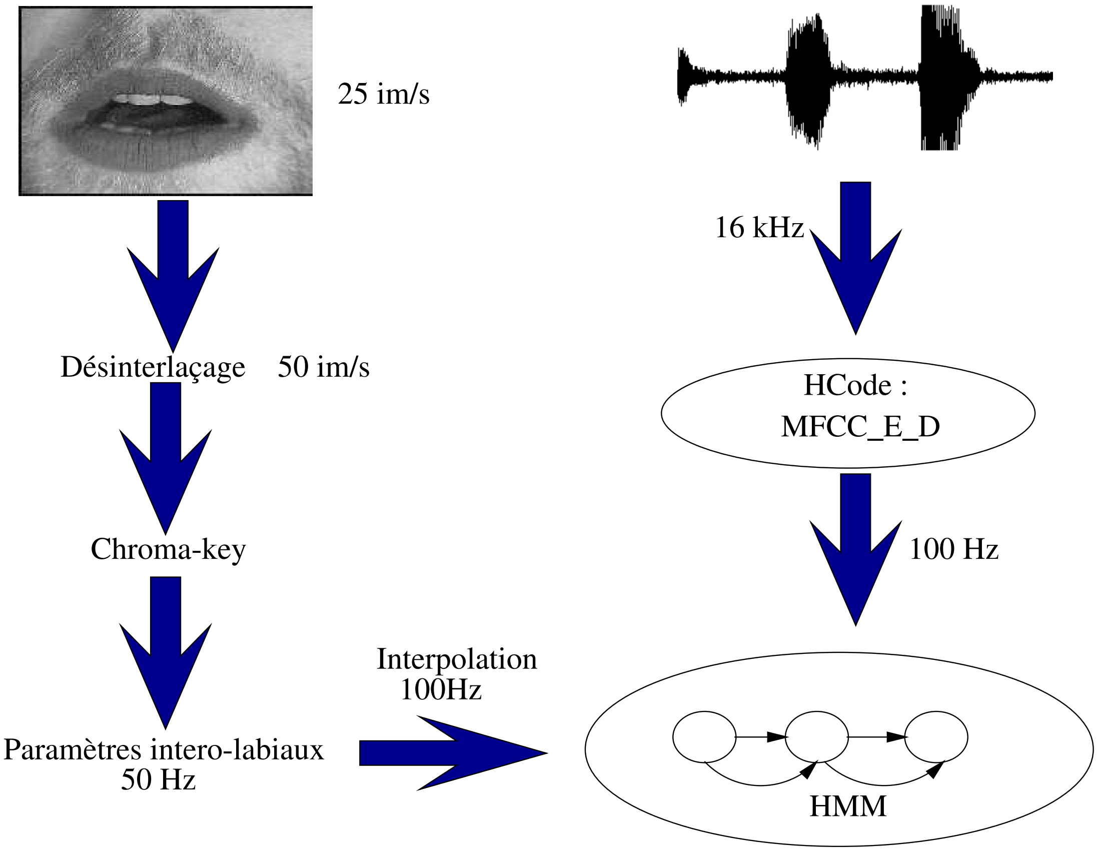

FIG. 3.4: Fonctionnement global

3.4.5 Pondération

Les modèles sur lesquels sont estimées les pondérations sont les modèles synchrones décrits en section 3.4.3. Les pondérations *optimales* sont déterminées avec la méthode décrite en chapitre 2 (p. 42) pour chaque lettre et chaque locuteur de la base AMIBE. Ces poids sont reportés dans le tableau 3.2.

	Silence	**A**	**B**	**C**	**D**	**E**
pj	0,43	0,48	0,43	0,51	0,51	0,59
jls	0,54	0,50	0,49	0,51	0,49	0,49
	F	**G**	**H**	**I**	**J**	**K**
pj	0,55	0,49	0,47	0,58	0,52	0,50
jls	0,51	0,49	0,49	0,44	0,49	0,50
	L	**M**	**N**	**O**	**P**	**Q**
pj	0,50	0,53	0,49	0,49	0,61	0,50
jls	0,49	0,50	0,51	0,49	0,49	0,49
	R	**S**	**T**	**U**	**V**	**W**
pj	0,49	0,57	0,55	0,50	0,50	0,52
jls	0,49	0,50	0,51	0,50	0,50	0,50
	X	**Y**	**Z**			
pj	0,54	0,54	0,49			
jls	0,51	0,50	0,50			

TAB. 3.2: Valeurs des poids acoustico-labiaux optimisés

Il serait hasardeux d'essayer de corréler ces poids avec des résultats provenant d'expériences de perception. Leurs valeurs sont en effet extrêmement liées au fonctionnement de notre approche statistique. Remarquons simplement que l'on ne s'écarte que très peu des valeurs utilisées durant l'apprentissage $(0,5)$. Ceci est un résultat attendu puisque le corpus de développement ne diffère que très peu de celui utilisé pour l'estimation des autres paramètres du système.

Il faut ajouter que le simple fait d'avoir estimé ces 27 nouveaux paramètres (un poids acoustico-labial par modèle) dans nos modèles constitue un biais expérimental. Ce biais semble cependant négligeable comparé au nombre de paramètres total à estimer (24692).

3.5 Résultats

3.5.1 Influence de la précision des paramètres labiaux

Lors de la mise au point de notre système d'acquisition, nous avons fait le choix de la facilité de mise en œuvre et d'utilisation aux dépends de la précision des paramètres. Il paraît donc tout naturel de quantifier l'influence de cette précision sur des résultats de reconnaissance.

Nous utilisons ici les modèles unimodaux visuels dans les conditions décrites dans la section 3.3. Le locuteur choisi est jls, nous pourrons ainsi parler de précision en millimètres.

Nous ajoutons à chaque valeur du paramètre selectioné (largeur et/ou hauteur) un nombre aléatoire compris entre $+x$ et $-x$. Nous substituons ensuite l'aire intero-labiale par le produit de la hauteur et de la largeur. Cette opération est effectuée pour toutes les données, d'apprentissage et de test. Nous pouvons alors calculer un taux de reconnaissance labiale pour chaque précision affectée à un paramètre donné, les autres paramètres restant inchangés. Nous obtenons de cette façon le tableau 3.3 et la figure 3.5 pour les variations de précision en dixième de millimètre. Rappelons que quelle que soit la colonne, la reconnaissance s'effectue sur l'ensemble des paramètres.

Précision	largeur	hauteur	largeur et hauteur
± 0.0 mm	44	44	44
± 0.5 mm	47	44	45
± 1.0 mm	47	46	42
± 1.5 mm	49	46	41
± 2.0 mm	44	42	41
± 2.5 mm	49	44	41
± 3.0 mm	48	41	36
± 3.5 mm	43	37	38
± 4.0 mm	48	38	36
± 4.5 mm	44	38	33

TAB. 3.3: Influence de la précision des paramètres labiaux sur les résultats de la lecture labiale (pourcentages de reconnaissance correcte)

La première chose que l'on constate est une amélioration des résultats lorsque l'on ajoute un peu d'imprécision dans la mesure des paramètres labiaux. Ce fait peut paraître suprenant mais il est provient d'un phénomème relativement simple.

La variance, calculée sur un corpus de trop faible taille, est sous-estimée.

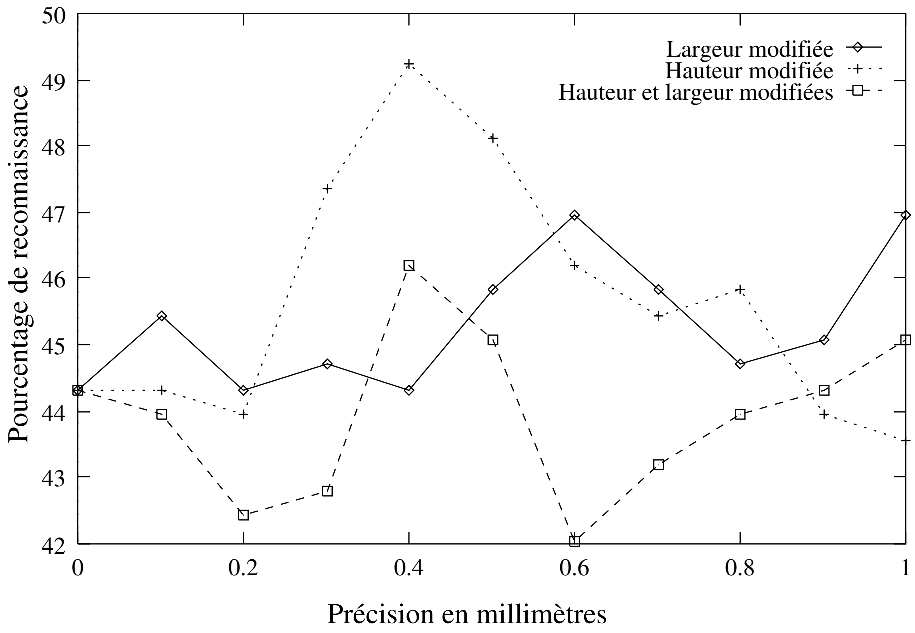

FIG. 3.5: Influence de la précision (en dixième de mm) des paramètres labiaux sur les résultats de la lecture labiale

Or, en ajoutant de l'imprécision sur un paramètre, on augmente la valeur de sa variance. L'estimation est donc plus proche des valeurs trouvées dans le corpus de test. Toutefois, l'amélioration constatée ne constitue au maximum qu'une réduction de 10% du nombre d'erreurs.

En revanche, lorsque la précision devient trop faible, le manque de fiabilité des données dégrade considérablement les résultats. Cette dégradation peut largement dépasser l'effet bénéfique de l'augmentation des variances.

Nous pouvons tout de même noter que dans ces conditions, la dégradation des performances du système n'apparaît que pour des précisions supérieures à ± 1mm et qu'une précision de $\pm 4,5$mm sur l'ensemble des paramètres n'augmente le nombre initial d'erreurs que de 20%.

On peut aussi constater que la hauteur intero-labiale semble être un paramètre nécessitant plus de précision que la largeur. Ceci est évidemment à mettre en relation avec l'espace de variation de ces deux paramètres.

3.5.2 Influence du bruit sur les modèles synchrones

Un bruit de foule (fréquences vocales) est ajouté au signal acoustique avec différents niveaux de rapport signal sur bruit. Pour chaque niveau de bruit, nous effectuons les phases d'apprentissage et de test. Cette expérimentation a été réalisée uniquement sur les modèles synchrones et avec le locuteur jls.

Aucune pondération n'est effectuée dans ce test et cela explique que le taux de reconnaissance soit plus faible pour le modèle acoustico-labial que pour le modèle purement acoustique lorsque qu'aucun bruit n'a été rajouté.

Il s'agit ici de vérifier sur un exemple une assertion relativement simple : Lorsque la source acoustique perd de sa fiabilité, l'ajout de paramètres visuels même moins fiables permet de compenser en partie cette perte (voir tableau 3.4).

Ce fait est confirmé dans de nombreux travaux, réalisés dans des laboratoires différents, avec différents locuteurs, corpus et langues. C'est pourquoi nous n'étendrons pas cette expérience aux autres modèles et locuteur.

Signal/Bruit	Labial	Acoustique	Acoustico-labial
sans bruit	42	88	86
6 dB	-	74	78
0 dB	-	53	63

TAB. 3.4: Influence du bruit sur les modèles synchrones

3.5.3 Modèles produits

Pour chaque locuteur et pour chaque type de modèle, nous avons effectué une série de tests. Les modèles sont évalués pour chaque itération d'apprentissage en contexte, ceci afin de mettre en évidence la variabilité des résultats en fonction d'un unique paramètre que nous avons fixé arbitrairement. Il est évident que cette variabilité est due au manque de données.

Pour une itération donnée, le modèle acoustico-labial produit est construit à partir des modèles acoustiques et labiaux entraînés jusqu'à ce niveau.

Les résultats sont rapportés dans les tableaux 3.5 et 3.6, figures 3.6 et 3.7. Ces résultats ont été obtenus avec des pondérations fixées arbitrairement.

Ces résultats font apparaître une amélioration du modèle produit sur le modèle synchrone pour le locuteur jls. En revanche, pour le locuteur pj, si le modèle produit obtient le meilleur résultat sur l'ensemble des itérations, ses performances sont très proches de celles du modèle synchrone. Ce dernier obtient d'ailleurs de meilleurs résultats pour certaines itérations.

Ceci peut indiquer une différence de quantité d'asynchronie présente pour chaque locuteur, mais aussi une différence de comportement du modèle produit en présence de données acoustiques moins fiables que pour le locuteur jls. Il faut néanmoins se garder, en l'absence de corpus suffisant, d'attacher trop d'importance aux résultats dans leur valeur absolue.

TAB. 3.5: Résultats pour le locuteur jls

Itération	Acoustique	Labial	Bimodal synchrone	Bimodal produit
0	89	36	90	90
2	90	39	92	93
4	91	40	91	93
6	91	40	90	93
8	91	39	91	93
10	91	39	92	93
12	91	38	92	93
14	91	39	92	93
16	90	38	92	93
18	90	38	92	93

TAB. 3.6: Résultats pour le locuteur pj

Itération	Acoustique	Labial	Bimodal synchrone	Bimodal produit
0	74	22	76	77
2	73	24	77	78
4	74	24	77	77
6	75	24	77	76
8	74	24	77	77
10	74	26	76	76
12	73	26	77	76
14	71	25	77	77
16	71	27	77	76
18	71	28	77	77

3.5.4 Pondération

Les résultats des différents modèles, avec une pondération égale ou optimisée sont résumés dans le tableau 3.7. L'entrainement des modèles en contexte a été réalisé en 6 itérations. Dans les précédentes expérimentations,

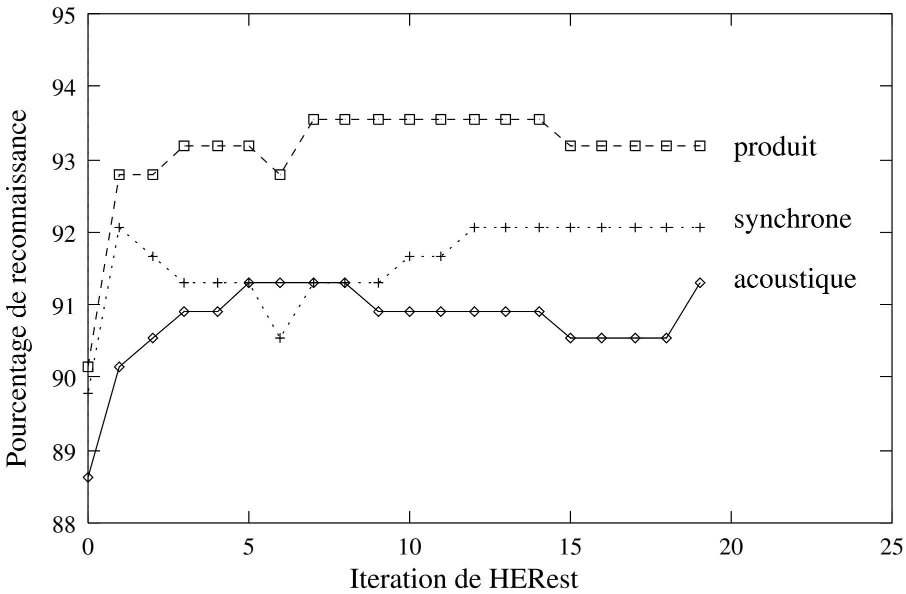

FIG. 3.6: Résultats pour le locuteur jls

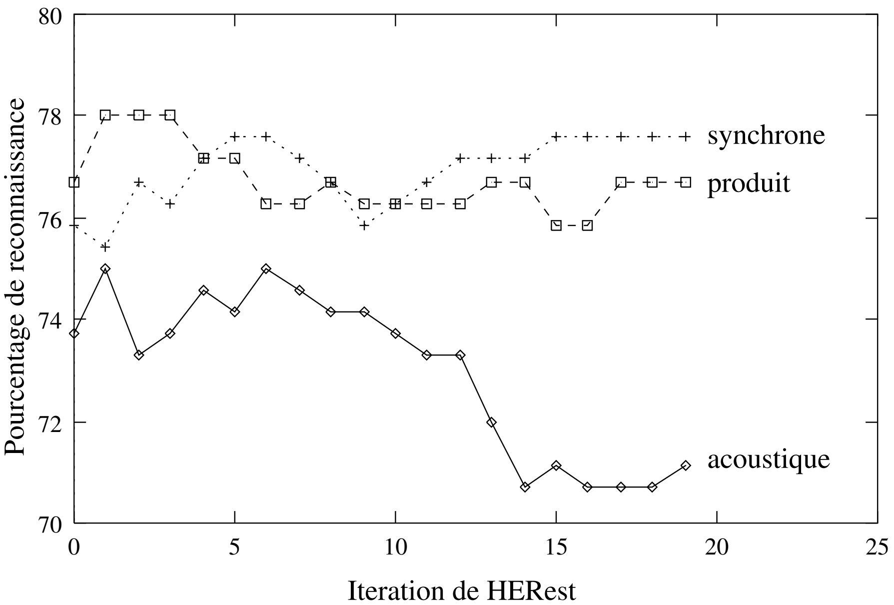

FIG. 3.7: Résultats pour le locuteur pj

nous avions fixé de façon empirique une valeur de pondération partagée par tous les modèles. Dans le cadre de la base AMIBE, cette approche empirique, bien que moins rigoureuse, reste d'une efficacité certaine. Les conditions naturelles d'élocution et d'acquisition acoustique expliquent les relativement faibles taux de reconnaissance pour le locuteur pj.

	taux de reconnaissance	
	locuteur pj	locuteur jls
Acoustique	75 %	91 %
Labial	25 %	40 %
Bimodal synchrone (poids égaux)	78 %	86 %
Bimodal produit (poids égaux)	78 %	93 %
Bimodal synchrone (poids optimisés)	79 %	93 %
Bimodal produit (poids optimisés)	79 %	93 %

TAB. 3.7: Résultats des différents systèmes suivant leur pondération acoustico-labiale

Une fois de plus, l'amélioration décrite ci-dessus n'est pas significative et sa valeur absolue est très certainement liée au locuteur et au corpus utilisé.

3.5.5 Matrices de confusion

Il faut tout d'abord préciser qu'il est très difficile d'interpréter correctement une matrice de confusion dans le cadre de la parole continue. En effet, une erreur de classification peut entraîner une erreur de segmentation et vice-versa.

Cependant, il peut être intéressant de voir ce que signifient réellement les améliorations obtenues dans la section précédente.

Locuteur jls

Nous sommes ici dans le cas où les données, aussi bien acoustiques que labiales sont de très bonne qualité. Le tableau 3.8 représente la matrice de confusion des lettres lorsque seules les données acoustiques sont prises en compte. Les erreurs qui apparaissent étaient attendues : confusion entre B et D, entre J, G et I, entre T, P et V. Certaines erreurs de classification peuvent paraître choquantes : Elles sont en réalité provoquées par des erreurs de segmentation.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Sup
A	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
B	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	.	1	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
E	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
H	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
I	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J	.	.	.	.	.	.	2	.	1	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
M	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	1	.	.
N	.	.	.	.	.	.	.	.	.	.	.	1	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.
O	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	2	.	.	.	.	.	.	.
Q	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.
S	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.
T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.
U	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.
V	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	1	.	8	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	3	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.
Z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.
Ins	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.

TAB. 3.8: Matrice de confusion acoustique pour le locuteur jls

Le tableau 3.9 montre que lorsque l'on rajoute des données labiales, les confusions qui disparaissent sont en plus grand nombre que celles qui apparaissent. Les confusions les plus notables ainsi évitées sont entre les lettres B et D et entre les lettres P, T et V.

Locuteur pj

Les données acoustiques étant plus bruitées, la segmentation devient plus ardue. Le tableau 3.10 fait apparaître des problèmes attendus : confusion entre les lettres A, K, H, entre B, D, P, V, T et entre M et N. D'autre confusions sont difficilement interprétables, il s'agit en fait surtout de problèmes de segmentation.

Nous voyons sur le tableau 3.11 les *faibles* améliorations apportées par un modèle bimodal produit à pondérations optimales. Certaines ambiguïtés (entre M et N, entre B, D, P et V) sont levées, mais d'autres apparaissent. En effet, plus les données sont bruitées, plus la représentativité du corpus d'apprentissage prend de l'importance.

3.6 Conclusion

Il nous a paru nécessaire de vérifier expérimentalement la validité de nos modèles théoriques. Nous avons vu malheureusement que les bases de données

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Sup
A	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
B	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
C	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
E	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
H	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
I	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
J	.	.	.	.	.	.	3	.	1	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
M	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	1	.	.
N	.	.	.	.	.	.	.	.	.	.	.	1	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.
O	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.
P	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.
Q	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.
S	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.
T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.
U	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.
V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.
X	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	8	.	.	1
Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.
Z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.
Ins	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	

TAB. 3.9: Matrice de confusion bimodale pour le locuteur jls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Sup
A	2	.	.	.	.	.	.	1	.	.	3	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	1
B	.	3	.	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	5	.	.	.	.	1
C	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.
D	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	2	.	1	.	.	.	.	.
E	.	.	.	.	9	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
G	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.
H	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
I	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.
J	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.
K	.	.	.	.	.	1	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	.	.	.	.	.	.	.	.	.	6	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	.	.	.	.	.	.	.	.	.	1	6	1	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	.	.	.	.	.	.	.	.	.	.	1	1	6	.	.	.	.	.	.	.	.	.	.	.	.	.
O	.	.	.	.	1	.	.	.	.	.	.	.	1	.	6	.	.	.	.	.	.	.	.	.	.	.	1
P	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	1	.	.	.	.	.	.	.
Q	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	2	.	.	.	.	.	.
R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.	1
S	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.
T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.
U	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	1
V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	6	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.
Z	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	6	.
Ins	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	.

TAB. 3.10: Matrice de confusion acoustique pour le locuteur pj

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Sup
A	2	.	.	.	.	.	.	1	.	.	3	.	2	.	.	.	.	1	.	.	.	.	.	.	.	.	.
B	.	4	.	.	.	.	.	.	.	.	.	2	.	.	.	1	.	.	.	.	1	3	.	.	.	.	1
C	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.
D	.	.	.	1	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	2	.	1	.	.	.	.	.
E	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1
F	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G	.	.	.	1	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.
H	.	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
I	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
K	.	.	.	.	.	.	.	.	.	.	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	.	.	.	.	.	.	.	.	.	5	1	2	.	.	.	.	.	.	.	.	1	.	.	.	1
M	.	.	.	.	.	1	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	.	.	.	.	.	.	.	.	.	.	.	1	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.
O	.	.	.	.	1	.	.	.	1	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	1
P	.	.	.	.	.	.	.	.	2	.	.	1	.	.	.	2	.	.	.	1	.	2	.	.	.	.	.
Q	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.
R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	1	.	.	.	.	.	.
S	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.
T	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	8	.	1	.	.	.	.	.
U	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9	.	.	.	.	.	1
V	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.
W	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.	.	.	.
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	11	.	.	.
Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	12	.	.
Z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	.
Ins	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.

TAB. 3.11: Matrice de confusion bimodale pour le locuteur pj

disponibles ne nous permettaient pas d'opter pour une approche fortement expérimentale de ces problèmes.

Nous avons, lors de l'élaboration de notre protocole de test cherché à limiter les biais expérimentaux. Il nous paraît néanmoins toujours hasardeux de comparer expérimentalement des méthodes différentes sur des corpus d'aussi faible taille.

En revanche, en construisant une base de données audiovisuelle pour le français, nous avons pu confirmer certaines conclusions obtenues dans d'autres travaux, pour d'autres locuteurs et d'autres langues.

En effet, le fait que les mouvements labiaux constituent un apport de qualité et de robustesse aux systèmes de reconnaissance de la parole est un point de vue partagé par l'ensemble des chercheurs dans ce domaine.

D'autre part, nos différentes approches des phénomènes d'asynchronie et du problème de la pondération ont donné de bons résultats. Les améliorations obtenues, bien que non-significatives ne sont pas négligeables : passer d'un taux de reconnaissance acoustique de 91% à un taux de reconnaissance acoustico-labial de 93% signifie une réduction de 22% du nombre total d'erreurs.

Chapitre 4

Reconnaissance du locuteur

RÉSUMÉ

Ce chapitre décrit l'approche que nous avons adoptée pour traiter les problèmes qui se posent en vérification multimodale du locuteur. Le système mis au point consiste en deux classifieurs, un premier fondé sur les caractéristiques visuelles des lèvres, le second sur des paramètres acoustiques. Un module de suivi des contours labiaux nous permet d'extraire une part d'information visuelle contenue dans le visage du locuteur. Deux types d'information sont considérés : les paramètres relatifs à la forme et ceux relatifs aux intensités.

Une méthode pour normaliser et projeter les différentes modalités dans un espace commun est présentée. Des expérimentations en vérification du locuteur ont été réalisées pour chaque modalité ainsi que pour leur fusion. Nous observons que les performances du système multimodal sont supérieures à celles de chaque sous-système et permettent une réduction du taux de fausse acceptation de $2,3\%$ à $0,5\%$ par rapport au sous-système acoustique.

4.1 Introduction

Il n'existe aucune personne qui soit strictement identique à une autre si l'on considère l'ensemble de ses caractéristiques individuelles. En revanche, si nous considérons qu'une personne ne change pas d'identité sur une période donnée, ses caractéristiques changent, quant à elles, continuellement. Les indices succeptibles de permettre une identification de la personne sont sujets à deux types de variations : intra et inter individu.

Il va de soi que les indices ayant la plus faible variation intra-individu et la plus forte variation inter-individu sont les plus pertinents pour un système de vérification d'identité. Cependant, ces deux types de variations peuvent être influencés par différents facteurs. Parmi ces facteurs, le temps qui sépare deux apparitions d'une même personne est succeptible d'augmenter la variation intra-individu.

D'autre part, un nombre considérable de caractéristiques, physiques ou cognitives, peuvent être utilisées pour mener à bien cette identification. Des informations sur ces caractéristiques individuelles peuvent être portées par un grand nombre de modalités. Parmi celles-ci, les modalités visuelle et auditive présentent un intérêt certain : celui de ne pas forcément nécessiter de collaboration ou d'efforts de la part de la personne subissant le processus d'identification.

4.2 Position du problème

Dans le cadre de la vérification automatique d'identité, la probabilité de réussite d'un acte d'imposture dépend du nombre d'indices et de modalités pris en compte par le système (Acheroy et al., 1996). Cependant, si les modalités visuelle et auditive ont été indépendamment étudiées avec beaucoup d'intérêt dans le domaine de la vérification d'identité, ce n'est que très récemment que sont apparues des études sur leur combinaison.

Brunelli et Falavigna (1995) ont précédemment décrit une approche bimodale de l'identification d'une personne. Leur système est fondé sur les caractéristiques visuelles contenues dans une image statique du visage et sur les paramètres acoustiques du signal de parole. Les performances de ce système multimodal sont supérieures à chacun de ses sous-systèmes unimodaux.

Cependant, dans un tel système, le mouvement des lèvres provoque un apport de variabilité intra-individu dans le processus d'identification visuelle. Une prise en compte de ces informations labiales serait pourtant succeptible d'une part, de réduire la variabilité intra-individu et d'autre part, d'augmenter la variabilité inter-individus. En effet, le fait que l'information labiale temporelle n'est pas seulement corrélée au contenu linguistique du message mais porte également une information sur l'identité du locuteur a été largement ignoré jusqu'à maintenant.

Récemment, Luettin et al. (1996b) ont proposé d'intégrer cette nouvelle modalité dans un système de reconnaissance de la personne sous la forme de paramètres labiaux spatiaux-temporels. Dans ce chapitre, nous étendons cette approche en considérant la combinaison des deux modalités, acoustique et labiale, pour un système de vérification du locuteur.

Il est important de noter que les méthodes décrites au cours des deux premiers chapitres ont été élaborées pour répondre à des problèmes assez différents de ceux évoqués à présent. En conséquence, notre démarche n'est pas d'appliquer directement ces méthodes au domaine de la vérification d'identité, mais plutôt de tirer profit des réflexions menées au cours des trois premiers chapitres afin de mettre au point un processus de fusion acoustico-labiale, pertinent pour ce nouveau domaine de recherche.

4.3 La base de données M2VTS

La base de données audiovisuelle M2VTS a été enregistrée à l'UCL (Université Catholique de Louvain) (Pigeon et Vandendorpe, 1997). Elle est constituée des enregistrements de 37 locuteurs (hommes et femmes) ayant prononcé en français les chiffres de zéro à neuf. Un *enregistrement* est une séquence de dix chiffres, prononcés de façon continue et dans l'ordre croissant (voir figure 4.1).

Cinq sessions d'enregistrement ont été effectuées pour chaque locuteur, à une semaine d'intervalle, afin de prendre en compte la variabilité intra-individu dans les deux modalités. Les images contiennent la tête de la personne dans son intégralité. Elles ont été filmées à une fréquence de 25 Hz. Nous avons divisé cette base en trois ensembles :

- les trois premières sessions devant être utilisées comme données d'apprentissage
- la quatrième comme ensemble de *validation*
- la cinquième comme ensemble de test

La cinquième session est la plus difficile à traiter. Elle diffère des autres dans les variations du visage (tête inclinée, non rasée), de la voix (faible RSB) ou encore des imperfections de la prise de vue (mauvaise mise au point, différents facteurs de grossissement). Cette session, utilisée en tant que base de test, permet un processus d'évaluation proche des conditions *réeles* de fonctionnement. La figure 4.2 montre les images du visage de 3

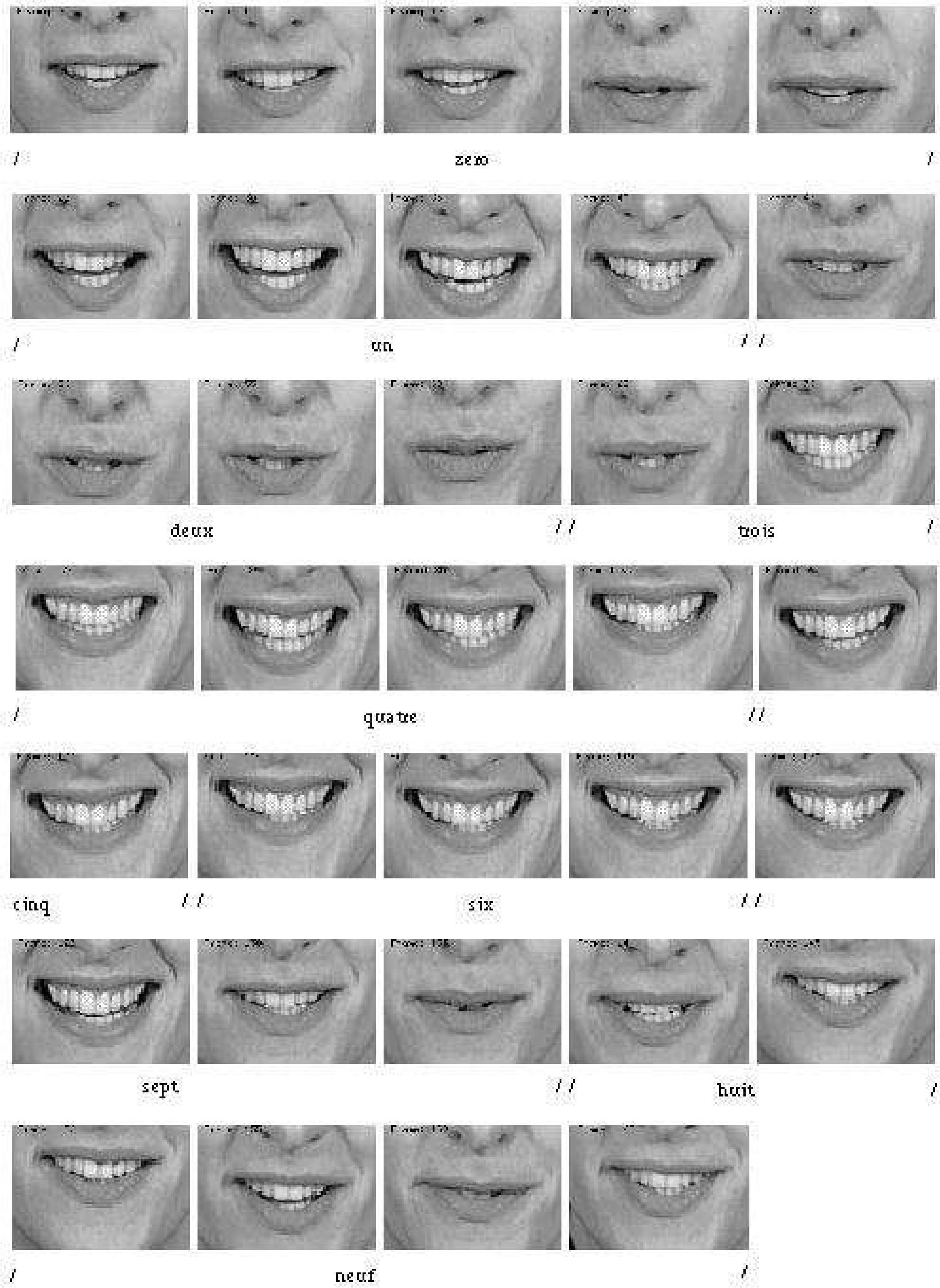

FIG. 4.1: Séquence d'images extraites de la base M2VTS dans la même session et pour le même locuteur

locuteurs sur les 5 sessions. En se référant à la section 3.1 (chapitre 2, pp. 50-53), on peut remarquer la grande qualité de cette base audiovisuelle pour ce domaine d'application (nombre de locuteurs, parole continue, variation intra-locuteurs, fréquence des données visuelles).

4.4 Extraction des paramètres labiaux

La réalisation d'un système d'extraction de paramètres labiaux en conditions réelles est un travail considérable qui se situe dans un domaine assez différent de celui du traitement automatique de la parole. Cependant, la qualité d'un tel système détermine en grande partie les performances du sous-système de vérification labiale et du système acoustico-labial complet. Le système que nous utilisons est issu des travaux de Luettin et al. (1996a). Une description très précise des paramètres visuels utilisés peut-être trouvée dans ce dernier document. Nous nous limitons ici a un bref descriptif.

Nous nous intéressons aux changements faciaux dus à la production de la parole et analysons par conséquent seulement la région buccale du visage du locuteur. Les approches classiques en reconnaissance du visage reposent soit sur des caractéristiques géométriques soit sur les intensités du visage complet ou de ses composantes (Chellappa et al., 1995).

Nous combinons ici les deux approches tout en émettant l'hypothèse que les contours labiaux et la distribution de niveaux de gris autour de la zone buccale constituent des indices importants sur l'identité d'un locuteur. Durant le processus de production de la parole, les contours labiaux se déforment et les intensités de la zone d'intérêt varient en fonction de la forme labiale, de la protrusion et de l'apparition des dents et de la langue.

Ces informations visuelles, statiques et temporelles, ne nous renseignent pas uniquement sur les caractéristiques physiques d'une personne, mais aussi sur sa manière de parler.

4.4.1 Le modèle de lèvres

Un modèle déformable est utilisé pour décrire les contours labiaux internes et externes et un modèle de niveaux de gris pour décrire les valeurs de l'intensité en ces points. Ce dernier doit décrire les changements d'intensité autour des lèvres du locuteur et par conséquent, se déformer en suivant les contours labiaux. Une approche fondée sur les *modèles de forme actifs* (Cootes et al., 1994) est mise en oeuvre pour la localisation, le suivi et la paramétrisation des lèvres sur une séquence d'images représentant le locuteur.

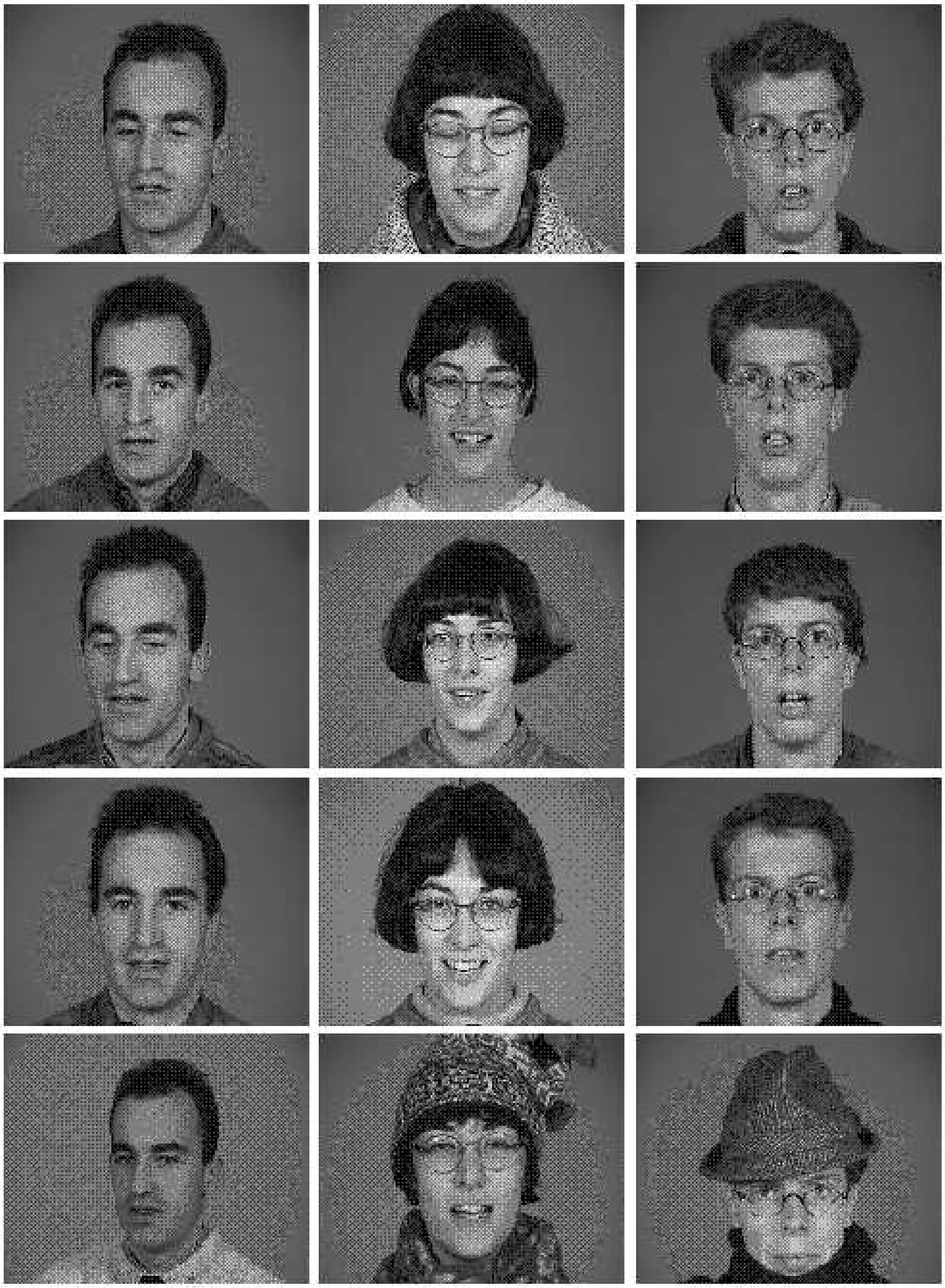

FIG. 4.2: Images extraites de la base M2VTS, les 5 sessions sont représentées de haut en bas

Les *modèles de forme actifs* sont des modèles déformables qui représentent un contour par un ensemble de points. Les principaux modes de déformation sont obtenus par *analyse en composante principale (ACP)* sur un ensemble d'apprentissage. Ceci permet d'obtenir un ensemble réduit de paramètres qui décrivent la forme labiale.

Les coefficients affectés aux modes de déformation qui décrivent la configuration du modèle lorsqu'il est appliqué à l'image des lèvres sont utilisés comme paramètres de forme.

Un modèle de niveaux de gris est utilisé pour représenter les valeurs d'intensité autour de la zone buccale. Il décrit les profils à une dimension des niveaux de gris. Ces profils sont centrés sur les points du modèle de forme et sont perpendiculaires au contour. Leur ensemble constituent un vecteur de profil global. Une analyse en composantes principales est alors réalisée, afin d'obtenir les principaux modes de variation de niveaux de gris.

La distance préalablement définie entre un modèle et une image est minimisée en utilisant l'algorithme du *simplexe*. Les contours labiaux sont déterminés lors de la convergence de cette fonction de minimisation.

Les coefficients permettant de mettre en correspondance le modèle et l'image serviront de paramètres d'intensité.

4.4.2 Suivi des contours labiaux

Les expérimentations ont été réalisées sur la base de données M2VTS. Celle-ci est composée d'images en couleurs, ici converties en niveaux de gris.

Les exemples des trois premières sessions sont utilisés pour construire le modèle de lèvres. Ce dernier nous permet d'effectuer le suivi des contours labiaux sur toutes les séquences d'images des cinq sessions, soit plus de 20000 images (voir section 4.3, 77).

Il est important d'évaluer les performances de l'algorithme de suivi, ce qui a été réalisé par inspection visuelle des résultats de la détection (Luettin et al., 1996a). Toutefois, cette tâche étant très subjective et laborieuse, nous nous limiterons ici à l'évaluation directe des performances en reconnaissance du locuteur. Il faut par conséquent noter que les erreurs de détection des lèvres et de classification ne seront pas différenciées dans ces résultats. Des exemples d'extraction des contours labiaux sont montrés en figure 4.3.

4.4.3 Identification visuelle du locuteur

En dehors de la vérification de la qualité du système de suivi de contours labiaux, il est également important d'obtenir un premier aperçu du pouvoir discriminant de ces nouveaux paramètres dans le cadre de l'identification

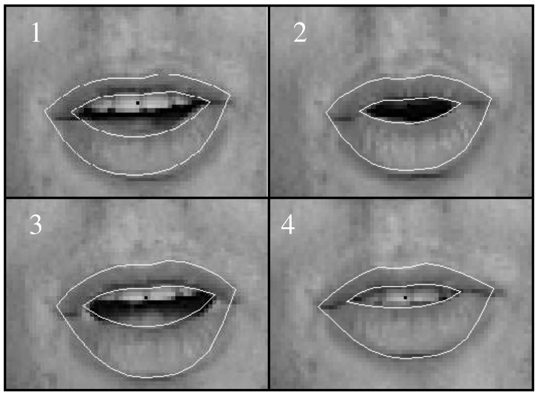

FIG. 4.3: Exemple de suivi de coutours labiaux

du locuteur. Afin de faciliter l'évaluation de cette paramétrisation du film de la région buccale nous proposons, dans un premier temps, un système d'identification du locuteur d'une très grande simplicité.

Pour ces premiers tests, les caractéristiques labiales d'une personne sont représentées par une mixture de distribution Gaussienne de probabilité. Cette modélisation est indépendante du texte prononcé. L'information temporelle portée par le mouvement des lèvres est modélisée non pas par des changements d'états, mais par l'ajout des derivées dans le vecteur d'observation. Ce dernier est composé des paramètres de forme, d'intensité ou de la combinaison des deux. Les composantes sont invariantes à l'échelle, à la translation et à la rotation. L'invariance à l'illumination est, bien entendu, uniquement vérifiée pour les paramètres de forme.

Le vecteur de forme comprend 14 composantes, le vecteur d'intensité en comprend 10. Pour certaines expérimentations, les dérivées du premier (Δ) et second ($\Delta\Delta$) ordre de ces paramètres seront prises en considération.

Les trois premières sessions sont utilisées en apprentissage, la quatrième pour le test. Les résultats sont reportés dans le tableau 4.1. Ces taux d'identification correcte sont calculés de manière classique : pour chaque locuteur, une erreur est comptabilisée si le modèle de plus forte vraisemblance n'est pas celui qui correspond à la référence. Les références des locuteurs ayant été réalisées à partir des 3 premières sessions d'enregistrement de la base M2VTS et la cinquième étant réservée pour l'évaluation finale, seulement 36 tests ont pu être réalisés pour le calcul de chaque taux.

Paramètres	1 distribution	2 distributions	3 distributions
forme	68	68	57
forme + Δ	70	65	62
forme + Δ + $\Delta\Delta$	73	59	54
intensité	78	78	86
intensité + Δ	81	86	84
intensité + Δ + $\Delta\Delta$	84	86	81
forme + intensité	86	*92*	78
forme + int. + Δ	89	*92*	84
forme + int. + Δ + $\Delta\Delta$	89	81	78

TAB. 4.1: Taux d'identification labiale correcte en fonction des paramètres utilisés

4.4.4 Interprétation

Un phénomène classique apparaît ici : l'augmentation du nombre de paramètres, qu'il s'agisse de la taille du vecteur d'observation ou du nombre de distributions gaussiennes, peut faire chuter les taux d'identification correcte. En effet, l'ensemble d'apprentissage ne nous permet pas d'estimer correctement un trop grand nombre de paramètres.

On peut néanmoins remarquer qu'une tendance se dégage, sur la base des meilleurs taux obtenus quel que soit le nombre de distributions utilisées :

- Les paramètres d'intensité semblent porter plus d'information dépendante de la personne que les paramètres de forme.

- Les dérivées de premier et second ordre sont susceptibles d'améliorer les performances.

Enfin, les meilleurs taux sont obtenus en utilisant la totalité des paramètres et de leurs dérivées. En effet, si l'ajout des accélérations ne permet pas d'obtenir les meilleures performances, il est très probable que cela soit dû a un rapport défavorable entre nombre de données et nombre de paramètres à estimer.

Ceci tend a confirmer l'hypothèse selon laquelle l'image des lèvres, mais aussi leur mouvement, contiennent une information non négligeable sur l'identité d'une personne.

Ces premières expérimentations et résultats sont donc encourageants, les meilleurs taux indiquant 92% d'identification correcte. Il faut cependant se garder de conclusions trop hâtives au vu du faible nombre de tests réalisés.

4.5 Vérification du locuteur

4.5.1 Protocole de test

Dans ces expérimentations, nous utilisons deux types de modèles. Un premier représente les *clients*, c'est-à-dire les personnes référencées et autorisées à accéder au système. Ces modèles sont entraînés avec les trois séquences de chiffres (trois sessions d'acquisition) correspondant à chaque client.

Un second modèle, appelé modèle *du monde*, résume les caractéristiques d'un nombre maximal de locuteurs référencés ou non dans le système. Son apprentissage est réalisé avec 500 locuteurs pour l'acoustique et 36 pour le visuel. Le score d'un modèle client pouvant être plus ou moins élevé en fonction du texte considéré, l'utilisation d'un modèle du monde va permettre la normalisation nécessaire.

L'ensemble de validation est utilisé pour le calcul des fonctions de normalisation et de projection. Il nous permet en outre de réaliser l'estimation de la pondération acoustico-labiale et des seuils de rejet acoustiques et labiaux.

Le sujet 37 n'est utilisé qu'en tant qu'imposteur, prétendant à l'identité des 36 clients. D'autre part, chaque client sera également considéré comme imposteur des 35 autres.

Ceci devait nous permettre, pour de futurs tests, d'utiliser la méthode dite du *leave-one-out* (Ney, 1995b). Cette dernière a été adoptée par les membres du projet M2VTS afin d'augmenter le nombre de tests sans avoir à augmenter la taille de la base de données. Néanmoins, les travaux décrits ici ayant été réalisés à un stade précoce du projet, nous n'avons pas pu mettre en œuvre cette méthode. Une conséquence notable est que pour 36 tests d'imposture, le modèle du monde visuel contient les références de l'imposteur.

Enfin, la vérification est effectuée en mode dépendant du texte ; elle est basée sur la séquence complète des 10 chiffres. Pour chaque chiffre prononcé, nous calculons la vraisemblance du modèle client et du modèle monde. Nous obtenons ainsi une vraisemblance *client* $L_c(O)$ et *monde* $L_w(O)$ sur l'ensemble du signal de parole, silences exceptés. La différence entre le rapport de ces deux scores et le seuil (t) est projetée dans l'intervalle [0, 1] par le biais d'une fonction sigmoïdale (Genoud et al., 1996) :

$$S(c) = \frac{1}{1 + \exp\left(-\left(\frac{L_c(O)}{L_w(O)} - t\right)\right)} \tag{4.1}$$

Si cette valeur $S(c)$ est inférieure ou égale à 0.5 le locuteur est rejeté, il est accepté dans le cas contraire.

Plusieurs méthodes peuvent être utilisées pour trouver un seuil de décision *a priori* et ce, en se référant à divers critères : Taux d'erreur égaux, Méthode de Furui (Furui, 1994), etc. Néanmoins, en regard du faible nombre de données disponibles pour chaque locuteur, nous avons choisi un seuil indépendant du locuteur, calculé par dichotomie sur les résultats obtenus avec l'ensemble de validation.

La valeur du seuil permettant d'obtenir le plus faible taux cumulé d'erreurs de fausse acceptation et faux rejet sera utilisée pour le test final. Le choix de ce critère d'estimation est issu des réflexions menées au cours du chapitre 2.

4.5.2 Vérification acoustique du locuteur

Dans nos expérimentations, les suites de mots sont connues[1]. Ceci nous permet de segmenter les phrases en chiffres, au moyen d'un système de reconnaissance de la parole fondé sur les MMC. Afin de détecter les frontières des mots, ce système met en correspondance la suite des modèles correspondant à la suite de chiffres donnée avec le signal de parole (algorithme de Viterbi appliqué à l'alignement). Ces modèles ont été préalablement entraînés avec la base *Polyphone* de l'IDIAP (Chollet et al., 1995). Chaque modèle de chiffre utilisé pour l'alignement résume les caractéristiques d'entre 110 et 200 exemples de prononciation par 835 locuteurs.

À partir cette segmentation obtenue sur la base M2VTS, un modèle est entraîné pour chaque chiffre et chaque client. Nous obtenons de cette faqn, 10 modèles mono-locuteurs pour chaque personne de la base M2VTS. Les 10 modèles du monde (un modèle multi-locuteur pour chaque chiffre) sont quant à eux entraînés avec 300 exemples provenant de 500 locuteurs, provenant de la base *Polyphone*.

Dans la phase de vérification proprement dite, les deux types de modèles (client et monde) sont similaires dans leur structure : Les paramètres acoustiques utilisés sont 13 coefficients CCPL (Coefficients Cepstraux de Prédiction Linéaire) auquels sont ajoutées leurs dérivées et accélérations (39 composantes au total). Ces modèles ont une topologie *gauche-droite* et leur nombre d'états, dépendant de la durée de prononciation du chiffre, varie typiquement entre 2 et 7. À chaque état correspond une distribution Gaussienne à matrice de variance-covariance diagonale.

Quand un test d'accès est réalisé, le signal est tout d'abord segmenté en chiffres. Le protocole de test défini en section 4.5.1 est alors appliqué.

Sur l'ensemble des données de tests et avec les seuils fixés sur l'ensemble de validation, nous obtenons un taux de fausse acceptation de $2,3\%$ et de

[1] Dans une application, il peut être demandé à la personne de lire un texte sur un écran, par exemple.

faux rejet de 2, 8%. Le taux d'identification correcte sur les 36 locuteurs est de 97, 2% (voir figure 4.4 ainsi que les tableaux 4.2 et 4.3, p. 89).

Cependant, il est important de préciser que seulement 36 tests ont été conduits pour le calcul des taux d'identification correcte et de faux rejet. En revanche, 1332 (36×37) tests ont été effectués pour le taux de fausse acceptation.

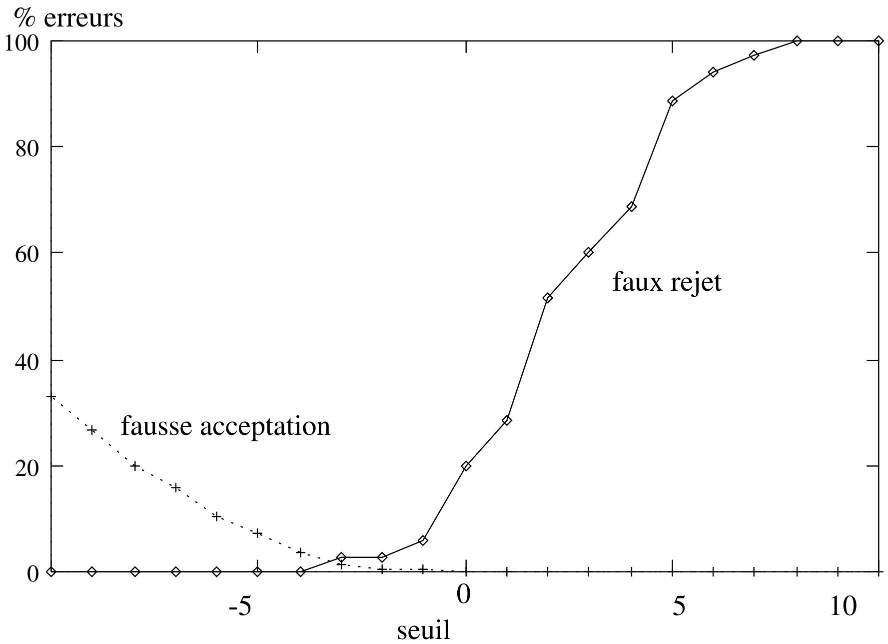

FIG. 4.4: Résultats de la vérification acoustique sur l'ensemble de validation, en fonction de la valeur de seuil t (définition 4.1)

4.5.3 Vérification labiale du locuteur

Nous considérons que la segmentation temporelle des données labiales est la même que celle obtenue pour le signal acoustique. En effet, si les paramètres labiaux peuvent améliorer les résultats de la reconnaissance de la parole (voir tableau 3.7 , p. 70), ils sont en eux-même très difficiles à segmenter (voir figure 3.1, p. 55). Si les données relatives aux mouvements des lèvres peuvent se révéler utiles pour la segmentation de la parole, il semble que ce soit essentiellement dans un environnement acoustique bruité (Mak et Allen, 1994).

Rappelons que la partie acoustique de la base de donnée M2VTS a été enregistrée dans un environnement protégé. D'autre part, nos modèles acoustiques étant entraînés sur une base de données beaucoup plus importante que cette dernière, ils sont par conséquent beaucoup plus fiables que leurs équivalents labiaux. Dans ces conditions, le fait d'utiliser comme segmentation temporelle des données labiales, celle obtenue pour le signal acoustique semble être un choix approprié.

L'obtention de scores est réalisée de la même façon que pour les données acoustiques, excepté le fait que le modèle labial du monde n'est entraîné que sur les 36 locuteurs de la base M2VTS. Les données labiales sont échantillonnées à une fréquence quatre fois inférieure aux paramètres acoustiques. Dans le but d'obtenir une qualité d'estimation comparable, il est donc nécessaire de construire des modèles labiaux plus simples que leurs correspondants acoustiques.

Par conséquent, nous avons fixé leur nombre d'états à 1 ou 2 et le nombre de composantes de leurs vecteurs à 25 : 14 paramètres de forme, 10 d'intensité et le facteur d'échelle.

Sur l'ensemble de test, nous avons ainsi obtenu un taux de fausse acceptation de $3,0\%$, de faux rejet de $27,8\%$ et d'identification correcte de $72,2\%$ (voir figure 4.5 et tableau 4.3, p. 90).

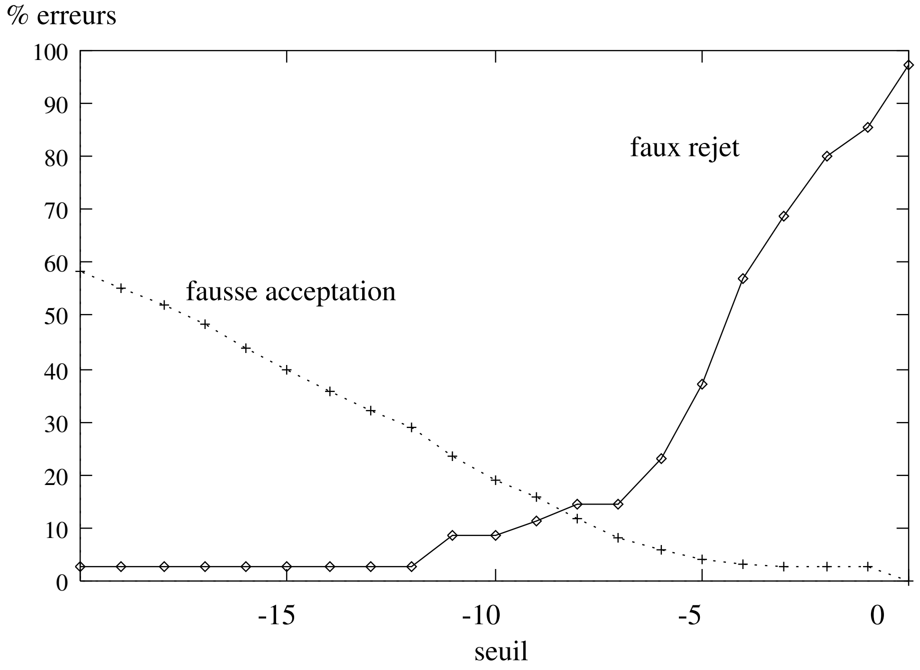

FIG. 4.5: Résultats de la vérification labiale sur l'ensemble de validation, en fonction de la valeur de seuil t (définition 4.1)

4.5.4 Vérification acoustico-labiale

Nous nous focalisons ici, tout comme au cours du chapitre 2, sur le problème de la pondération et de son estimation. Nous n'étudions pas différentes règles de fusion mais une seule, simple, mais pouvant mener à des décisions d'acceptation ou de rejet différentes suivant la valeur de la pondération.

Le score acoustico-labial est calculé comme étant la somme pondérée des scores acoustiques et labiaux. Ces derniers ont été normalisés comme décrit précédemment. Le processus de fusion est caractérisé par l'utilisation de valeurs individuelles de seuils pour chaque modalité et par la projection des deux scores dans un même intervalle.

Cette normalisation est un point critique dans la construction d'un schéma d'intégration. Cependant, les deux modalités possèdent encore, à ce niveau, des degrés de fiabilité différents. Il est par conséquent nécessaire de pondérer la contribution de chacune des sources d'information en fonction de leur fiabilité relative.

Si lors d'un test d'accès, une personne obtient un score acoustique S_a et un score labial S_l, son score acoustico-labial sera de $\alpha S_a + (1 - \alpha)S_l$. Il est important de rappeler que S_a et S_l sont issus d'une projection après seuillage dans l'intervalle $[0, 1]$ et sont par conséquent porteurs de la décision d'acceptation ou de rejet. La fusion pondérée est ainsi à la fois effectuée sur les scores et sur les seuils de rejets. Autrement dit, un locuteur peut être accepté sur la modalité acoustique ($S_a > 0,5$), rejeté par la modalité labiale ($S_l < 0,5$) et définitivement rejeté par combinaison des deux scores ($\alpha\ S_a + (1 - \alpha)S_l < 0,5$).

Pour les raisons évoquées au cours du chapitre 2, nous cherchons directement la valeur de α qui minimise le taux cumulé d'erreurs (FA + FR) sur l'ensemble de validation. Cependant, dans le cadre de ces expérimentations, nous n'avons qu'un seul poids à estimer qui, de surcroît, n'influe pas sur la segmentation du signal. Une simple recherche dichotomique sur l'ensemble de validation peut ainsi, efficacement remplacer la méthode fondée sur le *simplexe* décrite en section 2.4.2 (p. 39).

En utilisant un poids α de $0,86$, calculé sur l'ensemble de validation, nous obtenons sur l'ensemble de test un taux de fausse acceptation de $0,5\%$, de faux rejet de 3% et d'identification correcte de 100%. Il est important de rappeler ici que les sessions d'apprentissage et de test sont séparées au minimum d'une semaine et au maximum de quatre semaines.

La figure 4.6 montre les effets de la pondération sur les résultats du système acoustico-labial, les seuils d'acceptation étant optimalement fixés pour chaque modalité sur l'ensemble de validation. Les tableaux 4.2 et 4.3

résument ces résultats.

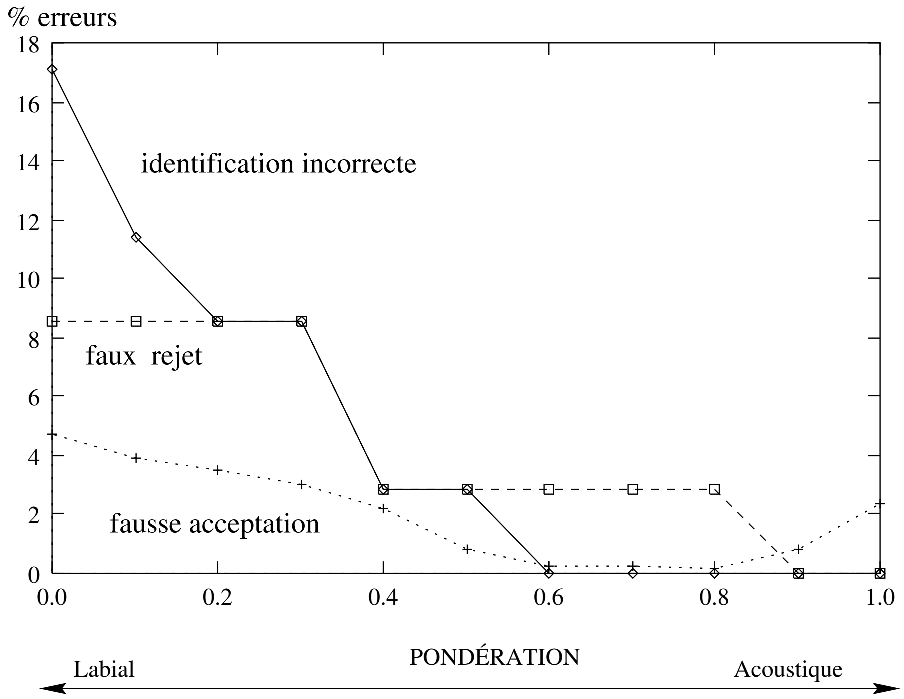

FIG. 4.6: Résultats de la vérification acoustico-labiale sur l'ensemble de validation en fonction de la valeur de pondération

Type de modèle	ID	FA	FR
Acoustique	100	2,5	0
Labial	82	4,9	9
Bimodal	100	0,6	0
Nombre de tests	36	1332	36

ID : Identification correcte, FA : Fausse acceptation, FR : Faux rejet.

TAB. 4.2: Résultats sur l'ensemble de validation

4.5.5 Processus de fusion

Il existe de nombreuses possibilités pour combiner les décisions et il est nécessaire de les expérimenter dans ce cadre précis. Néanmoins, dans un premier temps, il était opportun d'avoir une idée de l'amélioration que l'on peut attendre de la prise en compte des paramètres labiaux dans ce domaine.

Type de modèle	ID	FA	FR
Acoustique	97	2,3	3
Labial	72	3,0	28
Bimodal	100	0,5	3
Nombre de tests	36	1332	36

ID : Identification correcte, FA : Fausse acceptation, FR : Faux rejet.

TAB. 4.3: Résultats sur l'ensemble de test

Dans le but de créer un processus de fusion le plus possible guidé par le type des données à intégrer, nous avons introduit la fusion aux différents niveaux de ce système.

Dans une première étape, l'apprentissage et le décodage labial utilisent la segmentation produite par les modèles acoustiques. La première normalisation est réalisée avec le ratio des vraisemblances du client et du monde pour chacune des deux modalités. Une seconde est caractérisée par la recherche séparée d'une projection optimale dans l'intervalle $[0, 1]$.

À ce niveau d'intégration, les deux types de scores sont normalisés mais nous savons que chaque modalité possède son propre degré de fiabilité. La dernière étape de la fusion est la recherche de la pondération optimale à appliquer aux deux sources d'information. Un aperçu global du système de vérification acoustico-labial est représenté en figure 4.7.

4.6 Conclusion

Sur la base de techniques existantes dans le domaine de la reconnaissance de formes (Luettin et al., 1996a) et de la vérification du locuteur (Genoud et al., 1996), nous avons construit et expérimenté un modèle d'intégration audiovisuelle pour la vérification automatique de l'identité.

La partie acoustique de la base M2VTS peut paraître de faible taille en regard d'autres bases acoustiques de parole. Cependant, elle constitue à l'heure actuelle, une des plus grandes bases de données pour la vérification audiovisuelle de l'identité.

De plus, le suivi de contours labiaux en conditions réelles est un champ de recherche assez récent et il est réputé comme étant un problème complexe. En dépit de ces conditions difficiles, les résultats obtenus sont très encourageants (taux d'identification labiale correcte de 92%, voir page 83).

Le nombre de tests effectués est assez faible, en comparaison d'autres expérimentations de vérification acoustique du locuteur. Cependant, la réduc-

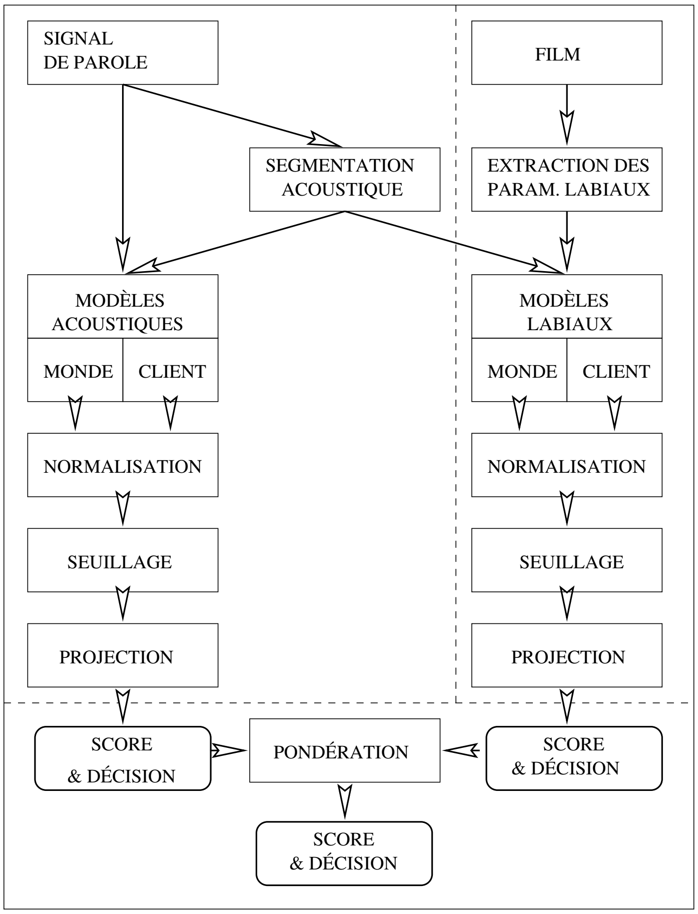

FIG. 4.7: Schéma de la vérification acoustico-labiale du locuteur.

tion d'erreurs de fausse acceptation constatée pour le système multimodal (0,5% contre 2,3% pour le système acoustique) suggère qu'il y a complémentarité entre ces deux sources d'information. En faisant l'hypothèse que les acceptations correctes suivent une distribution binomiale, l'intervalle de confiance à 99% des taux de fausse acceptation est de $[0,2\%, 1,26\%]$ pour le système multimodal et de $[1,5\%, 3,6\%]$ pour le système acoustique.

L'utilisation supplémentaire des informations visuelles portées par les lèvres peut mener à l'amélioration des performances d'un système de vérification de l'identité basé sur la parole. Dieckmann et al. (1997) ont également expérimenté des paramètres labiaux dynamiques en combinaison avec des paramètres acoustiques et rejoignent notre avis à travers leurs expériences. Cependant, il nous faudra attendre la parution d'un plus grand nombre d'études afin de pouvoir quantifier avec précision cette amélioration.

CONCLUSION ET PERSPECTIVES

Mes travaux en reconnaissance automatique de la parole ont été réalisés au LIA dans le cadre du projet AMIBE. Mes toutes premières expérimentations, caractérisées par le traitement de l'information audiovisuelle dans sa globalité (système synchrone et non pondéré) ont fait apparaître les enjeux d'une telle approche de la reconnaissance de la parole : dans un environnement bruité, l'utilisation d'informations visuelles compensait la perte de fiabilité du système acoustique. Cependant, ces résultats ont été tempérés par la relative simplicité du processus d'intégration qui provoquait le phénomène suivant : le mouvement des lèvres, portant moins d'information que le signal acoustique, s'apparentait d'une certaine façon à du bruit et faisait par conséquent légèrement chuter les taux de reconnaissance.

J'ai alors montré que l'on pouvait tirer parti d'une gestion des phénomènes d'anticipation et de rétention dans un système fondé sur les modèles de Markov cachés. Cette gestion a permis d'atteindre un autre but : l'amélioration des performances d'un système de reconnaissance de la parole dans un environnement acoustique protégé. La méthode développée pour gérer ces phénomènes d'asynchronie présente d'autres avantages, non moins importants. Elle permet tout d'abord la fusion de deux modèles de topologies différentes : nous pouvons donc choisir la topologie optimale pour chaque partie du vecteur d'information et pour chaque unité de reconnaissance. D'autre part, la recherche des différentes asynchronies possibles est effectuée durant le décodage et par conséquent le nombre de données nécessaires à la phase d'apprentissage ne subit pas d'augmentation. Enfin, elle ne requiert aucune modification des algorithmes classiques utilisés et peut donc être appliquée à n'importe quel système probabiliste.

La deuxième étape de ce travail est constituée par le traitement d'un problème crucial en fusion de données ou de décision : les deux flux d'informations ayant une fiabilité différente suivant le contexte observé, il était par conséquent nécessaire d'étudier le problème de la pondération. J'ai donc proposé une approche pour estimer un poids acoustico-labial différent pour chaque unité de reconnaissance. Les expérimentations menées dans ce cadre montrent une certaine supériorité du système à pondérations optimales sur un système équivalent mais à pondérations égales.

Enfin, le signal de parole, qu'il soit acoustique ou visuel, ne porte pas uniquement un message linguistique mais aussi une information sur l'identité du locuteur. Il m'est donc apparu intéressant d'étudier le problème de l'identification et de la vérification du locuteur sur de la parole audiovisuelle. Ces travaux ont été réalisés dans le laboratoire de l'IDIAP et dans le cadre du projet européen M2VTS. Sur la base de techniques existantes dans le domaine de la reconnaissance de formes et de la vérification du locuteur, j'ai alors construit et expérimenté un modèle d'intégration audiovisuelle pour la vérification automatique de l'identité. La partie acoustique de la base M2VTS peut paraître de faible taille en regard d'autres bases acoustiques de parole. Cependant, elle est très certainement la plus grande base de données pour la vérification d'identité audiovisuelle disponible à l'heure actuelle. De plus, l'extraction de paramètres labiaux en conditions réelles est un champ assez nouveau du domaine du traitement automatique d'images et il est réputé comme étant un problème complexe. En dépit de ces conditions difficiles, les résultats obtenus sont très prometteurs et encourageants.

La supériorité d'un système multimodal sur ses équivalents unimodaux, provient de la complémentarité des deux sources d'information acoustique et labiale. Ceci semble en effet se vérifier dans les deux cadres principaux du traitement automatique de la parole. Cette complémentarité n'est pas, à mon avis, la seule source d'amélioration. Pour répondre aux questions posées en avant-propos, je me permettrai un point de vue un peu plus général : dans mes propres travaux comme dans ceux menés par d'autres équipes et dans d'autres domaines (parole unimodale, reconnaissance de visage), force est de constater que le simple fait de considérer l'information à traiter comme étant composée de différents flux d'informations enrichit la problématique et peut, par voie de conséquence, améliorer le processus de classification.

Perspectives

Cependant, de nombreux travaux restent à accomplir après cette première approche :

Tout d'abord, je n'ai que très peu abordé les problèmes liés à la reconnaissance de formes. Il est cependant clair qu'ils prennent une place importante dans un système de traitement audiovisuel de l'information et que de nombreux progrès peuvent être réalisés dans ce domaine.

Nous avons vu que le problème de l'asynchronie ne peut être résolu de façon complète. En partant de ce constat, il serait souhaitable de trouver le compromis optimal entre temps de calcul et importance de la prise en compte de l'asynchronie. De plus, le produit de MMC permet la fusion de deux modèles de topologies différentes. Rechercher une topologie spécialement adaptée à chacune des deux modalités peut donc se révéler d'un grand intérêt.

Une autre source d'amélioration pourrait être l'utilisation de critères de pondération plus nombreux et plus variés : environnement sonore, phonèmes concernés, traits acoustiques ou visuels observés, identité du locuteur, etc. Cette pondération pourrait en outre être plus précise au niveau temporel, par exemple, dépendante de l'état du modèle concerné. Enfin, il peut être profitable d'appliquer cette méthode dans les processus d'intégration d'autres sources d'informations ou de connaissances. Une des perspectives les plus prometteuses est en effet l'estimation d'une pondération optimale entre modèles acoustiques et modèles linguistiques.

En ce qui concerne la vérification d'identité, nous n'avons pris en compte que la partie labiale de l'information visuelle. La combinaison du modèle de lèvres avec les autres caractéristiques du visage apparaît comme un prolongement naturel de ces travaux. Dans ce domaine également, la vérification de la synchronisation entre signal acoustique et labial permettrait d'atteindre un niveau supplémentaire de sécurité.

Enfin, en dehors des recherches qu'il reste à mener dans ces différents domaines, un pas décisif sera franchi par l'obtention de bases de données audiovisuelles de plus grande taille. Ceci nous permettrait d'obtenir des modèles et des paramètres mieux estimés mais aussi des résultats plus significatifs.

Bibliographie

Abry, C. et Boë, L.-J. (1986). Laws for lips, *Speech Communication* **5**(1) : 97–104.

Abry, C. et Lallouache, M. T. (1995). Modelling lip constriction anticipatory behaviour for rounding in french with the movement expansion model, *Proceedings of the International Congress of Phonetic Sciences*, Vol. 4, pp. 152–155.

Acheroy, M., Beumier, C., Bigün, J., Chollet, G., Duc, B., Fischer, S., Genoud, D., Lockwood, P., Maitre, G., Pigeon, S., Pitas, I., Sobottka, K. et Vandendorpe, L. (1996). Multi-modal person verification tools using speech and images, *Proceedings of the European Conference on Multimedia Applications*, pp. 747–761.

Adjoudani, A. et Benoît, C. (1995). Audio-visual speech recognition compared across two architectures, *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Vol. 2, pp. 1563–1566.

Adjoudani, A. et Benoît, C. (1996). On the integration of auditory and visual parameters in a HMM-based ASR, *in* D. Storck et M. Hennecke (eds), *Speechreading by Humans and Machines*, Vol. 150 of *NATO ASI Series*, Springer-Verlag, pp. 461–473.

Adjoudani, A. (1997). *Reconnaissance automatique de la parole audiovisuelle*, PhD thesis, INPG-Grenoble.

Alissali, M., Deléglise, P. et Rogozan, A. (1996). Asynchronous integration of visual information in an automatic speech recognition system, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, pp. 34–37.

Bahl, L., Brown, P., de Souza, P. et Mercer, R. (1986). Maximum mutual information estimation of hidden markov models parameters, *International Conference on Acoustics, Speech and Signal Processing*, Tokyo, pp. 49–52.

Baker, J. (1975). Stochastic modelling for automatic speech understanding, *in* D. Reddy (ed.), *Speech Recognition*, Academic Press, New York, pp. 512–542.

Besacier, L. et Bonastre, J.-F. (1997). Subband approach for automatic speaker recognition : Optimal division of the frequency domain, *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 196–202.

Bigün, E. S., Bigün, J., Duc, B. et Fischer, S. (1997). Expert conciliation for multi modal person authentication systems by bayesian statistics, *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 291–300.

Bocchieri, E. L. et Wilpon, J. G. (1993). Discriminative features selection for speech recognition, *Computer Speech and Language* **7** : 229–246.

Bourlard, H. et Dupont, S. (1996). A new ASR approach based on independant processing and recombination of partial frequency bands, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, USA, pp. 426–429.

Bregler, C., Hild, H., Manke, S. et Waibel, A. (1993). Improving connected letter recognition by lipreading, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 557–560.

Bregler, C. et Konig, Y. (1994). Eigenlips for robust speech recognition, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 669–672.

Bridle, J. (1989). Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition, *in* F. Fogelman-Soulie et J. Herault (eds), *Neuro-Computing : Algorithms, Architectures and Applications*, NATO ASI Series in Systems and Computer Science, Springer.

Brooke, N. M., Tomlinson, M. J. et Moore, R. (1994). Automatic speech recognition that includes visual speech clues, *Proceedings of the Institute of Acoustics*, Vol. 16, pp. 15–22.

Brooke, N. M. (1996). Using the visual component in automatic speech recognition, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 4, Philadelphia, pp. 1656–1659.

Brugnara, F., de Mori, R., Giuliani, D. et Omologo, M. (1992). A family of parallel hidden markov models, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 377–380.

Brunelli, R. et Falavigna, D. (1995). Person identification using multiple cues, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **17**(10) : 955–966.

Cathiard, M.-A., Lallouache, M. T., Mohamadi, T. et Abry, C. (1995). Configurational vs. temporal coherence in audiovisual speech perception, *Proceedings of the International Congress of Phonetic Sciences*, Vol. 3, pp. 218–221.

Chellappa, R., Wilson, C. L. et Sirohey, S. (1995). Human and machine recognition of faces : A survey, *Proceedings IEEE* **83**(5) : 705–740.

Chibelushi, C. C., Deravi, F. et Mason, J. S. (1993a). Voice and facial image integration for person identification, *in* R. Damper, W. Hall et J. W. Richards (eds), *Multimedia Technologies and Future Applications*, ISBN 0-7273-1320-7, Pentech Press, London, pp. 155–161.

Chibelushi, C. C., Mason, J. S. et Deravi, F. (1993b). Integration of acoustic and visual speech for speaker recognition, *Proceedings of the European Conference on Speech Communication and Technology*, Vol. 1, pp. 157–160.

Chollet, G., Cochard, J., Constantinescu, A. et Langlais, P. (1995). Swiss french polyphone and polyvar : Telephone speech databases to study intra and inter speaker variability, *Technical report*, IDIAP, Martigny, Suisse.

Chow, Y.-L. (1990). Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-BEST algorithm, *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 701–704.

Cootes, T. F., Hill, A., Talor, C. J. et Haslam, J. (1994). Use of active shape models for locating structures in medical images, *Image and Vision Computing* **12**(6) : 355–365.

Cosi, P., Caldognetto, E. M., Ferrero, F., Dugatto, M. et Vagges, K. (1996). Speaker independent bimodal phonetic recognition experiments, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, pp. 54–57.

Cosi, P., Caldognetto, E. M., Vagges, K., Mian, G. A. et Contolin, M. (1994). Bimodal recognition experiments with recurrent neural networks, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 553–556.

Daniels, R. W. (1978). *An Introduction to Numerical Methods and Optimization Techniques*, Elsevier North-Holland, Inc.

del Álamo, C. M., Gil, F. J. C., de la Torre-Munilla, C. et Hernández-Gómez, L. (1995). Codebook weights adaptation for discriminative training of SCHMM-based speech recognition systems, *Proceedings of European Conf. on Speech Communication and Technology*, pp. 93–96.

Dieckmann, U., Plankensteiner, P. et Schamburger, R. (1997). SESAM : A biometric person identification system using sensor fusion, *Pattern Recognition Letters* **18**(9) : 827–833.

Duc, B., Maître, G., Fischer, S. et Bigün, J. (1997). Person authentication by fusing face and speech information, *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 311–318.

Falavigna, D. et Brunelli, R. (1994). Person recognition using acoustic and visual cues, *Proceeding of the ESCA Workshop on Automatic Speaker Recognition, identification and Verification*, pp. 71–74.

Finn, K. E. et Montgomery, A. A. (1988). Automatic optically-based recognition of speech, *Pattern Recognition Letters* **8**(3) : 159–164.

Furui, S. (1994). An overview of speaker recognition technology, *Proceedings of the ESCA Workshop on Automatic Speaker Recognition Identification Verification*, Martigny, Suisse, pp. 1–9.

Gales, M. et Young, S. (1992). An improved approach to the hidden markov model decomposition of speech and noise, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 233–236.

Genoud, D., Bimbot, F., Gravier, G. et Chollet, G. (1996). Combining methods to improve speaker verification decision, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 3, Philadelphia, USA, pp. 1756–1759.

Goldschen, A. J., Garcia, O. N. et Petajan, E. D. (1994). Continuous optical speech recognition by lipreading, *Proceedings of the 28th Conference on Signals, Systems and Computer*, pp. 572–577.

Graf, H. P., Chen, T., Petajan, E. D. et Cosatto, E. (1995). Locating faces and facial parts, *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, Vol. 1, Zurich - Suisse, pp. 41–46.

Hernando, J., Ayarte, J. et Monte, E. (1995). Optimization of speech parameter weighting for CDHMM word recognition, *Proceedings of European Conf. on Speech Communication and Technology*, pp. 105–108.

Hernando, J. (1997). Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1267–1270.

Jacob, B. et Senac, C. (1996). Hidden markov models merging acoustic and articulatory information to automatic speech recognition, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 4, Philadelphia, pp. 2313–2315.

Jelinek, F. (1976). Speech recognition by statistical methods, *Proceedings of the IEEE*, Vol. 64, pp. 532–556.

Jourlin, P. (1996a). Asynchronie dans les systèmes de reconnaissance de la parole basés sur les hmm, *XXIes Journées d'Étude de la Parole*, Vol. 1, Avignon - France, pp. 351–354.

Kabré, H. (1995). Audiovisual speech recognition using the fuzzy shape filters model, *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Vol. 1, pp. 307–310.

Kittler, J., Li, Y. P., Matas, J. et Ramos-Sánchez, M. U. (1997). Combining evidence in multimodal personal identity recognition systems, *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 327–334.

Lallouache, M. T. (1991). *Un poste "visage-parole" couleur*, PhD thesis, INPG-Grenoble.

Luettin, J., Thacker, N. A. et Beet, S. W. (1996a). Locating and tracking facial speech features, *Proceedings of the International Conference on Pattern Recognition*, Vienne - Autriche, pp. 652–656.

Luettin, J., Thacker, N. A. et Beet, S. W. (1996b). Speaker identification by lipreading, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, pp. 62–65.

Mak, M. et Allen, W. (1994). Lip-motion analysis for speech segmentation in noise, *Speech Communication* **14**(3) : 279–296.

Matthiews, I., Bangham, J. A. et Cox, S. (1996). Audiovisual speech recognition using multiscale nonlinear image decomposition, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, pp. 38–41.

Meier, U., Wolfgang, H. et Duchnowski, P. (1996). Adaptative bimodal sensor fusion for automatic speechreading, *International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 833–837.

Montacié, C., Caraty, M. J., André-Obrecht, R., Boë, L.-J., Deléglise, P., El-Bèze, M., Herlin, I., Jourlin, P., Lallouache, T., Leroy, B. et Méloni, H. (1995). Applications multimodales pour interfaces et bornes evoluées, *Ecole Thématique : Traitement automatique de la parole : Fondements et Perspectives*, Vol. 1, Marseille - France, pp. 155–164.

Movellan, J. R. et Chadderdon, G. (1995). Channel separability in audio-visual integration of speech : A bayesian approach., *in* D. G. Stork et M. E. Hennecke (eds), *Speechreading by Humans and Machines, Models, Systems and Applications*, Vol. 150, Springer-Verlag, Berlin, pp. 473–488.

Nelder, J. A. et Mead, R. (1965). The downhill simplex method, *Computer Journal* **7** : 391–398.

Ney, H. (1995a). On the probabilistic interpretation of neural network classifiers and discriminative training, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(2) : 107–119.

Ney, H. (1995b). On the estimation of "small" probabilities by leaving-one-out, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **17**(12) : 1202–1212.

Niles, L., Silverman, H. et Bush, M. (1990). Neural networks, maximum mutual information training, and maximum likelihood training, *International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, NM, pp. 493–496.

Normandin, Y., Cardin, R. et de Mori, R. (1994). High-performance connected digit recognition using maximum mutual information estimation, *IEEE Trans. on SAP* **2**(2) : 299–311.

Otani, K. et Hasegawa, T. (1995). The image input microphone – a new nonacoustic speech communication system by media conversion from oral motion images to speech, *IEEE Journal on Selected areas in Communications* **13**(1) : 42–48.

Petajan, E. D., Bischoff, B. J., Bodoff, D. A. et Brooke, N. M. (1987). An improved automatic lipreading system to enhance speech recognition, *Technical Report TM 11251-871012-11*, AT & T.

Petajan, E. D. (1984). *Automatic Lipreading to Enhance Speech Recognition*, PhD thesis, University of illinois.

Pigeon, S. et Vandendorpe, L. (1997). The m2vts multimodal face database (release 1.00), *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 403–409.

Potamianos, G., Cosatto, E., Graf, H. P. et Roe, D. B. (1997). Speaker independent audio-visual database for bimodal ASR, *Proceedings of the Audio-Visual Speech Processing workshop*, ESCA ISSN # 1018 4554, pp. 65–68.

Robert-Ribes, J., Schwartz, J.-L. et Escudier, P. (1995b). Auditory, visual and audiovisual vowel representation : Experiments and modelling, *Proceedings of the International Congress of Phonetic Sciences*, Vol. 3, pp. 114–121.

Robert-Ribes, J. (1995a). *Modèles d'Intégration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique de voyelles*, PhD thesis, INPG-Grenoble.

Rogina, I. et Waibel, A. (1990). Learning state-dependent stream weights for multi-codebook HMM speech recognition systems, *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 349–352.

Rogozan, A., Deléglise, P. et Alissali, M. (1997). Adaptative determination of audio and visual weights for automatic speech recognition, *Proceedings of the Audio-Visual Speech Processing workshop*, ESCA ISSN # 1018 4554, pp. 61–64.

Silsbee, P. L. et Su, Q. (1996). Audio-visual sensory integration using hidden markov models, *in* D. Storck et M. Hennecke (eds), *Speechreading by Humans and Machines*, Vol. 150 of *NATO ASI Series*, Springer-Verlag, pp. 489–497.

Silsbee, P. L. (1994). Sensory integration in audiovisual automatic speech recognition, *Conference Record of the Twenty-Eighth Asilomar Conference on Signal, Systems and Computers*, pp. 561–565.

Sonoda, Y., Mori, K. et Kuriyama, T. (1990). Articulatory characteristics of lip shape during the production of japanese, *Proceedings of the International Conference on Spoken Language Processing*, Vol. 1, pp. 441–444.

Su, Q. et Silsbee, P. L. (1996). Robust audiovisual integration using semicontinuous hidden markov models, *Proceedings of the International Conference on Speech and Language Processing*, Vol. 1, Philadelphia, pp. 42–45.

Tibrewala, S. et Hermansky, H. (1997). Subband-based recognition of noisy speech, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1255–1258.

Tomlinson, M. J., Russel, M. J. et Brooke, N. M. (1996). Integrating audio and visual information to provide highly robust speech recognition, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 821–824.

Tomlinson, M. J., Russel, M. J., Moore, R. K., Buckland, A. P. et Fawley, M. A. (1997). Modelling asynchrony in speech using elementary single-signal decomposition, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 1247–1250.

Varga, A. et Moore, R. (1990). Hidden markov model decomposition of speech and noise, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 845–848.

Viterbi, A. et Omura, J. (1979). *Principles of Digital Communication and Coding*, McGraw-Hill, Tokyo.

Vo, M. T., Houghton, R., Yang, J., Bub, U., Meier, U., Waibel, A. et Duchnowski, P. (1995). Multimodal learning interfaces, *Proceedings of ARPA Spoken Language Technology Workshop*, Vol. 1.

Wagner, T. et Dieckmann, U. (1994). Multi-sensorial inputs for the identification of persons with synergic computers, *Proceedings of the IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 287–291.

Watanabe, T. et Kohda, M. (1990). Lip-reading of japanese vowels using neural networks, *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, pp. 1373–1376.

Wilpon, J. G., Lee, C.-H. et Rabiner, L. R. (1991). Improvements in connected digit recognition using higher order spectral and energy features, *Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 349–352.

Wolff, G. J., Prasad, K. V., Stork, D. G. et Hennecke, M. (1994). Lipreading by neural networks : Visual preprocessing, learning and sensory integration, *in* J. D. Cowan, G. Tesauro et J. Alspector (eds), *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 1027–1034.

Wu, J. T., Tamura, S., Mitsumoto, H., Kawai, H., Kurosu, K. et Okazaki, K. (1991). Neural network vowel-recognition jointly using voice features and mouth shape image, *Pattern Recognition* **24**(10) : 921–927.

Young, S., Woodland, P. et Byrne, W. (1993). *HTK : Hidden Markov Model Toolkit V1.5*, Entropics Research Laboratories Inc.

Yuhas, B. P., Goldstein, M., Sejnowski, T. J. et Jenkins, R. E. (1990). Neural network models of sensory integration for improved vowel recognition, *Proceedings of the IEEE* **78**(10) : 1658–1668.

Annexe A

Résultats en reconnaissance de la parole audio-visuelle

TAB. A.1: Résultats en reconnaissance de la parole audio-visuelle

Ref.	Langue	Voc.	nb tests/unité	Type	Loc.	Methode	A	V	AV
(Bregler et al., 1993)	D	26 Lettres	?(20)	Cont.	Mono1	TDNN	88.8	31.6	93.2
(Bregler et al., 1993)	D	26 Lettres	?(60)	Cont.	Mono2	TDNN	97.0	46.9	97.2
(Su et Silsbee, 1996)	?(USA)	22 a-C-a	4	Iso.	Mono	HMM	89.0	?	91.0
(Cosi et al., 1996)	I	? V-C-V	45	Iso.	Multi	RNM	?	?	71.1
(Tomlinson et al., 1996)	?(GB)	10 chiffres	30	Cont	Mono	HMM	92.7	23.0	98.7
(Brooke, 1996)	-	-	-	-	-	-	-	-	-
(Kabré, 1995)	F	10 Voyelles	?(17)	Iso	Mono	FSFM	?	?	95.0
(Adjoudani et Benoît, 1995)	F	54 VCVCV	<9	Iso	Mono	HMM	75.0	85.0	98.0
(Alissali et al., 1996)	F	26 Lettres	10	Cont	Mono	HMM	?	?	96.1
(Jacob et Senac, 1996)	F	26 Lettres	10	Cont	Mono1	HMM	90.1	40	96.1
(Jacob et Senac, 1996)	F	26 Lettres	10	Cont	Mono2	HMM	45.8	?	74.0
(Jourlin, 1996a)	F	26 Lettres	10	Cont	Mono1	HMM	90.9	38.2	93.5
(Jourlin, 1996a)	F	26 Lettres	10	Cont	Mono2	HMM	73.2	24.5	78.0
(Matthiews et al., 1996)	GB	26 Lettres	<3	Iso	4Mono	HMM	78.0	50.0	85.0
(Goldschen et al., 1994)	USA	150 phrases	1	Iso	Mono	HMM	?	?	25.3

Annexe B

Projets relatifs au traitement bimodal de la parole

B.1 Le projet AMIBE

Le but du projet GDR-PRC Communication Homme-Machine AMIBE[1] (Applications Multimodales pour Interface et bornes évoluées), coordonné par C. Montacié, est d'expérimenter la notion d'interfaces Homme-Machine multimodale dans un cadre multi-utilisateurs, non pour prendre en compte toute l'activité mentale de l'utilisateur, mais pour fiabiliser la communication parlée qui est le mode de communication le plus naturel et spontané.

AMIBE propose de résoudre en partie 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 sont :

- le suivi du contour des lèvres pour la distinction de silence/parole et la localisation d'événements caractéristiques,
- la localisation et le déplacement de la source sonore pour le débruitage de la parole,
- une vérification fiable du locuteur permettant le choix du dictionnaire de références le plus adapté en fonction de la typologie du locuteur.

Cette interface multimodale, à laquelle sera associée un module de reconnaissance des contours du visage, servira également à une vérification permanente de l'identité de l'utilisateur. Ce type d'interface multimodale nous a semblé très adapté à l'accèes sécurisé d'un nombre limité d'utilisateurs à des informations personnelles ou confidentielles. Les modes utilisés, Parole et Image,

[1] cette page est extraite de http ://www-laforia.ibp.fr/PAROLE/montacie/amibe.html

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. Les laboratoires impliqués dans ce projet sont l'ICP, l'INRIA, l'IRIT, le LAFORIA, le LIA et le LIUM.

B.2 Le projet M2VTS

B.2.1 Objectifs

Le premier but du projet M2VTS[2] (Multimodal Verification for Teleservices and Security Applications) est d'aborder les questions de l'accés sécurisé à des services locaux ou centralisés dans un environement multimédia. L'objectif principal est d'étendre les champs d'applications de services en réseau en ajoutant de nouvelles fonctionalités, permises par des systèmes de vérification automatique combinant des stratégies multimodales (accès sécurisé basé sur la parole, l'image et autres types d'information).

Les objectifs sont également de montrer que les limitations des technologies unimodales (reconnaissance de la parole, vérification du locuteur...) peuvent être surmontées en se basant sur des décisions multimodales (combinaison ou fusion des ces différentes technologies) et trouver des applications importantes dans le champ émergeant des interfaces évoluées pour les téléservices.

Les buts principaux de ce projet sont par conséquent :

- D'implémenter et de valider des schémas d'accès sécurisé dans des services existant basés sur la voix.
- De développer de nouveaux services de sécurité en exploitant les technologies émergeantes de reconnaissance basée sur la parole et l'image.
- De procurer des services sécurisés sur des réseaux non sécurisés (tels que PSTN, ISDN, LAN).
- De développer de nouveaux services pour des applications de sécurité (vérification d'alarme et controle d'accès)

B.2.2 Participants

- Matra Communication (France)

[2]cette page est une traduction de
http ://www.uni-stuttgart.de/SONAH/Acts/AC102.html

- Compagnie Européenne de Télésécurité (France)
- Cerberus AG (Suisse)
- École Polytechnique Fédérale de Lausanne (Suisse)
- IDIAP : Institut Dalle molle d'Intelligence Artificielle Perceptive (Suisse)
- Institute of Microtechnology University of Neuchatel (Suisse)
- Université Catholique de Louvain (Belgique)
- Renaissance (Belgique)
- University of Surrey (Royaume-uni)
- Aristotle University of Thessaloniki (Grèce)
- Unidad Tecnica Auxiliar de la Policia (Espagne)
- Banco Bilbao Vizcaya (Espagne)
- Universidad Carlos III (Espagne)
- Ibermática S.A. (Espagne)

B.3 Le projet VIDAS

B.3.1 Objectifs

Le premier but du projet VIDAS[3] (Video Assisted Audio Coding and Representation) est d'améliorer la qualité subjective des communications audio-visuelles à de très faibles débits dans le but de donner l'opportunité aux personnes malentendantes, par exemple, d'utiliser les videophones. Pour atteindre cet objectif, il est proposé d'utiliser l'information audio pour améliorer la qualité video.

Une première approche est de tester cette amélioration globale de la communication audio-visuelle sur les videophones existant. L'objectif principal du projet est d'extraire l'information du signal de parole et d'appliquer un traitement sur l'image de la région buccale pour réaliser la synchronisation labiale. De cette façon, au travers du post-traitement du signal, il sera possible d'effectuer la conversion fréquencielle basée sur l'information acoustique.

B.3.2 Participants

- Matra Communication (France)
- ANPEDA (France)

[3]cette page est une traduction de
http ://media.it.kth.se/SONAH/Acts/AC057.html

- IRISA (France)
- Modis SPA (Italie)
- University of Genova DIST (Italie)
- AFA Centro REUL FIADDA (Italie)
- German Aerospace Research Establishment (Allemagne)
- Linkoping University (Suède)
- Universitat Polytecnica Catalunya (Espagne)
- Ecole Polytechnique Fédérale de Lausanne (Suisse)
- University of Geneva UniG (Suisse)

LISTE DES PUBLICATIONS

Revue internationale

Jourlin, P., Luettin, J., Genoud, D. et Wassner, H. (1997e). Acoustic-labial speaker verification, *Pattern Recognition Letters* (18) 9 pp. 853-858.

Congrès internationaux avec actes et comité de lecture

Jourlin, P., El-Bèze, M. et Méloni, H. (1995a). Bimodal Speech Recognition, *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, Vol. 1, Zurich, Suisse, pp. 320–325.

Jourlin, P., El-Bèze, M. et Méloni, H. (1995b). Integrating visual and acoustic information in a speech recognition system based on HMM, *Proceedings of the International Congress of Phonetic Sciences*, Vol. 4, Stockholm - Suède, pp. 288–291.

Jourlin, P. (1996b). Handling the disynchronization phenomena with HMM in connected speech, *Proceedings of the VIII European Signal Processing Conference*, Vol. 1, Trieste - Italie, pp. 133–136.

Jourlin, P., Luettin, J., Genoud, D. et Wassner, H. (1997a). Acoustic-labial speaker verification, *Audio- and Video-based Biometric Person Authentication*, ISBN 3-540-62660-3, Springer-Verlag, Berlin, pp. 319–326.

Jourlin, P., Luettin, J., Genoud, D. et Wassner, H. (1997c). Integrating Acoustic and Labial Information for Speaker Identification and Verification, *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1599-1602.

Jourlin, P. (1997d). Word-Dependent Acoustic-Labial Weights in HMM-Based Speech Recognition, *Proceedings of the ESCA/ESOP Workshop on Audio-Visual Speech Processing*, ISSN 1018-455, pp 69-72.

Jourlin, P. (1997f). Estimating Acoustic-labial Weights in Connected Speech Recognition Systems Based on HMM, *Proceedings of IEEE Int. Conf. on Systems, Man, and Cybernetics.*.

Congrès francophone avec actes et comité de lecture

Jourlin, P. (1996a). Asynchronie dans les systèmes de reconnaissance de la parole basés sur les HMM, *XXIes Journées d'Étude de la Parole*, Vol. 1, Avignon - France, pp. 351–354.

Écoles

Montacié, C., Caraty, M., Obrecht, R. A., Boë, L., P.Deléglise, El-Bèze, M., Herlin, I., Jourlin, P., Lallouache, T., Leroy, B. et Méloni, H. (1995). Applications multimodales pour interfaces et bornes evoluées, *École Thématique : Traitement automatique de la parole : Fondements et Perspectives*, Vol. 1, Marseille - France, pp. 155–164.

Chollet, G., Bigün, J., Duc, B., Fischer, S., Jourlin, P., Luettin, J., Maître, G. et Verlinde, P. (1997b). Vérification Multimodale de l'Identité d'une Personne, *École de printemps NSI'97*, Aussois, France.

Index des Auteurs

Remerciements et responsabilités

Sans le support d'un nombre considérable de gens, il m'aurait été absolument impossible de suivre ce chemin qui amène quelqu'un à produire un tel document. Je vais donc ici partager ma responsabilité en les dénonçant publiquement :

Merci à ma grand-mère, qui me soutient encore, malgré ses 84 ans et mes 84 kilos. Merci à mes parents qui n'ont jamais perdu l'espoir de faire quelque chose de moi. Merci à ma soeur pour sa gentillesse inconditionnelle. Merci à Guillaume, Corinne et Alex, pour leur compagnie dans l'université buissonnière. Merci à Karine.

Merci à Ghislain pour les joies du travail en binôme. Merci à Georges pour la découverte du Var et de ses vignobles.

Merci à Steph pour notre longue cohabitation pacifique. Merci à Philou pour ses renseignements généreux. Merci à Thierry pour la protection contre la débauche. Merci à Fred à Felipe pour l'apprentissage du dur métier de conférencier international. Merci à Jeff pour nos discussions constructives. Merci à Pascal pour les collaborations fructueuses. Merci à Loïc pour son assiduité à mes cours de Volley. Merci à Régis pour sa bonne humeur inébranlable. Merci à Thierry pour les cours de médecine sportive. Merci à Laurence et Vilou pour leur compagnie pendant les courses. Merci à Jocelyne pour les voyages. Merci à Corine.

Merci à Jean-luc de m'avoir accepté dans son laboratoire. Merci à Gilbert de m'avoir accepté dans son équipe. Merci à Jürgen pour les modèles déformables et la vie nocturne de Martigny-Ville. Merci à Gilles pour les voyages à Martigny-Bourg. Merci à Cédric pour les excursions à Martigny-Combes. Merci à Georg pour les rencontres des Alpes. Merci à Tomas pour l'Absolut. Merci à Doms pour les modèles du monde. Merci à Hubert pour les scripts. Merci à Sacha pour les virées à Lausanne. Merci à Guillaume pour sa patience.

Merci à Pierrot-Yves pour les soirées guitare et les petits déjeuners au

Guronzan. Merci à Laurent pour le coup de fil salutaire. Merci à Séverine pour les danses latines. Merci à Sylvie pour les danses tropicales. Merci à Fabienne pour les danses folkloriques. Merci à Philippe pour s'en être sorti et moi avec. Merci à Linda-Chantal pour son amour.

Thanks to CUED and CUCL people for their welcome. Thanks to SWP. Thanks to Miles for his moral support in public houses. Thanks to Martin, Jana, Sylvia, Donnla, Olivia, James for their constructive discussions in tea breaks.

Et enfin, un grand merci à ceux que j'ai cité comme à ceux que j'ai oublié, pour ne pas m'en tenir rigueur...