

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

وزارة التعليم العالي و البحث العلمي

UNIVERSITE BADJI MOKHTAR - ANNABA
BADJI MOKHTAR-ANNABA UNIVERSITY



جامعة باجي مختار - عنابة

Faculté des Sciences de l'Ingéniorat
Département d'Informatique

Année : 2015/2016

THESE

Présentée en vue de l'obtention du diplôme de
Doctorat 3^{ième} cycle

Reconnaissance automatique de la parole en milieu réel bruité par fusion audiovisuelle

Filière : Informatique

Spécialité: Reconnaissance des Formes et Intelligence Artificielle

par

Amina Makhlouf

Devant le jury:

Laskri Mohamed-Tayeb	Professeur à l'Université Badji Mokhtar-Annaba	(Président)
Bensaker Bachir	Professeur à l'Université Badji Mokhtar-Annaba	(Directeur)
Kazar Okba	Professeur à l'Université de Biskra	(Examineur)
Dib Lynda	Professeur à l'Université Badji Mokhtar-Annaba	(Examineur)
Lazli Lilia	Maître de Conférence à l'Université Badji Mokhtar	(Invitée)

« Vous ne pouvez pas énoncer une idée nouvelle autrement qu'avec des mots anciens, ceux que vous avez à votre disposition. Il va donc falloir un temps de travail pour comprendre ce que vous venez de faire. C'est ce que Bachelard appelait la 'la refonte épistémologique'. »

La science contemporaine est-elle moderne ?

Jean-Marc Lévy-Leblond, 1999.

Remerciements

Je remercie en premier lieu Allah qui m'a donné à la fois le courage, la volonté, et la patience afin d'élaborer cette thèse de recherche scientifique.

Je tiens particulièrement à remercier Monsieur le Professeur **Bensaker Bachir** et Madame **Lazli Lilia** Maître de Conférences qui sont les instigateurs de mon sujet de thèse et qui m'ont soutenu tout au long de ce travail. Monsieur Bensaker a dirigé ma thèse et m'a aidé dans toutes les démarches relatives à celle-ci. Madame Lazli m'a encadrée tout au long de mon parcours universitaire; pour son encouragement, ainsi que son soutien tout au long de la thèse. Je la remercie pour tout son aide. Son enthousiasme et sa patience ont beaucoup facilité et agrémenté mon travail. elle a été toujours disponible pour répondre aux questions que je lui posais. Ses remarques m'ont permis de faire progresser ce travail.

Je remercie Monsieur le Professeur **Laskri Med Tayeb** de l'Université Badji Mokhtar-Annaba, de m'avoir fait l'honneur de présider le jury de ma soutenance.

Un grand merci également aux membres du jury de soutenance qui m'ont fait l'honneur de bien vouloir évaluer mon travail. Je suis particulièrement reconnaissante aux examinateurs **Kazar Okba**, Professeur à l'Université de Biskra et **Dib Lynda**, Professeur à l'Université Badji Mokhtar-Annaba.

Et bien sûr, ceux sans qui je ne serais rien: mes parents, mes sœurs, mon mari, ma famille et ma belle famille et tous mes amis d'enfance qui me supportent et soutiennent depuis toujours.

Enfin, je remercie toutes les personnes (nombreuses) que je n'ai pas citées et qui, à un moment ou à un autre, m'ont donné l'envie et la force de continuer.

Résumé

La présence de bruit de fond et des conditions variables (environnement, réverbération, types de microphones) peut affecter significativement la qualité de la reconnaissance automatique de la parole (RAP). Cette thèse présente un système de reconnaissance audiovisuelle de la parole qui est un domaine de recherche qui a connu un intérêt grandissant durant ces dernières années. Notre contribution s'axe sur la vérification de ces deux conditions, c'est-à-dire la modélisation de la perception audiovisuelle de la parole en vue d'une implémentation logicielle, et de l'extraction des informations les plus pertinentes. Notre étude a été au centre d'une recherche pluridisciplinaire: de la psychologie cognitive aux techniques de traitement d'images couleurs, nous nous sommes investis dans le domaine de la paramétrisation des lèvres, le traitement du signal et la reconnaissance automatique des formes.

D'autre part, les modèles de Markov cachés (HMM) sont à l'origine de la majorité des avancées récentes en reconnaissance de la parole discrète ainsi continue. Ces modèles gèrent les distorsions temporelles du signal de parole en s'appuyant sur des densités de probabilité pour modéliser les distorsions en fréquence. Une technique de combinaison des probabilités a posteriori des états d'un HMM connaissant un vecteur de paramètres acoustiques ainsi visuels est également proposée. Afin d'améliorer l'estimation des probabilités a posteriori, les probabilités obtenues avec différents modèles acoustiques et visuels sont fusionnées. Pour combiner les probabilités de manière cohérente, les deux modèles doivent avoir la même topologie.

En partant donc de cette idée, des systèmes audiovisuels permettant l'enregistrement simultané des flux visuels et du flux acoustique ont été développés, en utilisant les HMM combinés avec les Algorithmes génétiques (GA), et respectant successivement les modèles suivants : fusion des données acoustiques et visuelles par identification directe (ID), et fusion des résultats acoustiques et visuelles après identification séparée (IS).

Afin d'évaluer l'ensemble des approches proposées, deux bases de données contenant chacune des vidéos avec une langue différente (arabe et anglaise) ont été construites et utilisées. Pour la caractérisation des images, et les signaux acoustiques deux approches basées sur l'utilisation de la transformée en cosinus discrète (DCT), et la méthode RASTA-PLP, respectivement, ont été mises en œuvre.

Nos résultats expérimentaux montrent qu'il existe en effet des informations dans la modalité visuelle utile pour la reconnaissance de la parole. Nos expériences ont aussi montré une grande possibilité d'améliorer la performance et la robustesse de notre système de reconnaissance audiovisuel proposé qui utilise la méthode hybride HMM/GA comparé avec les méthodes classiques utilisées dans la littérature.

Mots-clés: parole audiovisuelle, lecture labiale, paramétrisation, modèle de markov cachés, algorithme génétique, vision, signaux acoustique, hybridation HMM/GA

Abstract

The presence of background noise and varying conditions (environment, reverberation microphone types) can significantly affect the quality of automatic speech recognition (ASR). This thesis presents an audiovisual speech recognition system which is a research domain that has seen a growing interest during these last years. Our contribution is centered on the verification of these two conditions, i.e. the perception modeling of the audiovisual speech for a software implementation, and the extraction of the most pertinent information. Our study was the center of a pluridisciplinary research: cognitive psychology to the techniques of color image processing, we are invested in the field of lips parameterization, Signal processing and the automatic pattern recognition.

Furthermore, the Hidden Markov Models (HMM) are the origin of the majority of recent advances in the continuous and discrete speech recognition. These models support the temporal distortions of the speech signal based on the probability density for modeling the distortions frequency. Combination of a posteriori probabilities of states of a HMM given a feature frame is also proposed. In order to better estimate such a posteriori probabilities, probabilities obtained with several acoustic and visual models are fused. For the sake of consistency, the topology of the two models has to be equivalent.

Based on this idea, audiovisual systems that allow the simultaneous recording of the visual and acoustic stream has been developed, by using the HMM combined with the Genetic Algorithms (GA), according to data fusion for direct integration (DI) and result fusion for separate integration (SI).

In order to evaluate all of the proposed approaches, two databases, each containing videos using a different language (Arabic and English) were constructed and used. For the characterization of images, and the acoustic signals two approaches based on the use of the discrete cosine transform (DCT), and the RASTA-PLP method, respectively, have been implemented.

Our experimental results show that there is in fact useful information in the visual modality for speech recognition. Our experiments have also shown a great possibility to improve the performance and robustness of our proposed AVASR using the hybrid HMM/GA method compared with traditional methods in the literature.

Keywords: audiovisual speech, lip-reading, parameterization, hidden Markov model, genetic algorithm, vision, acoustic signals, hybrid HMM/GA

Table des matières

Remerciements	i
Résumé	ii
Abstract	iii
Table des matières	iv
Table des illustrations	viii
Liste des figures.....	viii
Liste des tableaux	x
Introduction	1
1. Contexte et cadre de recherche.....	1
2. Plan de la thèse	2
Première partie: Etat de l'art	1
Les lèvres et la production de la parole	5
1.1 Architecture et fonctionnement de l'appareil vocal	5
1.1.1 L'appareil vibrateur.....	5
1.1.2 Le résonateur	7
1.2 L'anatomie des lèvres.....	10
1.2.1 Les tissus	10
1.2.2 Les muscles des lèvres.....	11
1.2.3 Classification fonctionnelle des muscles labiaux	13
1.3 Repères phonétiques.....	14
1.3.1 Acoustique et articulation.....	14
1.3.2 Des sons et des lèvres	15
1.3.3 La coarticulation : cibles en contexte	17
1.4 La parole audiovisuelle et ses applications en communication	18
1.4.1 La bimodalité intrinsèque de la parole	18
1.4.2 L'intelligibilité de la parole audiovisuelle.....	20
1.4.3 Perspectives pour la communication homme-machine	22
1.4.3.1 Reconnaissance automatique de la parole audiovisuelle	22
1.4.3.2 Codage spécifique de la parole : la norme MPEG4.....	23
1.4.3.3 Le rôle de la biométrie.....	23
1.5 Conclusion.....	24
La reconnaissance visuelle de la parole	27
2.1 Influence de l'angle de vue.....	28

2.2	Visage complet ou indices visuels ?	29
2.3	Localisation et suivi de visages	30
2.3.1	Localisation de visages	31
2.3.1.1	Approches couleur	32
2.3.1.2	Approches statistiques	36
2.3.2	Localisation de la bouche	39
2.3.2.1	Approches couleur	40
2.3.2.2	Approches statistiques	43
2.3.2.3	Approche par corrélation avec des patrons	45
2.3.2.4	Approches mouvement	46
2.3.2.5	Autres approches	47
2.4	Conditions « naturelles » (écologiques)	49
2.5	Comparaison image-modèle	51
2.6	Corpus existants	52
2.7	Conclusion	53
De la reconnaissance acoustique à la reconnaissance bimodale de parole.....		54
3.1	Définition de la parole	54
3.2	Le signal de la parole	55
3.2.1	Redondance du signal	55
3.2.2	Variabilité du signal	55
3.2.3	Les effets de coarticulation	56
3.3	Extraction des paramètres	56
3.3.1	Énergie du signal	57
3.3.2	Coefficients MFCC	58
3.3.3	Taux de passage par zéro	60
3.3.4	Autres paramétrisations du signal	60
3.3.5	Dérivées première et seconde	61
3.4	Réduction de l'espace de représentation	61
3.5	Les modes de fonctionnement d'un système de reconnaissance	62
3.6	La reconnaissance bimodale de la parole	63
3.6.1	Les modèles d'intégration audio-visuelle de la parole	64
3.6.1.1	Modèle ID	65
3.6.1.2	Modèle IS	66
3.6.1.3	Modèle RD	69
3.6.1.4	Modèle RM	70
3.6.2	Éléments du choix d'une architecture : théoriques et expérimentaux	71

3.6.3	Etudes comparatives.....	72
3.6.3.1	ID vs. IS.....	72
3.6.3.2	RD vs. RM.....	73
3.7	Conclusion.....	74
Deuxième partie : Approches proposées		58
Moteur de reconnaissance GA/HMM		77
4.1	Modèles de Markov Cachés	77
4.1.1	Définition.....	77
4.1.2	Utilisation et algorithmes	79
4.1.2.1	Evaluation et l'algorithme de Forward.....	79
4.1.2.2	Décodage et l'algorithme de Viterbi	81
4.1.3	Différents types de modèles HMM	84
4.1.4	Résumé	85
4.2	Les algorithmes génétiques	86
4.2.1	Principe des algorithmes génétiques	86
4.2.2	Description détaillée.....	88
4.2.2.1	Codage des données	88
4.2.2.2	Génération aléatoire de la population initiale.....	88
4.2.2.3	Évaluation.....	89
4.2.2.4	Gestion des contraintes.....	90
4.2.2.5	Principes de sélection	90
4.2.2.6	Opérateur de Croisement.....	91
4.2.2.7	Opérateur de mutation	93
4.2.2.8	Partage (Sharing).....	94
4.2.2.9	Critères d'arrêt de l'algorithme	95
4.2.3	Avantages et désavantages des algorithmes génétiques	95
4.3	Moteur de reconnaissance GA/HMM	95
4.4	Conclusion.....	97
Description du système proposé.....		98
5.1	Architecture de système de reconnaissance par fusion audiovisuelle	99
5.1.1	Traitement visuel.....	100
5.1.1.1	Détection de visage.....	100
5.1.1.2	Localisation de la bouche	104
5.1.1.3	Extraction des paramètres visuels	105
5.1.1.3.1	Découpage de l'image.....	106

5.1.1.3.2	Extraction de caractéristiques.....	106
5.1.2	Traitement acoustique	109
5.1.2.1	Analyse RASTA-PLP.....	109
5.1.2.2	La quantification vectorielle.....	110
5.1.3	Moteur de reconnaissance GA/HMM.....	112
5.1.4	La fusion audiovisuelle.....	112
5.1.4.1	Fusion des paramètres	113
5.1.4.2	Fusion des scores.....	113
5.2	Conclusion.....	114
Réalisation.....		115
6.1	Architecture général du système de reconnaissance.....	115
6.2	Base de données utilisée.....	118
6.2.1	Les bases de données audiovisuelle arabe.....	118
6.2.2	La base de données CUAVE	120
6.3	Validation du système	120
6.4	Traitement des données audiovisuelles	121
6.4.1	Séparation audiovisuelle.....	121
6.4.2	Données visuels.....	121
6.4.3	Données acoustiques	123
6.5	Modélisation par GA/HMM.....	125
6.5.1	Résultats obtenus et discussion	125
6.5.1.1.	Expérimentations avec des bruits sonore et visuel additifs	125
6.5.1.2.	Expérimentations avec un bruit réel	127
6.6	Conclusion.....	130
Conclusion et perspectives.....		131
7.1	Conclusion.....	131
7.2	Perspectives.....	131
Annexe A		133
A.1	Environnement de développement: MATLAB R2013a.....	133
A.2	Structure et fonctionnement du logiciel.....	135
Bibliographie.....		137
Notations		146
Publications réalisées au cours de la thèse		147

Table des illustrations

Liste des figures

Figure 1.1 – Vue schématique de l'appareil vocal, dans le plan sagittal médian.....	6
Figure 1.2 – Vue schématique antérieure du larynx (à gauche). Vue laryngoscopique des cordes vocales (à droite).....	7
Figure 1.3 – Structures de la langue, détails des muscles extrinsèques (plan sagittal médian, vue de droite).....	9
Figure 1.4 – Aspect schématique des lèvres (d'après Zemlin, 1968).	11
Figure 1.5 – Les muscles de la face (d'après Bouchet et Cuilleret 1972).....	12
Figure 1.6 – Le conduit vocal et les 8 lieux d'articulation principaux.....	15
Figure 1.7 – Les réalisations articulatoires et les mouvements labiaux correspondant (d'après Abry 1980).	17
Figure 1.8 – Comparaison de l'intelligibilité de la parole bimodale en condition bruitée en ajoutant successivement les lèvres, le mouvement de la mâchoire puis tout le visage du locuteur (Benoît et al., 1996).....	21
Figure 1.9 – Schéma de principe de la reconnaissance automatique de la parole.....	23
Figure 2.1 – Image couleur en entrée (a), pixels candidats pour appartenir au visage et localisation.	33
Figure 2.2 – Détecteur de visage de Hunke et Duchnowski basé sur la couleur (FCC) : (a) Image couleur à analyser et région utilisée pour entraîner le modèle (IFCC) de couleur du visage, (b) Sortie du FCC : en blanc, les zones de « non-visage », d'après (Duchnowski et al. 1995; Hunke and Waibel 1994).	34
Figure. 2.3 – Une scène complexe (a) et sa classification en tons « peau » (b), d'après (Senior 1999).....	34
Figure. 2.4 – Localisation du visage sur le corpus M2VTS, d'après (Wark and Sridharan 1998).....	35
Figure. 2.5 – Localisation de différentes régions de visage (a) automatiquement (b) en utilisant l'approche « template matching », d'après (Brunelli and Poggio 1993).	38
Figure. 2.6 – Localisation des lèvres en utilisant la teinte H, d'après (Coianiz et al. 1996). ...	41
Figure. 2.7 – Localisation des lèvres en utilisant le quotient Q, d'après (Wark and Sridharan 1998).....	42
Figure. 2.8 – Détection des lèvres d'après (Liew et al. 1999).	43
Figure. 2.9 – Détection des lèvres d'après (Rao and Mersereau 1995).	44
Figure. 2.10 – Détection des lèvres d'après (Wojdel and Rothkrantz 2001a; Wojdel and Rothkrantz 2001b).....	45
Figure 3.1 – Schéma de calcul des MFCC.	59
Figure 3.2 – Schémas de calcul les paramètres PLP et LPC.....	61
Figure 3.3 – Le noyau d'un processus d'intégration audio-visuelle dans la perception de la parole (d'après Schwartz et al. (1998)).	65
Figure 3.4 – Modèle à identification directe.	65

Figure 3.5 – Modèle à identification séparée.	67
Figure 3.6 – Modèle d'intégration basé sur la maximisation des produits des probabilités conjointes (D'après Adjoudani (1998)).	67
Figure 3.7 – Méthode de sélection du meilleur candidat acoustique ou visuel (D'après Adjoudani (1998)).	68
Figure 3.8 – Architecture d'intégration audiovisuelle par pondération (D'après Adjoudani (1998)).	68
Figure 3.9 – Modèle à recodage dans la modalité dominante.	69
Figure 3.10 – Modèle à recodage dans la modalité motrice.	70
Figure 3.11 – Taxinomie des modèles d'intégration (d'après Robert-Ribès (1995)).	71
Figure 4.1 – HMM à 5 états dont 3 émetteurs.	78
Figure 4.2 – Trois types distincts de modèles HMM. Illustration avec un exemple de HMM à 4 état (d'après Rabiner et Juang 1993).	85
Figure 4.3 – Principe général des algorithmes génétiques.	87
Figure 4.4 – Slicing crossover.	92
Figure 4.5 – Slicing crossover à 2 points.	93
Figure 4.6 – Croisement barycentrique.	93
Figure 4.7 – Principe de l'opérateur de mutation.	94
Figure 4.8 – Méthode de représentation des chromosomes dans l'apprentissage des GA/HMMs.	96
Figure 5.1 – Phases de spécification d'un système d'intelligence artificielle utilisant des HMM.	98
Figure 5.2 – Système d'un AVASR mis en œuvre.	100
Figure 5.3 – Exemple de 4 caractéristiques de Haar. La somme des valeurs des pixels appartenant aux zones encadrées claires est soustraite à la somme des valeurs des pixels appartenant aux zones encadrées sombres pour obtenir la caractéristique de Haar. Chacune des quatre caractéristiques de Haar est représentée avec son cadre de détection respectif. ...	102
Figure 5.4 – Cascade de classifieurs forts. A chaque étage, uniquement les candidats classifiés positifs sont transmis à l'étage suivant.	104
Figure 5.4 – Découpage de l'image de l'histogramme.	106
Figure 5.5 – Exemple de fonctions de base de DCT qui forme le domaine fréquentiel.	108
Figure 5.6 – Parcours en zigzag d'une matrice de dimension 8×8	108
Figure 5.7 – Analyse RASTA PLP.	110
Figure 5.8 – Distribution de probabilités, un échantillon de points associés, et un découpage en nuages (clusters).	111
Figure 6.1 – Architecture générale du système proposé.	117
Figure 6.2 – quelques exemples de trames de notre base audiovisuelle AVARB.	119
Figure 6.3 – Exemples de trames de la base CUAVE.	120
Figure 6.4 – Un exemple de détection de visage : (a) image originale (b) détection de peau avec suppression de bruit (c) résultat de détection de visage.	121
Figure 6.5 – Exemples de la région de la bouche détectée à partir de : (a) la base AVARB (b) la base CUAVE.	122
Figure 6.6 – Le processus de sélection des coefficients DCT avec un échantillon à partir: (a) la base AVARB (b) la base CUAVE.	122

Figure 6.7 – Exemple d'un signal de parole du mot arabe "/ marhaban /" (a) son spectrogramme (b) et l'ensemble des caractéristiques spectrales RASTA-PLP (c).	124
Figure 6.8 – ROI avec bruit gaussien, l'écart type =(A) 0 (B) 15 (C) 30 (D) 50 et (E) 100. .	126
Figure 6.9 – La performance du système AVASR : (a) sous une fréquence des trames vidéo réduite (b) pour un bruit aléatoire gaussien.....	126
Figure 6.10 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD AVARB.....	128
Figure 6.11 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD CUAVE.....	129
Figure A.2 – Interface principale du logiciel.	135
Figure A.3 – Interface d'extraction des paramètres visuels.	136
Figure A.4 – Interface d'extraction des paramètres acoustiques.	136

Liste des tableaux

Table 2.1 – Scores d'identification obtenus par Summerfield (1979) dans cinq conditions de présentation des stimuli.	29
Table 6.1 – Notre deux corpus proposés de chiffres et commandes arabes.	119
Table 6.2 – paramètres GA pour l'entraînement du HMM pour l'audio seul: (a) base AVARB (b) base CUAVE.	127
Table 6.3 – paramètres GA pour l'entraînement du HMM pour le vidéo seul: (a) base AVARB (b) base CUAVE.	127

Introduction

L'utilisation de connaissances supplémentaires conjointement au signal de parole est une méthode classique pour améliorer les performances et la robustesse des systèmes de reconnaissance automatique de la parole. De nombreux travaux sur la perception de la parole ayant montré l'importance des informations visuelles dans le processus de reconnaissance chez l'homme, l'utilisation de données sur la forme et le mouvement des lèvres du locuteur semble être une voie prometteuse pour la reconnaissance automatique surtout en milieux sonores bruités.

Les êtres humains emploient l'information visuelle de façon subconsciente afin de comprendre les paroles, particulièrement dans des environnements bruyants, mais également quand les conditions acoustiques sont bonnes. Le mouvement des lèvres du locuteur apporte une série d'information importante, par exemple au sujet des articulations, ce qui est automatiquement intégré par le cerveau. L'effet McGurk (1976) en apporte la preuve en montrant que le cerveau, soumis à des stimuli auditifs et visuels inconsistants, perçoit un son différent de celui qui a été dit.

1. Contexte et cadre de recherche

L'objet de nos travaux de recherche concerne l'intégration des informations visuelles aux informations acoustiques en vue de leur exploitation pour la reconnaissance automatique de la parole. Si cette exploitation est fort séduisante, la problématique qu'elle soulève est cependant loin d'être simple. Tout d'abord, se pose la question du niveau d'intégration : est-ce le niveau des données ou bien celui des résultats. Puis il y a les phénomènes de décalage temporel entre la réalisation auditive et la réalisation visuelle d'un même phonème. Ensuite intervient le problème d'adaptation des contributions des modalités acoustique et visuelle selon leur fiabilité relative. Enfin se pose la question de la pertinence de l'utilisation, pour le traitement du signal visuel de parole, d'unités de décision spécifiques, nommées visèmes.

Reconnaissance automatique de la parole audio-visuelle (AVASR) a été lancée par Petajan (1984) et elle est encore une zone active de recherche. Cette thèse se positionne clairement dans le champ des systèmes AVASR, le système de reconnaissance proposé utilise les modèles de Markov cachés (Hidden Markov Model, HMM) comme moteur de reconnaissance combiné avec un algorithme génétique (Genetic Algorithm, GA) pour résoudre le problème de la convergence vers l'optimum local. Le point principal de ce travail

est basé sur la qualité de la modélisation des données (appelé observations) faites par HMM. Notre objectif est de proposer des algorithmes qui permettent d'améliorer cette qualité. Le critère utilisé pour quantifier la qualité de HMM est la probabilité qu'un modèle donné génère une observation donnée. Pour résoudre ce problème, nous utilisons comme nous l'avons déjà mentionné une hybridation génétique des HMM et nous proposons des méthodes de représentation d'un gène et la méthode pour l'évaluation des mesures de remise en forme des populations de chaque génération crée par algorithme génétique. L'expérience est menée afin d'évaluer chaque population et la précision de résultat d'inférence sur un ensemble de données audiovisuelles réelles.

Le traitement de la parole arabe est encore à ses débuts, la raison pour laquelle, nous avons pensé à l'application de la méthode hybride GA/HMM, ayant comme objectif la reconnaissance de la parole en mode multi-locuteur.

2. Plan de la thèse

Ce document est structuré en deux parties. La première partie établit plusieurs états de l'art sur les domaines abordés (chapitre 1, 2 et 3) tandis que la deuxième partie présente nos approches proposées.

Le premier chapitre donne une brève présentation de quelques éléments physiologiques sur la production de la parole et la paramétrisation des lèvres. Cette étude présente une description des muscles faciaux intervenant dans le processus de la parole. Nous nous décrivons aussi quelques propriétés intrinsèques de la perception de la parole bimodale afin de mieux comprendre ce processus différents modèles d'intégration audiovisuelle chez l'homme et dans la machine sont présentées.

Nous passons dans le chapitre 2 à une description détaillée des techniques d'extraction des informations visuelles des mouvements des lèvres, notamment celles basées sur le traitement vidéo, ainsi que notre méthode de calcul des paramètres labiaux basée sur un maquillage préalable des lèvres.

Par la suite dans le chapitre 3, Nous nous consacrons a une revue de l'état de l'art dans le domaine du développement des systèmes de reconnaissance visuelle et audiovisuelle.

Chapitre 4 définit le principe et le fonctionnement de notre système AVASR proposé en utilisant la méthode hybride GA/HMM.

Puis dans le chapitre 5, nous décrivons le principe et le fonctionnement de notre système de reconnaissance de la parole audiovisuelle proposé.

Le dernier chapitre (chapitre 6) présente les résultats de nos tests sur les deux modèles d'intégration couramment utilisés dans la littérature (précoce et tardive) en insistant sur notre architecture d'intégration originale, basée sur une pondération des canaux en fonction de leur fiabilité, estimée par la dispersion des meilleurs candidats. La dernière partie de ce manuscrit est dédiée à la description technique de notre système électronique d'extraction des paramètres labiaux en temps réel et à l'évaluation de ses performances dans une application de lecture labiale automatique.

Enfin, nous concluons par un bilan de nos travaux de recherche et nous proposons quelques perspectives d'amélioration associées aux différentes réalisations.

Première partie: Etat de l'art

Les lèvres et la production de la parole

1

Le téléphone et la radio prouvent la capacité d'une parole purement auditive à transmettre avec efficacité une communication langagière. Néanmoins, la perception humaine tire aussi profit de l'information visuelle apportée par le visage du locuteur notamment lorsque les conditions acoustiques sont dégradées. C'est cette bimodalité intrinsèque, et le gain d'intelligibilité qu'elle apporte, qu'explore l'étude de la parole audiovisuelle. Mise en évidence pour la communication humaine, elle ouvre de nouvelles perspectives pour la communication avec et par la machine.

Bien que la communication orale engage l'ensemble du visage du locuteur, les lèvres occupent une place privilégiée : elles fournissent une source visuelle d'information pour la perception de la parole et, étant toujours identifiables, se prêtent à une analyse automatique. La capture automatique des mouvements labiaux (ou la biométrie) tend à doter l'ordinateur de paramètres intelligibles et indépendants pour contrôler des visages synthétiques parlants ou bien identifier le message énoncé par une reconnaissance audiovisuelle automatique. Les difficultés technologiques résident dans la complexité de ces mouvements et la variabilité intra- et inter- locuteurs.

1.1 Architecture et fonctionnement de l'appareil vocal

Cette section, qui rappelle l'architecture et les principes généraux de fonctionnement de notre appareil vocal s'appuie sur les ouvrages suivants : (Le Huche 2001) et (Boite *et al.* 2000). Une vue schématique de notre appareil vocal est proposée à la figure 1.1.

1.1.1 L'appareil vibrateur

L'air est la matière première de la voix. Si le fonctionnement de notre appareil vocal est souvent comparé à celui d'un instrument de musique, il doit être décrit comme celui d'un instrument à vent. En effet, en expulsant l'air pulmonaire à travers la trachée, le système respiratoire joue le rôle d'une soufflerie. Il s'agit du « souffle phonatoire » produit, soit par l'abaissement de la cage thoracique, soit dans le cadre de la projection vocale par l'action des muscles abdominaux.

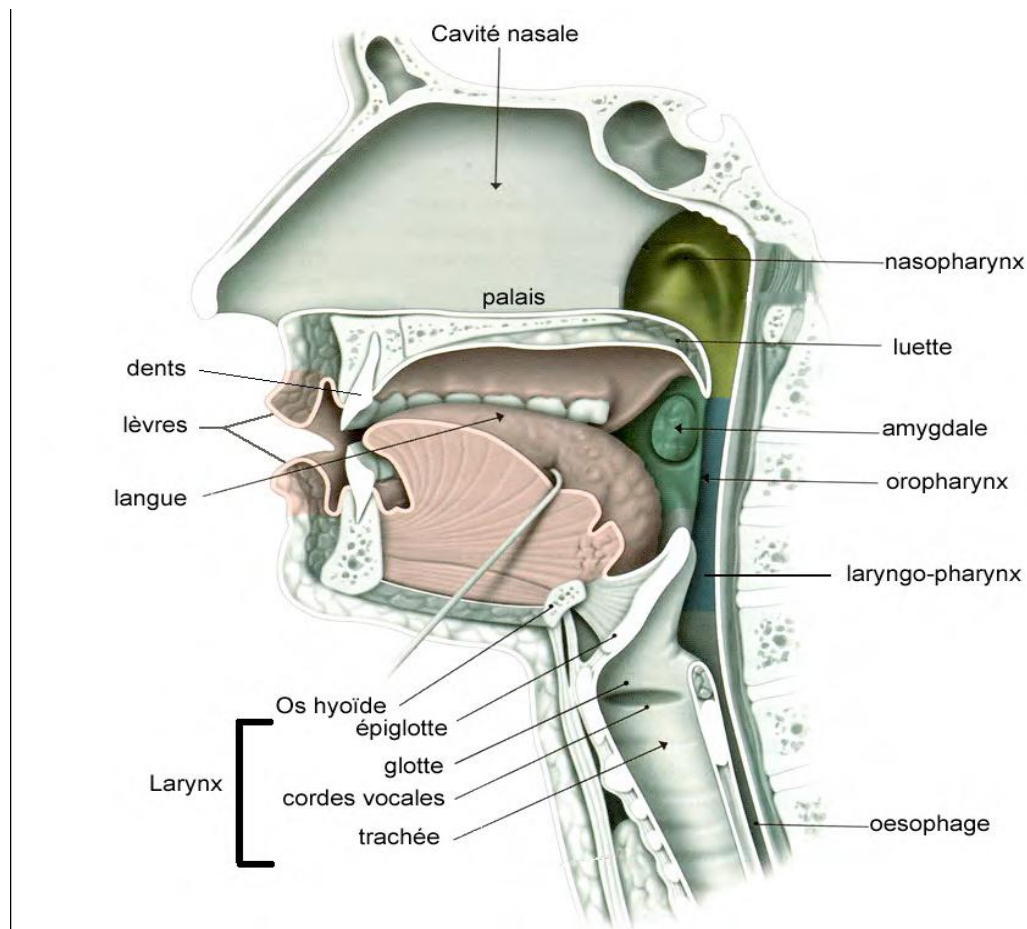


Figure 1.1 – Vue schématique de l'appareil vocal, dans le plan sagittal médian.

L'extrémité supérieure de la trachée est entourée par un ensemble de muscles et de cartilages mobiles qui constituent le larynx. Le plus important est le cartilage thyroïde qui forme le relief de la pomme d'Adam. Le larynx se trouve au carrefour des voies aériennes et digestives, entre le pharynx et la trachée, et en avant de l'œsophage. Les plis vocaux, communément nommés « cordes vocales » sont deux lèvres symétriques (structures fibreuses) placées en travers du larynx. Ces lèvres se rejoignent en avant et sont plus au moins écartées l'une de l'autre sur leur partie arrière (structure en forme de V); l'ouverture triangulaire résultante est nommée glotte. Les structures du larynx et des plis vocaux sont illustrés à la figure 1.2. Le larynx et les plis vocaux forment notre « appareil vibreur ».

Lors de la production d'un son qualifié de « non-voisé » (ou sourd), comme c'est le cas, par exemple, pour les phonèmes [s] ou [f], les plis vocaux sont écartés et l'air pulmonaire circule librement en direction des structures en aval.

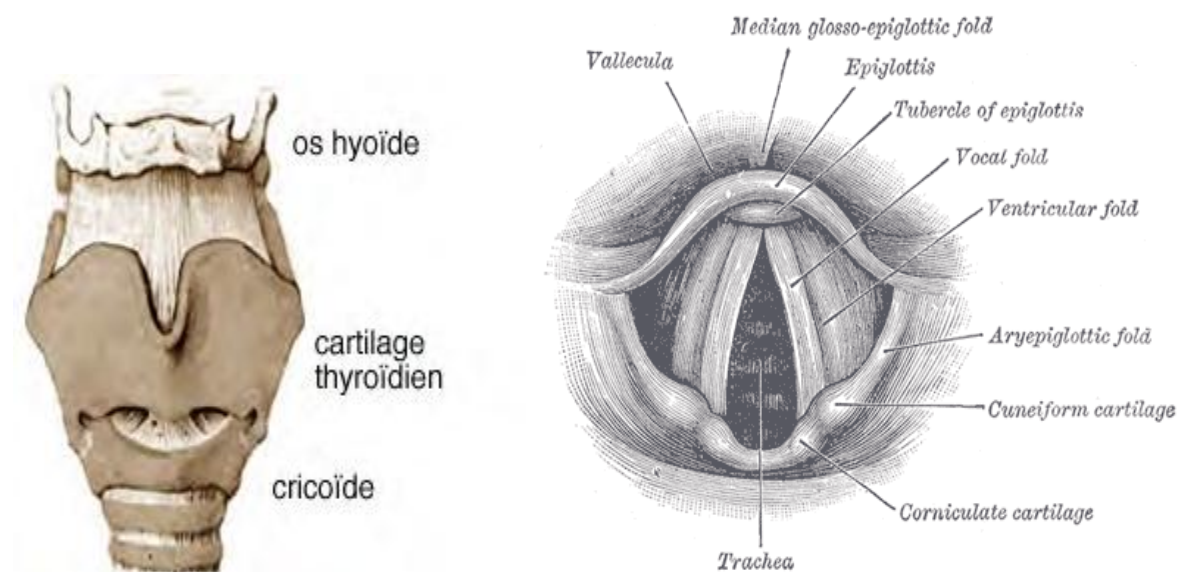


Figure 1.2 – Vue schématique antérieure du larynx (à gauche). Vue laryngoscopique des cordes vocales (à droite).

En revanche, lors de la production d'un son voisé (ou sonore), comme c'est le cas, par exemple, pour les phonèmes [z], [v] et pour les voyelles, les plis vocaux s'ouvrent et se ferment périodiquement, obstruant puis libérant par intermittence le passage de l'air dans le larynx. Le flux continu d'air pulmonaire prend ainsi la forme d'un train d'impulsions de pression ; nos « cordes vocales vibrent ». Le dernier élément principal de notre appareil vibrateur est l'épiglotte. Lors de la déglutition, cette dernière agit comme un clapet qui se rabat sur le larynx, conduisant les aliments vers l'œsophage en empêchant leur passage dans la trachée et les poumons (« fausse route »).

1.1.2 Le résonateur

L'air pulmonaire, ainsi modulé par l'appareil vibrateur, est ensuite appliqué à l'entrée du conduit vocal. Ce dernier est principalement constitué des cavités pharyngiennes (laryngopharynx et oropharynx situés en arrière-gorge) et de la cavité buccale (espace qui s'étend du larynx jusqu'aux lèvres). Pour la réalisation de certains phonèmes, le voile du palais (le velum) et la luette qui s'y rattache, s'abaissent, permettant ainsi le passage de l'air dans les cavités nasales (fosses nasales et rhinopharynx ou nasopharynx). Ces différentes cavités forment un ensemble que nous qualifierons ici de « résonateur ». Si l'appareil vibrateur peut être décrit comme le lieu de production de « la voix », le résonateur apparaît alors comme le lieu de naissance de « la parole ». Il abrite en effet des organes mobiles,

nommés articulateurs, qui en modifiant sa géométrie et donc ses propriétés acoustiques, mettent en forme le son laryngé (ou son glottique) en une séquence de sons élémentaires. Ces derniers peuvent être interprétés comme la réalisation acoustique d'une série de phonèmes, unités linguistiques élémentaires propres à une langue. Les articulateurs principaux sont la langue, les lèvres, le voile du palais et la mâchoire (maxillaire inférieur).

L'articulateur principal de la cavité buccale est la langue. Intervenant dans la mastication et la déglutition, la langue est également l'organe du goût. S'étendant sur une longueur d'une dizaine de centimètres environ, cet organe complexe et hautement vascularisé est composé d'un squelette, de muscles et d'une muqueuse. Son squelette est qualifié d'ostéofibreux ; il est constitué de l'os hyoïde, situé au dessus du larynx, sur lequel se fixe la membrane hyoglossienne, d'une hauteur d'un centimètre environ, et le septum lingual, lame fibreuse à l'origine de la dépression visible sur toute la longueur de la langue. Son mouvement est contrôlé par dix sept muscles, dont huit paires de muscles agonistes/antagonistes. Quatre paires de muscles extrinsèques (muscles qui prennent naissance à l'extérieur de la langue) servent notamment à sa protrusion, sa rétraction, sa dépression ou son élévation.

La langue est usuellement décrite comme un ensemble de deux structures au comportement distinct, la racine (ou base), fixée à l'os hyoïde, et le corps, plus mobile. Ce dernier se décompose également en deux parties, le dos et la pointe de la langue, nommée apex. L'organisation du système musculaire de la langue ainsi que ses principales structures sont illustrées à la figure 1.3. Le rôle de la langue dans la phonation est déterminant, notamment pour la production des voyelles, caractérisée par le libre passage de l'air dans le résonateur. La phonétique articulatoire décrit le système vocalique d'une langue (classification des voyelles) précisément à l'aide de deux critères qui décrivent la configuration de la langue dans la cavité buccale. Le premier est le « lieu d'articulation » ; « avant » ou « arrière », il localise la masse de la langue et qualifie ainsi les voyelles produites d'« antérieures », de « centrales » ou de « postérieures ». Le second critère est « l'aperture » ; il décrit l'espace de résonance ménagé entre la langue et le palais (fermé ou ouvert), qualifiant ainsi les voyelles produites de « hautes » ou « basses ». La langue joue également un rôle important pour l'articulation des consonnes, dont le mode de production est, à l'inverse des voyelles, caractérisé par l'obstruction du passage de l'air dans le résonateur. Dans ce cas, le « lieu d'articulation » localise cette obstruction. Pour produire une consonne dite « dentale » ([t], [d], [n]), la pointe de la langue crée cette obstruction en se rapprochant des dents.

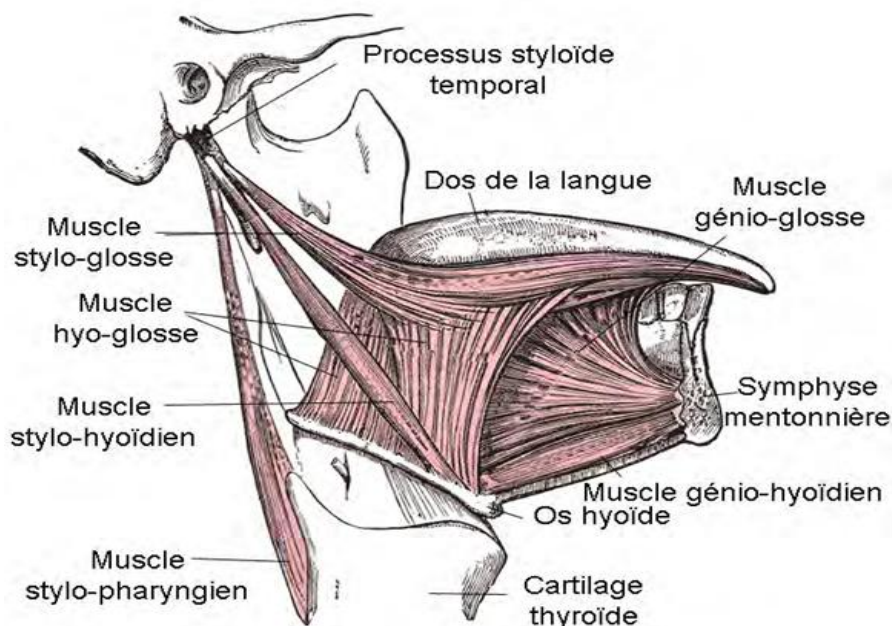


Figure 01.3 – Structures de la langue, détails des muscles extrinsèques (plan sagittal médian, vue de droite).

Dans le cas des consonnes « inter-dentales » ([th] comme *thin*, [dh] comme *then*), la langue dépasse les dents et vient s'appuyer directement sur les incisives. Pour les consonnes dites « alvéolaires » ([s], [z] ou la consonne liquide [l] mais également [t], [d], [n]), et « postalvéolaires » ([ch] comme *church*, [jh] comme *judge*, [sh] comme *she*, [zh] comme *azure*), elle se déplace respectivement vers les alvéoles (creux de l'os alvéolaire dans lequel est enchâssée une dent) et vers la partie antérieure du palais (à la juxtaposition avec le palais dur). Pour une consonne dite « palatale » ([j] comme *ye*, catégorisée également comme une semi-voyelle), l'organe articulateur est le dos de la langue, l'obstruction ayant lieu au niveau du palais dur. Pour une consonne vélaire ([k], [g], [ng] comme *parking*), la partie postérieure du dos de la langue se bombe et se rapproche du palais mou. Enfin, pour une consonne uvulaire ([r] comme *Paris* en français), le lieu d'articulation se situe au niveau de la luette.

Les lèvres constituent l'autre articulateur majeur de la cavité buccale. Elles permettent la production des consonnes « bilabiales » (rapprochement des lèvres inférieures et supérieures, [p], [b], [m]) et des consonnes « labio-dentales » ([f], [v], rapprochement de la lèvre inférieure avec les dents). Elles interviennent également dans le cadre de la production vocalique en apportant la notion d'arrondissement des voyelles. Enfin, la réalisation acoustique de certains phonèmes nécessite parfois deux lieux d'articulation, impliquant à la fois la langue et les lèvres ; c'est le cas notamment de la consonne « labio-velaire » [w] (comme *who*).

Le dernier articulatoire du résonateur est le voile du palais qui permet, lorsqu'il s'abaisse, de mettre en parallèle les cavités buccale et nasale. Il intervient notamment dans la production des consonnes nasales [m], [n] et [ŋ] en les différenciant respectivement des groupes de consonnes ([p], [b]), ([t], [d]), et ([k], [g]), qui présentent la même configuration linguale et labiale. Enfin, l'abaissement du voile du palais permet, en langue française notamment, la formation des voyelles nasales [ɔ] (*on*), [ɛ] (*hein*), [œ] (*un*), [ɑ] (*an*).

Au regard de ces principaux résultats issus de la phonétique articulatoire, la réalisation acoustique d'un phonème dépend principalement des configurations de la langue, des lèvres et du voile du palais mais également de l'activité des cordes vocales. Lorsque ces dernières doivent être retirées, dans le cadre notamment du traitement chirurgical du cancer du larynx, les mécanismes de la phonation sont profondément modifiés.

1.2 L'anatomie des lèvres

1.2.1 Les tissus

D'après les données anatomiques présentées dans (Abry 1980), les lèvres forment deux replis musculaires, recouverts d'une membrane, qui circonscrivent l'orifice de la cavité buccale. Ces replis supérieur et inférieur sont indépendants et se réunissent à leurs extrémités pour former les commissures labiales. La face externe des lèvres est recouverte par de la peau et la face interne par de la muqueuse composée de cellules disposées comme des pavés (l'épithélium). Les muscles se trouvent directement sous la peau.

La ligne entre la peau et la muqueuse dessine dans sa partie supérieure et, au centre, une courbe concave dénommée « arc de Cupidon ». Elle délimite une zone de transition, dite vermillon. Celle-ci se caractérise par sa haute teneur en un liquide semi-fluide qui augmente la transparence du tissu, à tel point qu'on aperçoit la teinte rouge de la couche vasculaire sous-jacente. C'est cette caractéristique qui fait ressortir la couleur des lèvres par rapport au reste de la peau. La zone de vermillon de la lèvre supérieure montre, en son milieu, une protubérance : le tubercule.

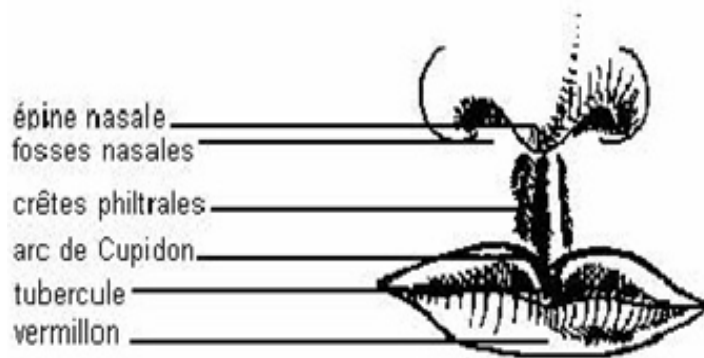


Figure 01.4 – Aspect schématique des lèvres (d'après Zemlin, 1968).

A l'intérieur de la bouche, la muqueuse de teinte rosée rejoint les arcades alvéolo-dentaires. L'espace incurvé, ainsi délimité, forme les gouttières vestibulaires. Dans leurs parties médianes, les gouttières vestibulaires supérieure et inférieure présentent un repli muqueux : le frein de la lèvre. Celui-ci est nettement plus proéminent pour la lèvre supérieure.

1.2.2 Les muscles des lèvres

Les muscles des lèvres font partie des muscles faciaux. Ils ont tous la particularité de présenter une insertion mobile cutanée. C'est cette caractéristique qui rend possible les différentes combinaisons d'expression du visage et la souplesse des mouvements en production de la parole. Le muscle essentiel des lèvres est l'orbiculaire des lèvres qui opère comme un sphincter annulaire. Autour de celui-ci, rayonnent les autres muscles de la face dont les fibres s'imbriquent directement avec celles de l'orbiculaire.

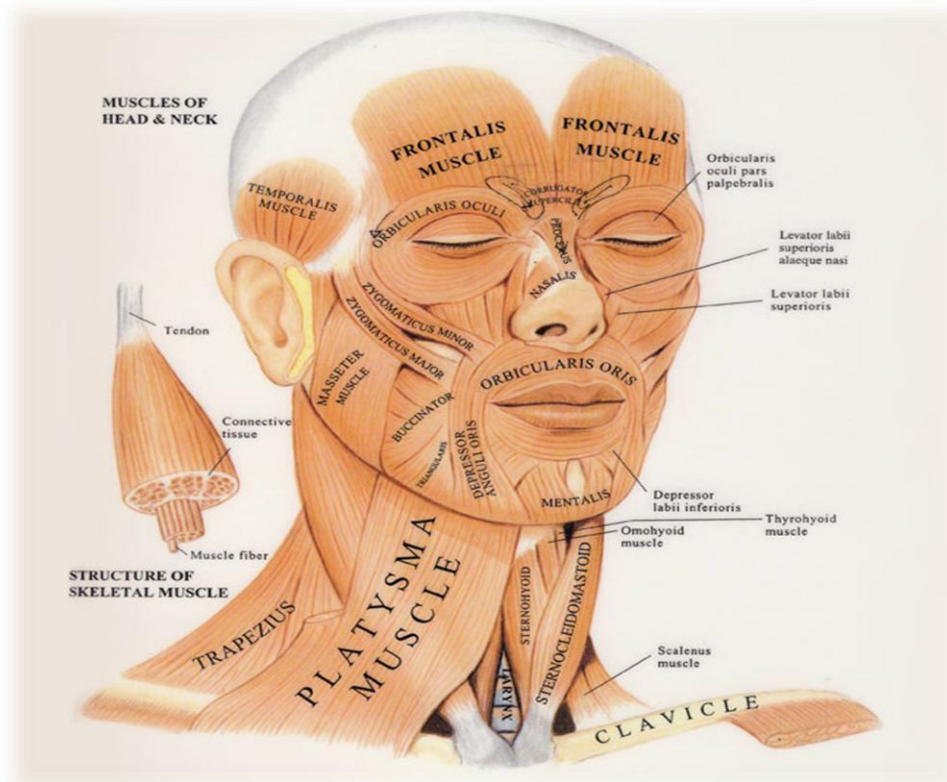


Figure 01.5 – Les muscles de la face (d'après Bouchet et Cuilleret 1972).

Les classifications courantes dénombrent douze muscles pour les lèvres (Zemlin 1968 ; Hardcastle 1976) :

- l'orbiculaire des lèvres (orbicularis oris),
- le canin (levator anguli oris),
- le buccinateur (buccinator),
- les muscles de la houppe du menton (mentalis),
- la carré du menton (quadratus labii inferioris, ou depressor labii inferioris),
- le releveur superficiel de l'aile du nez et de la lèvre (levator labii superioris alaeque nasi),
- le releveur profond (levator labii superioris),
- le petit zygomatique (zygomaticus minor),
- le petit zygomatique (zygomaticus minor),
- le grand zygomatique (zygomaticus major),
- le risorius,
- le triangulaire des lèvres (depressor anguli oris).
- le peaucier du cou (muscle platysma).

1.2.3 Classification fonctionnelle des muscles labiaux

En complément d'études anatomiques, des mesures par électromyographie ont permis de dresser une classification des muscles labiaux suivant les mouvements qu'ils génèrent. Cette classification suit celle de (Hardcastle 1976), reprise dans (Abry 1980). Elle présente les tendances générales observées chez plusieurs sujets.

Muscles assurant l'occlusion labiale

Par contraction l'orbiculaire accole les lèvres supérieures et inférieures en abaissant la lèvre supérieure et en tirant la lèvre inférieure vers le haut. Le mouvement de la lèvre inférieure est fortement dépendant de la mâchoire. Le canin et le triangulaire peuvent aussi intervenir pour fermer les lèvres.

Muscles assurant la protrusion des lèvres

La protrusion correspond à un mouvement poussant les lèvres vers l'avant, s'accompagnant d'un rapprochement des lèvres et des commissures. C'est aussi une des fonctions principales de l'orbiculaire. La houppette du menton contribue à faire basculer la lèvre inférieure.

Muscles assurant l'arrondissement des lèvres

L'arrondissement correspond à une forme de lèvres obtenue en rapprochant les commissures. Ce geste s'oppose à l'étirement. Bien que l'arrondissement s'obtienne par une contraction de l'orbiculaire, ce geste ne s'accompagne pas forcément d'une protrusion. Des muscles comme le buccinateur ou le risorius peuvent limiter l'action de l'orbiculaire.

Muscles élévateurs de la lèvre supérieure

Comme leur nom l'indique, les releveurs supérieurs et profonds de la lèvre sont attachés à cette fonction. Du fait de leur insertion, c'est essentiellement la partie centrale de la lèvre supérieure qui est relevée.

Muscles abaisseurs de la lèvre inférieure

La lèvre inférieure est tirée vers le bas par le carré du menton. Ce muscle peut être aidé par la mâchoire. De même, le triangulaire peut aussi intervenir pour abaisser la lèvre inférieure.

Muscles étirant les commissures

Le buccinateur entre en action pour étirer les commissures. Cette activité est antagoniste à celle de protrusion de l'orbitaire ou de la houppe du menton.

Muscles abaisseurs des commissures

La fonction principale du triangulaire est d'abaisser les commissures. Cette fonction s'accompagne d'un abaissement de la lèvre inférieure.

Muscles élevateurs des commissures

L'insertion du canin est située sur les commissures dont il assure l'élévation. Le relèvement de la lèvre inférieure qui s'accompagne est limité par l'action antagoniste du carré du menton. Le grand zygomatique intervient aussi pour le relèvement.

En conclusion, les lèvres sont commandées par des couples agonistes / antagonistes de muscles permettant ainsi un contrôle fin par équilibre des forces. Cette habileté est mise en œuvre dans la production de la parole pour un contrôle géométrique précis de la cavité buccale, rentrant directement en compte dans la génération des sons.

1.3 Repères phonétiques

1.3.1 Acoustique et articulation

Les différents sons de la parole sont produits par la manière dont l'air, expulsé par les poumons, s'écoule à travers le conduit vocal. La forme du conduit et les caractéristiques de cet écoulement déterminent directement l'onde sonore en sortie. Le passage de l'air s'effectue selon deux passages partant du larynx, l'un débouchant dans la cavité nasale, et l'autre vers la bouche puis les lèvres. Dans le larynx, les cordes vocales peuvent être mises en vibration par la conjugaison d'une pression transglottique et de la contraction des effecteurs laryngés. On parle alors de son voisé. A l'inverse, on parle de son non voisé dans le cas où les cordes vocales ne vibrent pas. Le passage de l'air à travers la cavité nasale est commandé par l'ouverture du voile du palais pour la production des sons dits nasals. Le voile du palais est fermé pour les sons dits oraux pour lesquels l'air est intégralement expulsé par la cavité buccale.

L'air s'écoule dans la cavité buccale de trois manières : libre, rétrécie ou arrêtée. Le cas libre correspond à la production des voyelles. Sauf contrôle explicite (chuchotement par

exemple), il s'accompagne généralement d'une vibration des cordes vocales pour accroître l'énergie de l'onde. La position de la langue et la forme des lèvres modifient alors la géométrie (et donc les résonances) du conduit vocal, donnant le timbre de l'onde sonore. Les cas d'écoulement rétréci ou arrêté correspondent à la production des consonnes. Le son est alors généré par le bruit des turbulences créées par le rétrécissement (constriction) ou la brusque explosion qui suit une fermeture complète du passage de l'air (occlusion). La phonétique caractérise la production d'une consonne selon son mode et lieu d'articulation. Le mode d'articulation spécifie la manière dont s'écoule l'air et s'il s'accompagne d'un voisement. Le lieu d'articulation indique l'endroit de rapprochement maximal des parois le long du conduit vocal. La figure 1.6 indique les 8 lieux d'articulation principaux identifiés en phonétique.

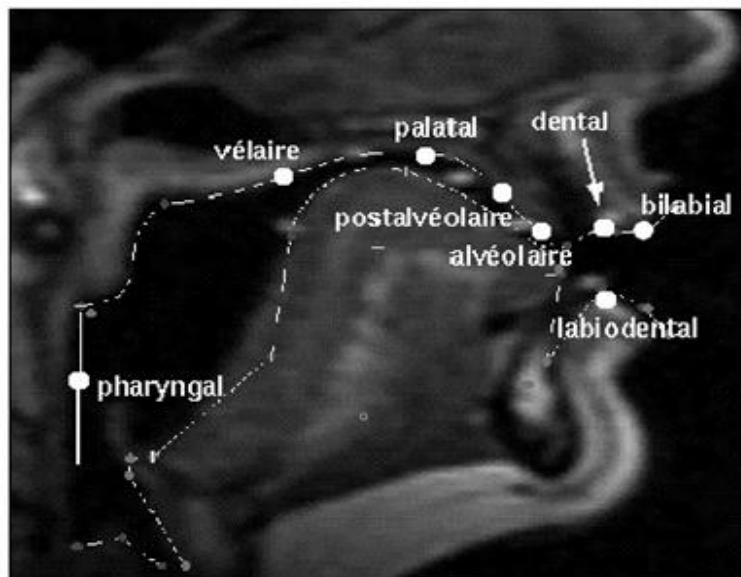


Figure 1.6 – Le conduit vocal et les 8 lieux d'articulation principaux.

1.3.2 Des sons et des lèvres

En maintenant stables et non ambiguës les différences entre les sons articulés, une représentation sensible (acoustique et visuelle) du code phonologique peut être mise en commun entre celui qui parle et celui qui écoute, d'où la mise en place d'une communication.

L'ensemble fini des sons d'une langue suggère un ensemble fini d'articulations pour les produire, donnant pour les lèvres un jeu de formes « cibles » ou prototypiques de l'articulation. Les lèvres n'assurent pas à elles seules la production distinctive de tous les sons : la production de /p/, /b/ et /m/, par exemple, implique dans les trois cas une même occlusion

bilabiale, les sons se distinguant par leur mode d'articulation (respectivement non voisé, voisé et nasal).

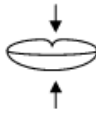
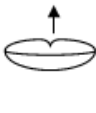

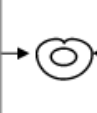
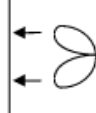
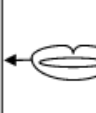

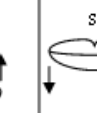
Se basant à la fois sur les observations phonétiques et l'activité des muscles labiaux, Gentil et Boë ont regroupé les formes labiales des sons du Français en six classes articulatoires (Abry 1980) :

- voyelles arrondies (/y/, /u/, /o/, /O/, ...), caractérisées par un arrondissement de la forme des lèvres, le but étant de réduire l'aire interne (l'arrondi est plus ou moins marqué selon la voyelle faisant une distinction entre des arrondies fermées telle /u/ et ouvertes comme /o/),
- voyelles non arrondies (/i/, /e/, /E/, /a/, ...), par opposition aux précédentes, où les commissures sont écartées et la forme des lèvres plus étirée,
- occlusives bilabiales, caractérisées par une fermeture complète des deux lèvres (/p/, /b/, /m/),
- constrictives labiodentales, caractérisées par un rapprochement de la lèvre inférieure et des dents de la mâchoire supérieure (/f/, /v/),
- constrictives post-alvéolaires à projection labiale, caractérisées par un arrondissement des lèvres s'accompagnant d'une protrusion et un relèvement de la lèvre supérieure (/ʃ/, /ʒ /),
- constrictives alvéolaires, caractérisée par un étirement des commissures (/s/, /z/).

Globalement, les formes de lèvres se distinguent donc par les traits d'arrondissement (opposé à étirement), d'ouverture (opposé à fermeture) et de protrusion. De même, la plupart des manuels de phonétique distinguent 3 degrés de liberté pour mesurer l'articulation labiale : étirement, aperture et protrusion (Ladefoged 1979). L'étirement correspond à la largeur de l'aire interne : elle discrimine les formes arrondies des étirées lorsque les lèvres ne sont pas complètement fermées. L'aperture correspond à la hauteur entre les lèvres supérieure et inférieure : cette mesure caractérise les occlusions. La protrusion désigne l'avancement du pavillon : on retient généralement cette mesure pour séparer les voyelles arrondies des étirées.

Gentil et Boë ont dressé un récapitulatif des différents mouvements labiaux, et des muscles les générant, requis dans la production des classes articulatoires citées.

Chapitre 1. Les lèvres et la production de la parole

Réalisation	Fermeture des lèvres	Élévation de la lèvre sup.	Abaissement de la lèvre inf.	Arrondissement des lèvres	Protrusion des lèvres	Rétraction des commissures	Élévation des commissures de la lèvre inf.	Élévation des commissures de la lèvre sup.
COVS. p, b, m 1.phase fêr. 2.phase ouv.	 O.O.(p)	 L.L.S.(p) L.A.O.(s)	 D.L.I.(a) D.L.I.(p) D.A.O.(s)	 M.(s)			 L.A.O.(s)	 D.A.O.(s)
f, v	O.O.I.(p)	Zyg Min(p) L.L.S.(s)				Buc.(s)	L.A.O.(s) Zyg Maj(s)	
ʃ, ʒ		Zyg Min(s) L.L.S.(s)	D.L.I.(s)		O.O.I.(p) M.(s) Plat.(s)			
s, z						Buc.(p) Ris.(s)		
VOY. arrondies fermées y, u				O.O.(p)	M.(s) Plat.(s)	Buc.(a)		D.A.O.(s)
arrondies fermées ø, u, œ, ɔ		Zyg Min(s) L.L.S.(s)	D.L.I.(s)	O.O.(p)	M.(s)			
Etirées i, e						Buc.(p) Ris.(s) Zyg Maj.(s)		D.A.O.(s)

Liste des Abréviations

Buc. = Buccinator D.A.O. = Depressor Anguli Oris D.L.I. = Depressor Labii Inferioris L.A.O. = Levator Anguli Oris L.L.S. = Levator Labii Superioris

M. = Mentalis O.O. = Orbicularis Oris O.O.L. = Orbicularis Oris Inferior Plat. = Platysma Ris. = Risorius

Zyg. Maj. = Zygomaticus Major Zyg. Min. = Zygomaticus Minor (a) = Action antagoniste (p) = Action protagoniste (s) = Action synergique
--

Figure 01.7 – Les réalisations articulatoires et les mouvements labiaux correspondant (d'après Abry 1980).

1.3.3 La coarticulation : cibles en contexte

Les six classes labiales précédentes, et les trois degrés de liberté qui les distinguent, caractérisent des situations où les sons prononcés sont complètement isolés. Comme il a été évoqué plus haut, la production de la parole ne suit pas un fonctionnement idéal où une séquence de formes labiales traduit directement au niveau visuel la séquence du code phonologique initial. Cette approche fut celle des tout premiers systèmes de synthèse visuelle de la parole. A chaque phonème (unité de son) on associe une forme labiale prédéfinie (« key frame »). On crée ensuite une animation pour n'importe quel texte en juxtaposant les formes

clés des phonèmes. Si cette approche peut faire « illusion » (elle est encore largement utilisée dans l'industrie du dessin animé), elle ne recouvre cependant pas le caractère continu de la production de la parole. D'abord, la biomécanique musculaire imprime par nature des transitions continues entre les différentes formes de lèvres. De plus, au cours de la séquence des sons produits, les articulations consécutives s'influencent mutuellement par des phénomènes d'anticipation et de rétention motrice. On parle de coarticulation pour désigner ces phénomènes (Whalen, 1990).

Les études sur la géométrie labiale rassemblées dans (Abry 1980) mettent en évidence ce problème de coarticulation pour le Français sur un cas particulier. Le cadre de travail s'appuie sur la mesure géométrique du maintien de la séparation des voyelles arrondies et étirées (/y/ vs /i/) dans un contexte consonantique « assimiland » de constrictives protruses /S/ ou étirées /z/. Pour illustrer l'importance de la coarticulation, il est montré par exemple que, sur 6 locuteurs prononçant une syllabe /Si/, la protrusion pour l'articulation du /S/ se répercute sur la voyelle /i/ et ne permet plus à elle seule de distinguer géométriquement la voyelle /i/ de la voyelle /y/ prise dans un contexte similaire /Sy/.

1.4 La parole audiovisuelle et ses applications en communication

Cette section dresse un bilan des études qui ont mis en évidence la bimodalité, auditive et visuelle, de la parole et le gain en intelligibilité qu'elle apporte dans la communication parlée.

1.4.1 La bimodalité intrinsèque de la parole

La perception audiovisuelle de la parole ne procède pas d'une simple juxtaposition des modalités mais découle de notre sensibilité à rechercher et percevoir la cohérence entre les phénomènes acoustiques et visuels liés à la production de la parole (Dodd and Campbell, 1987 ; Massaro 1987 ; Cathiard 1989). La sensibilité à la cohérence audiovisuelle se manifeste dès le plus jeune âge, avant même l'acquisition du langage. Kuhl and Meltzoff (1982) ont présenté à des enfants de 4 à 5 mois deux visages d'une même personne prononçant deux séquences différentes de parole accompagnées de la bande son correspondante à une seule des deux. Il a été observé que les enfants étaient davantage attirés par le visage prononçant ce qu'ils entendaient.

Ce mécanisme de fusion semble de plus être relativement précoce dans la perception bimodale : c'est ce que révèle une célèbre illusion connue sous le nom de « l'effet McGurk » (McGurk and McDonald 1976). Dans cette illusion, des sujets à qui on présente une séquence

vidéo où un visage prononce /ga/, synchronisée avec une séquence audio /ba/, perçoivent avec certitude un troisième stimulus /da/. Cette illusion a été observée dans plusieurs langues et même chez des enfants (Burnham and Dodd, 1996). Cette fusion est très robuste aux conditions externes puisqu'elle persiste même lorsque les sujets sont prévenus de l'effet. Ce mécanisme résiste aussi à une désynchronisation de plusieurs dizaines de millisecondes entre les deux sources.

Le montage inverse (stimuli visuel /ba/ et acoustique /da/) ne donne cependant pas la même illusion : il est perçu comme une succession rapide /bga/ des deux stimuli qui sont ainsi perçus séparément (effet de streaming). Lors de l'effet McGurk, les perceptions de ces deux stimuli sont intégrées en une perception audiovisuelle unique, prenant le dessus sur chacune des deux modalités séparées. Cet effet suggère l'existence d'une représentation audiovisuelle autonome pour la perception de la parole, intégrant les deux sources d'information avant tout décodage phonétique séparé dans l'une ou l'autre des modalités. Un manque de cohérence entre ces deux sources peut donc entraîner une perception erronée de la réalité.

De manière naturelle l'interaction entre les perceptions auditive et visuelle de la parole opère en coopération dans les trois situations suivantes :

- localisation et focalisation de l'attention sur un locuteur particulier dans un environnement où d'autres parlent en même temps (effet « cocktail-party »),
- redondance entre les informations acoustique et visuelle lorsque les deux modalités sont bien perçues, entraînant un gain d'intelligibilité systématique quel que soit la qualité de décodage dans chaque canal,
- complémentarité entre les informations acoustique et visuelle lorsque du bruit ambiant dégrade la perception auditive pure.

Summerfield (1987) a comparé les réponses de sujets pour la reconnaissance de séquences comportant des consonnes en contexte vocalique (VCV), en condition auditive seule et en condition visuelle seule. L'arbre de confusion des réponses auditives montre une organisation globalement inverse de son équivalent visuel : ce qui est bien perçu acoustiquement ne l'est pas visuellement et vice versa. Notamment, les résultats montrent un discernement visuel entre /p/, /t/ et /k/ plus efficace qu'en acoustique. A l'inverse une forte confusion visuelle entre /p/, /b/ et /m/, tout trois caractérisé par une même fermeture bilabiale, disparaît au niveau acoustique. Walden et al (1977) ont rapporté des résultats similaires avec des sujets spécialement entraînés à la lecture labiale. Une des propositions de Summerfield (1989) sur cette complémentarité est d'associer les articulateurs visibles (lèvres, dents et

langue) à la production des sons de fréquence élevée, sons provoqués par des mouvements rapides comme lors de certaines consonnes occlusives. Ils correspondent acoustiquement à des turbulences de faible intensité sonore dont la sensibilité au bruit acoustique est alors corrigée par l'information visuelle apportée par leur articulation. A l'inverse, la position des articulateurs non visibles (langue, vélum, larynx) produisent des sons constants, de forte intensité, à des fréquences basses caractéristiques notamment du mode d'articulation (nasal ou oral) et des voyelles.

On peut aussi expliquer cette complémentarité à travers les résultats présentés par Fant (1973) : la résonance de la cavité arrière (non visible) correspond généralement au premier formant, alors que le second formant correspond plutôt à la cavité avant. Si le premier formant présente une bonne stabilité, le second varie davantage. La vision des lèvres, auxquelles il est lié, renforce alors la stabilité de la perception.

Au delà de la reconnaissance de phonèmes isolés, la continuité des transitions entre les réalisations articulatoires d'une séquence d'unités phonologiques fait apparaître des phénomènes de coarticulation. Ce dernier est une conséquence directe des contraintes de production propre à la nature continue de la parole. Les gestes articulatoires, programmés pour la réalisation d'un phonème « cible », peuvent être anticipés avant et persister après la réalisation (Whalen 1990). Affectant à la fois les réalisations acoustiques et visuelles, les phénomènes de coarticulation sont largement exploités dans la perception audiovisuelle de la parole. Dans une expérience où des sujets devaient simplement deviner la voyelle finale dans des séquences /zizi/ et /zizy/ tronquées, Escudier et al. (1990) ont montré que des sujets identifiaient le /y/ de /zizy/ sur une photo du visage prise environ 80 ms avant l'instant où ils étaient capables de l'identifier auditivement sur des séquences acoustiques tronquées de forme générale /ziz/. Ces résultats montrent que, de manière naturelle, la perception auditive et visuelle peuvent intégrer et exploiter d'une manière cohérente des désynchronisations entre vision et audition pour la reconnaissance d'une même unité phonologique. Ces phénomènes de coarticulation font partie prenante de la parole audiovisuelle.

1.4.2 L'intelligibilité de la parole audiovisuelle

La lecture labiale chez certains déficients auditifs prouve la capacité du visage d'un locuteur à porter de l'information linguistique. Cette faculté se retrouve chez des sujets ne présentant aucune perte auditive. Bien sûr, la perception auditive reste alors prépondérante sur la perception visuelle tant que le signal acoustique est suffisamment clair. Par contre, en présence de bruit, l'information visuelle contribue de manière significative à augmenter

l'intelligibilité du signal de parole par effet à la fois de redondance et de complémentarité. La bimodalité intrinsèque de la perception de la parole a été illustrée à travers de nombreuses expériences d'intelligibilité en milieu acoustiquement dégradé (Sumbly et Pollack 1954 ; Neely 1956 ; Binnie et al. 1974 ; Erber 1975 ; Summerfield 1979, 1989 ; Benoît et al. 1996).

La figure 1.8 montre des scores d'identification d'un vocabulaire de 18 mots sans signification, du type VCVCV, en fonction du rapport signal sur bruit. La courbe inférieure représente les scores avec l'audio seul, la courbe intermédiaire représente les scores avec l'audio et une image seuillée des lèvres du locuteur, la courbe supérieure représente les scores obtenus avec le signal acoustique et le visage complet du locuteur (Benoît et al. 1996). Ces résultats illustrent le rôle prépondérant des lèvres dans la perception visuelle de la parole. Il n'est pas suffisant puisque la vision des lèvres seules excluent l'information apportée par la mâchoire, la pointe de la langue et tout le mouvement du visage en général.

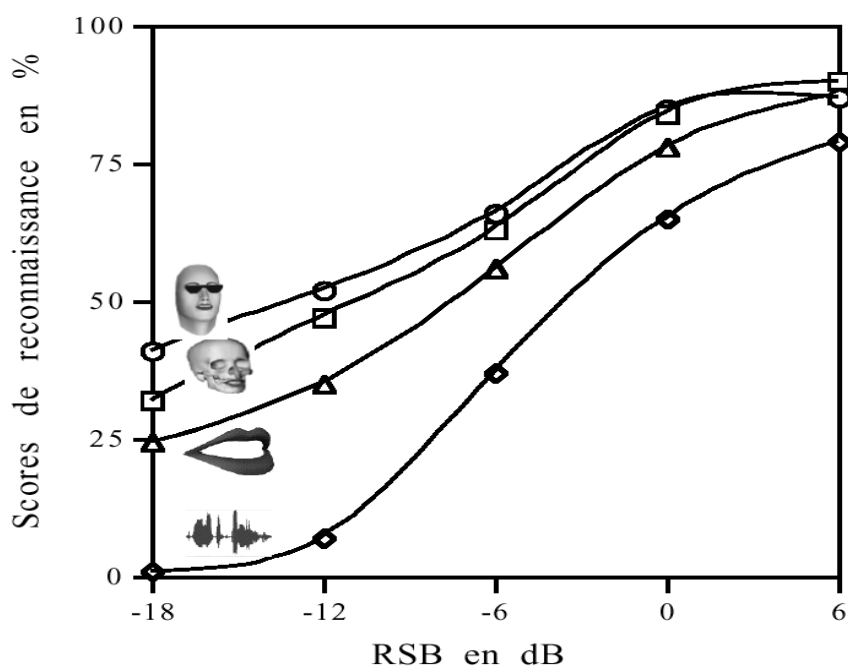


Figure 01.8 – Comparaison de l'intelligibilité de la parole bimodale en condition bruitée en ajoutant successivement les lèvres, le mouvement de la mâchoire puis tout le visage du locuteur (Benoît et al., 1996).

Le gain d'intelligibilité apporté par le visuel a été observé dans d'autres situations où la difficulté de compréhension est liée non pas à la dégradation des conditions acoustiques mais à la complexité linguistique du message. Dans une étude menée par Reisberg et al (1987), il est apparu que la compréhension orale d'un passage de la Critique de la Raison Pure (Kant, 1787) était améliorée lorsque le visage du locuteur prononçant le texte était présenté aux sujets.

1.4.3 Perspectives pour la communication homme-machine

L'essor exceptionnel du multimédia et des réseaux informatiques lance aux technologies de la parole un défi d'humanisation dans la communication avec et par la machine. La production et la perception de la parole humaine étant bimodale par nature, son exploitation par la machine à travers des personnages synthétiques audiovisuels parlants ou des systèmes de reconnaissance automatique peut rendre la communication avec celle-ci plus humaine et donc plus conviviale. Pour ces deux types d'applications, l'analyse automatique des mouvements labiaux fournit une source pertinente de paramètres.

La plate-forme « canonique » de télécommunication constituée de caméras, d'un canal de transmission à haut débit et de moniteurs vidéo permet de connecter des interlocuteurs sur deux modalités. Telle est l'approche classique de la visioconférence. Outre le fait que ce mode de communication ne laisse aucune chance à la machine d'intervenir ni sur la représentation du communicant (possibilités de substitution par un clone virtuel), ni sur le contenu du message (reconnaissance et interactions homme-machine), il interdit la connexion entre participants ne s'exprimant pas dans la même modalité (communication avec une personne handicapée). Indépendamment des problèmes technologiques liés au transport des informations (notamment vidéo) à une cadence temps réel, ces limitations expliquent sans doute les échecs relatifs des systèmes de visioconférences auprès du grand public. Par contre, l'engouement pour la réalité virtuelle et ses applications connaît un développement exceptionnel. Si l'animation des mouvements corporels des personnages de synthèse atteint aujourd'hui des degrés impressionnants, l'équivalent pour les mouvements de parole présente un retard technologique important.

1.4.3.1 Reconnaissance automatique de la parole audiovisuelle

Comme il a été observé et mesuré pour l'intelligibilité de la parole humaine en milieu bruyé, l'information visuelle permet d'envisager un gain en robustesse pour les systèmes de reconnaissance automatique de la parole. En effet, le problème majeur des systèmes purement acoustique réside dans leur sensibilité à différentes sources de bruit rencontrées en situation réelle d'application : dégradation du signal, confusion avec d'autres signaux de parole ambiants, bruit environnant... Plusieurs études ont montré qu'en ajoutant des paramètres optiques aux paramètres acoustiques habituels les scores de reconnaissance augmentaient de manière significative (Petajan 1984 ; Waibel and Lee 1990 ; Bregler et al. 1993 ; Rogozan et al. 1996; Luetin 1997).

A l'ICP (Institut de la Communication Parlée), les mêmes paramètres labiaux géométriques utilisés pour la synthèse visuelle ont servi de paramètres optiques pour les systèmes de reconnaissance audiovisuelle. Le système développé par Adjoudani et Benoît (1995) a montré en particulier la capacité à fusionner les informations auditives et visuelles de telle sorte que, comme pour l'homme, les scores audiovisuels dépassent les résultats des systèmes ne prenant en entrée qu'une seule des deux modalités, et ce quelque soit le niveau de rapport signal sur bruit. En effet tous les travaux dans ce domaine ont le même schéma de principe (voir figure 1.9) : extraction des paramètres audio et vidéo, intégration audiovisuelle de ces données, puis le système de reconnaissance a proprement parlé.

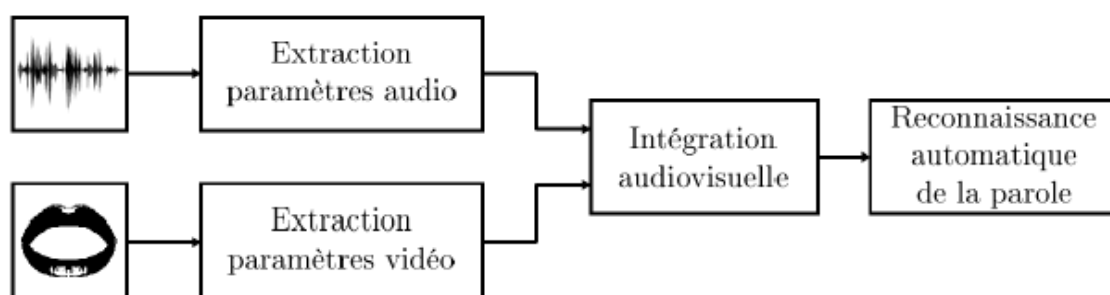


Figure 01.9 – Schéma de principe de la reconnaissance automatique de la parole.

1.4.3.2 Codage spécifique de la parole : la norme MPEG4

L'intérêt de ces applications de télécommunication a fait émerger la nécessité de prendre en compte la parole audiovisuelle (et son codage) comme un objet spécifique. Les travaux menés dans le cadre de la norme MPEG4 (1999, <http://drogo.cselt.stet.it>) visent à donner une spécification stable pour le codage numérique des informations audiovisuelles. Le visage humain en particulier est décrit par un ensemble de points géométriques (Facial Animation Parameters, FAP). Dans l'optique de véhiculer à la fois parole et émotions à travers la modalité visuelle, la région des lèvres bénéficie d'un surcroît de détails. En se focalisant sur la communication langagière, l'ensemble des résultats présentés dans cette thèse s'inscrivent dans cet enjeu technologique de codage optimisé des signaux humains.

1.4.3.3 Le rôle de la biométrie

Les applications de synthèse et de reconnaissance audiovisuelle ont démontré la validité des approches pour la communication homme-machine. Elles s'appuient, à l'ICP en

particulier, sur l'extraction précise de paramètres géométriques labiaux obtenus grâce à un maquillage bleu et un fort éclairage (Lallouache 1991). Ces paramètres ont prouvés leur pertinence pour représenter une information visuelle de parole. Si les conditions de mesure garantissent une excellente précision, elles s'opposent à une utilisation « conviviale ». Or, les applications de telles techniques audiovisuelles visent justement à améliorer la convivialité de la communication avec la machine. En particulier, un des arguments de la reconnaissance audiovisuelle automatique s'appuie sur la robustesse au bruit d'une telle approche la destinant donc à une utilisation en environnement « réel ». Un maquillage systématique rentre en contradiction avec cette argumentation. Une labiométrie sans maquillage s'impose donc comme l'étape suivante pour rendre réellement accessible un tel mode de communication avec la machine.

L'état de l'art dans le domaine montre que, par sa complexité, le défi d'une labiométrie sans maquillage a d'abord intéressé la recherche en vision par ordinateur. En effet, les mouvements labiaux suivent des déformations complexes qui imposent nécessairement d'avoir recours à des techniques élaborées. Néanmoins, ces déformations tendent à suivre des degrés de liberté identifiables et en faible nombre lorsque le contexte est contraint par un but de production de la parole.

1.5 Conclusion

Les lèvres fournissent les paramètres les plus fiables pour la reconnaissance visuelle de la parole puisqu'elles portent à la fois une part importante d'information et qu'elles sont toujours présentes et clairement identifiables. Un articulatoire comme la langue ne présente pas autant de facilité d'accès à partir d'une séquence vidéo.

L'aperçu de l'état de l'art montre que la labiométrie sans maquillage a d'abord fourni un défi technologique pour la vision artificielle. Du traitement de la couleur à l'extraction de paramètres visuels, toutes les étapes sont complexes. Il ressort que l'on ne peut envisager de résoudre que par des techniques d'apprentissage l'immense variabilité des conditions d'éclairage, des mouvements labiaux d'un locuteur et des différences entre locuteurs. De plus, il est nécessaire d'intégrer à la fois un traitement sur la couleur et la forme dans une approche à la fois orientée image et modèle. L'utilisation d'une information comme le gradient spatial d'une image se révèle largement insuffisante.

Le but des méthodes classiques de suivi de contour s'inscrit dans une optique de reconnaissance de formes et vise à retrouver l'allure exacte des contours. Cette tâche est mal définie lorsque le contraste de couleur entre les régions à segmenter est faible. Elle nécessite

alors un apport d'information par des contraintes sur un modèle de contour pour régulariser le problème.

Toutes les méthodes proposées se positionnent suivant un compromis entre contraintes au niveau local ou global. Les contraintes locales se limitent souvent à respecter des conditions de continuité du contour (au premier et second ordre). Elles laissent beaucoup de liberté à la description géométrique mais présentent de ce fait des problèmes de stabilité, le modèle de contour ayant la possibilité de se fixer sur n'importe quelle limite de régions. A l'inverse, les contraintes globales imposent des propriétés géométriques de haut niveau (contours décrits en termes d'ellipse, d'arc de parabole, ...) pour limiter les variations de forme du modèle à la topologie propre du contour suivi. Les paramètres de contrôle de la forme étant plus réduits, la recherche est stabilisée. Elle évite les frontières parasites mais perd la précision de description des méthodes locales. Les limitations de formes imposées par les méthodes globales peuvent être telles qu'elles ne sont plus en mesure de représenter la forme réelle à suivre et ainsi d'assurer une convergence correcte.

Le débat reste ouvert quant au choix des méthodes pour le suivi des contours labiaux. Aucune ne s'est encore imposée. La faiblesse du contraste entre peau et lèvres exclut une utilisation unique des méthodes locales. Les méthodes globales actuelles ne résolvent pas le compromis entre une description géométrique suffisamment précise et un contrôle sur peu de paramètres.

Le problème réside dans le fait que les paramètres des modèles doivent contrôler directement toute la variation géométrique de la forme labiale. En séparant caractérisation géométrique et contrôle articulatoire, nous montrons dans cette thèse que, pour un locuteur particulier, il est possible de définir un modèle à la fois précis au niveau géométrique et de le commander ensuite par seulement trois paramètres, représentatifs de toute la variation articulatoire du locuteur. Ainsi, utilisé dans un cadre de suivi de contour, notre approche résout les deux exigences de précision et de stabilité.

Enfin, au delà du défi de vision artificielle, on retiendra de la section sur la parole audiovisuelle qu'il ne faut pas perdre d'esprit le but premier d'une labiométrie : extraire des paramètres visuels qui, comme les paramètres issus du « bleu », portent de manière pertinente une information de parole. C'est précisément ce codage de « l'objet de parole » que nous visons par notre approche articulatoire de la labiométrie.

La reconnaissance visuelle de la parole

2

La première difficulté rencontrée pour l'obtention des informations visuelles utilisables pour la reconnaissance audiovisuelle de la parole est celle de la localisation de la zone à étudier. Cette zone se situe, en général, vers bas du visage, voire plus exactement la bouche seule. Cette difficulté n'apparaît pas pour les systèmes fournissant directement des mesures, mais elle se posait déjà de façon très simplifiée dans les systèmes où le locuteur est préparé à être filmé pour extraire des informations visuelles. En effet, le maquillage ou les pastilles utilisées sont choisis pour être aisément repérables, ce qui facilite d'autant la localisation de ces zones marquées.

Pour simplifier le problème quand le locuteur n'est pas préparé, il est possible de recourir à des dispositifs spécifiques pour le filmer (casques-caméra), ce qui permet d'assurer le cadrage voulu, voire un éclairage contrôlé et constant. Si l'on ne dispose pas de tels dispositifs ou que l'on vise un cadre applicatif plus libre, ou le recours à de tels dispositifs n'est pas envisageable, une première phase consistera alors nécessairement à localiser le(s) locuteur(s) dans l'image, puis assez souvent, à délimiter plus précisément la zone d'étude (la bouche). Une fois la zone d'intérêt (ROI : Region Of Interest) déterminée, il faudra en extraire les informations utilisables pour la reconnaissance de parole. Dans ce contexte deux approches sont fréquemment rencontrées dans la littérature du domaine:

- Approche modèle : Dans ce cas on cherche à extraire les informations de type mesures de distances et de surfaces comparables à celles que l'on extrayait avec préparation du locuteur. Cependant, il est extrêmement difficile d'atteindre la qualité des mesures effectuées avec préparation du locuteur, pour lesquelles les erreurs sont très faibles. Sans préparation, dans des conditions que nous qualifierons par la suite de naturelles, on ne pourra, dans l'état actuel de la recherche, qu'obtenir des mesures fortement entachées d'erreurs que nous qualifierons d'estimations pour ne pas les confondre avec les mesures précises que l'on obtenait avec préparation.
- Approche image: Pour ce type d'approche, l'information visuelle est dérivée plus ou moins directement des valeurs de niveaux de gris (voire de couleur) des pixels de

- l'image de la région de la bouche. Dans ce cas l'utilisation de mesures fait perdre une information visuelle importante, notamment la présence ou l'absence de la langue et des dents quand la bouche est ouverte ou fermée.

Dans ce chapitre, nous présenterons dans un premier temps les techniques utilisées pour localiser le visage et assurer son suivi, puis, nous passerons en revue des méthodes permettant de localiser plus précisément la bouche et le type d'informations visuelles (image ou modèle) qu'on peut extraire, ainsi que les méthodes permettant cette extraction, dans certains cas, quand le locuteur n'est pas préparé. Enfin, nous finirons ce chapitre par une présentation des principaux corpus de parole audiovisuelle présentant des locuteurs non-maquillés.

2.1 Influence de l'angle de vue

Dans les tests de perception visuelle de la parole, nous trouvons qu'il y a des auteurs choisissent de présenter leurs stimuli visuels sous des angles de vue différents. Ceci prouve en quelque sorte que l'information visuelle perçue dépend en partie de ce facteur de visibilité. Ce dernier a été l'objet de plusieurs études, parmi lesquels (Neely 1956; Larr 1959; Nakano 1961; Berger et al. 1971; Erber 1974; Cathiard 1988, 1994; Adjoudani 1998).

A l'exception de l'étude de (Adjoudani 1998), utilisant des paramètres extraits des contours des lèvres, toutes ces études, s'appuient sur des tests perceptifs. Dans ces études, trois vues ont été comparées : la vue de face, la vue de profil et la vue de 3/4. De ces comparaisons, nous pouvons conclure que :

- la vue de face apporte plus d'information que la vue de profil, à l'exception de certains cas spécifiques concernant la classification des traits labiaux de protrusion et d'étirement (Cathiard 1988, 1994), ou la vue de profil peut être plus efficace que la vue de face.
- La vue de 3/4 est globalement équivalente à la vue de face.

Dans le cas du code LPC (Langage Parlé Complété), ou la main et les lèvres doivent être simultanément visibles, la vue de 3/4 poserait des problèmes de visibilité notamment pour la forme de la main. De même, la vue de profil ne peut permettre la visibilité complète des positions de la main ni des formes. De plus, elle est, en général, moins efficace que les deux autres vues. Il reste donc la vue de face qui, a priori, semble la plus appropriée au cas du code LPC.

2.2 Visage complet ou indices visuels ?

Percevoir le visage d'un locuteur apporte bien un gain d'intelligibilité en perception de la parole. Mais quelles sont les parties qui contribuent le plus à ce gain ? Pour répondre à cette question, rappelons d'une part que dans la majorité des expériences décrites au chapitre 1, notamment celles sur la perception visuelle de la parole, le visage complet (et dans certains cas les épaules et la tête) était présenté aux sujets testés. D'autre part, des études ont montré que la région de la bouche transmettait la plus grande partie de l'information visuelle de parole. D'autres études allaient jusqu'à suggérer de se contenter seulement des lèvres.

Dans cette section, nous présentons les résultats de quelques études comparant différentes conditions de présentation des stimuli visuels. Summerfield (1979) a comparé les gains d'intelligibilité de différents types d'information visuelle. Il a présenté à 10 sujets (âgés de 15 à 27 ans) des stimuli audiovisuels produits par un locuteur anglais sous forme de phrases, mélangés avec d'autres signaux de parole, dans cinq conditions différentes: (i) signal acoustique seul, (ii) signal acoustique+ le visage du front à la mandibule, (iii) signal acoustique + les lèvres seules, (iv) signal acoustique + 4 points lumineux placés autour des lèvres sur les coins et sur les intersections de l'axe de symétrie avec les lèvres supérieure et inférieure, (v) et signal acoustique + un cercle dont le diamètre varie selon l'amplitude du signal acoustique non bruitée. Sous ces différentes conditions les sujets devaient identifier les phrases testées et les noter sur papier. Les résultats obtenus dans cette expérience sont présentés par la table 2.1.

Condition	Audio seul	Audio + visage complet	Audio + lèvres	Audio + 4 points	Audio + cercle
Pourcentage moyen (%)	22.7	65.3	54	30.7	20.8
Ecart type	8.59	19.7	14.5	16.2	10

Table 02.1 – Scores d'identification obtenus par Summerfield (1979) dans cinq conditions de présentation des stimuli.

De ces résultats nous pouvons tirer quelques constats intéressants. Tout d'abord, les deux informations visuelles dans les conditions (iv) et (v) ne semblent apporter aucune information aidant à comprendre les phrases bruitées. Les différences entre ces deux conditions et la condition (i) sont en effet, selon l'auteur, non significatives. Ensuite, il est évident que la

présentation de l'image complète ou de l'image des lèvres est bénéfique pour la compréhension du message. Dans les deux conditions, les scores d'identification augmentent en moyenne de plus de 31% par rapport aux scores dans la condition audio seule. Et enfin, les lèvres seules portent une information importante mais restent encore inférieures à celle portée par le visage complet. Ces deux derniers constats ont été confirmés par d'autres études (Le Goff et al. 1995, 1996; Adjoudani et al. 1994).

Globalement, le visage complet est l'indice visuel qui apporte le plus d'information visuelle. Les lèvres portent une grande partie de l'information visuelle équivalente en quantité à peu près aux deux tiers de celle transmise par le visage complet. L'étude de Summerfield (Summerfield, 1983) a porté sur les conditions de présentation des indices visuels pour que l'information visuelle contribue plus pertinemment à la perception audiovisuelle de la parole. Ainsi, il suggérait les conditions suivantes :

- une distance de 1,5m,
- une luminance suffisante,
- le corps et les bras visibles aussi,
- pas de moustache ni de barbe sur le visage,
- et un maquillage des lèvres pour augmenter le contraste.

2.3 Localisation et suivi de visages

Comme nous le verrons par la suite, nous avons été amenés à enregistrer un corpus de parole audiovisuelle et avons choisi de cadrer le locuteur en limitant la prise de vue à la zone de la bouche. Cette prise de vue nous a semblé intéressante car elle permet de disposer d'une bonne résolution au niveau de la bouche et d'en détecter les mouvements même s'ils sont réduits. Cependant, le choix de filmer en gros plan la région des lèvres n'est pas neutre. Il impose d'effectuer une localisation approximative de la bouche de façon automatique et fiable, puis son suivi, non seulement dans des conditions de laboratoire, mais également pour des environnements plus variables, ce qui nous a amené à une étude bibliographique de faisabilité. En effet, la localisation de visages est le sujet de nombreuses études car les applications à ces recherches sont nombreuses : en plus de la reconnaissance automatique de parole audiovisuelle qui est notre principal centre d'intérêt, ces recherches s'appliquent à la reconnaissance automatique du locuteur et, plus généralement, à la vérification d'identité à partir du visage sans que le sujet ne parle (domaine de la biométrie).

À l'exception des travaux de (Shdaifat et al. 2001), qui localisent directement la bouche d'un locuteur dans une image, la localisation automatique de la région de la bouche se décompose généralement en deux étapes : dans un premier temps, le visage est localisé dans l'image, puis une localisation plus précise de la bouche est effectuée sur ce visage. Pour localiser les visages, deux types d'approches sont utilisées : des approches globales qui considèrent le visage comme un tout ayant une « apparence » particulière, et des approches par éléments qui détectent un certain nombre d'éléments du visage dans l'image, pour le localiser.

Dans cette section, nous aborderons tout d'abord la question de la localisation de visages à travers des deux approches précédentes, puis nous passerons en revue quelques systèmes de suivi.

2.3.1 Localisation de visages

La localisation de visages dans une image revient généralement à étiqueter les points de l'image suivant deux classes : le(s) visage(s) et le reste de l'image (qui n'est pas nécessairement uniforme). Dans tous les travaux que nous avons rencontrés pendant notre étude bibliographique, à l'exception de (Dai and Nakano 1996) et de (Yang and Waibe 1996), qui traitent des images contenant trois visages, ainsi que dans (Senior 1999) où, grâce à la multi-résolution, des visages d'échelles différentes peuvent être localisés, cette tâche est ramenée à une segmentation de l'image en deux zones : le visage et le fond, les images traitées ne contenant qu'un seul visage. Ceci peut sembler être une limite, mais dans la pratique, les images sur lesquelles il est possible d'étudier les mouvements des lèvres du locuteur rentrent généralement dans ce cadre contraint.

Plusieurs approches ont été étudiées : (Benoît et al. 1998) les séparaient en deux catégories principales, celles utilisant la couleur, et celles reposant sur la détection d'éléments du visage. Cette catégorisation peut être légèrement affinée : nous proposons d'étudier le fonctionnement de méthodes de détection de visages reposant dans un premier temps sur une utilisation de la couleur avec des contraintes définies a priori par les auteurs, puis définies statistiquement. Par la suite, nous examinerons quelques approches reposant sur la détection d'éléments faciaux. Enfin, nous verrons brièvement que l'information dynamique (mouvement) peut également être utilisée. Nous constaterons à cette occasion que de nombreux systèmes utilisent une combinaison des différentes approches.

2.3.1.1 Approches couleur

Dans cette première partie, nous allons passer en revue quelques méthodes de localisation de visages utilisant l'information couleur sous des formes variées et basées sur des critères a priori. Les chercheurs faisant appel à ces méthodes utilisent un espace couleur particulier permettant de faire ressortir l'information de teinte et déterminent des valeurs de seuils pour séparer les zones de peau du reste, empiriquement, à partir d'exemples.

Sobottka et Pitas (1996) utilisent l'espace de représentation couleur (H, S, V) et segmentent l'image en régions en la « filtrant » (passe-bande) en fonction des informations de teinte (H) et de saturation (S). Les pixels i retenus ont une saturation telle que $0.23 \leq S_i \leq 0.68$, et une teinte telle que $0^\circ \leq H_i \leq 50^\circ$. Des régions sont formées, puis combinées à partir des points candidats. Ce premier « filtrage » laisse passer de nombreux faux-positifs. Le visage ayant une forme approximativement elliptique, pour déterminer la zone la plus vraisemblable, des ellipses sont utilisées pour diminuer à nouveau le nombre de zones (de visage) candidates. Enfin, des éléments faciaux (yeux et bouche, décrits par les auteurs comme des zones sombres) sont recherchés en utilisant l'information d'intensité. En fonction des éléments trouvés et de leurs positions relatives à l'intérieur de la région candidate, le visage et la position de ces éléments seront localisés.

Ramos Sánchez (2000), de façon relativement similaire, utilise l'information couleur pour localiser le visage en approximant sa forme par une ellipse (voir figure 2.1). L'espace couleur utilisé est le plan de chromaticité ($r; v$) qui correspond à l'espace (R, V, B) normalisé par l'intensité totale (R + V + B) :

$$r = \frac{k.R}{R+V+B}, \quad v = \frac{k.V}{R+V+B}, \quad b = \frac{k.B}{R+V+B} \quad (2.1)$$

où le facteur $k = 3$ pour Ramos Sánchez qui divise les composantes couleur par la moyenne des trois composantes $\frac{R+V+B}{3}$, alors que généralement $k = 1$ (division par la somme des composantes (R + V + B)). La troisième composante normalisée b n'est pas utilisée car elle est redondante et peut se déduire des deux autres :

$$r + v + b = k. \quad (2.2)$$

Dans cette représentation, les points du visage se regroupent dans une zone réduite du plan (r, v), et la décision d'appartenance ou non au visage est faite suivant un critère de

distance à une valeur centrale. L'auteur indique avoir testé un modèle générique de la couleur de la peau construit à partir de 100 images de différents sujets de la base XM2VTSDB (Messer et al. 1999), mais que les résultats étaient « assez logiquement » moins précis qu'en utilisant des modèles de la couleur spécifiques aux locuteurs.

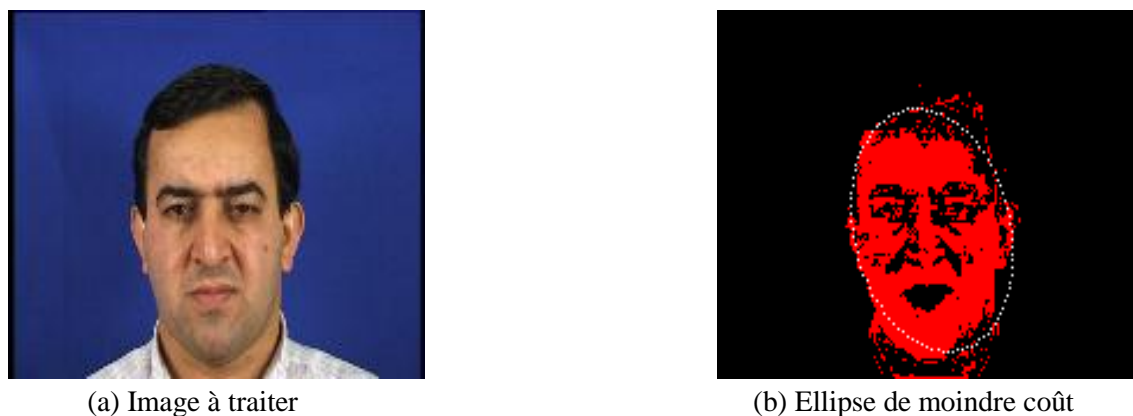


Figure 02.1 – Image couleur en entrée (a), pixels candidats pour appartenir au visage et localisation.

Duchnowski, dans des travaux plus anciens (Duchnowski et al. 1995), proposait déjà d'utiliser la couleur dominante des visages pour les localiser, grâce à un classificateur de couleur de visages basé sur les travaux de Hunke (Hunke 1994; Hunke and Waibel 1994), le FCC (Face Color Classifier, voir figure 2.2). Pour déterminer si un pixel de l'image a une couleur qui correspond à la peau du visage ou non, un modèle général de la couleur de visages (GFCC) a été obtenu en utilisant une image contenant des portions de peau de 30 visages de différentes couleurs (asiatiques, noirs et blancs). Les valeurs (R; V;B) des pixels de l'image ont été projetées dans le plan de chromaticité (r ; v) et un histogramme 2D a été calculé pour mesurer la fréquence d'occurrence de chaque couleur. Les occurrences les plus élevées se regroupent dans une portion réduite du plan (r ; v) et un rectangle est déterminé autour de cette zone (l'auteur ne précise pas comment). Pour la classification, les pixels i à l'intérieur du rectangle, c'est-à-dire ceux pour lesquels $r_{min} \leq r_i \leq r_{max}$ et $v_{min} \leq v_i \leq v_{max}$ où (r_{min}, v_{min}) sont les coordonnées du coin supérieur gauche du rectangle et (r_{max}, v_{max}) celles du coin inférieur droit, sont considérés comme appartenant au visage et les autres comme appartenant au fond. Ceci fournit de nombreux faux-positifs qui peuvent être éliminés en utilisant le mouvement (les zones immobiles peuvent être éliminées), puis, pour les faux-positifs restants, l'information géométrique (forme des objets), modélisée à l'aide de réseaux de neurones, est utilisée pour éliminer par exemple les mains et bras et ne conserver que les bons candidats. Après une première détection avec le modèle général GFCC, un modèle de la couleur du visage

individuel (IFCC) est calculé et utilisé. Il peut être ré-estimé régulièrement pour rendre la détection du visage robuste aux changements de l'environnement.

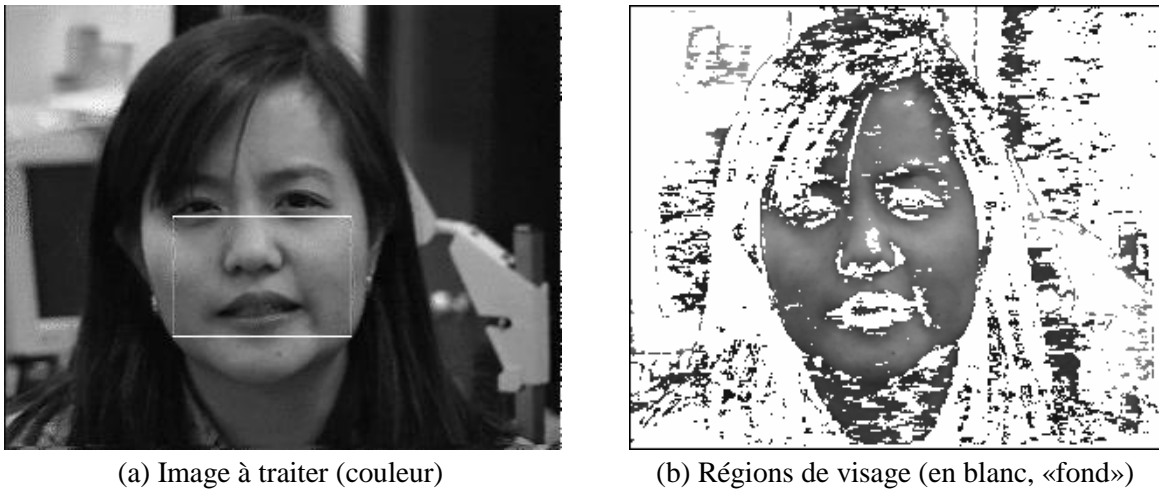


Figure 02.2 – Détecteur de visage de Hunke et Duchnowski basé sur la couleur (FCC) : (a) Image couleur à analyser et région utilisée pour entraîner le modèle (IFCC) de couleur du visage, (b) Sortie du FCC : en blanc, les zones de « non-visage », d'après (Duchnowski et al. 1995; Hunke and Waibel 1994).

Senior (Senior 1999; Neti and Senior 1999) utilise également une segmentation basée sur la couleur. Dans l'espace de représentation couleur (H, C, I), il utilise des seuils minimaux et maximaux sur ces trois composantes pour classifier les pixels comme « peau » ou « non-peau » (voir figure. 2.3). Il utilise notamment comme bornes pour la teinte $-90^\circ \leq H_i \leq 90^\circ$. Le calcul des bornes sur les autres composantes est détaillé dans (Senior 1999).

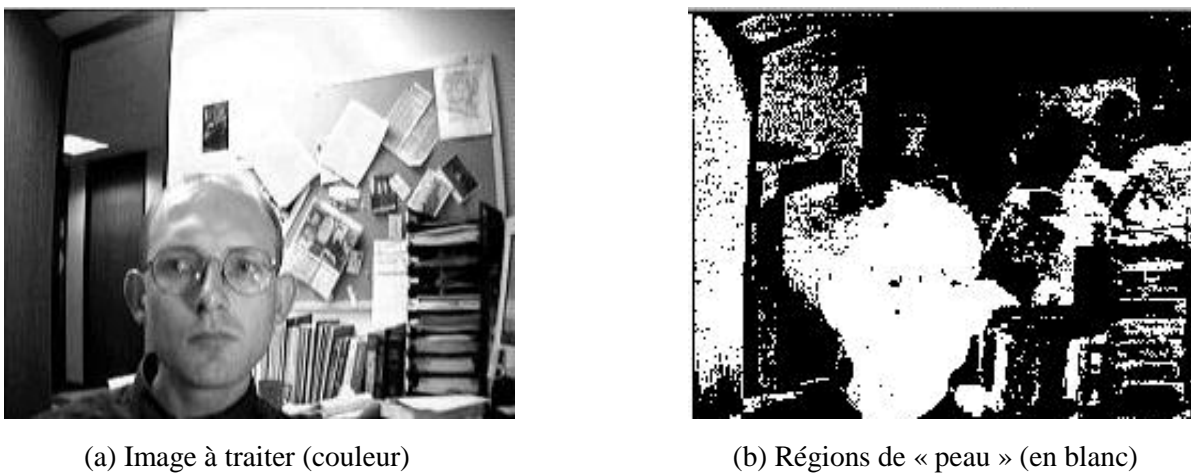


Figure. 2.3 – Une scène complexe (a) et sa classification en tons « peau » (b), d'après (Senior 1999).

Pour repérer plusieurs visages ou des visages de tailles différentes dans une image, Senior propose une approche multi-résolution en utilisant une pyramide d'images (l'image initiale ré-échantillonnée à des résolutions inférieures) et considère chaque zone rectangulaire de $m \times n$ pixels comme un candidat visage F . Les images de niveaux successifs dans la pyramide sont réduites d'un facteur de $\sqrt[3]{2}$ et la plus petite contient au moins $m \times n$ pixels. Chaque région F est évaluée en comparant à un seuil son nombre de pixels de « peau » selon les bornes utilisées dans l'espace (H, C, I). Quand des régions F sont retenues comme contenant un visage, elles sont évaluées de façon plus approfondie (scores), et la recherche peut encore être affinée en utilisant des ré-échantillonnages d'images intermédiaires ou des rotations légères de l'image.

(Wark and Sridharan 1998) utilisent la composante couleur quotient $Q = \frac{R}{V}$ proposée par (Chiou and Hwang 1996) pour la détection des lèvres (voir section 2.3.2), pour localiser le visage du locuteur dans les images du corpus M2VTS (Pigeon and Vandendorpe 1997). Plus précisément, les valeurs Q_i de chaque pixel i sont telles que :

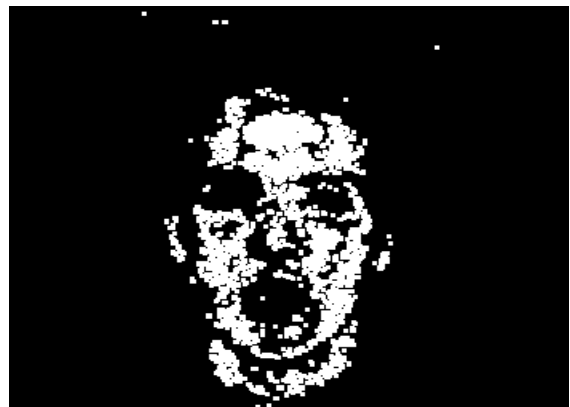
$$Q_{bas} \leq Q_i \leq Q_{haut} \quad (2.3)$$

Si Q_i est comprise entre ces deux bornes, le pixel i appartient au visage, sinon, il fait partie du « fond » (qui est uniforme dans M2VTS).

Les auteurs ont déterminé manuellement à partir d'exemples, les valeurs des seuils $Q_{bas} = 1.2$ et $Q_{haut} = 1.45$ et ces valeurs semblent convenir pour les 37 locuteurs du corpus M2VTS. Les pixels solitaires du « fond » étiquetés à tort comme faisant partie du visage sont supprimés à l'aide d'une opération morphologique (ouverture). L'application de ce traitement à une image de M2VTS (Figure. 2.4a), est illustrée dans la figure 2.4b.



(a) Image à traiter (couleur)



(b) Régions de visage (en blanc)

Figure. 2.4 – Localisation du visage sur le corpus M2VTS, d'après (Wark and Sridharan 1998).

(Dai and Nakano 1996) utilisent l'espace de représentation couleur (Y, I, Q) qui s'obtient par combinaison linéaire à partir des valeurs de base (R, V, B) comme suit :

$$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} 0.30 & 0.59 & 0.11 \\ 0.60 & -0.27 & -0.32 \\ 0.21 & -0.52 & 0.31 \end{pmatrix} \begin{pmatrix} R \\ V \\ B \end{pmatrix}. \quad (2.4)$$

Dans cet espace, la composante I varie de $I = 150$ (rouge) à $I = -150$ (cyan) en passant par $I = 0$ en l'absence de couleur dominante (pixels gris). Les auteurs construisent des images de la composante I en laissant inchangés les pixels i de l'image pour lesquels $1 \leq I_i \leq 50$. Les pixels ayant des valeurs dépassant le seuil ($I_i > 50$) sont ramenés à zéro. Les auteurs n'indiquent pas le traitement réservé aux valeurs négatives, mais on peut supposer qu'elles sont également ramenées à 0. Les images sont ensuite filtrées (moyennées) et le visage est repéré par simple seuillage de cette image. De façon plus précise, ce travail (Dai and Nakano 1996) étudie la localisation de visages à faible résolution (typiquement 20×20 pixels) dans des scènes complexes, en utilisant des textures (SGLD : Space Gray-Level Dependence matrix). L'utilisation de la couleur est vue par les auteurs comme un prétraitement qui a pour but de supprimer les zones qui pourraient par la suite être détectées à tort comme des visages par la SGLD. Un point faible de ce travail, souligné par les auteurs eux-mêmes, est qu'il est dédié à la teinte de peau asiatique et qu'en l'absence de tests pour d'autres types de couleur de peau, il n'est pas possible de mesurer sa généralité.

2.3.1.2 Approches statistiques

L'approche statistique pour la localisation de visages consiste à se baser sur un échantillon (des images exemples) que l'on souhaite représentatif, pour modéliser l'apparence d'un visage. L'approche peut être directe à partir d'exemples sans a priori, ou indirecte, en choisissant un espace de représentation intermédiaire sur lequel on réalise l'apprentissage statistique (Yang 2007). Dans ce second cas, la principale différence entre l'approche statistique et les travaux reposant sur une approche couleur précédemment évoqués est l'utilisation de bornes a posteriori, apprises à partir de données et non a priori, réglées « manuellement » par le concepteur du système.

(Rao and Mersereau 1995) proposent une approche statistique non-supervisée fondée sur la segmentation d'un objet et du fond. Une première estimation de la position de l'objet doit le contenir intégralement, ou être contenue intégralement dans l'objet, puis des ré-estimations successives des modèles de l'objet et du fond sont faites jusqu'à convergence. Pour le cas

particulier de la localisation de visages, l'objet visage est approximé par une ellipse (sans rotation). Les auteurs proposent également d'utiliser cette méthode pour segmenter les lèvres du reste du visage, ceci sera abordé plus en détail dans la section 2.3.2. L'approximation initiale est réalisée en utilisant un modèle du visage et du fond appris sur une seule image d'un autre sujet. Ce modèle est utilisé sur l'image à segmenter. Un seuil élevé assure que l'estimation initiale est entièrement contenue dans le visage à localiser. Puis les modèles du visage et du fond sont ré-estimés en fonction de la zone trouvée sur l'image de ce nouveau sujet. La zone initiale est modifiée en fonction de ces nouvelles estimations du visage et du fond. Une bonne localisation du visage est obtenue après quelques itérations. Pour la modélisation, un mélange de deux gaussiennes (2 GMM) avec matrice de covariance complète est utilisé pour chaque modèle (« visage » et « fond »). Cette technique n'est utilisable qu'avec des images ne présentant qu'un seul visage, sinon la convergence n'est pas assurée. De plus, selon les auteurs, le résultat dépend de façon importante de l'initialisation, et pour utiliser cette technique sur des locuteurs quelconques exposés à des éclairages différents, il faudrait constituer un modèle général de l'apparence d'un visage.

(Brunelli and Poggio 1993) localisent tout d'abord les yeux en utilisant la corrélation entre l'image à analyser et une imagerie d'œil droit et gauche. La bouche, le nez et les sourcils sont ensuite localisés en utilisant le gradient spatial horizontal et vertical ainsi que les connaissances anthropométriques standard a priori (voir figure. 2.5a). Les auteurs proposent également, dans cet article, d'utiliser la corrélation d'images modèles des yeux, du nez et de la bouche avec l'image (template matching), pour localiser ces différents éléments (voir figure. 2.5b). Les résultats obtenus en termes de reconnaissance d'identité sont de l'ordre de 90% en repérant les éléments avec le gradient spatial et de l'ordre de 100% avec l'approche « template matching ». Cependant la corrélation est plus coûteuse en temps de calcul que l'utilisation du gradient spatial.

Enfin, (Malasné et al. 2002) suivent des visages en temps réel avec une approche connexionniste, à l'aide de dispositifs électroniques dédiés (des FPGA). Un apprentissage supervisé de l'apparence est effectué avec des images des visages de deux sujets en basse résolution (40×32), sous-échantillonnés quatre fois horizontalement (10×32), avec un réseau de neurones. Les sujets sont ensuite correctement localisés (dans le meilleur des cas à 98,2%), dans quatre séquences de 256 images. Notons toutefois que ces images sont filmées avec la même caméra dans une pièce avec peu de variation de luminosité.

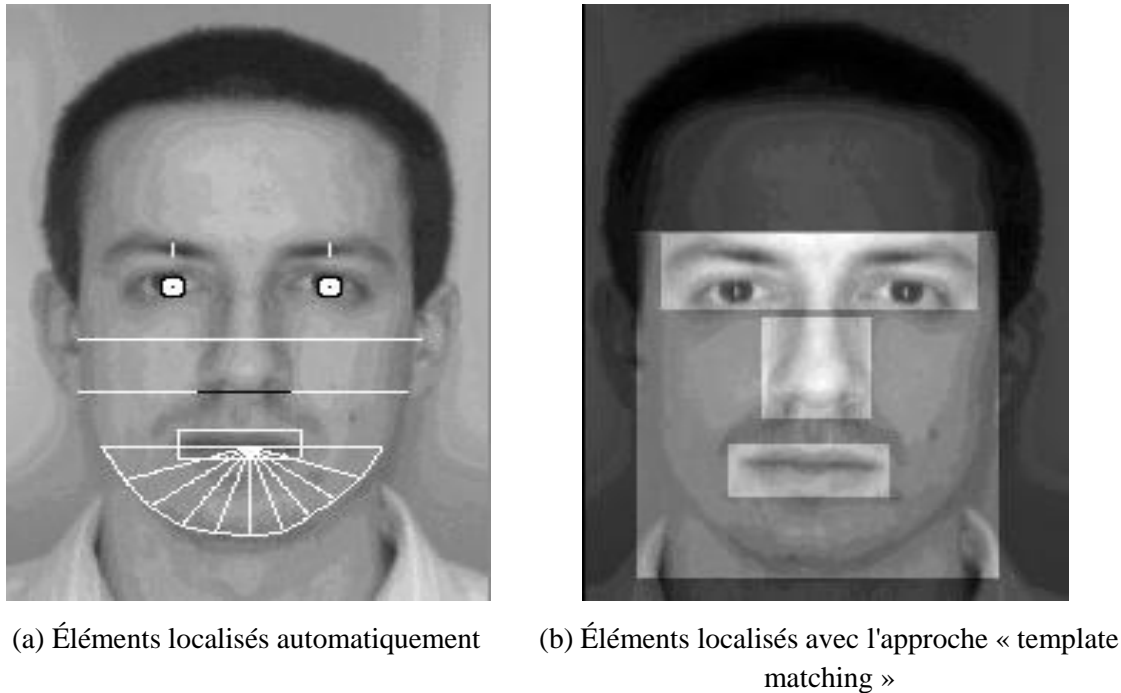


Figure. 2.5 – Localisation de différentes régions de visage (a) automatiquement (b) en utilisant l'approche « template matching », d'après (Brunelli and Poggio 1993).

Dans ce travail nous nous n'intéressons pas à la mise au point d'un système de localisation et de suivi de visages. Cette étude bibliographique avait pour but de déterminer la faisabilité, d'une part de la localisation approximative de la zone contenant la bouche (bas du visage), et d'autre part du suivi en temps réel d'un locuteur préalablement localisé. Une recherche bibliographique montre qu'on peut presque supposer qu'il est envisageable d'obtenir des images où la bouche du locuteur est toujours cadrée de manière identique, même si le locuteur bouge. Toutefois, si un certain nombre des techniques précédemment exposées sont utilisables dans le cadre que nous souhaitons étudier où le locuteur n'est pas préparé, le fond non obligatoirement uniforme, l'éclairage naturel et les problèmes d'ombre, les performances que l'on est susceptible d'atteindre risquent de diminuer. En effet, les approches par éléments peuvent être sensibles à un fond non-uniforme qui pourra créer de nombreux faux candidats. Les approches couleur peuvent également voir leurs performances diminuer si l'on ne contrôle pas l'éclairage comme l'explique Hunke (1994).

Cependant, même diminuées, les performances de localisation et de suivi de visage devraient rester suffisantes. Les approches utilisant un apprentissage statistique de la couleur (ou plus généralement de l'apparence globale) du visage et une détection d'éléments à l'intérieur de ce visage nous semblent les mieux adaptées. Le système de (Senior 1999) par

exemple a été utilisé avec succès par (Neti and Senior 1999; Potamianos et al. 2000) dans un cadre d'utilisation proche de celui que nous souhaitons étudier.

Comme nous l'avons signalé au début de ce chapitre, deux types d'informations sont extraits d'images de locuteurs non maquillés, pour la reconnaissance automatique de parole audiovisuelle : des informations « image » de bas niveau et des informations « modèle » de haut niveau. En réalité, il existe également des travaux adoptant une approche mixte qui extraient des images, des informations sur les valeurs de niveaux de gris de pixels le long de segments (profils) déterminés en utilisant des modèles.

Nous allons présenter dans cette section le type d'informations visuelles qui sont utilisées en lecture labiale automatique ou en AVASR dans les systèmes adoptant une approche « image » (section 2.2.2), puis dans les systèmes adoptant une approche « modèle » (section 2.2.3) et enfin dans les systèmes adoptant une approche mixte (section 2.2.4). La grande majorité de ces travaux nécessite d'avoir préalablement localisé la bouche de façon assez précise pour réduire l'étendue des images à traiter, et nous allons donc commencer par présenter comment cette localisation précise peut être obtenue dans la partie suivante.

2.3.2 Localisation de la bouche

Pour localiser approximativement la bouche d'un locuteur, connaissant la position de son visage dans l'image, il est possible d'utiliser les connaissances anthropométriques : de manière simplifiée, la bouche se situe dans la moitié inférieure du visage. Cependant, la qualité de la localisation du visage, et par la même occasion de la bouche, variera en fonction des techniques utilisées et de l'environnement considéré. Elle ne sera pas toujours parfaite, de plus, il existe des différences physiques intra-locuteur importantes. Si l'on envisage la création d'un système multi-locuteur, il faudra prévoir de s'y adapter. Pour toutes ces raisons, que l'on souhaite adopter une approche « modèle » ou une approche « image », il sera souvent nécessaire de localiser précisément la bouche. Pour l'approche « image », la zone localisée (ROI) délimitera l'image à utiliser, tandis que pour l'approche « modèle », le fait de restreindre la zone d'étude permet de limiter le nombre de minima locaux potentiels qui pourraient rendre la localisation du modèle inefficace. Notons également qu'en utilisant un dispositif d'acquisition comme un casque-caméra, même si la localisation du visage n'est plus nécessaire, le même problème de localisation précise des lèvres peut se poser.

Globalement, dans de nombreux cas, les équipes ayant également travaillé sur la localisation de visages, se proposent d'utiliser le même type d'approche pour la localisation

des lèvres. Pour les approches utilisant la couleur, il est possible de travailler sur un modèle de la couleur des lèvres comme il était possible de travailler sur un modèle de la couleur de la peau. Pour les approches statistiques, on peut tenter d'effectuer un apprentissage de l'apparence des lèvres comme pour le visage.

Nous allons donc présenter dans cette partie des techniques utilisées pour localiser finement la bouche. Certaines servent à définir la ROI utilisée pour les approches « image ». D'autres visent à détecter précisément les contours des lèvres pour calculer par la suite des paramètres labiaux géométriques (mesures de distances) ou de surfaces. Pour passer en revue les différentes possibilités, nous allons suivre un plan comparable à celui utilisé pour la localisation de visages en commençant par les approches couleur et statistique, en continuant avec celle utilisant la corrélation avec des patrons (template matching) et en terminant par l'utilisation de l'information temporelle.

2.3.2.1 Approches couleur

(Coianiz et al. 1996) propose d'utiliser l'information de teinte H de l'espace de représentation couleur ($H; S; L$) pour localiser les lèvres dans des images de bas de visage (du nez jusqu'au menton, voir figure 2.6a). Ils justifient leur choix par le fait que la teinte est peu sensible aux variations d'éclairement et que le contour externe des lèvres est difficile à localiser sur des images en niveaux de gris, ce qui rend hasardeuse l'utilisation du gradient spatial. Plus précisément, pour faire ressortir les zones à dominante rouge, l'angle de teinte H_i de chaque pixel i est tout d'abord décalé de $\frac{2\pi}{3}$ pour que le rouge corresponde uniquement à un angle de $H_0 = \frac{2\pi}{3}$ au lieu des deux valeurs de 0 et de 2π . La teinte est alors filtrée à l'aide d'un filtre parabolique centré sur le rouge. La teinte filtrée HF_i de chaque pixel s'obtient avec :

$$HF_i = 1 - \frac{(H_i - H_0)^2}{w^2} \text{ si } |H_i - H_0| \leq w \text{ et } HF_i = 0 \text{ sinon,} \quad (2.5)$$

où $w = \frac{1}{8} \times 2\pi = \frac{\pi}{4}$, permet d'indiquer la sélectivité du filtre. L'image filtrée peut être bruitée et l'auteur propose d'utiliser un filtrage passe bas (moyennage) pour faire disparaître les pixels aberrants (voir figure 2.6b). Pour enfin repérer la bouche, un sous-échantillonnage de l'image, puis un seuillage simple est utilisé : les pixels de niveaux de gris $HF_i * 255 \geq 244$ sont considérés comme les lèvres (voir figure 2.6c).

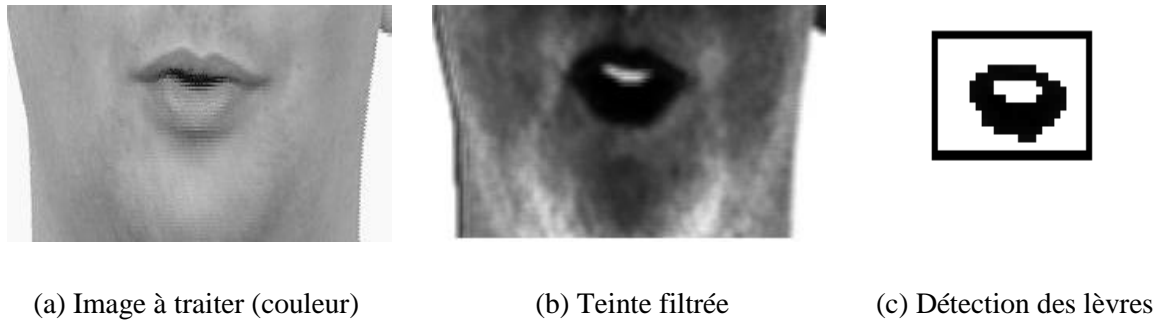


Figure. 2.6 – Localisation des lèvres en utilisant la teinte H, d'après (Coianiz et al. 1996).

(Vogt 1996; Vogt 1997) propose également d'utiliser l'espace de représentation couleur (H, S, I). Il utilise une combinaison de critères déterminés « manuellement » à partir d'images exemples, sur les composantes teinte H et saturation S. Ceci est codé dans une LUT (Look-Up Table), qui convertit l'image à analyser en une image permettant d'extraire les lèvres. Cette image est filtrée (Sobel) pour détecter les contours. Le contour externe des lèvres est finalement localisé à l'aide d'un modèle des lèvres (polygone) qui est placé sur l'image de contours (voir figure 2.6c).

(Chan et al. 1998) utilise également les informations de teinte H et de saturation S, mais calculées sur l'image sous-échantillonnée huit fois. Des seuils haut et bas sur les composantes H et S permettent de déterminer les pixels de lèvres. La plus grande zone de pixels de lèvres connectés est utilisée comme première estimation de la bouche.

Pour localiser les lèvres dans l'espace (R; V; B), Chiou et Hwang (1996) proposent d'utiliser le quotient $Q = \frac{R}{V}$ et d'appliquer un simple seuillage haut et bas de la valeur de ce quotient (voir eq. 2.3). Les pixels compris entre les bornes Q_{bas} et Q_{haut} appartiennent aux lèvres et les autres au fond. Notons que le locuteur est éclairé à l'aide d'une lampe de 60 Watts et que les auteurs indiquent que le système est dépendant du locuteur.

(Wark and Sridharan 1998) utilisent cette approche pour plusieurs locuteurs, les valeurs des seuils à $Q_{bas} = 1.7$ et $Q_{haut} = 2.0$, pour la détection de la région des lèvres dans le visage sur l'ensemble des images du corpus M2VTS (Pigeon and Vandendorpe 1997).

Pour la localisation préalable du visage (Wark and Sridharan 1998) utilisent cette même approche (voir section 2.1.1). Une fois la position approximative de la bouche détectée, de nouveaux seuils $Q_{bas} = 1.5$ et $Q_{haut} = 2.2$, sont utilisés (figure 2.7b), puis des opérations morphologiques (une ouverture suivie d'une fermeture, figure 2.7c) sont effectuées pour affiner la localisation et extraire le contour externe. (Gurbuz et al. 2001b; Gurbuz et al. 2001a;

Gurbuz et al. 2002) utilisent également l'approche proposée par (Chiou and Hwang 1996), en ajoutant une étape de filtrage pour diminuer le bruit dans l'image binaire obtenue à la place des opérations morphologiques proposées par (Wark and Sridharan 1998).

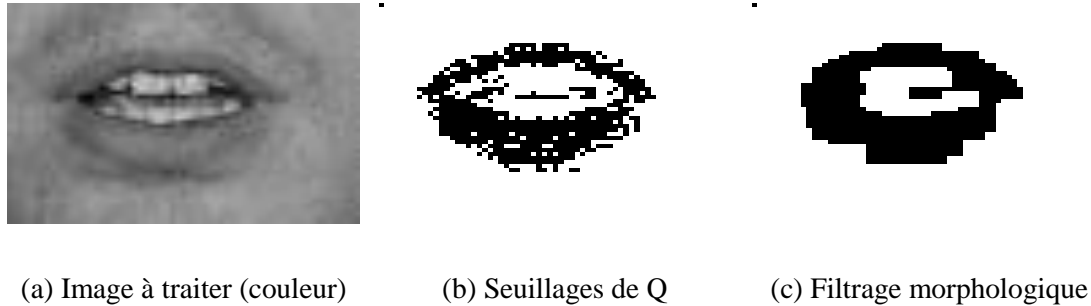


Figure. 2.7 – Localisation des lèvres en utilisant le quotient Q , d'après (Wark and Sridharan 1998).

Liew et al. (1999) proposent d'utiliser les espaces couleur (L, A, B) et (L, U, V) de la CIE (commission internationale de l'éclairage). Plus précisément, chaque pixel est représenté par un vecteur de dimension 7 :

$$\{A, B, U, V, \text{hue}_{ab}; \text{hue}_{uv}; \text{chroma}_{uv}\} \quad (2.6)$$

avec $\text{hue}_{ab} = \arctan\left(\frac{B}{A}\right)$, $\text{hue}_{uv} = \arctan\left(\frac{V}{U}\right)$, et $\text{chroma}_{uv} = \sqrt{U^2 + V^2}$.

Les auteurs proposent d'utiliser l'agrégation floue (« fuzzy clustering ») en fixant le nombre de classes à deux. Pour éviter des erreurs liées à l'apparition sur certaines images des dents (une troisième classe), les auteurs proposent de les masquer en utilisant un seuillage (la valeur du seuil est déterminée « manuellement » à partir d'exemples) sur la chrominance qui est relativement constante pour les dents quelque soit le sujet. Les régions de faible luminance L sont également masquées en raison de l'instabilité de leur chrominance. Les résultats présentés montrent que cette approche permet d'efficacement encadrer la région de la bouche, mais les résultats finaux pour le contour interne ne semblent pas particulièrement probants (voir figure 2.8c). En revanche, la carte d'appartenance floue aux deux régions (voir figure 2.8b) semble être une information plus facilement exploitable que la segmentation finale.

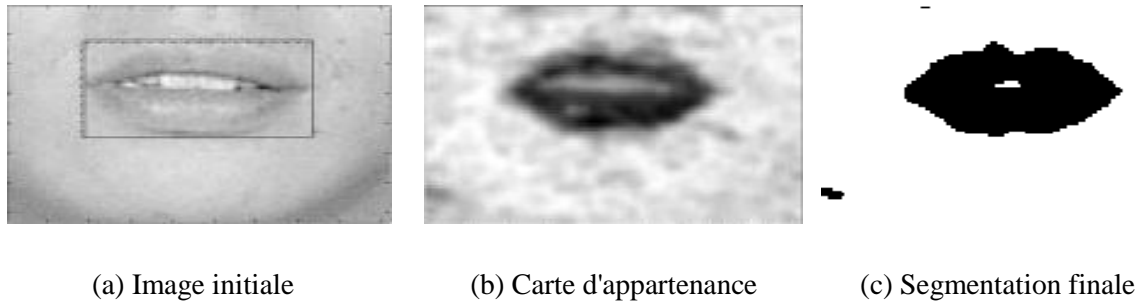


Figure. 2.8 – Détection des lèvres d'après (Liew et al. 1999).

2.3.2.2 Approches statistiques

Pour les approches statistiques, comme nous l'avons déjà évoqué dans la section 2.3.2.1, l'espace de représentation (couleur) idéal pour séparer les lèvres du reste du visage sera déterminé statistiquement à partir d'exemples, au lieu d'être déterminé a priori.

Pour la localisation de la bouche dans le visage, (Rao and Mersereau 1995) proposent d'utiliser la même approche statistique que celle qu'ils adoptent pour localiser le visage dans une scène complète (voir section 2.3.1.2). Le modèle de la bouche est constitué de deux arcs de parabole contenus dans un rectangle. Les modèles statistiques d'apparence de la bouche et du fond sont appris sur une seule image étiquetée manuellement. Les résultats préliminaires obtenus sur une séquence d'un locuteur unique semblent corrects, voir figure 2.9. On peut notamment remarquer sur cette illustration que l'intérieur de la bouche ouverte est correctement reconnu, mais aucun résultat où les dents sont visibles n'est présenté, ce qui limite l'évaluation d'une telle approche. Enfin, les auteurs indiquent que le contour interne pourrait également être détecté par cette méthode en considérant comme « objet », l'intérieur de la bouche et comme « fond », les lèvres.

Pour la localisation précise du contour externe des lèvres, (Chan et al. 1998) utilise une transformation linéaire des composantes (R, V, B) de chaque pixel i :

$$C_i = \alpha \cdot R_i + \beta \cdot V_i + \gamma \cdot B_i . \quad (2.7)$$

Les coefficients de pondération α , β et γ sont choisis statistiquement, comme dans (Kaucic and Blake 1998), pour maximiser la différence entre les pixels de bouche et de peau du locuteur, sur des images représentatives du problème à traiter, étiquetées manuellement.

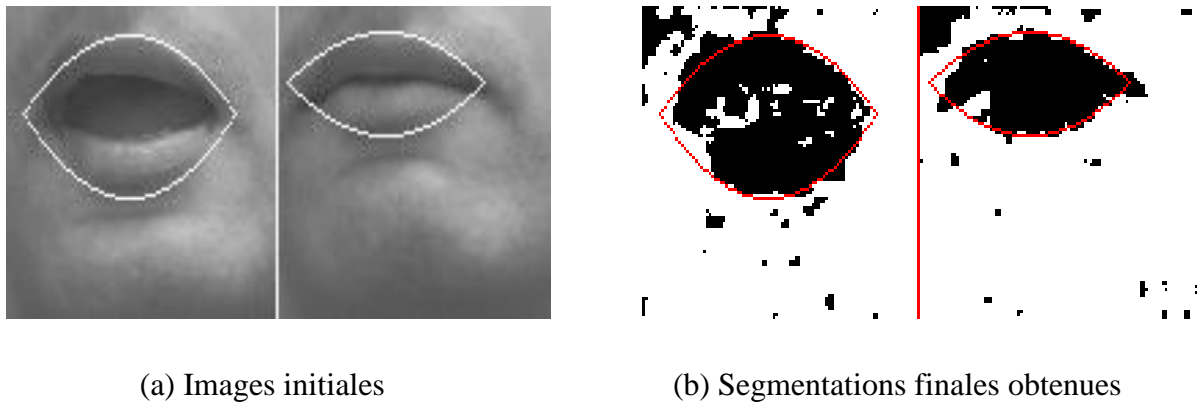
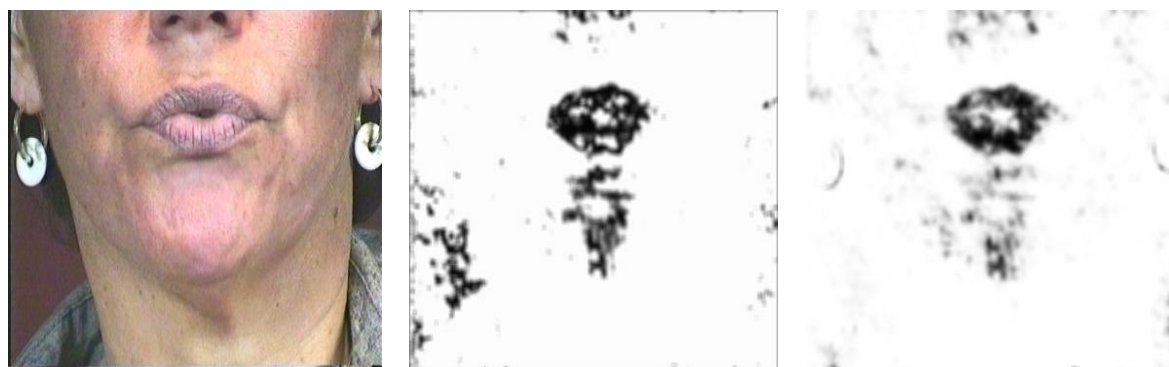


Figure. 2.9 – Détection des lèvres d'après (Rao and Mersereau 1995).

À partir de l'image composite C (voir figure 2.10c), le contour externe des lèvres est recherché en utilisant un modèle de forme spécifique au locuteur, la multi-résolution (des sous-échantillonnages successifs de l'image) et le gradient spatial. Revéret (1999), ainsi que (Nefian et al. 2002), utilisent également une image composite C. Les coefficients α , β et γ sont obtenus par analyse discriminante linéaire utilisant des images du visage et de la bouche segmentées manuellement. Une image binaire des lèvres est ensuite obtenue par seuillage et permet la détection du contour externe des lèvres.

(Wojdel and Rothkrantz 2001a; Wojdel and Rothkrantz 2001b) repèrent les lèvres en utilisant soit l'approche couleur proposée par (Coianiz et al. 1996), soit une approche statistique basée sur l'utilisation d'un réseau de neurones d'architecture très simple $R_{3,5,1}$. Les auteurs indiquent que dans certaines conditions, l'approche de Coianiz ne permet pas de segmenter efficacement les lèvres du reste de l'image et proposent deux alternatives. La première réside dans l'utilisation conjointe de la teinte filtrée et de l'intensité filtrée, dans les deux cas à l'aide d'un filtre parabolique qu'il est préférable d'adapter aux images à traiter. La position centrale (équivalent du paramètre H_0 de l'eq. 2.5) et la sélectivité du filtre (w) doivent alors être réglées et il faudra choisir comment utiliser conjointement les informations de teinte et d'intensité filtrées. Les auteurs proposent d'effectuer de manière automatique les réglages en demandant aux utilisateurs de leur système de désigner (à l'aide de la souris) leurs lèvres sur la première image acquise de leur visage. La seconde alternative réside dans l'utilisation de la zone marquée par l'utilisateur pour étiqueter l'image et entraîner un réseau de neurones à la tâche de classification entre les classes « lèvres » et « non-lèvres ». Le perceptron multicouches utilisé contient trois entrées pour les valeurs R, V et B de chaque pixel, une couche cachée de cinq nœuds et une sortie comprise dans l'intervalle $[0,1]$ indiquant si le pixel couleur en entrée appartient plutôt aux lèvres (valeurs proches de 0) ou au reste (valeurs

proches de 1). Les résultats de classification obtenus à l'aide du modèle neuronal sont, d'après les auteurs, légèrement supérieurs à ceux obtenus avec la teinte (qui est plus bruitée), comme l'illustre la figure 2.10.



(a) Image initiale

(b) Teinte filtrée

(c) Sortie du réseau de neurones

Figure 2.10 – Détection des lèvres d'après (Wojdel and Rothkrantz 2001a; Wojdel and Rothkrantz 2001b).

Enfin, (Luetin et al. 1996a; 1996b ; 1996c; 1996e; 1996f; Luetin and Thacker 1997) détectent précisément les contours interne et externe des lèvres à l'aide de modèles de la forme et de l'apparence des lèvres appris statistiquement à partir d'images étiquetées manuellement sur le corpus Tulips1 (Movellan 1995). Il utilise des images en niveaux de gris et extrait, à partir des contours matérialisés par des polygones, le profil en niveaux de gris perpendiculaire au contour, pour chacun des sommets de ses polygones. Les profils correspondants à tous les points de contour sont alors concaténés et les vecteurs globaux ainsi obtenus pour de nombreuses images, sont analysés par l'Analyse en Composantes Principales (ACP) pour obtenir l'apparence moyenne de la bouche ainsi que ses principales variations d'apparence. La localisation de la bouche se fait par minimisation du modèle de la forme et de l'apparence des lèvres. Signalons également que cette même approche est utilisée sur le corpus M2VTS (Pigeon and Vandendorpe 1997) dans (Luetin 1997a; 1997b; Luetin and Dupont 1998; 2000). Les images couleurs de ce corpus sont converties en niveaux de gris pour être utilisées.

2.3.2.3 Approche par corrélation avec des patrons

Nous avons rencontré une approche où, à l'instar des travaux de (Brunelli and Poggio 1993) qui repèrent différents éléments du visage en recherchant le point de meilleure mise en

correspondance d'imagettes de ces éléments sur l'image, la bouche était localisée de façon relativement précise par une approche « template matching ».

(Shdaifat et al. 2001) localisent directement la bouche sur une image présentant un visage complet avec un fond non-uniforme, en utilisant la corrélation entre une image de « bouche moyenne » et l'image à analyser. Dans un premier temps, les auteurs constituent par inspection visuelle, des classes des différentes formes de bouche susceptibles d'être rencontrées (visèmes). Puis des images représentatives de ces cinq visèmes sont moyennées pour obtenir une image de « bouche moyenne » utilisée pour localiser la bouche sur l'image. Les auteurs reconnaissent que des éléments du visage autres que la bouche peuvent être détectés à tort (yeux notamment) et proposent de raffiner la recherche en calculant la corrélation entre des imagettes des commissures droite et gauche de la bouche, du même locuteur, et les zones de l'image à analyser où le coefficient de corrélation dépasse un seuil. Les commissures sont ainsi localisées et leur position sert de référence pour normaliser l'image en rotation et en échelle. L'image de la zone de la bouche normalisée est finalement comparée aux images des cinq visèmes pour sa classification. Des expérimentations de cette méthode ont été effectuées pour quatre locuteurs, et les taux de classification correcte obtenus varient de façon très importante selon le locuteur et la généralisation de ces travaux mono-locuteur à un cadre multi-locuteurs ne nous semble pas évidente.

2.3.2.4 Approches mouvement

(Leroy and Herlin 1995; Leroy et al. 1996a), dont nous avons déjà évoqué les travaux dans la section sur la localisation de visage (section 2.3.1), propose d'utiliser le gradient spatiotemporel (voir figure 2.10), calculé sur une trentaine d'images, pour détecter la position de la bouche. Plus précisément, la bouche est définie dans l'approche de Leroy comme la zone de fort gradient spatio-temporel la plus basse située le long de la médiatrice du segment des yeux. Selon l'auteur, la localisation de la bouche n'est pas très précise et dépend du mouvement qu'elle a eu pendant la séquence d'images étudiées.

Broun et al. (2002) utilisent également la différence inter-images combinée à la couleur pour localiser la bouche d'un sujet en train de parler. Ils se distinguent de (Liévin and Luthon 1999), en utilisant l'accumulation des différences inter-images sur une séquence de 30 images. Les différences inter-images sont calculées pixel à pixel sur la composante rouge, puis elles sont sommées et seuillées pour obtenir une image binaire faisant ressortir les zones en mouvement. Cette observation de mouvement est combinée (opérateur ET), avec une image

obtenue à l'aide de seuils haut et bas de la teinte et de la saturation. L'image-produit obtenue fait ressortir les zones en mouvement dont la teinte et la saturation correspondent à celles des lèvres.

Enfin, signalons que (Mase 1991 ; Pentland and Mase 1989) effectuent un calcul de flot optique sur des images contenant les lèvres d'un locuteur. L'information de mouvement ne sert pas, dans ces travaux, à localiser les lèvres, mais bien à étudier leurs mouvements, ou plus exactement à mesurer le mouvement dans quatre fenêtres : les deux premières contiennent les moitiés haute et basse de la bouche, c'est-à-dire les lèvres supérieures et inférieure et les deux restantes les moitiés gauche et droite de la bouche. (Gray et al. 1997b) compare d'ailleurs cette approche par flot optique à d'autres approches dynamiques pour la reconnaissance de parole visuelle.

2.3.2.5 Autres approches

(Matthews et al. 1996a) évoque la possibilité de localiser la région des lèvres dans une image de visage en utilisant des transformations morphologiques simples, mais sans donner plus de détails. Une fois que l'on a localisé précisément les lèvres, il est possible d'extraire les informations visuelles. Dans la plupart des travaux que nous avons rencontrés, ces informations sont exclusivement labiales. Deux types bien distincts d'informations sont extraites des images: des informations de bas niveau extraites par des transformations des valeurs de niveaux de gris des pixels de l'image et des informations de haut niveau correspondant à des mesures obtenues à l'aide de modèles.

(Gray et al. 1997a) utilisent le corpus Tulips1 (Movellan 1995), qui contient 934 images en niveaux de gris. Chaque image est normalisée en translation, échelle et rotation (dans le plan image) grâce à l'étiquetage réalisé par (Luettin et al. 1996f), puis les parties gauche et droite de l'image sont rendues symétriques. Les images résultantes sont de résolution 87×65 et différentes stratégies de réduction de la dimension (5655) de ces vecteurs visuels sont étudiées : l'analyse en composantes principales en retenant les 50 premiers vecteurs propres (PCA 50), l'analyse en composantes indépendantes (ICA 50), ainsi que d'autres approches par PCA et ICA locales. Les résultats suggèrent que l'utilisation des approches locales est plus efficace que les approches globales (Gray et al. 1997a).

Matthews et al. (1996a) calculent à partir d'images de la zone des lèvres de 80×60 , obtenues en cadrant manuellement la bouche dans des images de visage complet de résolution 376×288 , la transformation morphologique « sieve ». Cette transformation crée des triplets

{échelle, amplitude, position} appelés granules. Les informations d'amplitude et de position ne peuvent être utilisées car elles rendraient le système dépendant des variations dans l'environnement dont il est souhaitable d'être indépendant. En revanche, l'information d'échelle est relativement robuste aux variations d'éclairément et peut être utilisée. Pour réduire la taille du vecteur d'observation, l'histogramme de l'information d'échelle est calculé. On obtient ainsi un vecteur de dimension 60 (hauteur de l'image en entrée). La dimension du vecteur est divisée par deux en moyennant deux à deux les coefficients successifs. L'image initiale est alors représentée par un vecteur de dimension 30 qui est utilisé directement ou après réduction à 10 coefficients par projection sur les 10 principaux axes obtenus par ACP. Dans (Harvey et al. 1997), la même approche est utilisée, mais le vecteur histogramme de dimension 60 est projeté directement sur les 20 principaux axes obtenus par ACP. D'autres variantes sont également testées dans cet article, mais les performances rapportées en terme de lecture labiale automatique sont nettement moins élevées.

Pour (Lee and Kim 2001), des images couleur de la région de la bouche de résolution 320×240 sont utilisées en début de traitement. Ces images sont sous-échantillonnées (160×120), puis converties en niveaux de gris. L'histogramme des images est normalisé, puis la zone la plus sombre est considérée comme étant l'intérieur de la bouche. Cette zone permet de calculer la largeur l de la bouche et d'obtenir la région d'intérêt (ROI) en utilisant $1; 1 * l$ comme largeur de ROI. Les auteurs ré-échantillonne la ROI pour obtenir une image de 64×64 qui est ensuite sous-échantillonnée à 16×16 pixels. Les auteurs utilisent une transformée en cosinus discret (DCT) puis une ACP sur ces images de 16×16 , ainsi que sur ces images symétrisées (8×16) et ont obtenu 80, 90 et 95% de la variance totale avec 7, 15 et 23 vecteurs propres au lieu de 9, 23 et 47 vecteurs propres sans symétrisation. Ceci les amène à conclure qu'il est intéressant d'utiliser la symétrie des lèvres car ceci permet même d'améliorer les scores de RAP AV en éliminant les problèmes d'illumination non uniforme.

Sur le corpus AT&T (Potamianos et al. 1997 ; Potamianos and Graf 1998a), effectue une transformée en ondelettes discrètes (DWT) de l'image de la zone de la bouche sous-échantillonnée sur 16×16 pixels. Quinze coefficients ainsi que leur dérivées et accélérations sont utilisés comme vecteurs visuel.

Dans (Potamianos et al. 2001a; Potamianos et al. 2001b; Neti et al. 2000), les auteurs calculent leurs vecteurs d'observation visuelle à partir d'images sur le corpus IBM Viavoice™ (Neti et al. 2000). La position de la bouche est estimée en suivant l'approche décrite dans (Senior 1999) (voir section 2.3.1), à partir d'images contenant le visage complet. La zone d'intérêt est extraite et sous-échantillonnée dans une image de 64×64 pixels.

Une DCT est appliquée à cette image et les 24 coefficients de plus forte énergie sont retenus pour former le vecteur visuel statique. Pour obtenir le vecteur d'observation visuelle final, une interpolation linéaire est utilisée pour modifier la cadence des vecteurs de 60 à 100 Hz, puis 15 vecteurs statiques consécutifs sont concaténés (7 avant + 7 après). Les vecteurs de dimension $15 \times 24 = 360$ sont réduits à 41 dimensions par projection après LDA+MLLT. Le vecteur visuel final est alors concaténé au vecteur acoustique de dimension 60 obtenu suivant un procédé similaire pour former l'observation audiovisuelle. Ce dernier vecteur (de dimension 101) subit également une réduction de dimension par LDA+MLLT, pour finalement atteindre 60 coefficients.

2.4 Conditions « naturelles » (écologiques)

Enfin, la dernière catégorie que nous allons évoquer est celle des systèmes qui ne supposent aucune préparation du locuteur et qui ne nécessitent pas non plus d'équipement ou de posture spécifique : l'acquisition des images est effectuée à l'aide d'une caméra qui filme le locuteur de face.

Ce sont les systèmes les plus « libres » du point de vue de l'utilisateur, mais ce sont également ceux pour lesquels l'extraction des paramètres labiaux est la plus problématique. Aux difficultés déjà rencontrées dans les systèmes sans préparation du locuteur, mais avec prise de vue ou dispositif d'acquisition particulier présentés dans la section précédente, viennent s'ajouter les problèmes de cadrage et d'éclairage : l'éclairage peut ne pas être optimal et le locuteur peut se déplacer pendant qu'il parle, ce qui peut également faire varier l'éclairage.

Les systèmes de ce type peuvent être utilisés dans des cadres applicatifs plus vastes que les systèmes présentés dans la partie précédente. Si de tels systèmes atteignaient un bon niveau de fiabilité, ils seraient même utilisables dans la plupart des situations, dans la mesure où la prise de vue de face est très largement répandue dans l'existant et relativement facile à obtenir pour de nouvelles applications. En télévision par exemple, la vue de face est utilisée pour les journaux télévisés, mais également pour d'autres types d'émission. Dans le cas d'indexation par le texte d'archives audiovisuelles ayant un canal acoustique dégradé, il serait envisageable d'employer un tel système de AVASR. Pour des applications comme la dictée vocale audiovisuelle ou l'interaction homme-machine audiovisuelle, la vue de face semble également un choix envisageable. Quant à la lecture labiale automatique à distance effectuée à l'insu du locuteur (espionnage) comme celle effectuée par l'ordinateur HAL du film de science

fiction de Kubrick « 2001, l'odyssée de l'espace » (Kubrick 1968) (voir également (Stork 1997)), il est fort peu vraisemblable que l'on atteigne ce niveau de performance avant de très nombreuses années (s'il est possible de les atteindre un jour). En effet, même dans des conditions favorables, le canal visuel porte une information moindre que le canal acoustique et une application de lecture labiale grand vocabulaire n'est pas à l'ordre du jour. De plus, pour un tel type d'application, il sera difficile d'obtenir une image d'une résolution suffisante pour être utilisée, car certains mouvements labiaux ont une amplitude de l'ordre de quelques millimètres comme l'indique (Lallouache 1991) en précisant que les systèmes d'extraction de paramètres doivent fournir des mesures dont la précision doit être de l'ordre du demi-millimètre !

Comme pour tous les systèmes évoqués précédemment, il faut pouvoir gérer la grande variabilité intra-locuteur d'apparence et de forme de la bouche pendant la production de parole, mais la tâche d'extraction de paramètres devient largement plus complexe qu'avec les autres systèmes utilisant l'image du locuteur, car le gradient spatial entre les lèvres et la peau peut être quasiment inexistant, en particulier pour la lèvre inférieure¹⁷. Si l'on n'emploie pas des méthodes robustes, la détection de ce contour risque d'être très hasardeuse. Si l'éclairage n'est pas constant, l'intensité moyenne de l'image variera. Ceci peut se corriger pour partie en effectuant une normalisation comme le propose (Vanegas et al. 1998), mais si l'éclairage n'est pas uniforme ou s'il y a des ombres portées, la normalisation globale risque de ne pas être satisfaisante et il faudra s'orienter vers des techniques plus sophistiquées comme celles proposées par (Gouet and Montesinos 2002 ; Pinel et al. 2001), ou enfin par (Basso et al. 2001). Si le locuteur est mobile, de possibles problèmes de cadrage pourront se poser : ceci pourra amener à cadrer une zone plus large du visage du locuteur et ajoutera potentiellement des minima locaux (nez, fond) dans les recherches de contours. Si de plus, l'éclairage arrive du dessus, il est vraisemblable que des ombres portées apparaissent (sous le nez et la bouche), ce qui peut réduire le gradient spatial entre la lèvre inférieure et la peau, et augmenter encore la difficulté de localisation du contour externe de la lèvre inférieure. Dans le cas le plus défavorable, éclairage artificiel du dessus et éclairage externe variable avec un locuteur mobile, des conditions qui sont pourtant celles de nombreux postes de travail, toutes les sources d'erreurs s'ajoutent et il faudra des modèles très robustes pour extraire les paramètres labiaux avec une qualité suffisante pour qu'ils soient utilisables pour l'AVASR. Il n'y a pas à notre connaissance de systèmes qui aient été évalués dans des conditions aussi défavorables. En pratique, les différents systèmes qui ont été présentés dans ce chapitre ont été bâtis ou testés à partir de corpus et il n'y a pas de corpus enregistré dans ces conditions. Le seul corpus

qui corresponde à une lumière variable est, à notre connaissance, celui que nous avons enregistré pour les besoins de nos recherches en utilisant la lumière solaire ambiante, mais l'éclairage y est diffus et il n'y a d'ombres très marquées.

L'évaluation de chaque système étant dépendante de son corpus de test, il nous semble utile de présenter rapidement les corpus de parole audiovisuelle existants.

2.5 Comparaison image-modèle

Les deux approches "modèle" et "image" ont toutes les deux des avantages et des inconvénients. En dépit des différences évidentes entre ces deux approches, une caractéristique qu'elles partagent toutes les deux est le besoin éventuel d'une intervention manuelle. En effet, on peut intervenir manuellement pour étiqueter des données ou définir une région d'intérêt (d'habitude c'est la région de lèvres). Cependant, l'utilisation de l'une ou l'autre dépend globalement de la difficulté de la méthode, de sa robustesse et de la pertinence de la paramétrisation visuelle résultante.

Par ailleurs, il existe dans la littérature peu d'études comparant les deux approches. Nous présentons ci-dessous trois études les comparant :

(Brunelli and Poggio 1993) comparent les performances obtenues par deux techniques automatiques pour la reconnaissance du visage, à partir d'images prises en vue frontale. La première technique, qu'on peut qualifier d'approche "image", s'appuie sur le calcul d'un ensemble de paramètres géométriques à partir de l'image du visage. La seconde technique est fondée sur une adaptation d'un modèle du visage sur l'image réelle (Template Matching). La comparaison entre ces deux techniques nous semble intéressante même si l'objet à traiter dans l'étude était le visage et non pas seulement la bouche. Elle peut nous livrer certains aspects utiles pour fonder des arguments sur l'utilisation de ces techniques. Les auteurs ont obtenu, en terme de reconnaissance, des performances supérieures en utilisant la seconde technique ("template matching").

(Matthews et al. 1998) comparent deux techniques différentes pour caractériser les formes de la bouche pour la reconnaissance visuelle de la parole (lecture labiale automatique). La première technique extrait les paramètres requis pour adapter un modèle actif de forme (Active Shape Model, ASM) aux contours des lèvres. La seconde utilise des paramètres dérivés d'une analyse spatiale multi-échelle (Multiscale Spatiale Analysis, MSA) de la région de la bouche. Les résultats semblent avantager l'analyse spatiale multi-échelle. Ils montrent que cette technique est plus robuste, rapide et plus précise. En effet, dans les tests de

reconnaissance avec des locuteurs multiples et utilisant seulement les données visuelles, la précision de reconnaissance des lettres est de 45% pour la méthode MSA et de 19% pour ASM. Pour reconnaître des digits, la précision est la même pour les deux méthodes (77%). Cette performance relativement faible de l'ASM peut être expliquée par l'incorporation de connaissances a priori dans la méthode qui peuvent être inexactes. Le fait de représenter le contour des lèvres par un modèle simple semble être aussi trop limité pour diffuser des informations plus précises. En général, l'ASM est confronté comme toutes les techniques de l'approche "modèle" à des erreurs de modélisation et de capture.

Matthews et al. (2001) comparent, dans une tâche de reconnaissance audio-visuelle continue à large vocabulaire, quatre techniques différentes de paramétrisation visuelle. Trois de ces techniques appartiennent à l'approche "image". Il s'agit de la transformée en cosinus discrète (DCT), la transformée en ondelettes discrète (DWT) et l'analyse en composante principale (ACP). Ces trois méthodes nécessitent de localiser la région de la bouche. La quatrième technique, utilisant l'approche modèle active d'apparence (AAM), tente de modéliser le visage entier par un modèle déformable de l'apparence du visage et inclut un algorithme de capture. Il est évident a priori qu'utiliser le visage entier devrait être bénéfique. Le visage entier peut inclure des caractéristiques visuelles supplémentaires qui pourraient être utiles et bénéfiques à la reconnaissance. Toutefois, les résultats obtenus dans un test de reconnaissance visuelle de mots semblent contredire cette évidence. Les résultats expérimentaux montrent que les performances des méthodes de l'approche "image" sont meilleures (en taux d'erreurs : autour de 59% pour les trois méthodes "image" vs. 64% pour l'AAM). La méthode AAM est probablement désavantagée par les problèmes que rencontrent toute méthode de l'approche "modèle", à savoir les erreurs d'apprentissage du modèle.

En résumé, ces quelques comparaisons donnent un petit avantage à l'approche "image". Ceci dit, comme nous l'avons évoqué précédemment, l'approche "modèle" dépend beaucoup des algorithmes employés pour l'apprentissage du modèle. Une amélioration de ces algorithmes et l'incorporation de connaissances a priori qui rendent mieux compte de la structure de déformation de l'objet considéré, augmentera probablement la robustesse de cette approche.

2.6 Corpus existants

Un corpus est un ensemble de données qui doivent être représentatives de « l'objet » scientifique à étudier. De façon générale, un tel ensemble de données peut servir à tester et valider (ou invalider !) des modèles (a priori ou a posteriori) ou à les adapter pour qu'ils

fonctionnent sur une « vérité terrain ». Dans le cas des modèles statistiques a posteriori, appris à partir de données, le corpus sert également à construire les modèles et il est alors très nettement préférable de scinder le corpus en une portion servant à l'entraînement, le corpus d'apprentissage, et une autre, disjointe, servant à l'évaluation que l'on nommera corpus de test. L'une des principales difficultés matérielles auxquelles les chercheurs en parole audiovisuelle sont confrontés est alors la taille des corpus. Notons également que plus le corpus d'apprentissage sera représentatif du problème à résoudre, plus les performances des modèles entraînés avec devraient être élevées dans des conditions réelles. Il semble alors important de limiter les contraintes imposées au locuteur et sur le contrôle de l'éclairage pour enregistrer des corpus dans des conditions que nous qualifierons par la suite de « naturelles ».

2.7 Conclusion

Nous avons rappelé dans ce chapitre, que l'information visuelle est d'un bénéfice important dans le domaine de la reconnaissance audio-visuelle de la parole. Elle est un vecteur d'information nécessaire et essentiel dans la compréhension, même partielle, de la parole chez les personnes sourdes. Elle porte une partie complémentaire de l'information de parole perçue par les utilisateurs de ce code. La présentation des informations visuelles doit être optimale pour une reconnaissance maximale des gestes visuels. En d'autres termes, dans quelles conditions de présentation et de visibilité du visage, un système de reconnaissance peut-il percevoir (reconnaître) un maximum d'information de parole ?

Le chapitre suivant est d'ailleurs consacré à la description du signal de parole et nous présenterons les différents problèmes posés lors de son traitement, ainsi les principales méthodes d'analyse du signal de parole pour extraire les paramètres acoustiques qui seront fournis au système de reconnaissance.

De la reconnaissance acoustique à la reconnaissance bimodale de parole 3

Le son est un élément majeur permettant à l'être humain d'appréhender son environnement. Il est également, par le biais de la parole, le vecteur naturel de la communication humaine. Présent dans de nombreux documents multimédias, il est, de ce fait, porteur d'une information précieuse pour leur compréhension.

Le problème de la reconnaissance de la parole est un domaine d'études actif depuis le début des années 50. Actuellement les modèles les plus utilisés en reconnaissance de la parole sont les modèles de Markov cachés (HMM) et les réseaux de neurones.

La reconnaissance automatique de la parole peut être basée directement sur une comparaison de formes nouvelles avec des références des mots à reconnaître, ou bien sur l'identification d'un ensemble d'unités élémentaires (phonèmes, diphones, syllabes). Dans le premier cas, il s'agit d'une reconnaissance dite globale (approche retenue dans ce travail), dans le second cas d'une reconnaissance dite analytique.

Dans ce chapitre, nous donnons une définition rapide de la parole. Nous présentons ensuite les grands principes de la reconnaissance automatique de la parole, avant de nous intéresser aux méthodes bimodale de la RAP.

3.1 Définition de la parole

La parole est le mode de communication privilégié pour l'espèce humaine. Il est la représentation sonore d'un langage et est produit par le système vocal.

La parole, comme représentation d'un langage, est constituée d'unités linguistiques, les mots. Pour décrire la représentation sonore de ces unités linguistiques, on utilise des phonèmes. Un phonème peut être défini comme la plus petite unité sonore distinctive que l'on peut obtenir par segmentation de la parole. Pour produire un phonème, le système vocal adapte sa configuration : débit de l'air, tension des cordes vocales et forme du conduit vocal. Les phonèmes sont classifiés en trois familles :

- les voyelles sont produites par les vibrations des cordes vocales. Ce sont des sons qui sont souvent considérés comme quasi-périodiques et pour une configuration quasi

statique du conduit vocal. Elles peuvent être nasales ou orales selon que l'air passe par la cavité nasale ou la cavité buccale ;

- les consonnes sont elles produites par occlusion totale (consonnes occlusives) ou partielle (consonnes fricatives, latérales ou vibrantes) du conduit vocal. Elles peuvent être non voisées — il n'y alors pas de vibration des cordes vocales et le son est essentiellement produit par un bruit (bruit de friction, d'explosion ou de relâchement) — ou au contraire voisées — elles sont alors produites aussi par vibration des cordes vocales. Les consonnes sont habituellement considérées comme des transitions rapides entre deux voyelles, avec donc une géométrie du conduit vocal qui varie rapidement. On peut donc dire que la caractérisation essentielle des consonnes c'est la nature du son, dans leur cas, un son de type « bruit » ou contenant un bruit ;
- les semi-voyelles ont des sons de type voyelle — vibration des cordes vocales et sans bruit — mais générés pendant une évolution rapide de la géométrie du conduit vocal. Leur son ne peut donc pas être considéré comme quasi-statique.

3.2 Le signal de la parole

Le signal de la parole n'est pas un signal ordinaire. Il est le vecteur d'un phénomène complexe : la communication parlée. La reconnaissance de la parole pose de nombreux problèmes aux chercheurs depuis 1950 (Allegre 2003). D'un point de vue mathématiques, il est difficile de modéliser le signal de parole, compte tenu de sa variabilité. Nous allons ici tenter de mettre en évidence quelques caractéristiques importantes du signal non stationnaire afin de faire ressortir les problèmes posés lors de son traitement (Haton 2006).

3.2.1 Redondance du signal

Le signal de parole est extrêmement redondant. Cette grande redondance lui confère une robustesse à certains types de bruits. De nombreuses recherches sont menées afin de rendre les systèmes de reconnaissance robustes aux bruits, mais les performances humaines sont encore loin d'être atteintes.

3.2.2 Variabilité du signal

Le signal de parole possède une très grande variabilité. Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution peut varier, la durée du

signal est alors modifiée. Toute altération de l'appareil phonatoire peut modifier la qualité de l'émission (exemple : rhume, fatigue...). De plus, la diction évolue dans le temps. La voix est modifiée au cours des étapes de la vie d'un être humain (enfance, adolescence, âge adulte...).

La variabilité interlocuteur est encore plus accentuée. La hauteur de la voix, l'intonation et l'accent diffèrent selon le sexe, l'origine sociale, régionale ou nationale. Un exemple pertinent de cette variabilité apparaît lorsque nous comparons la voix d'un locuteur originaire du Nord avec celle d'un locuteur originaire du sud de l'Algérie. Enfin, la parole est un moyen de communication où de nombreux éléments entrent en jeu, tels que le lieu, l'émotion du locuteur, la relation qui s'établit entre les locuteurs (stressante ou amicale). Ces facteurs influencent la forme et le contenu du message. L'acoustique du lieu (milieu protégé ou environnement bruyant), la qualité du microphone, les bruits de bouche, les hésitations, les mots hors vocabulaire sont autant d'interférences supplémentaires sur le signal de parole.

3.2.3 Les effets de coarticulation

La production parfaite d'un son suppose un positionnement précis des organes phonatoires. Le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis subissent alors l'influence de ceux qui les précèdent ou les suivent. Ces effets de coarticulation est un facteur de variabilité supplémentaire important du signal de parole.

3.3 Extraction des paramètres

Dans un système de RAP, les paramètres acoustiques permettant de décrire le signal de parole sont généralement définis sur une échelle d'information de niveau local. Le signal continu de parole est fourni en entrée du système de RAP après une conversion sous la forme d'échantillons sonores. Une suite de vecteurs représentatifs, appelés vecteurs acoustiques ou vecteurs d'observation, est alors retournée en sortie du module de paramétrisation acoustique.

Les paramètres acoustiques définis pour la représentation acoustique du signal de parole devraient respecter les critères de (Deviren 2004):

- pertinence. Les paramètres acoustiques doivent représenter de manière précise le signal de parole. Leur nombre doit cependant rester limité afin de conserver un coût de calcul raisonnable lors de leur exploitation dans les modules de calcul des paramètres acoustiques et de reconnaissance des formes.

- discrimination. Les paramètres acoustiques doivent représenter de manière caractéristique les différents éléments représentatifs des unités linguistiques afin de les rendre facilement distinctes.
- robustesse. Les paramètres acoustiques doivent résister aux effets perturbateurs liés aux distorsions du signal de parole émis (Milner and Darch 2011).

Dans le processus de traitement du signal acoustique d'un système de RAP, un découpage du signal de parole analysé retourne une séquence de segments d'échantillons sonores appelés trames. La durée de ces trames est choisie de telle sorte que le signal de parole est considéré stationnaire (Boite et al. 2000). Cette segmentation permet alors d'extraire les propriétés locales du signal de parole. Le continuum de parole est donc représenté par une suite de vecteurs d'observation calculés sur des trames du signal de courte durée par exemple de l'ordre de 20 ms, par fenêtre glissante asynchrone ou synchrone au pitch (Young et al. 2006). Les vecteurs d'observation peuvent représenter le signal de parole sous la forme de différents types de coefficients qui constituent les paramètres acoustiques.

Ces paramètres sont choisis pour être le plus utile à la représentation du signal de parole dans l'objectif de décrire le message linguistique. Se basant sur l'analyse des caractéristiques physiologiques de l'oreille (Dallos 1973), de nombreux types de paramètres acoustiques sont utilisés dans la littérature pour la RAP (Davis and Melmerstein 1980; Eyben et al. 2010).

Parmi les principaux types de paramètres exploités dans les systèmes de RAP, on peut distinguer :

3.3.1 Énergie du signal

Après la phase de numérisation et surtout de quantification, le paramètre intuitif pour caractériser le signal ainsi obtenu est l'énergie. Cette énergie correspond à la puissance du signal. Elle est souvent évaluée sur plusieurs trames de signal successives pour pouvoir mettre en évidence des variations. La formule de calcul de ce paramètre est :

$$E(\text{fenêtre}) = \sum_{n \in \text{fenêtre}} |n|^2 \quad (3.1)$$

Il existe des variantes de ce calcul. L'une des plus utilisées réalise une simple somme des valeurs absolues des amplitudes des échantillons pour alléger la charge de calcul, les variations restant les mêmes. D'autres, comme celle de (Taboada et al. 1994) proposent la modification suivante du calcul intégrant une normalisation par rapport au bruit ambiant.

$$E(\text{fen\^etre}) = \log \left(\sum_{n \in \text{fen\^etre}} \frac{|n|^2}{R} \right) \quad (3.2)$$

Dans cette \^equation, R est la valeur moyenne de l'\^energie du bruit. Le r^esultat de ce calcul tend vers 0 lorsque la portion consid^er^ee est une zone o\^u il n'y a que le bruit de fond. Tout le probl^eme de cette variante r^eside dans l'estimation du facteur de normalisation R .

3.3.2 Coefficients MFCC

Le principe de calcul des MFCC (Mel-scaled Frequency Cepstral Coefficients) est issu des recherches psychoacoustiques sur la tonie et la perception des diff^erentes bandes de fr^equences par l'oreille humaine.

Un vecteur acoustique MFCC est form^e de coefficients cepstraux obtenus \^a partir d'une r^epartition fr^equentielle selon l'\^echelle de Mel (Bogert et al. 1963) (voir figure 3.1). L'utilisation d'\^echelles de fr^equences non-lin^eaires, telles les \^echelles de Mel (Stevens et al. 1937) ou Bark (Zwicker 1961), permettent une meilleure repr^esentation des basses fr^equences qui contiennent l'essentiel de l'information linguistique pour la majeure partie du signal de parole. La correspondance entre les valeurs de fr^equences en Hertz F_{Hertz} et en Mel F_{Mel} est calcul^ee par (O'Shaughnessy 1987) :

$$F_{mel} = 2.595 \cdot \log \left(1 + \frac{F_{Hertz}}{700} \right) \quad (3.3)$$

Par ailleurs, il est possible de calculer des coefficients cepstraux \^a partir d'une r^epartition fr^equentielle lin^eaire sans utiliser une \^echelle de Mel mais en conservant la r^epartition lin^eaire des \^echelles de fr^equences. Ces coefficients sont alors appel^es LFCCs (*Linear Frequency Cepstral Coefficients*) (Rabiner and Juang 1993).

Afin de s^eparer la source spectrale de la r^eponse fr^equentielle, l'op^eration de m^ethode cepstrale se base sur la propri^ete du logarithme qui permet de transformer un produit en addition. Une transform^ee discr^ete en cosinus (*Discret Cosinus Transform*, DCT) permet ainsi d'obtenir les N coefficients cepstraux d^esir^es (Ahmed et al. 1974). Consid^erant f la fonction de transformation spectrale, le k^{me} coefficient cepstral $C(k)$ est donc obtenu par :

$$C(k) = \sqrt{\frac{2}{N}} \sum_{i=1}^N f(i) \cdot \cos \left(\frac{\pi k}{N} (i - 0.5) \right) \quad (3.4)$$

Cette analyse a pour avantages un nombre réduit de coefficients par vecteur acoustique et un faible indice de corrélation entre ces différents coefficients. Les coefficients MFCCs sont réputés plus robustes que ceux issus d'une analyse spectrale (Lockwood et al. 1992).

Les coefficients de type MFCC sont souvent associés à la valeur d'énergie contenue dans la trame de signal de parole appelée sous le terme de coefficient $C(0)$ (Young et al. 2006). De surcroît, l'utilisation des dérivées premières et secondes de ces coefficients fournit de l'information utile sur la dynamique du signal de parole. En effet, l'information complémentaire apportée par le filtrage temporel introduit par les dérivées des coefficients MFCCs permet une plus grande robustesse des paramètres acoustiques dans les systèmes de RAP face à l'usage des seuls coefficients MFCCs statiques (Yang et al. 2007). Dans ces conditions, ces paramètres acoustiques prennent souvent la forme de vecteurs de 39 coefficients formés par les 12 premiers coefficients MFCCs, l'énergie $C(0)$ (et leurs dérivées premières et secondes).

Cette information complémentaire apporte toutefois un complément utile dans la classification de certaines consonnes (Liu et al. 1997). Par ailleurs, il est possible de ré-synthétiser un message intelligible sur de la parole propre à partir d'une analyse des seuls coefficients MFCCs, c'est-à-dire à partir des spectres et cepstres en échelle de Mel (Demuyck et al. 2004). Donc dans le cas de parole propre, un signal d'excitation basé sur une analyse du pitch est utilisé pour cette opération de re-synthèse (Collen et al. 2007). Dans ce cas, l'information initiale de phase n'est alors pas utile. Par contre, dans le cas d'un signal de parole bruitée, les informations de phase et de résolution spectrale fine sont très utiles pour la bonne reconnaissance des composantes du message linguistique (Murty and Yegnanarayana 2006).

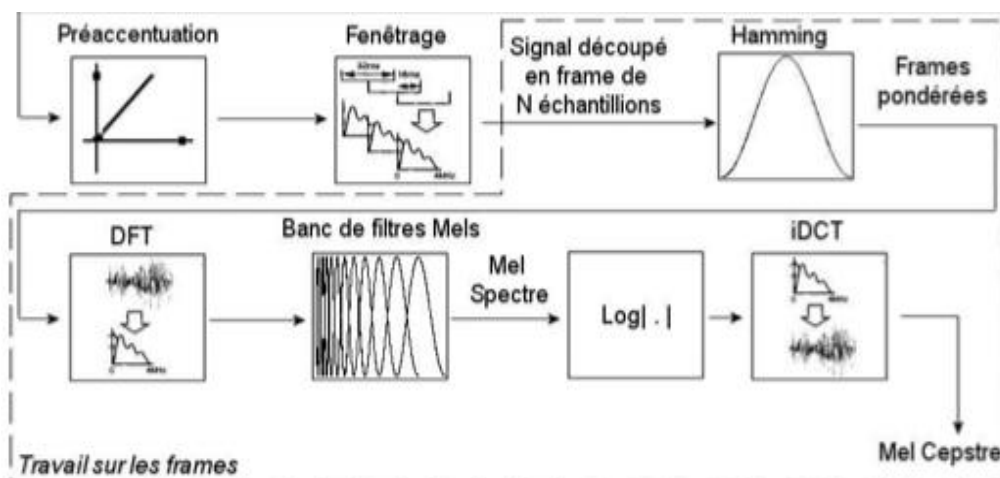


Figure 3.1 – Schéma de calcul des MFCC.

3.3.3 Taux de passage par zéro

Le taux de passage par zéro (*zero crossing rate* en anglais) représente le nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la valeur centrale de l'amplitude (généralement zéro). Il est fréquemment employé pour des algorithmes de détection de section voisée/non voisée dans un signal. En effet, du fait de sa nature aléatoire, le bruit possède généralement un taux de passage par zéro supérieur à celui des parties voisées.

Le comptage du nombre de passages par zéro est très simple à effectuer. Dans un premier temps, il faut enlever le décalage d'amplitude (*offset* en anglais), produit par la majorité des matériels d'acquisition, pour centrer le signal autour de zéro. Ensuite, pour chaque trame, il suffit de dénombrer tous les changements de signe du signal. Pour éliminer certains phénomènes parasites, (Taboada et al. 94) ont proposé une méthode nommée le *band-crossing*. Un seuil d'amplitude S permet de définir une zone autour du zéro de largeur $2xS$ au sein de laquelle les oscillations ne sont pas prises en compte. La formule du *band-crossing* pour chaque fenêtre analysée est donc :

$$bcr(\text{fenêtre}) = \sum_{n \in \text{fenêtre}} |f(n) - f(n-1)| \text{ avec } f(n) = \begin{cases} 1 & \text{si } n > S \\ f(n-1) & \text{si } -S \leq n \leq S \\ -1 & \text{si } n < -S \end{cases} \quad (3.5)$$

Cette mesure se montre très intéressante, dans le cadre d'une détection de parole en amont d'un système de reconnaissance, pour la détection de fricative en fin de signal à reconnaître ou d'attaque de plosive.

3.3.4 Autres paramétrisations du signal

Nous n'énumérerons pas tous les types de paramètres employés dans le domaine de la recherche en parole car il y en a énormément et ce n'est pas le propos de notre thèse. Pourtant, il est à noter que d'autres approches plus proches de l'audition humaine, telles les modèles d'oreille, ont été étudiées. De plus, le lecteur trouvera des informations sur différents paramètres très largement utilisés pour le codage LPC (*Linear Predictive Coding*) présent dans la norme GSM, pour les PLPs (*Perceptual Linear Predictive*) et pour les RASTA-PLP, version approfondie des PLP (Laprie 2000). Cette liste ne se veut pas exhaustive mais permet d'avoir un aperçu des différents paramètres qu'il est possible d'extraire d'un signal de parole.

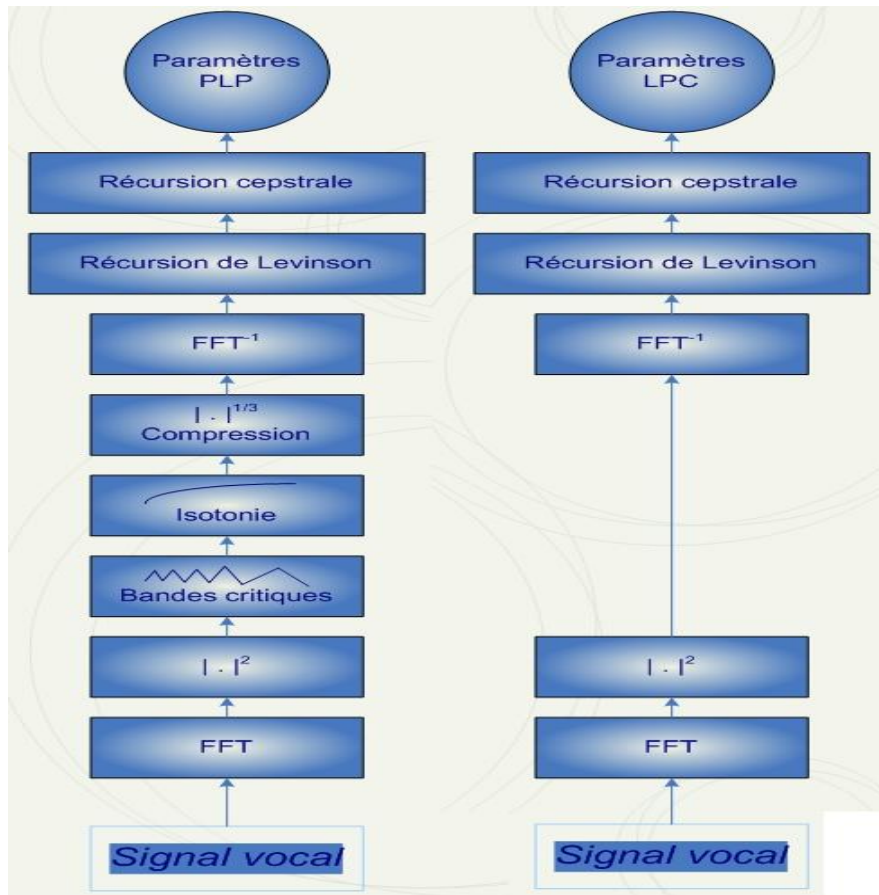


Figure 3.2 – Schémas de calcul les paramètres PLP et LPC.

3.3.5 Dérivées première et seconde

Le but final de l'extraction des paramètres est de modéliser la parole, un phénomène très variable. Par exemple, même si elle a de l'importance, la simple valeur de l'énergie n'est pas suffisante pour donner toute l'information portée par ce paramètre. Il est donc souvent nécessaire de recourir à des informations sur l'évolution dans le temps de ces paramètres. Pour cela, les dérivées première et seconde sont calculées pour représenter la variation ainsi que l'accélération de chacun des paramètres. Même si la robustesse de la représentation obtenue est accrue, cela implique aussi de multiplier par 3 l'espace de représentation.

3.4 Réduction de l'espace de représentation

Comme nous venons de le voir, l'espace de représentation du signal est souvent de taille conséquente, généralement de plusieurs dizaines de paramètres. Il est donc important de ne garder que des paramètres discriminants. La méthode majoritairement utilisée, de nos jours, est l'analyse discriminante linéaire, LDA pour *Linear Discriminant Analysis* en anglais. Cette

technique s'apparente à l'analyse en composantes principales (ACP). Elle permet l'obtention de paramètres considérés comme discriminants en appliquant une transformation linéaire de l'espace d'entrée de taille n vers un espace de taille réduite q ($q < n$). L'application de cet algorithme maximise la séparation des classes qui sont affectées à chaque vecteur acoustique et ainsi améliore la robustesse de la représentation. Ils ont d'ailleurs montré que l'utilisation d'une telle analyse permet de pallier certaines catégories de bruits.

3.5 Les modes de fonctionnement d'un système de reconnaissance

Un système de reconnaissance peut être utilisé sous plusieurs modes (Hlaoui 1999):

- **Dépendant du locuteur (monolocuteur)**

Dans ce cas particulier, le système de reconnaissance est configuré pour un locuteur spécifique. C'est le cas de la plupart des systèmes de reconnaissance de parole disponibles sur le marché. Les principaux systèmes de dictée vocale actuels possèdent une phase d'apprentissage recommandée avant toute utilisation (voire même une adaptation continue des paramètres au cours de l'utilisation du logiciel) afin d'effectuer une adaptation des paramètres à la voix de l'utilisateur.

- **Pluri-locuteur (ou multi-locuteur)**

Le système de reconnaissance est élaboré pour un groupe restreint de personnes. Le passage d'un locuteur à un autre du même groupe se fait sans adaptation.

- **Indépendant du locuteur**

Tout locuteur peut utiliser le système de reconnaissance.

- **Elocution**

Le mode d'élocution caractérise la façon dont on peut parler au système. Il existe quatre modes d'élocution distincts :

- **Mots isolés :**

Chaque mot doit être prononcé isolément, c'est à dire précédé et suivi d'une pause.

- **Mots connectés :**

Le système reconnaît des séquences de quelques mots sans pause volontaire pour les séparer (exemple : reconnaissance de chiffres connectés ou de nombres quelconques...).

- **Parole continue lue :**

C'est le discours usuel, si ce n'est que les textes sont lus.

- **Parole continue spontanée :**

C'est le discours usuel, sans aucune contrainte.

La reconnaissance de mots isolés fonctionne relativement bien de nos jours pour différentes langues. De bons résultats ont été publiés par de nombreux laboratoires. Généralement, de tels outils de reconnaissance de parole sont utilisés pour un vocabulaire de commande correspondant à des actions spécifiques et simples (gestion de menus...).

Le premier mode d'élocution sera abordé lors de cette étude. Les expériences décrites dans ce travail ont été effectuées sur de la parole bruitée.

3.6 La reconnaissance bimodale de la parole

Afin de rendre les interfaces en parole naturelle plus fiables, une solution est d'augmenter les modalités pouvant être perçues par la machine en « ouvrant les yeux aux machines ». Se pose alors le problème d'intégrer des informations de nature différente : acoustique et visuelle. C'est précisément cette intégration d'informations hétérogènes, acoustiques et visuelles, en vue de leur exploitation pour la RAP.

Nous abordons dans cette partie l'intégration audiovisuelle selon le point de vue de la théorie de l'information. Ensuite nous expérimentons quelques modèles d'intégration selon que celle-ci intervient dans le système de RAP au niveau numérique par identification directe ou bien au niveau symbolique après identification séparée ou encore au niveau numérique et symbolique selon un schéma hybride ID+IS. Les traitements acoustiques et visuels utilisés dans les systèmes développés selon ces trois stratégies sont également décrites.

Dans les systèmes audiovisuels de RAP, il s'agit d'interpréter des images en plus des signaux de parole usuels pour identifier un message oral. Cette interprétation doit exploiter les points de vue acoustique et visuel pour produire des résultats de reconnaissance plus performants et plus fiables. Ces points de vue peuvent se situer aussi bien au niveau des

données que des leurs traitements. L'intégration de ces points de vue bimodaux suit différents modèles sans couvrir cependant de manière complète les modes d'interaction formulés précédemment.

3.6.1 Les modèles d'intégration audio-visuelle de la parole

Nous avons vu précédemment comment la parole peut être considérée comme bimodale. De nombreuses études ont été menées pour rendre compte de la manière avec laquelle interagissent les deux modalités audition et vision pour la compréhension de la parole. Ces études menées tant par des psychologues, linguistes que par des ingénieurs, s'étendent sur plusieurs domaines allant de la cognition, aux sciences de l'ingénieur en passant par la neurophysiologie.

Ainsi, plusieurs modèles ont été proposés. Mentionnons par exemple, le célèbre modèle Fuzzy-Logical Model of Perception (FLMP) proposé par (Massaro 1987, 1998). Les premiers travaux se concentraient spécialement sur les architectures de fusion en considérant arbitrairement des représentations internes monomodales (représentation visuelle seule et auditive seule). Sur ces représentations, les différents travaux consistaient à appliquer un certain nombre de calculs afin de prédire la performance bimodale.

Dans ces études, le traitement de la représentation des informations des modalités est souvent négligé. Schwartz et al. (1998); Schwartz (2002), en croisant des modèles issus de la psycho-physique et de la fusion des capteurs, ont classé les modèles d'intégration audiovisuelle en quatre grandes architectures : (i) modèle à « Identification Directe » noté ID; (ii) modèle à « Identification Séparée » noté IS ; (iii) modèle à « Recodage dans la modalité Dominante » noté RD; et (iv) modèle à « Recodage commun des deux modalités sensorielles vers la modalité Motrice » noté RM.

Pour simplifier la compréhension du système d'intégration audio-visuelle dans la perception de la parole, nous pouvons le considérer comme une boîte qui a en entrée deux flux de nature différente (vision et audio) et en sortie une décision ou un code qui peuvent être de nature phonétique ou lexicale. Le schéma de la figure 3.3 illustre un tel système.

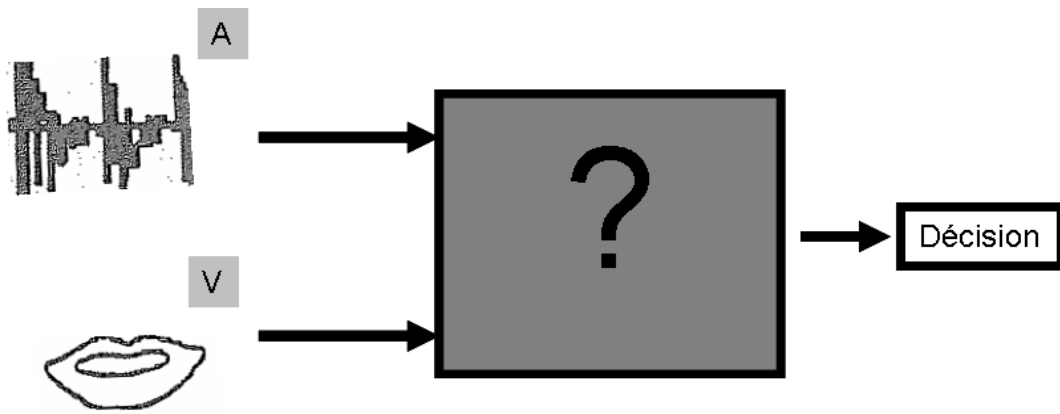


Figure 3.3 – Le noyau d'un processus d'intégration audio-visuelle dans la perception de la parole (d'après Schwartz et al. (1998)).

Dans la suite, nous survolerons rapidement les 4 architectures classiques de l'intégration audio-visuelle. En plus de les définir, nous donnerons des exemples réalisés pour chacune de ces architectures.

3.6.1.1 Modèle ID

Dans ce modèle, appelé aussi modèle données-vers-décision, les deux sources d'information sont injectées directement dans un classifieur bimodal qui effectue le traitement de l'information des deux modalités (figure 3.4). La classification se fait donc directement sans aucun niveau intermédiaire de mise en forme commune des données. Le classifieur prend une décision dans l'espace des caractéristiques bimodales, dans lequel des prototypes bimodaux ou des règles de décision bimodales ont été appris. Ce modèle est une extension du modèle « Lexical Access From Spectra » (LAFS) de Klatt (1979) vers « Lexical Access From Spectra and Face Parameters ».

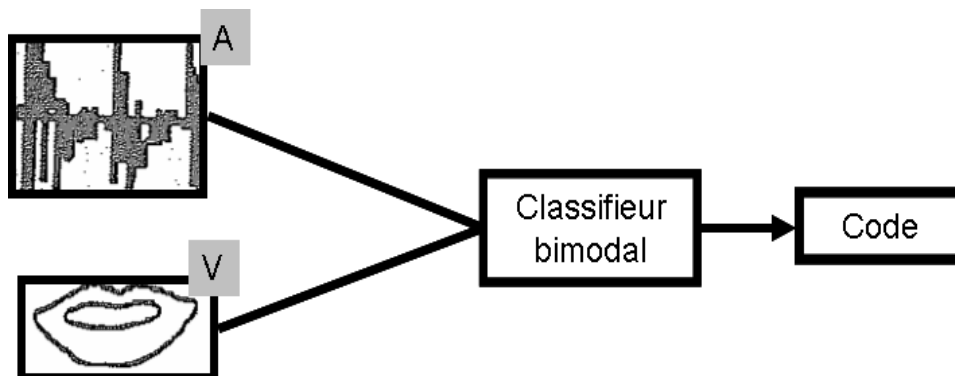


Figure 3.4 – Modèle à identification directe.

Benoît et al. (1996) ont implémenté le modèle d'identification directe pour la reconnaissance audio-visuelle et ont évalué les performances pour une grande plage de rapport signal sur bruit. Ils injectent un vecteur d'observation audiovisuel dans un processus de reconnaissance s'appuyant sur les chaînes de Markov Cachées (HMM). Le vecteur audiovisuel est obtenu en concaténant des paramètres acoustiques issus d'une analyse acoustique à six paramètres géométriques des lèvres et leur dérivée. Dans une structure semblable, l'implémentation de Teissier et al. (1999) du modèle ID implique un classifieur Gaussien dans un espace de six dimensions. Le vecteur d'entrée bimodal de ce classifieur est composé de six paramètres : trois paramètres acoustiques issus d'une analyse en Composantes Principales (ACP) et trois paramètres géométriques du contour interne des lèvres. Dans cette implémentation, un paramètre supplémentaire est ajouté dans le processus de fusion. Les deux flux d'entrée audio et vidéo sont pondérés. Ceci permet ainsi de contrôler les poids respectifs de chaque entrée conformément à leur efficacité pour la décision.

Potamianos et al. (2001c) ont proposé une technique de fusion des flux visuel et auditif en appliquant deux transformées l'une après l'autre. Ils utilisent tout d'abord une Analyse Discriminante Linéaire (ADL, en anglais LDA pour Linear Discriminant Analysis) pour réduire de façon discriminante les dimensions du vecteur concaténé des caractéristiques audiovisuelles. Puis, une Transformée Linéaire de Maximum de Vraisemblance (TLMV, en anglais MLLT pour Maximum Likelihood Linear Transform) est appliquée pour améliorer la modélisation des données.

Ces deux transformées sont aussi utilisées pour prendre en compte l'information dynamique dans les flux des données audio-visuelles avant la fusion. Les auteurs réalisent ainsi un schéma hiérarchique d'intégration audio-visuelle.

3.6.1.2 Modèle IS

Le modèle d'identification séparée (IS) est fondé sur ce que les psychologues cognitifs appellent « intégration tardive » du fait que l'intégration vient après la classification phonétique dans chaque voie sensorielle séparée par opposition au modèle ID qui est une intégration « précoce » car s'appliquant directement aux données. Dans le modèle IS, les informations visuelles et auditives sont traitées séparément chacune par un classifieur. Puis, la fusion des résultats des deux classifieurs dans un module d'intégration permet la reconnaissance du code (voir figure 3.5).

Le modèle IS est aussi appelé décision-vers-décision en référence à la caractéristique de base de la fusion qui est une fusion de décisions. Dans ce type de modèle, la fusion peut être

réalisée soit sur des valeurs logiques, à l’instar du modèle VPAM (Vision-Place, Audition-Manner) dans lequel chaque modalité est en charge d’un groupe spécifique de caractéristiques phonétiques (distinctives), soit par un processus probabiliste, comme dans le cas du modèle FLMP de Massaro (Massaro 1987, 1998).

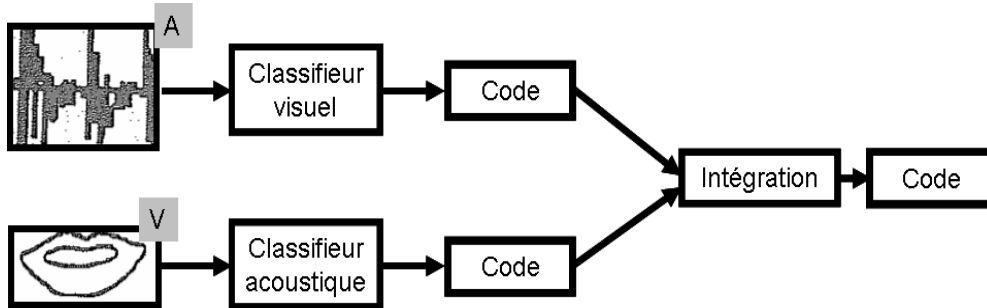


Figure 3.5 – Modèle à identification séparée.

Adjoudani et Benoît (1995) ont aussi implémenté le modèle IS dans leur système de reconnaissance audiovisuelle. Ils ont utilisé deux réseaux HMM acoustique et visuel séparés. Dans cette implémentation, chaque modèle HMM est entraîné avec des données visuelles ou acoustiques.

Les deux classifieurs fonctionnent ainsi indépendamment l’un de l’autre. En test, les vecteurs d’observations visuels ou acoustiques sont présentés séparément à l’entrée de chaque modalité. Les auteurs présentent ensuite trois méthodes pour le module d’intégration. La première, utilisée également dans d’autres études de reconnaissance de la parole audiovisuelle (Movellan and Chadderdon 1996), consiste à calculer le maximum des produits des probabilités conjointes des deux modalités. En d’autres termes, l’intégration s’appuie sur une sélection, pour chaque entité à reconnaître (phonème, syllabe, mot ...), d’un candidat qui maximise la vraisemblance dans les deux canaux. Le schéma synoptique de la figure 3.6 résume le processus d’intégration suivant ce principe.

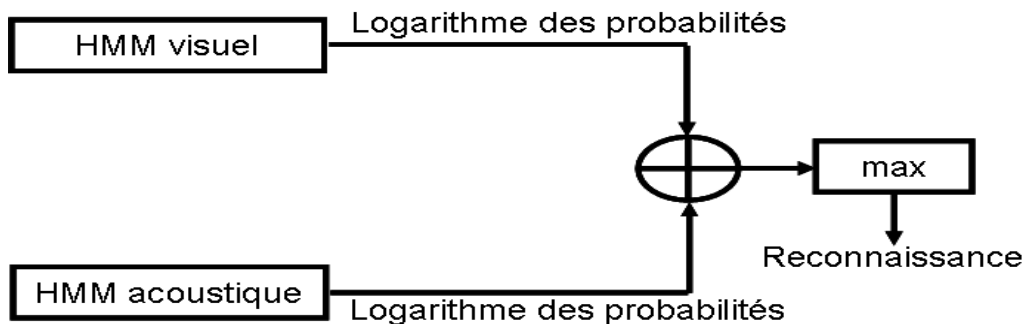


Figure 3.6 – Modèle d’intégration basé sur la maximisation des produits des probabilités conjointes (D’après Adjoudani (1998)).

La seconde méthode repose sur une sélection du meilleur candidat d'une des deux modalités acoustique ou visuelle selon son degré de certitude (ou confiance). Ce dernier est évalué à partir des probabilités de sortie de chaque modèle HMM et sert à commander un « interrupteur » qui sélectionne la voie ayant une plus grande certitude dans sa sélection. Le principe de cette méthode ne permet pas de fusionner les données provenant des deux canaux. De ce fait, cette méthode ne peut être considérée comme une architecture d'intégration. La figure 3.7 illustre le principe de cette dernière.

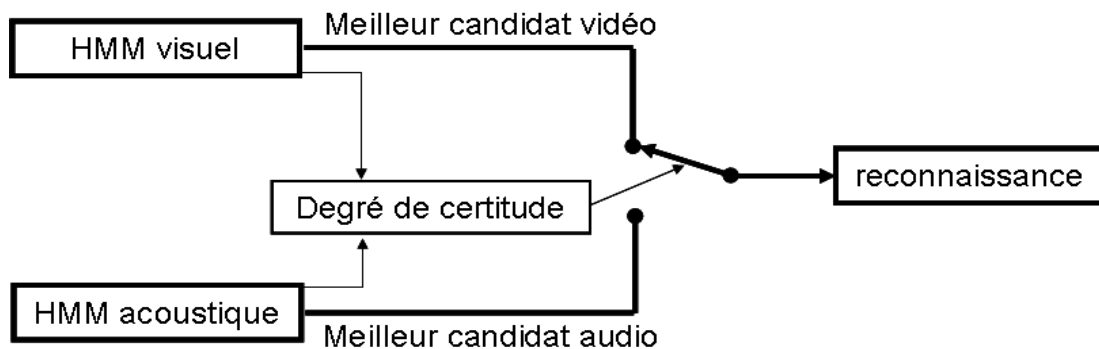


Figure 3.7 – Méthode de sélection du meilleur candidat acoustique ou visuel (D'après Adjoudani (1998)).

La troisième méthode consiste à intégrer les informations auditives et visuelles suivant une pondération de chaque modalité en fonction de l'indice de confiance (voir figure 3.6). Le principe de cette méthode est identique au principe de la première sauf qu'ici les probabilités sont pondérées. D'abord, un indice est estimé de la même façon que dans la seconde méthode, c'est-à-dire à partir des probabilités de sortie de chaque voie. Le résultat de cette estimation définit ensuite le coefficient normalisé de pondération. Puis, en maximisant le produit des probabilités pondérées, un candidat est sélectionné.

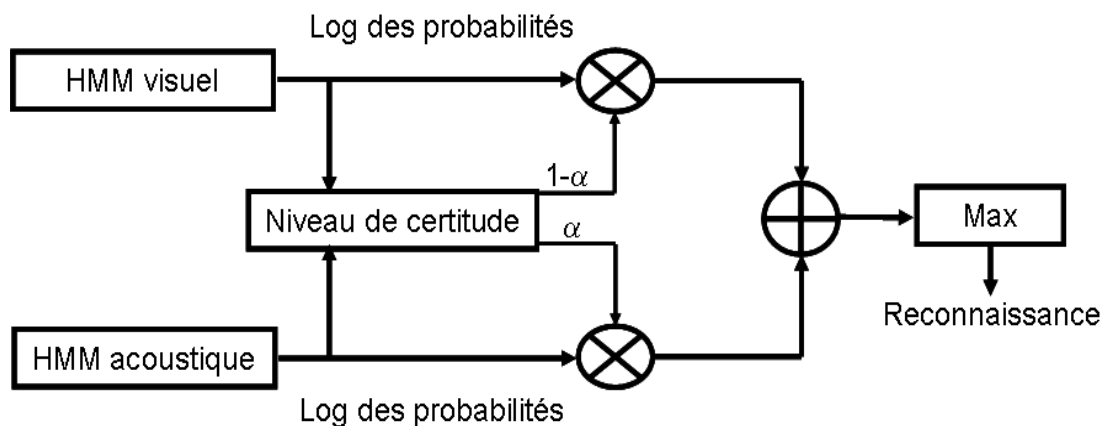


Figure 3.8 – Architecture d'intégration audiovisuelle par pondération (D'après Adjoudani (1998)).

3.6.1.3 Modèle RD

Dans ce type de modèle, les informations visuelles sont codées dans un format compatible avec les représentations de la modalité auditive qui est considérée comme la modalité dominante.

Un tel format peut être la fonction de transfert du conduit vocal. Cette fonction de transfert est estimée séparément par un module de traitement du signal et par les indices visuels à partir des deux entrées auditive et visuelle. L'estimation de la fonction de transfert peut être effectuée par exemple par association à partir de l'entrée visuelle et par un traitement cepstral à partir de l'entrée auditive. Les deux estimations sont ensuite fusionnées et l'ensemble ainsi obtenu est présenté à un classifieur phonétique (voir figure 3.9). Il s'agit d'une fusion précoce.

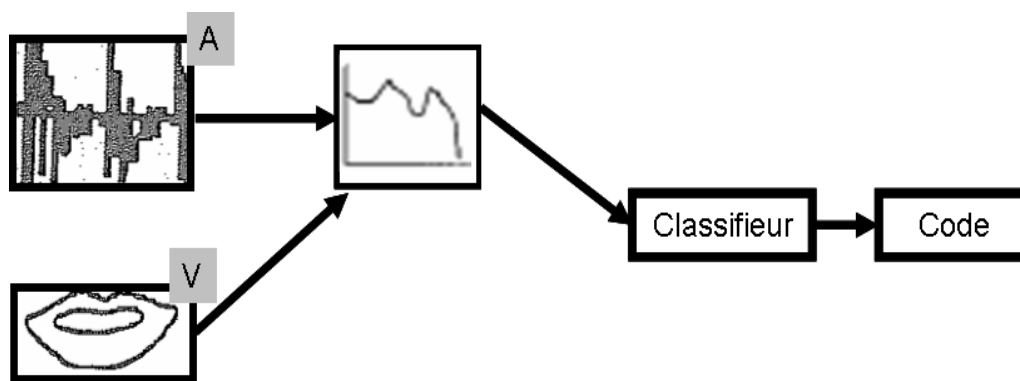


Figure 3.9 – Modèle à recodage dans la modalité dominante.

Le recodage des informations visuelles dans l'espace de la modalité acoustique (en un spectre acoustique) est fait grâce à un réseau de neurones. Le spectre estimé à partir des caractéristiques visuelles est combiné avec le spectre provenant de l'analyse acoustique pour finalement obtenir le spectre audiovisuel. La combinaison des deux spectres est réalisée en pondérant chaque entrée par un poids variant suivant le niveau de bruit de l'audio. Le spectre audiovisuel résultant alimente ensuite un deuxième réseau de neurones pour enfin identifier la voyelle produite. Cette implémentation a été adaptée par Robert-Ribes et al. (1996) aux voyelles du Français avec quelques différences. En effet, le classifieur audiovisuel employé par Robert-Ribes et al. (1996) est un classifieur gaussien tandis que le recodage de la modalité visuelle en une représentation auditive est réalisé par association utilisant des distances euclidiennes.

3.6.1.4 Modèle RM

Ce modèle est inspiré en partie de la théorie motrice de la perception de la parole proposée par Liberman et Mattingly (1985). Selon cette théorie, l'information phonétique est perçue par un module spécialisé dans la détection des gestes planifiés par le locuteur qui sont le fondement des catégories phonétiques. Dans ce type d'architecture, les deux entrées sont codées dans une nouvelle représentation commune dans l'espace moteur avant d'être classifiées. Dans ce modèle, le choix de l'espace moteur est crucial pour l'intégration. En général, les paramètres du conduit vocal sont les plus choisis comme représentation commune. Dans ce cas, à partir de chaque entrée, visuelle ou acoustique, les principales caractéristiques articulatoires sont estimées. Ensuite, la représentation finale est définie en additionnant les deux projections avec une certaine pondération et elle est fournie au classifieur pour la reconnaissance du code (voir figure 3.10).

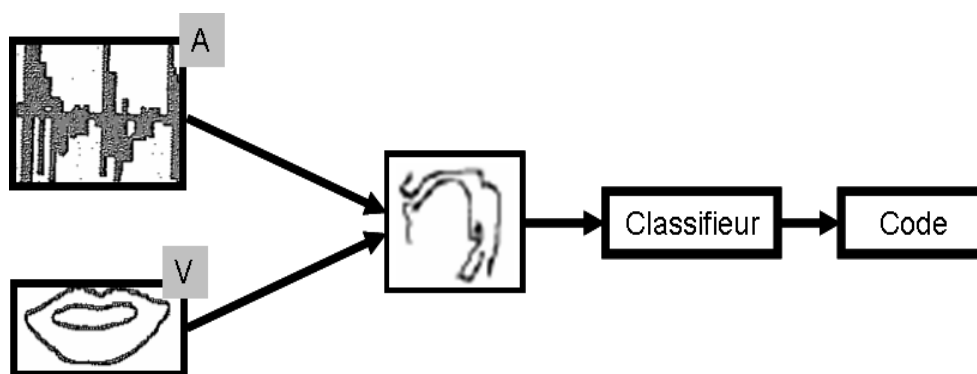


Figure 3.10 – Modèle à recodage dans la modalité motrice.

A notre connaissance, seuls Teissier et al. (1999) et Robert-Ribes et al. (1996) ont proposé une implémentation de ce type de modèle. Dans l'implémentation de Teissier et al. (1999), qui a pour objectif la reconnaissance de voyelles du Français, la transformation des deux entrées en représentation motrice est réalisée par des associations linéaires. Les auteurs ont choisi comme espace moteur des caractéristiques articulatoires représentées par trois paramètres qui fournissent les corrélats articulatoires des dimensions d'arrondissement, d'ouverture-fermeture et d'avant-arrière : les coordonnées horizontale et verticale, respectivement X et Y, du point le plus haut de la langue et l'étirement, noté A, du contour interne des lèvres. Le réglage des associateurs est obtenu en définissant ces trois paramètres pour chaque voyelle d'un corpus d'apprentissage.

Le paramètre A est mesuré directement sur l'entrée visuelle. Par contre, les auteurs ont utilisé comme coordonnées X et Y des valeurs prototypiques provenant d'un expert phonétique. La classification est ensuite réalisée de la même façon que pour le modèle RD, c'est-à-dire avec un classifieur Gaussien.

3.6.2 Eléments du choix d'une architecture : théoriques et expérimentaux

Dans une tâche de fusion de deux modalités, un des principaux problèmes réside dans le choix du modèle d'intégration le plus approprié. Suivant la perspective envisagée, modélisation des processus cognitifs ou reconnaissance de la parole, le modèle retenu doit rendre compte au mieux des données au niveau reconnaissance automatique. Dans ce sens, Robert-Ribès (1995) propose une taxinomie mettant en correspondance les 4 modèles d'intégration décrits précédemment avec les modèles généraux de la psychologie cognitive (figure 3.11). Cette taxinomie s'organise autour de 3 questions :

1. Peut-on considérer, en fonction de l'interaction entre les modalités, une représentation intermédiaire commune? Sinon, c'est un modèle ID à préconiser.
2. Dans le cas de l'existence d'une représentation intermédiaire, l'intégration est-elle tardive ou précoce pour accéder au code? Une intégration est tardive quand elle suit l'intervention d'un processus de décodage ; c'est-à-dire qu'il y'a d'abord extraction des informations auditives et visuelles, puis fusion (c'est le cas du modèle IS). Dans le cas où la fusion intervient au cœur du processus d'extraction de l'information, l'intégration est dite précoce.
3. Si l'intégration est précoce, quelle forme prend le flux commun des données après fusion? Plus précisément, existe-t-il une modalité dominante susceptible de fournir la représentation intermédiaire commune dans une architecture à intégration précoce (cas du modèle RD)? ou cette représentation est elle amodale (cas du modèle RM) ?

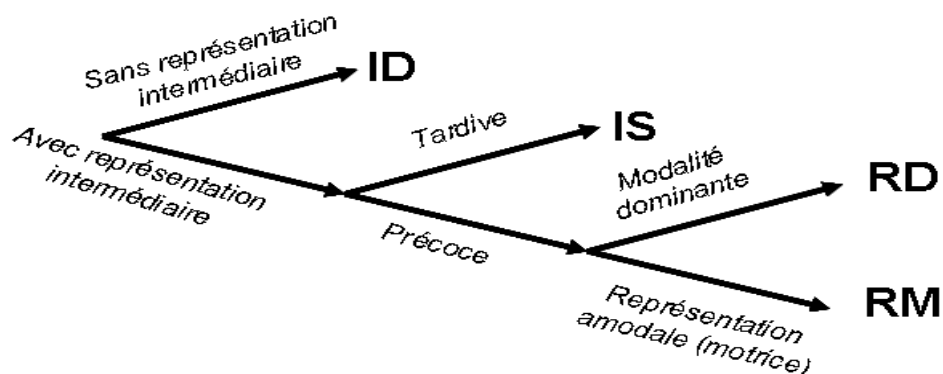


Figure 3.11 – Taxinomie des modèles d'intégration (d'après Robert-Ribès (1995)).

Parmi les 4 architectures, les modèles ID et IS sont ceux qui sont les plus fréquemment utilisés en reconnaissance de parole (Schwartz 2004). Les deux autres modèles sont très rarement implémentés et ceci malgré le fait qu'ils semblent être les plus pertinents au regard des données issues de la psychologie expérimentale. C'est précisément ces données qui ont conduit Schwartz et al. (1998) à privilégier le modèle RM.

3.6.3 Etudes comparatives

Dans cette sous-section nous passons en revue quelques études comparant les quatre architectures d'intégration.

3.6.3.1 ID vs. IS

Adjoudani (1998) rapporte plusieurs études menées dans le domaine de la reconnaissance audiovisuelle de la parole, parmi lesquelles Robert-Ribès (1995); Movellan et Chadderdon (1996), comparant les deux modèles IS et ID. Il conclut que la grande partie de ces études semblent avantager le modèle IS (Duchnowski et al. 1995; Robert-Ribès et al. 1996 ; Silsbee et Su 1996) tout en notant le statut quo entre ces deux modèles relevé dans d'autres études (Jourlin 1996 ; Silsbee et Su 1996). L'auteur a aussi procédé, en tenant compte des résultats de ces études comparatives, à un regroupement des avantages (\oplus) et des inconvénients (\ominus) de chacun de ces deux modèles.

Modèle ID

- \oplus Modèle facile à implémenter: l'observation bimodale peut se former à partir d'une concaténation des indices des deux modalités.
- \oplus Possibilité de pondérer chaque canal à condition de disposer d'un corpus d'apprentissage de taille importante (Silsbee et Su 1996).
- \ominus Modèle nécessitant un corpus de taille relativement grande par rapport au modèle IS (Jacob et Sénac 1996) car la taille des modèles à apprendre est plus importante.
- \ominus Nécessité d'une topologie identique des deux sources.
- \ominus Conservation de la coordination temporelle entre les deux modalités durant la fusion.
- \ominus Le problème de déphasage n'est pas géré.
- \ominus Apprentissage adapté à chaque niveau du Rapport Signal sur Bruit (RSB) de l'entrée acoustique (Silsbee et Su 1996).

Modèle IS

- ⊕ Nécessité d'un corpus moins important pour l'apprentissage que pour le modèle ID grâce au traitement séparé de chaque modalité.
- ⊕ Les deux modalités ne demandent pas forcément d'avoir la même architecture de reconnaissance.
- ⊕ Le modèle s'approche plus des hypothèses faites sur la perception audiovisuelle (Robert-Ribès 1995; Massaro 1996).
- ⊕ Capable de traiter l'asynchronie: par exemple dans le cadre d'un mot entre son état initial et final.
- ⊖ Le module d'intégration peut être complexe et dépendant du corpus.

Après avoir comparé les modèles IS et ID, Adjoudani (1998) a implémenté, comme nous l'avons vu précédemment dans la section précédente, ces deux modèles et en a comparé les performances dans une tâche de reconnaissance audiovisuelle de la parole avec un niveau de bruit variant sur l'entrée auditive. Les résultats obtenus montrent que malgré que le modèle ID améliore significativement les scores de reconnaissance quand l'entrée acoustique est bruitée (on passe de 3% en reconnaissance acoustique à 33% en audiovisuelle pour la condition d'un RSB acoustique de -6 dB), l'intégration reste encore non optimale. Par contre, avec une pondération de chaque canal par son degré de confiance, le modèle IS peut donner des résultats meilleurs.

Enfin, l'auteur conclut que la complémentarité audio/ vision est mieux exploitée en IS et ceci grâce au traitement séparé des deux modalités, même si dans ce cas la coordination audiovisuelle semble perdue mais peut être retrouvée à certains points d'ancrage. Inversement, le modèle ID exploite bien les covariations des entrées visuelle et auditive mais dans le cas où l'entrée auditive est bruitée la complémentarité entre l'entrée propre et l'entrée atténuée n'est pas aussi prise en compte à cause du traitement conjoint des deux sources.

3.6.3.2 RD vs. RM

Comme ces deux modèles sont peu utilisés dans la reconnaissance audiovisuelle de la parole, les comparaisons sont rares pour déterminer le plus performant des deux. Il est important de rappeler que la différence entre ces deux modèles est la nature de leur représentation commune au niveau de la fusion. Le modèle RD appliqué à la fusion en parole considère la modalité auditive comme dominante alors qu'elle peut ne pas l'être. De ce fait, la

complémentarité naturelle entre le son et l'image est difficilement exploitable dans ce modèle. Robert-Ribès (1995), l'un des rares à implémenter les modèles RD et RM, démontre que le modèle RM est mieux adapté que le modèle RD à la structure de l'information audiovisuelle et à la complémentarité audio-visuelle.

3.7 Conclusion

Ce chapitre qui porte un aperçu sur la reconnaissance automatique de la parole, a permis de dégager les caractéristiques du signal et l'identification de ses paramètres en vue de leur utilisation en reconnaissance vocale. Divers modes de fonctionnement ont été évoqué dans ce chapitre tel que le mode monolocuteur et le mode multilocuteur.

Dans ce chapitre, nous avons également décrit un ensemble de modèles d'intégration audiovisuelle. Cette intégration peut être réalisée avec quatre modèles basiques : ID, IS, RD et RM. Ces derniers peuvent être classifiés en deux grandes familles. La première famille, fusion de représentations, regroupe les modèles s'appuyant sur l'entraînement d'un seul classifieur appliqué sur un vecteur des représentations audio et visuelles concaténées, ou sur toute transformation sur ce vecteur (modèles ID, RM, RD). La seconde famille, fusion de décisions, regroupe des modèles reposant sur une fusion des sorties de deux classifieurs monomodal. A ces deux familles, une troisième famille, fusion hybride, peut être considérée, qui consiste à combiner deux modèles des deux familles précédentes. La comparaison entre les quatre modèles classiques semble plutôt favoriser les modèles ID et IS. Cependant, ces derniers ne peuvent être départagés.

Dans notre travail, nous nous intéressons à la reconnaissance de la parole arabe en utilisant les et les modèles de Markov cachés de type gauche-droit. Pour pallier les insuffisances des paradigmes utilisés dans le système proposé. Nous avons combiné les avantages des HMM et les algorithmes génétiques pour aboutir à un modèle hybride GA/HMM qui offre plus de performances que les paradigmes classiques.

Dans le chapitre qui suit, nous exposons le fonctionnement des méthodes mentionnées précédemment ainsi leurs modèle hybride proposé.

Deuxième partie : Approches proposées

Moteur de reconnaissance GA/HMM

4

Les modèles de Markov cachés (HMM) sont des outils statistiques permettant de modéliser des phénomènes stochastiques. Ces modèles sont utilisés dans de nombreux domaines (Cappé 2001) tels que la reconnaissance et la synthèse de la parole, la biologie, l'ordonnancement, l'indexation de documents, la reconnaissance d'images, la prédiction de séries temporelles, ... Pour pouvoir utiliser ces modèles efficacement, il est nécessaire d'en connaître les principes.

L'amélioration de l'apprentissage des HMM à l'aide de métaheuristique à base de population est l'objet de ce chapitre. Ce chapitre a donc pour objectif d'établir les principes, les notations utiles et les principaux algorithmes qui constituent la théorie des HMM.

A cet effet, nous commençons ce chapitre en définissant de que sont les HMM leur principes, et nous présentons les algorithmes classiques des HMM : *Forward*, *Backward* et de *Viterbi*.

4.1 Modèles de Markov Cachés

4.1.1 Définition

Un modèle HMM est défini comme un ensemble d'états, chacun d'entre eux associé à une distribution de probabilité (en général multidimensionnelle). Les transitions entre les états sont régies par un ensemble de probabilités appelées probabilités de transition. Dans un état particulier, un résultat ou observation peut être généré conformément à la distribution de probabilité associée. Par opposition à un modèle de Markov classique où l'état est directement observable par un observateur externe, dans un modèle HMM, l'état n'est pas directement observable et seulement des variables influencées par l'état le sont. Les états sont donc cachés, d'où le nom de modèle de Markov caché.

Un HMM (représenté dans la figure 4.1) est défini par :

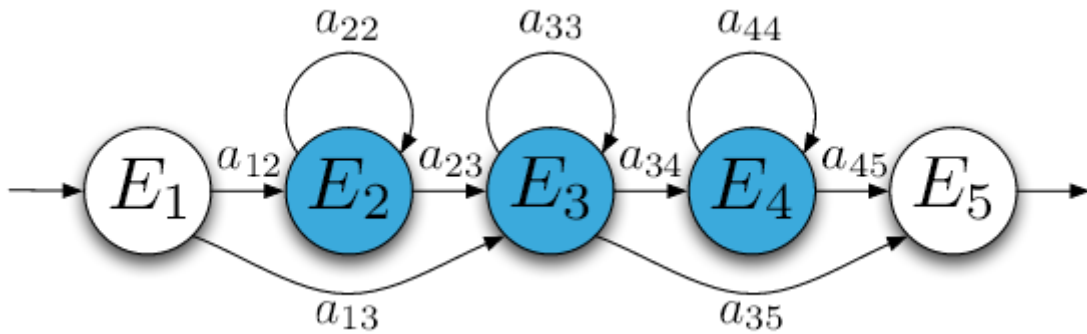


Figure 4.1 – HMM à 5 états dont 3 émetteurs.

- N : le nombre d'états du modèle. Les états seront notés x_i pour $1 \leq i \leq N$
- M : le nombre de symboles d'observation. Dans le cas où les observations sont continues, M est infini. Dans notre notation, les symboles d'observation de l'alphabet sont notés $Y = \{y_j\}$ pour $1 \leq j \leq M$.
- π : le vecteur de probabilités initiales des états. Concernant cet élément, un autre type de HMM utilise des états start et end et non une distribution d'états initiaux. Ce type d'HMM est notamment employé en bioinformatique.
- A : la matrice de transition où sont définies les probabilités de transition entre les états. Ces probabilités $A = \{a_{ij}\}$ sont définies comme :

$$a_{ij} = p(x_t = i | x_{t-1} = j), 1 \leq i, j \leq N \quad (4.1)$$

avec x_t désigne l'état courant à l'instant t . Les probabilités de transition a_{ij} doivent satisfaire les contraintes stochastiques :

$$a_{ij} \geq 0 \text{ et } \sum_{j=1}^N a_{ij}, 1 \leq i, j \leq N \quad (4.2)$$

- B : la matrice de confusion (ou matrice d'observation) contenant les probabilités d'observation (ou probabilités d'émission) $B = \{b_j(k)\}$ associées aux états. Ces probabilités sont définies comme :

$$b_j(k) = p(y_t = v_k | x_t = j), 1 \leq j \leq N, 1 \leq k \leq M \quad (4.3)$$

avec v_k dénote le $k^{\text{ème}}$ symbole d'observation dans l'alphabet, et y_t le vecteur de paramètres actuel (ou simplement observation actuelle) à l'instant t . Les probabilités d'observation satisfont aussi les contraintes stochastiques. Dans le cas d'observations continues, des densités de probabilités continues sont à utiliser.

Pour dénoter un modèle HMM le triplet $\lambda = (\pi, A, B)$ est généralement utilisé. Il est important de noter que chaque probabilité dans la matrice de transition (de confusion) est

indépendante du temps. En d'autres termes, les matrices ne changent pas dans le temps quand le système évolue. En pratique, ceci est l'une des suppositions les plus discutables des modèles de Markov à propos des processus réels.

Dans la théorie des HMMs, des hypothèses sont faites pour une docibilité mathématique et informatique :

- Hypothèse markovienne : concernant la définition des éléments de la matrice de transition A , la probabilité de transition vers un état ne dépend que de l'état actuel et non des états rencontrés précédemment. Ainsi, la séquence des états constitue une chaîne de Markov simple.
- Hypothèse de stationnarité : comme nous l'avons déjà évoqué, la matrice des probabilités de transition est indépendante de l'actuel temps, dans lequel les transitions prennent place.

Mathématiquement :

$$p(x_{t_1+1} = j | x_{t_1} = i) = p(x_{t_2+1} = j | x_{t_2} = i) \text{ pour tout } t_1 \text{ et } t_2, \quad (4.4)$$

- Hypothèse d'indépendance des sorties (observations) : l'observation courante est statiquement indépendante des observations précédentes. Mathématiquement, cette hypothèse peut être formulée pour un HMM λ par :

$$p(Y|x_1, x_2, \dots, x_t, \lambda) = \prod_{t=1}^T p(y_t | x_t, \lambda). \quad (4.5)$$

4.1.2 Utilisation et algorithmes

Une fois qu'un système est décrit comme un HMM, trois problèmes doivent être résolus. Les deux premiers sont des problèmes qu'on peut associer à la reconnaissance : détermination de la probabilité d'une séquence observée étant donné un HMM (c'est le problème de l'évaluation); et, étant donné un modèle HMM et une séquence d'observations, déterminer quelle séquence d'états cachés dans le modèle est la plus probable (c'est le problème de décodage). Le troisième problème est la génération d'un HMM étant donné une séquence d'observations (c'est le problème d'apprentissage).

4.1.2.1 Evaluation et l'algorithme de Forward

Ce problème se pose notamment quand nous avons, par exemple, plusieurs HMMs décrivant différents systèmes, et une séquence d'observations. Nous voulons ainsi connaître

quel est le HMM ayant la plus forte probabilité d'avoir généré cette séquence. En d'autres termes, pour un modèle $\lambda = (\pi, A, B)$ et une séquence d'observations $Y = y_1, y_2, \dots, y_T$, nous avons à calculer la probabilité $P(Y|\lambda)$. Un calcul de cette probabilité implique un nombre d'opérations de l'ordre de N^T . Heureusement, une autre méthode, ayant une complexité inférieure, existe. Cette méthode utilise une variable intermédiaire appelée variable "avant" ou forward; d'où le nom de l'algorithme Forward (ou "avant").

Algorithme Forward : Cet algorithme est utilisé pour calculer la probabilité d'une séquence d'observation de longueur T :

$$Y = y_1, y_2, \dots, y_T \quad (4.6)$$

avec chaque y est un élément de l'ensemble observable. La variable intermédiaire $\alpha_t(i)$ est définie comme la probabilité de la séquence d'observation partielle $Y^t = y_1, y_2, \dots, y_t \ t \leq T$, qui se termine à l'état i . Les probabilités intermédiaires (ou partielles) sont calculées de manière récursive en calculant premièrement ces probabilités pour tous les états à $t = 1$.

$$\alpha_1(j) = \pi(j) \cdot b_j(1), \text{ pour } 1 \leq j \leq N \quad (4.7)$$

Ensuite, pour chaque instant, $t = 2, \dots, T$, les probabilités partielles sont calculées pour chaque état par la relation récursive suivante :

$$\alpha_{t+1}(j) = \sum_{i=1}^N (\alpha_t(i) a_{ij}) b_j(t), \text{ pour } 1 \leq j \leq N, \ 1 \leq t \leq T - 1 \quad (4.8)$$

Avec cette relation, nous pouvons alors calculer la probabilité intermédiaire à l'instant T pour chaque état j , $\alpha_T(j)$. Et finalement, la somme de toutes les probabilités partielles à l'instant T fournit la probabilité requise :

$$p(Y|\lambda) = \sum_{i=1}^N \alpha_T(i) \quad (4.9)$$

Pour récapituler, chaque probabilité partielle (à l'instant $t > 2$) est calculée à partir de tous les états précédents. De façon similaire, nous pouvons définir une variable « arrière » ou backward $\beta_t(i)$ comme la probabilité de la séquence d'observation partielle $y_{t+1}, y_{t+2}, \dots, y_T$, étant donné que l'état courant est i . Pour calculer les $\beta_t(i)$, il existe aussi, comme pour les $\alpha_t(i)$, une relation récursive :

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_{ij} b_j(t+1), \text{ pour } 1 \leq i \leq N, 1 \leq t \leq T-1 \quad (4.10)$$

Avec

$$\beta_T(i) = 1, \text{ pour } 1 \leq i \leq N. \quad (4.11)$$

Si nous cherchions un lien entre les deux variables intermédiaires $\beta_t(i)$ et $\alpha_t(i)$, nous pouvons remarquer que :

$$\alpha_t(i)\beta_t(i) = p(Y, y_t = i | \lambda), \text{ pour } 1 \leq i \leq N, 1 \leq t \leq T. \quad (4.12)$$

Ainsi, la somme de ce produit donne une autre façon pour calculer la probabilité $p(Y|\lambda)$, tout en utilisant les probabilités forward et backward :

$$p(Y|\lambda) = \sum_{i=1}^N p(Y, y_t = i | \lambda) = \sum_{i=1}^N \alpha_t(i)\beta_t(i), \text{ pour } 1 \leq t \leq T \quad (4.13)$$

4.1.2.2 Décodage et l'algorithme de Viterbi

Le problème du décodage se pose quand, étant donné une série d'observations, nous avons à trouver la séquence la plus probable des états cachés d'un modèle HMM. Ce problème est d'autant plus intéressant que dans plusieurs cas, les états cachés du HMM représentent quelque chose de non observable directement. Pour déterminer la séquence des états cachés la plus probable, étant donné une séquence d'observations, $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T$ et un HMM $\lambda = (\pi, A, B)$, l'algorithme de Viterbi est le plus utilisé. Dans cette méthode, la séquence complète des états avec le maximum de vraisemblance est trouvée.

Algorithme de Viterbi : L'algorithme peut se résumer formellement de la façon suivante :

- Pour chacun des états, calcul par récurrence de la variable intermédiaire :

$$\delta_t(i) = \max p(x_1, x_2, \dots, x_{t-1}, x_t = i, y_1, y_2, \dots, y_{t-1} | \lambda) \quad (4.14)$$

Le maximum étant calculé sur toutes les séquences d'états possibles x_1, x_2, \dots, x_{t-1} . Ce calcul se fait de manière récursive en deux étapes :

- Initialisation :

$$\delta_1(j) = \pi(j) \cdot b_j(1), \text{ pour } 1 \leq j \leq N \quad (4.15)$$

- Relation récursive :

$$\delta_{t+1}(j) = b_j(t+1) \{ \max_{1 \leq i \leq N} \delta_t(i) a_{ij} \}, \text{ pour } 1 \leq j \leq N, 1 \leq t \leq T-1 \quad (4.16)$$

- Calcul de $\delta_T(i)$, $1 \leq j \leq N$, en utilisant cette dernière récursion et en retenant toujours un pointeur sur l'état « élu » dans une opération de maximisation.
- Détermination de l'état final du système ($t = T$) le plus probable :

$$i_t = \operatorname{argmax}_{1 \leq j \leq N} (\delta_T(i)) \quad (4.17)$$

- Suivi du chemin le plus probable en revenant en arrière, soit : Si on note :

$$\phi_t(i) = \operatorname{argmax}_{1 \leq j \leq N} (\delta_{t-1}(j)) \quad (4.18)$$

la séquence d'état la plus probable peut être trouvée par :

$$i_t = \phi_{t+1}(i_{t+1}) \quad (4.19)$$

Et en fin, la séquence i_1, i_2, \dots, i_T est la séquence la plus probable des états cachés pour la séquence d'observation considérée.

4.1.2.3 Apprentissage

Le troisième, et le plus difficile, problème associé aux HMMs est de prendre une séquence connue d'observations pour représenter un ensemble d'états cachés, et d'obtenir le HMM $\lambda = (\pi, A, B)$ qui est le modèle le plus probable décrivant ce qui est observé. En d'autres termes, dans plusieurs cas d'applications, le problème de l'apprentissage concerne la façon avec laquelle les paramètres du HMM sont ajustés, étant donné un ensemble d'observations (appelé ensemble d'apprentissage). Les paramètres du HMM à optimiser peuvent être différents d'une application à l'autre. De ce fait, il peut y avoir divers critères d'optimisation pour l'apprentissage, chacun d'entre eux étant choisi selon l'application considérée. Parmi ces critères, nous trouvons le critère du maximum de vraisemblance et de l'Information Maximum Mutuelle (MMI pour Maximum Mutual Information). Nous nous contentons ici de décrire un seul algorithme permettant de générer les paramètres d'un HMM à partir d'une séquence d'observations. Il s'agit de l'algorithme de Baum-Welch avec un critère de maximum de vraisemblance. Cet algorithme est aussi connu sous le nom de *Forward-Backward*.

- **Algorithme de Forward-backward** : Cet algorithme est utilisé quand les matrices A et B d'un HMM ne sont pas directement mesurables, comme c'est souvent le cas dans plusieurs applications réelles. Plus formellement, on considère une unique séquence d'observation $Y =$

$\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T$. Notre but est de trouver les paramètres $\lambda = (A, B)$ qui maximisent la probabilité de générer Y avec le modèle. Formellement, les calculs doivent maximiser la quantité :

$$Q(\lambda, \bar{\lambda}) = \sum_x p(x|Y, \lambda) \log\{p(Y, x, \bar{\lambda})\} \quad (4.20)$$

ou x désigne un état donné et $\bar{\lambda}$ le modèle estimé. Pour décrire l'algorithme nous avons à définir deux variables intermédiaires : $-\varepsilon_t(i, j) = p(x_t = i, x_{t+1} = j|Y, \lambda)$: la probabilité d'être dans l'état i à l'instant t et dans l'état j à l'instant $t+1$. $-\gamma_t(i) = p(x_t = i|Y, \lambda)$: la probabilité d'être dans l'état i à l'instant t étant donné la séquence d'observation et le modèle HMM. Ces deux variables peuvent être exprimées en fonction des variables forward, $\alpha_t(i)$ et backward, $\beta_t(i)$ définies précédemment. Pour résumer, l'algorithme peut être décrit de la façon suivante :

Initialisation : Des paramètres arbitraires pour le modèle sont choisis ; entre autre, les valeurs de π sont choisies aléatoirement tandis que les variables A et B sont initialisées. Par exemple, les valeurs de A sont fixées à priori et celles de B sont initialisées par une quantification vectorielle.

Itération :

- Les variables A et B sont placées à leurs valeurs de pseudo-comptes.
- Calcul des variables $\alpha_t(i)$ et $\beta_t(i)$ pour chaque état i , en utilisant respectivement les algorithmes forward et backward.
- En déduire les variables $\varepsilon_t(i, j)$ et $\gamma_t(i)$ en utilisant les expressions suivantes qui les lient aux variables forward et backward :

$$\varepsilon_t(i, j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(t+1)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(t+1)} \quad (4.21)$$

et

$$\gamma_t(i) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{i=1}^N \alpha_t(i) \beta_t(i)} \quad (4.22)$$

De ces deux expressions, il facile de remarquer que :

$$\gamma_t(i) = \sum_{j=1}^N \varepsilon_t(i, j) \quad (4.23)$$

- L'étape suivante consiste à actualiser les paramètres du HMM en utilisant ce qu'on appelle les *formules de ré-estimation* :

$$\bar{\pi} = \gamma_1(i), \text{ pour } 1 \leq i \leq N \quad (4.24)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^T \varepsilon_t(i,j)}{\sum_{t=1}^T \gamma_t(i)} \text{ pour } 1 \leq i \leq N, 1 \leq j \leq N \quad (4.25)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1}^T \varepsilon_t(i,j)}{\sum_{t=1}^T \gamma_t(i)} \text{ pour } 1 \leq i \leq N, 1 \leq k \leq M \quad (4.26)$$

L'algorithme est arrêté si le changement de la log-vraisemblance est inférieur à un seuil prédéfini ou si le nombre maximum d'itération est atteint.

4.1.3 Différents types de modèles HMM

Depuis le début de cette section, nous avons traité en général le modèle HMM en supposant qu'il est caractérisé par une matrice de transition des états pleine ; c'est-à-dire que les transitions peuvent s'effectuer à partir de n'importe quel état vers n'importe quel autre état. On parle ici de modèle ergodique. Un tel modèle est défini comme un HMM tel que tous les états sont accessibles à partir de n'importe quel autre état. Pour certaines applications, il est demandé d'imposer certaines contraintes sur la matrice de transition ; ce qui rend le modèle non ergodique.

Dans ce sens, la littérature nous donne deux exemples types de modèles non-ergodique largement employés (Rabiner and Juang 1993). Ces deux modèles sont appelés gauche-droite du fait que la séquence des états produisant la séquence d'observations doit toujours avancer de l'état le plus à gauche à l'état le plus à droite. Ils diffèrent par le fait qu'un est un simple gauche-droite dans lequel il y a qu'un seul chemin à travers les états, et l'autre est un parallèle gauche-droite dans lequel il y a plusieurs chemins. Un modèle gauche-droite (parallèle ou simple) impose une structure temporelle ordonnée pour le HMM dans laquelle l'état numéroté avec un numéro inférieur précède toujours l'état avec un numéro supérieur. La figure 4.2 illustre les trois structures HMM.

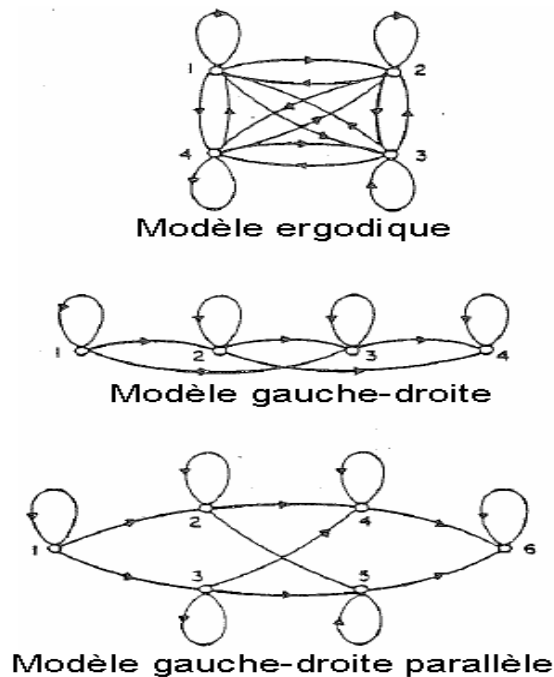


Figure 4.2 – Trois types distincts de modèles HMM. Illustration avec un exemple de HMM à 4 état (d'après Rabiner et Juang 1993).

4.1.4 Résumé

Le modèle de Markov caché est un outil statistique qui peut être défini quand les états d'un processus ne sont pas directement observables, mais sont indirectement et probabilistiquement observables comme un autre ensemble d'états. De tels modèles, appliqués dans des processus réels, imposent de résoudre trois problèmes :

- Evaluation : avec quelle probabilité un modèle donné génère-t-il une séquence d'observations donnée. L'algorithme forward résout efficacement ce problème.
- Décodage : quelle est la séquence d'états cachés la plus probable qui génère une séquence d'observations. L'algorithme de Viterbi résout ce problème.
- Apprentissage : comment optimiser (apprendre) les paramètres d'un modèle HMM à partir d'un échantillon donné de séquences d'observations. Ce problème peut être résolu en utilisant l'algorithme *forward-backward*.

Enfin, il est à noter un défaut habituel des modèles HMM qui concerne la sur-simplification associée à l'hypothèse markovienne ; c'est-à-dire qu'un état dépend seulement de ses prédécesseurs directs et que cette dépendance est indépendante du temps. Cependant,

les HMMs ont prouvé leur grande valeur dans des systèmes réels d'analyse et restent l'un des outils les plus utilisés en RAP.

4.2 Les algorithmes génétiques

4.2.1 Principe des algorithmes génétiques

Les AG proviennent de la modélisation de la théorie de l'évolution de C. Darwin (Darwin 1859). Les AG sont des algorithmes d'optimisation, ils font partie du cadre plus générale des métaheuristiques évolutionnaires comprenant, entre autres, les AG (Holland 1975), les stratégies d'évolution (Beyer 2001) et la programmation évolutionnaire (Fogel et al. 1966).

Une étude bibliographique très riche a été présentée par Kicinger dans (Kicinger et al. 2005). Cette étude regroupe des travaux de recherche récents dans le domaine de l'optimisation structurale par méthode évolutionnaire, en particulier par AG.

Initialement développés par Holland (1975), les AG sont devenus populaires à partir de la publication du livre « Genetic Algorithms in search, optimization and machine learning » de Goldberg (1989). La forme canonique d'un algorithme génétique est donnée par :

Initialiser la population d'individus : P_0
Evaluer les individus de la population P_0
 $t=0$
Répéter
Sélectionner les individus pour la production : $P_t^{parent} \subseteq P_t$
Croiser les individus de $P_t^{parent} : P_t^{enfant} - 1$
Muter les individus de $P_t^{enfant} : \tilde{P}_t^{enfant}$
Sélectionner les individus de $P_t^{parent} \cup \tilde{P}_t^{enfant}$ à conserver
Les individus sélectionnés forment de la population P_{t+1}
 $t=t+1$
Tant que condition d'arrêt non vérifiée

Le principe général du fonctionnement d'un algorithme génétique est représenté sur la figure 4.3 : on commence par générer une population d'individus de façon aléatoire. Pour passer d'une génération k à la génération $k+1$, les trois opérations suivantes sont répétées pour tous les éléments de la population k . Des couples de parents P_1 et P_2 sont sélectionnés en fonction de leurs adaptations. L'opérateur de croisement leur est appliqué avec une probabilité

P_c (généralement autour de 0.6) et génère des couples d'enfants C_1 et C_2 . D'autres éléments P sont sélectionnés en fonction de leur adaptation. L'opérateur de mutation leur est appliqué avec la probabilité P_m (P_m est généralement très inférieur à P_c) et génère des individus mutés P_0 . Le niveau d'adaptation des enfants (C_1, C_2) et des individus mutés P_0 sont ensuite évalués avant insertion dans la nouvelle population.

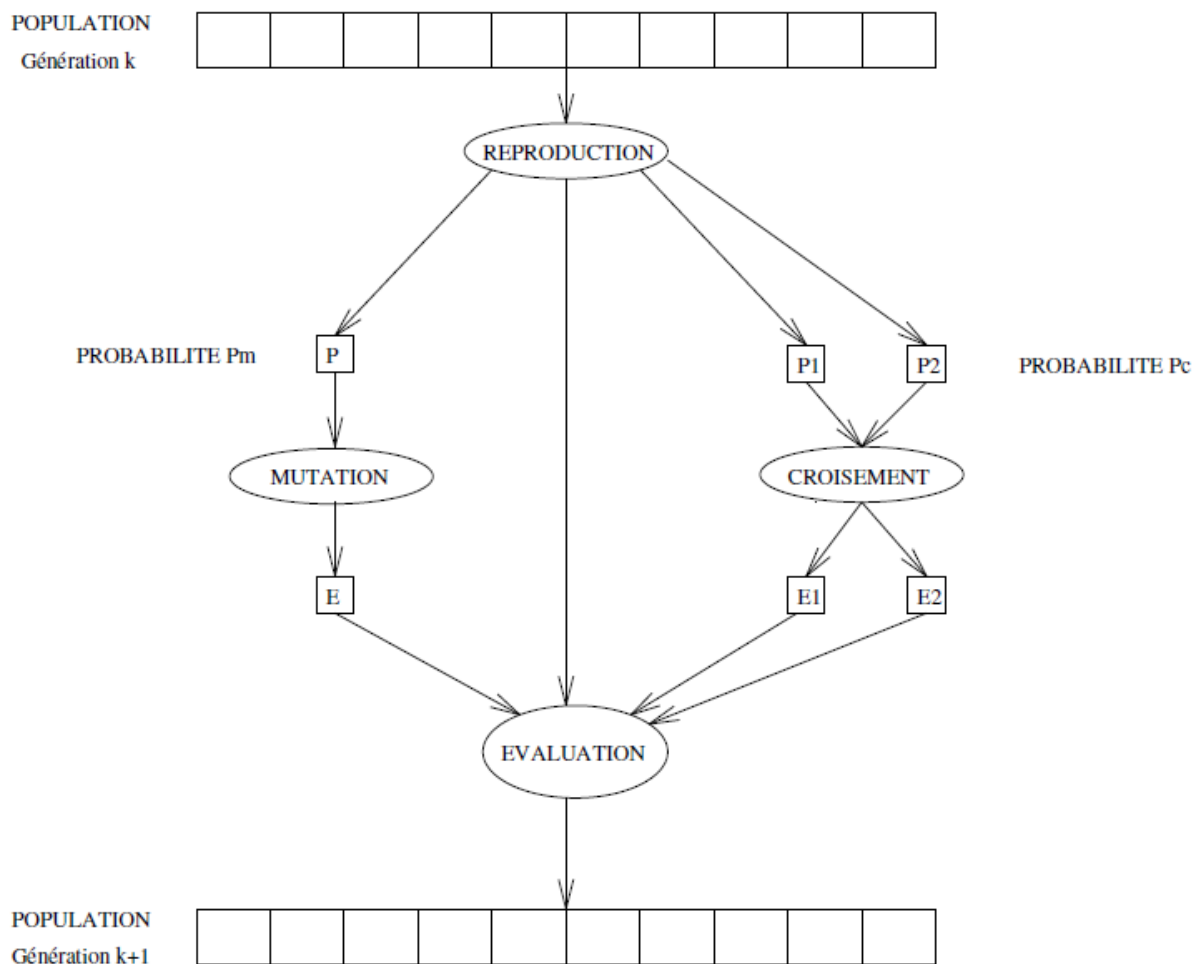


Figure 4.3 – Principe général des algorithmes génétiques.

Différents critères d'arrêt de l'algorithme peuvent être choisis :

- Le nombre de générations que l'on souhaite exécuter peut être fixé a priori. C'est ce que l'on est tenté de faire lorsque l'on doit trouver une solution dans un temps limité.
- L'algorithme peut être arrêté lorsque la population n'évolue plus ou plus suffisamment rapidement.

Nous allons maintenant détailler chacun de ces points.

4.2.2 Description détaillée

4.2.2.1 Codage des données

Historiquement le codage utilisé par les AG était représenté sous forme de chaînes de bits contenant toute l'information nécessaire à la description d'un point dans l'espace d'état. Ce type de codage a pour intérêt de permettre de créer des opérateurs de croisement et de mutation simples. C'est également en utilisant ce type de codage que les premiers résultats de convergence théorique ont été obtenus.

Cependant, ce type de codage n'est pas toujours bon comme le montrent les deux exemples suivants :

- deux éléments voisins en terme de distance de Hamming ne codent pas nécessairement deux éléments proches dans l'espace de recherche. Cet inconvénient peut être évité en utilisant un codage de Gray.
- Pour des problèmes d'optimisation dans des espaces de grande dimension, le codage binaire peut rapidement devenir mauvais. Généralement, chaque variable est représentée par une partie de la chaîne de bits et la structure du problème n'est pas bien reflétée, l'ordre des variables ayant une importance dans la structure du chromosome alors qu'il n'en a pas forcément dans la structure du problème.

Les AG utilisant des vecteurs réels (Goldberg 1991 ; Wright 1991) évitent ce problème en conservant les variables du problème dans le codage de l'élément de population sans passer par le codage binaire intermédiaire. La structure du problème est conservée dans le codage.

4.2.2.2 Génération aléatoire de la population initiale

Le choix de la population initiale d'individus conditionne fortement la rapidité de l'algorithme. Si la position de l'optimum dans l'espace d'état est totalement inconnue, il est naturel de générer aléatoirement des individus en faisant des tirages uniformes dans chacun des domaines associés aux composantes de l'espace d'état en veillant à ce que les individus produits respectent les contraintes (Michalewicz and Janikov 1991). Si par contre, des informations a priori sur le problème sont disponibles, il paraît bien évidemment naturel de générer les individus dans un sous-domaine particulier afin d'accélérer la convergence. Dans l'hypothèse où la gestion des contraintes ne peuvent se faire directement, les contraintes sont

généralement incluses dans le critère à optimiser sous forme de pénalités. Il est clair qu'il vaut mieux, lorsque c'est possible ne générer que des éléments de population respectant les contraintes.

4.2.2.3 Évaluation

A chaque solution, on associe une fonction performance « *fitness* » reliée à la valeur de la fonction objectif. Cette fonction décrit le mérite de l'individu qui est représenté par un chromosome. L'évaluation des individus en optimisation topologique des structures se fait par une méthode d'analyse numérique des structures, généralement la méthode des éléments finis.

La fonction performance est très importante pour un AG au même titre que le codage. En effet, pour que les AG se comportent bien, nous devons trouver une manière de formuler des fonctions performance ne comportant pas trop de maxima locaux et ne présentant pas de maximum local isolé. La construction de la fonction performance est évidente pour certains problèmes. Pour les problèmes de maximisation par exemple, la fonction mérite peut être égale à la fonction objectif. Par contre, pour les problèmes de minimisation, l'objectif est de trouver des solutions pour lesquelles la fonction objectif atteint des valeurs minimales. Dans ce cas, la fonction performance choisie est la réciproque de la fonction objectif. Dans tous les cas, l'AG cherche à maximiser la fonction performance qui, dans le cadre d'un problème de minimisation, prend la forme suivante :

$$Fitness(x_i) = \frac{1}{f(x_i)} \quad (4.27)$$

Où $f(x_i)$ représente la fonction objectif évaluée pour l'individu x_i .

Un choix classique de fonction objectif est la compliance. Ce choix se justifie pleinement dans le cadre d'une approche déterministe ou il est nécessaire de dériver pour pouvoir procéder à une analyse de sensibilité. Dans un AG, ou l'approche stochastique ne nécessite pas de dérivation, le choix de la compliance comme fonction objectif n'est pas aussi vital. Jakiela et ses collaborateurs (Jakiela 2000) ont posé le problème d'optimisation sous forme de maximisation de la raideur de la structure en supposant que la raideur est inversement proportionnelle au déplacement maximal ' δ_{max} ' de la structure :

$$Fitness = \frac{1}{|\delta_{max}|} \quad (4.28)$$

La raideur n'étant pas une grandeur différentiable, les méthodes déterministes ne sont opérationnelles pour maximiser mérite définie par (1-10). En revanche, ce n'est pas le cas pour les méthodes stochastique, telles que les AG, qui sont exemptées de l'analyse de sensibilité.

4.2.2.4 Gestion des contraintes

Un élément de population qui viole une contrainte se verra attribuer une mauvaise fitness et aura une probabilité forte d'être éliminé par le processus de sélection. Il peut cependant être intéressant de conserver, tout en les pénalisant, les éléments non admissibles car ils peuvent permettre de générer des éléments admissibles de bonne qualité. Pour de nombreux problèmes, l'optimum est atteint lorsque l'une au moins des contraintes de séparation est saturée, c'est-à-dire sur la frontière de l'espace admissible.

Gérer les contraintes en pénalisant la fonction fitness est difficile, un « dosage » s'impose pour ne pas favoriser la recherche de solutions admissibles au détriment de la recherche de l'optimum ou inversement. Disposant d'une population d'individus non homogène, la diversité de la population doit être entretenue au cours des générations afin de parcourir le plus largement possible l'espace d'état. C'est le rôle des opérateurs de croisement et de mutation.

4.2.2.5 Principes de sélection

A l'inverse d'autres techniques d'optimisation, les AG ne requièrent pas d'hypothèse particulière sur la régularité de la fonction objectif. L'AG n'utilise notamment pas ses dérivées successives, ce qui rend très vaste son domaine d'application. Aucune hypothèse sur la continuité n'est non plus requise. Néanmoins, dans la pratique, les AG sont sensibles à la régularité des fonctions qu'ils optimisent. Le peu d'hypothèses requises permet de traiter des problèmes très complexes. La fonction à optimiser peut ainsi être le résultat d'une simulation.

La sélection permet d'identifier statistiquement les meilleurs individus d'une population et d'éliminer les mauvais. On trouve dans la littérature un nombre important de principes de sélection plus ou moins adaptés aux problèmes qu'ils traitent. Dans le cadre de notre travail, les deux principes de sélection suivants ont été testés et évalués (Goldberg 1989):

- *Roulette wheel selection;*
- *Stochastic remainder without replacement selection;*

Le principe de *Roulette wheel selection* consiste à associer à chaque individu un segment dont la longueur est proportionnelle à sa fitness. On reproduit ici le principe de tirage aléatoire utilisé dans les roulettes de casinos avec une structure linéaire. Ces segments sont ensuite concaténés sur un axe que l'on normalise entre 0 et 1. On tire alors un nombre aléatoire de distribution uniforme entre 0 et 1, puis on « regarde » quel est le segment sélectionné. Avec ce système, les grands segments, c'est-à-dire les bons individus, seront plus souvent adressés que les petits. Lorsque la dimension de la population est réduite, il est difficile d'obtenir en pratique l'espérance mathématique de sélection en raison du peu de tirages effectués. Un biais de sélection plus ou moins fort existe suivant la dimension de la population.

La *Stochastic remainder without replacement selection* évite ce genre de problème et donne de bons résultats pour nos applications. Décrivons ce principe de sélection :

- Pour chaque élément i , on calcule le rapport r_i de sa fitness sur la moyenne des fitness.
- Soit $e(r_i)$ la partie entière de r_i , chaque élément est reproduit exactement $e(r_i)$ fois.
- La *roulette wheel selection* précédemment décrite est appliquée sur les individus affectés des fitness $r_i - e(r_i)$.

Compte-tenu du fait que des faibles populations seront utilisées par la suite, ce principe de sélection s'avèrera le plus efficace dans les applications pratiques et sera donc utilisé par la suite.

4.2.2.6 Opérateur de Croisement

Le croisement a pour but d'enrichir la diversité de la population en manipulant la structure des chromosomes. Classiquement, les croisements sont envisagés avec deux parents et génèrent deux enfants.

Initialement, le croisement associé au codage par chaînes de bits est le croisement à découpage de chromosomes (*slicing crossover*). Pour effectuer ce type de croisement sur des chromosomes constitués de M gènes, on tire aléatoirement une position dans chacun des parents. On échange ensuite les deux sous-chaînes terminales de chacun des deux chromosomes, ce qui produit deux enfants C_1 et C_2 (voir figure 4.4).

On peut étendre ce principe en découpant le chromosome non pas en 2 sous-chaînes mais en 3, 4, etc. (Bridges and Goldberg 1991) (voir figure 4.5). Ce type de croisement à découpage de chromosomes est très efficace pour les problèmes discrets. Pour les problèmes continus, un croisement « barycentrique » est souvent utilisé : deux gènes $P_1(i)$ et $P_2(i)$ sont

sélectionnés dans chacun des parents à la même position i . Ils définissent deux nouveaux gènes $C_1(i)$ et $C_2(i)$ par combinaison linéaire :

$$\begin{cases} C_1(i) = \alpha P_1(i) + (1 - \alpha)P_2(i) \\ C_2(i) = (1 - \alpha)P_1(i) + \alpha P_2(i) \end{cases} \quad (4.29)$$

ou α est un coefficient de pondération aléatoire adapté au domaine d'extension des gènes (il n'est pas nécessairement compris entre 0 et 1, il peut par exemple prendre des valeurs dans l'intervalle $[-0.5, 1.5]$ ce qui permet de générer des points entre, ou à l'extérieur des deux gènes considérés).

Dans le cas particulier d'un chromosome matriciel constitué par la concaténation de vecteurs, on peut étendre ce principe de croisement aux vecteurs constituant les gènes (voir figure 4.6) :

$$\begin{cases} \vec{C}_1(i) = \alpha \vec{P}_1(i) + (1 - \alpha)\vec{P}_2(i) \\ \vec{C}_2(i) = (1 - \alpha)\vec{P}_1(i) + \alpha \vec{P}_2(i) \end{cases} \quad (4.30)$$

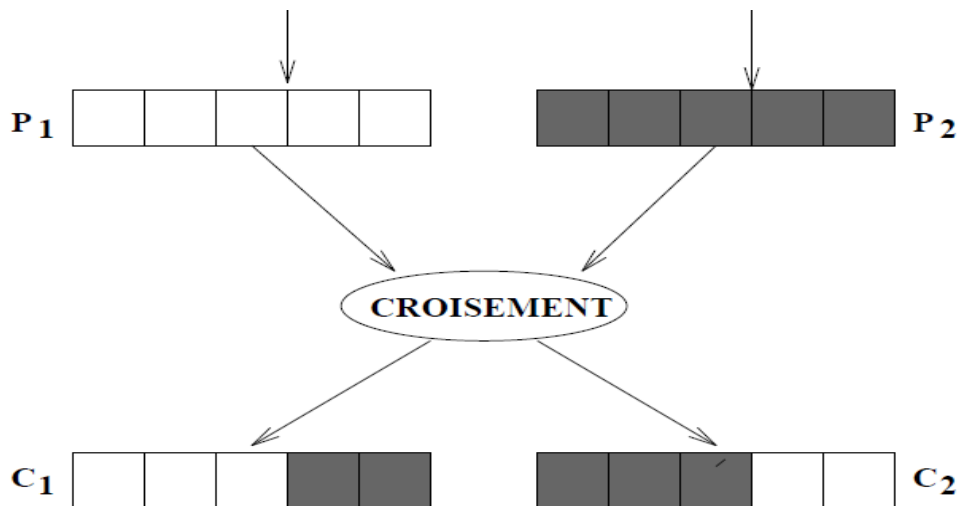


Figure 4.4 – Slicing crossover.

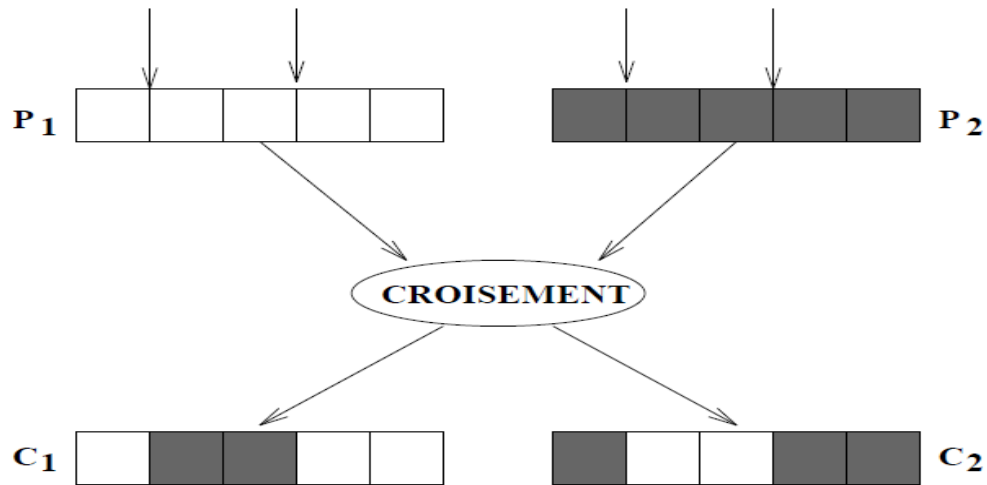


Figure 4.5 – Slicing crossover à 2 points.

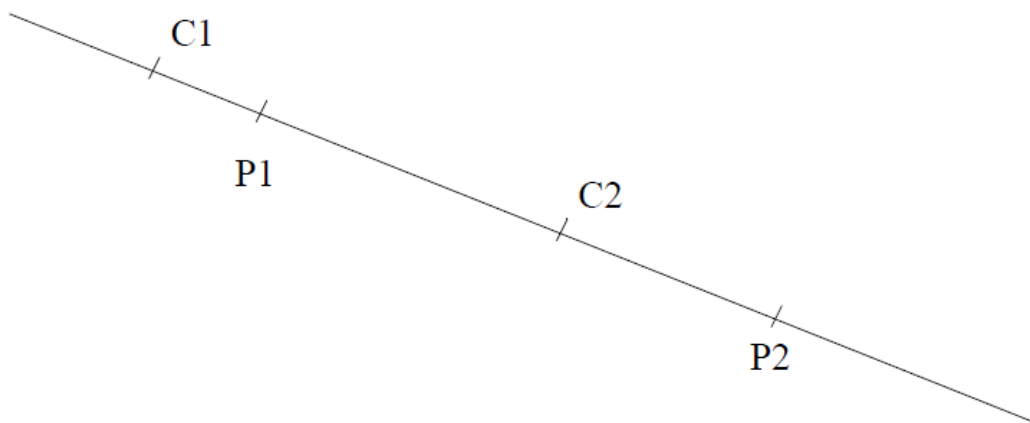


Figure 4.6 – Croisement barycentrique.

On peut imaginer et tester des opérateurs de croisement plus ou moins complexes sur un problème donné mais l'efficacité de ce dernier est souvent liée intrinsèquement au problème.

4.2.2.7 Opérateur de mutation

L'opérateur de mutation apporte aux AG la propriété d'ergodicité de parcours d'espace. Cette propriété indique que l'AG sera susceptible d'atteindre tous les points de l'espace d'état, sans pour autant les parcourir tous dans le processus de résolution. Ainsi en toute rigueur, l'AG peut converger sans croisement, et certaines implantations fonctionnent de cette manière. Les propriétés de convergence des AG sont donc fortement dépendantes de cet opérateur sur le plan théorique.

Pour les problèmes discrets, l'opérateur de mutation consiste généralement à tirer aléatoirement un gène dans le chromosome et à le remplacer par une valeur aléatoire (voir figure 4.7). Si la notion de distance existe, cette valeur peut être choisie dans le voisinage de la valeur initiale.

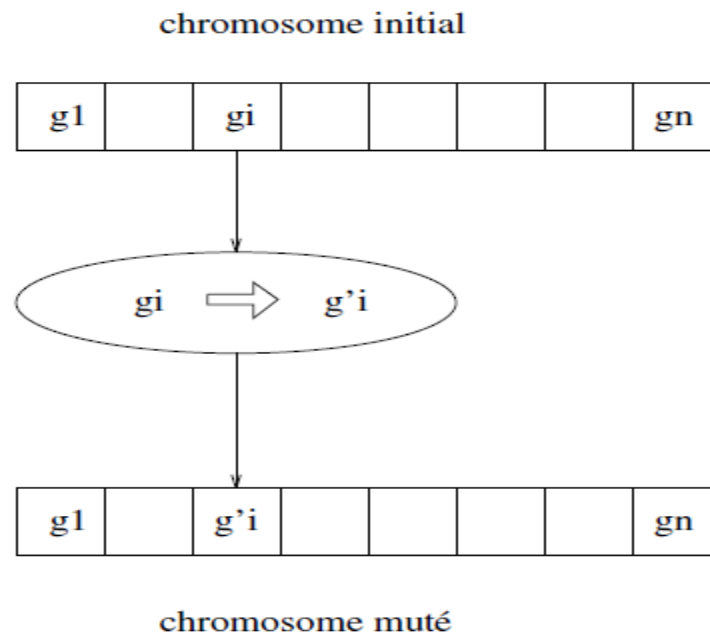


Figure 4.7 – Principe de l'opérateur de mutation.

Dans les problèmes continus, on procède un peu de la même manière en tirant aléatoirement un gène dans le chromosome, auquel on ajoute un bruit généralement gaussien. L'écart type de ce bruit est difficile à choisir a priori.

4.2.2.8 Partage (Sharing)

Le partage est un paramètre évolué des AG (Goldberg and Richardson 1987). Il est utilisé pour éviter le regroupement d'individus performants, et assurer une certaine diversité génétique dans la population. Le principe consiste à pénaliser les individus qui ont beaucoup de voisins proches en divisant leurs performances sur une fonction dite de partage. Cette dernière, dont la valeur est comprise entre 0 et 1, est calculée en fonction d'un paramètre qui mesure le degré de similarité entre les individus. La performance modifiée pour un individu x_i s'écrit de la manière suivante :

$$Fitness_{Sh}(x_i) = \frac{Fitness(x_i)}{\sum_{j=1}^n Sh(d(x_i, x_j))} \quad (4.31)$$

Sh (Sharing) est la fonction de partage de d est la distance entre les individus qui exprime le degré de similarité entre ces individus.

La technique de partage est souvent accompagnée par une technique de regroupement appelé « Clustering ». L'information fournie par la fonction de partage peut être utilisée pour éviter le croisement, inutile, entre les individus similaires.

4.2.2.9 Critères d'arrêt de l'algorithme

Le test d'arrêt joue un rôle très important dans le jugement de la qualité des individus. Il existe trois types:

- Arrêt de l'algorithme après un certain nombre de générations.
- Arrêt de l'algorithme lorsque le meilleur individu n'a pas été amélioré depuis un certain nombre de générations.
- Arrêt de l'algorithme lorsqu'il y a perte de diversité génétique.

Ces valeurs sont à paramétrer selon le temps disponible pour l'exécution de l'algorithme, la performance de la recherche de celui-ci et les conditions du problème à résoudre.

4.2.3 Avantages et désavantages des algorithmes génétiques

Les algorithmes génétiques présentent les avantages suivants : ce sont des méthodes robustes à l'initialisation (c'est-à-dire que leurs convergences ne dépendent pas de la valeur initiale), qui permettent de déterminer l'optimum global d'une fonctionnelle ou de s'en approcher, et qui sont parallélisables. En revanche leur inconvénient majeur réside dans le nombre important d'évaluations nécessaires et leur temps de convergence.

En revanche, les méthodes déterministes convergent rapidement vers un optimum. Cependant, elles ne sont pas aussi robustes à l'initialisation que les algorithmes génétiques, ce qui n'assure pas que l'optimum déterminé est un optimum global.

4.3 Moteur de reconnaissance GA/HMM

Dans ce travail, nous avons opté pour des modèles statistiques : les HMM qui se sont imposés comme une technique prédominante en reconnaissance de la parole ces dernières

années (Kwong and Chau 1997 ; Shing-Tai et al. 2010). Nous avons utilisé pour cette phase en commun entre la reconnaissance acoustique et visuelle N HMM de type gauche-droite.

L'algorithme de classification effectue une partition géographique d'un nuage de points (vecteurs acoustiques respectivement visuels) en différenciant classes en minimisant la distorsion moyenne de l'ensemble, on utilise pour cette étape la méthode de K-means la plus connue et la plus utilisée. La taille K du CodeBook est un paramètre crucial dont la valeur affecte en grande partie les performances des HMMs utilisés pour la reconnaissance, car on le considère que c'est le nombre des mixtures.

Dans la phase d'apprentissage en utilisant comme il est mentionné avant les HMM mais cette fois combinés avec les AG, cette algorithm va chercher à obtenir des HMM optimales (Patterson et al. 2002 ; Xue-ying et al. 2007 ; Goh et al. 2010), Le processus de formation d'un modèle $\lambda = (A, B, \pi)$ Pour les données de référence en utilisant une méthode hybride GA/HMM peut être tirée à partir du diagramme ci-dessous:

Premièrement on commence par créer une population de taille S , aléatoirement, d'une façon que chaque individu contient n chromosomes pour les probabilités initiales, $n \times n$ chromosomes contiennent les probabilités de transition et $n \times m$ chromosomes chacune contient la probabilité d'émission. Aucun individu n'est marqué « parent ». Le codage de chaque individu est comme suit :

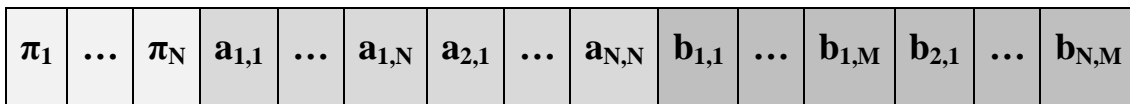


Figure 4.8 – Méthode de représentation des chromosomes dans l'apprentissage des GA/HMMs.

Après nous appliquons sur chaque HMM de la population non marqué « parent » l'algorithme de Baum-Welch à partir de l'observation O .

L'étape suivante est de calculer pour chaque individu de la population qui ne porte pas la marque « parent » la valeur de fitness (noté aussi la fonction objective) on utilise l'algorithme de Baum-Welch, et noter la valeur de probabilité de l'observation. Pour tous ceux qui portaient cette marque, l'enlever. Mathématiquement la fitness de n model est exprimé comme suit (Oudelha and Aïnon 2010):

$$f(\lambda_i) = \frac{P_n}{\sum_{i=1}^N P_i} \tag{4.32}$$

Où, P_n est la probabilité moyenne du model λ_i , N est le nombre des individus dans une population et M représente le nombre de vecteurs dans o_i .

La probabilité moyenne P_n est donc donnée par l'équation suivante:

$$P_n = \frac{\sum_{i=1}^M \log(P(o_i|\lambda_i))}{M} \quad (4.33)$$

Où $P(o|\lambda)$ est la probabilité de vraisemblance.

La troisième étape est de sélectionner parmi tous les individus de la population, un certain nombre $S' < S$, qui seront utilisés comme parents pour régénérer les $S - S'$ autres individus non retenus. La sélection se réalise suivant les meilleurs scores calculés à la phase 3. Chaque individu sélectionné est marqué « parent ».

Les opérations génétiques peuvent inclure croisement et par mutation. L'opération génétique est réalisée pour améliorer la technique de ré-estimation de Baum-Welch pour que les populations génétiques de cette opération résultent un modèle optimal (Pérez et al. 2007 ; Xue-ying et al. 2007 ; Oudelha and Aïnon 2010).

A la fin on termine par l'évaluation de la condition d'arrêt, Si le nombre d'itérations maximum n'est pas atteint, alors retourner à la deuxième étape, sinon aller à la dernière étape qui vas renvoyer la meilleure HMM parmi la population en cours.

Un tel classifieur est basé sur un critère de maximum de vraisemblance, il prend le mot à reconnaître comme étant une séquence d'observations discrètes (codes) produites par analyse et quantification vectorielle de la séquence de vecteurs de caractéristiques. Ce classifieur calcule la probabilité qui correspond à la probabilité d'obtenir la séquence par le modèle. Ces probabilités sont évaluées par la version logarithmique de l'algorithme de Viterbi. Finalement, le mot testé est affecté à la classe du mot K pour laquelle le modèle maximise la probabilité d'émission.

4.4 Conclusion

Les modèles de Markov cachés, présentés dans ce chapitre sont des techniques largement utilisées en reconnaissance de formes, et sont les plus utilisés en reconnaissance de la parole. Ils bénéficient d'algorithmes d'entraînement et décodage performants.

Dans le chapitre suivant, nous présentons, le principe et le fonctionnement de notre système de reconnaissance de la parole audiovisuelle proposé en utilisant la méthode hybride GA/HMM.

Description du système proposé

5

La spécification d'un système d'intelligence artificielle utilisant des HMM peut s'effectuer en trois phases distinctes, mais interagissantes entre elles (voir figure 5.1). La première phase, que nous nommerons prétraitement par la suite, consiste en l'ensemble des actions nécessaires à la transformation des données en séquences temporelles. La deuxième phase, dite d'apprentissage, consiste en la transformation de certaines des séquences construites en HMM, grâce à un algorithme d'apprentissage, tel que ceux décrits au chapitre précédent. La dernière phase, dite de post-traitement, consiste en l'utilisation des HMM produits en deuxième phase et de séquences produites par la première phase pour effectuer le traitement. Les traitements pouvant être réalisés par un tel système sont très variés : classification, segmentation, analyse, décision,...

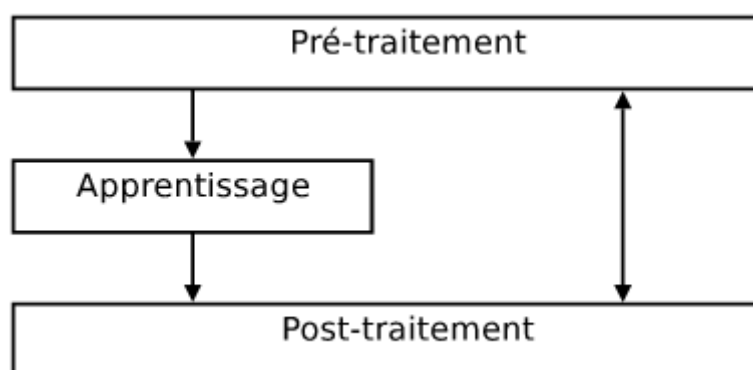


Figure 5.1 – Phases de spécification d'un système d'intelligence artificielle utilisant des HMM.

La phase d'apprentissage joue un rôle central au sein d'un tel système mais, en pratique, peu d'attention lui est accordée dans le cas des HMM. Dans de nombreuses applications des modèles sous optimaux sont utilisés avec succès. Cependant, ces applications s'appuient sur des principes théoriques qui ne sont valables que lorsque les modèles sont optimaux. Par conséquent, il est communément admis que des modèles optimaux permettraient, du moins en théorie, d'améliorer les performances du système d'intelligence artificielle.

La RAP s'applique à ce jour sur de nombreux signaux de qualité différente (fréquence d'échantillonnage, quantification, codage, conditions d'enregistrement). Nous rappelons que la parole est l'un des moyens les plus naturels par lequel des personnes communiquent. La RAP a pour objet la transformation du signal acoustique en une séquence de mots qui,

idéalement, correspond à la phrase prononcée par un locuteur. Les systèmes de reconnaissance qui utilisent comme entrée uniquement le signal acoustique atteignent leurs limites surtout dans des cas de situations environnementales bruitées donc réelles. Dans ces cas, l'intégration de l'information visuelle dans le système de reconnaissance peut constituer une voie de solution (Rogozan 1999). A cet effet nous nous intéressons à la mise en œuvre d'un système de reconnaissance intégrant conjointement les deux informations acoustique et visuelle de la parole se sont focalisés sur une interaction sensorielle de type fusion ou intégration. A ce niveau, reste posée la question du ou et comment cette fusion des modalités acoustique et visuelle se passe-t-elle chez l'homme. Pour répondre à cette question, il existe plusieurs modèles cognitifs qui diffèrent de par leur lieu d'intégration des informations en vue de leur intégration. La RAP audiovisuelle est née de l'idée que si l'homme exploite les informations provenant du visage du locuteur pour améliorer l'intelligibilité, la machine peut en faire autant, si d'une part le principe d'intégration des deux modalités est suffisamment bien connu, et si d'autre part les informations visuelles sont exploitées d'une façon optimale (Adjoudani and Benoît 1995).

Dans ce chapitre nous définissons les différentes méthodes que nous utiliserons par la suite dans la partie expérimentale.

5.1 Architecture de système de reconnaissance par fusion audiovisuelle

Le système AVASR comprend trois modules qui sont: le module de reconnaissance acoustique, le module de reconnaissance visuelle et le module de fusion.

Le module de reconnaissance acoustique utilise l'approche stochastique basée sur les modèles de Markov cachées (HMM) qui sont un type particulier des réseaux bayésiens. On processus générique est basé sur trois phases qui sont : la para métrisation du signal acoustique utilisant dans notre cas l'analyse log RASTA-PLP (RelAtive SpecTral Analysis-Perceptual Linear Predictive), l'apprentissage des modèles repose sur une recherche génétique d'un bon modèle parmi une population hétérogène des HMM (contenant différentes architectures) et une optimisation par un algorithme de gradient (Baum-Welch) et leur décodage sur l'algorithme de viterbi. Le module de reconnaissance visuelle utilise la même approche stochastique, il diffère uniquement par la phase de para métrisation basée elle sur la DCT (Discrete Cosine Transform).

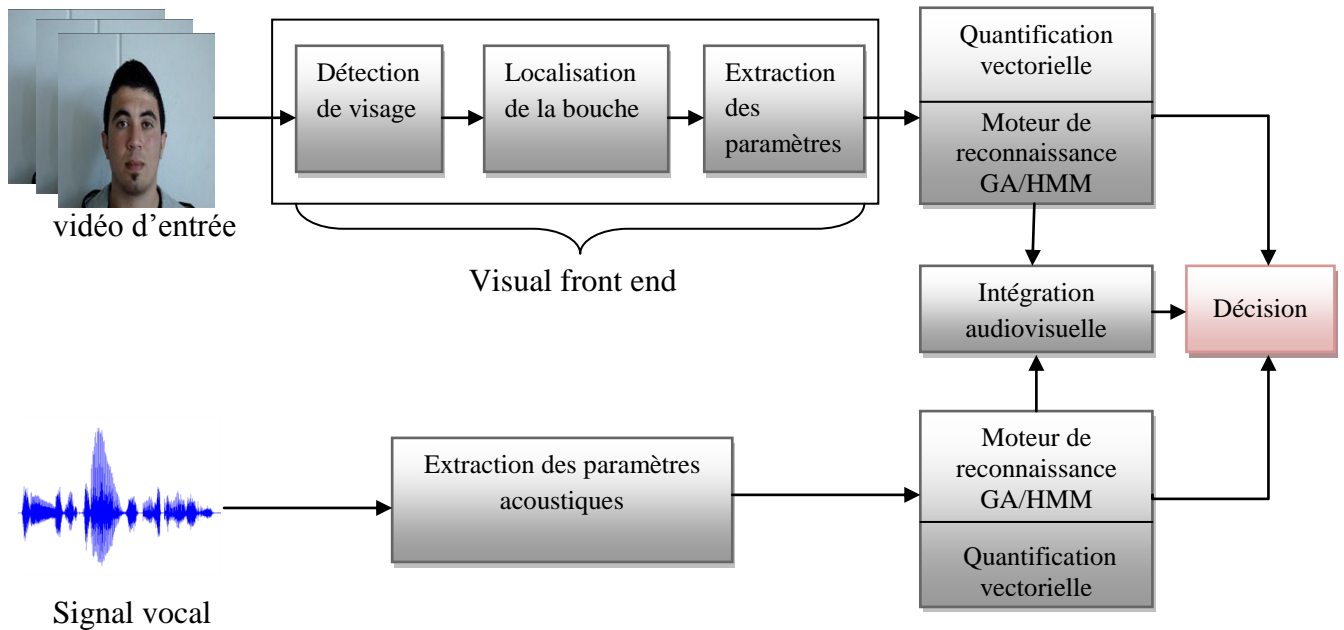


Figure 5.2 – Système d'un AVASR mis en œuvre.

La figure 5.2 présente les différentes étapes dans les processus d'apprentissage et de reconnaissance du système proposé. Chacun des éléments présents sur cette figure sera détaillée dans les prochaines sections.

5.1.1 Traitement visuel

Nous savons que les lèvres véhiculent la majeure partie de l'information visuelle utile pour la compréhension de la parole.

Les êtres humains emploient l'information visuelle de façon subconsciente afin de comprendre les paroles, particulièrement dans des environnements bruyants, mais également quand les conditions acoustiques sont bonnes. Le mouvement des lèvres du locuteur apporte une série d'information importante. L'effet McGurk (McGurk and MacDonald 1976) apporte la preuve en montrant que le cerveau, soumis à des stimuli auditifs et visuels inconsistants, perçoit un son différent de celui qui a été dit.

5.1.1.1 Détection de visage

La détection des visages pose le problème de la localisation des visages présents dans une image d'entrée. Idéalement, la détection fourni aussi leurs dimensions pour un éventuel traitement ultérieur.

Tous les AVASR nécessitent l'identification et le suivi de la ROI, qui peut être soit seulement la bouche, ou une région plus vaste, comme tout le visage. Cela commence généralement par localisation de visage du locuteur, en utilisant un algorithme de détection de visage.

Une avancée majeure dans le domaine a été réalisée par (Viola and Jones 2001). Ces derniers ont proposé une méthode basée sur l'apparence ("Appearance-based methods") rapide et robuste. La renommée de cette approche se base essentiellement sur trois contributions:

- **Algorithme de Viola & Jones**

Comme nous avons déjà mentionnés Viola et Jones ont proposé une méthode basée sur l'apparence ("Appearance-based methods") robuste et tournant à 15 fps pour des images de 384 x 288 pixels sur un pc Intel Pentium III 700Mhz. Ce fut la première méthode en temps réel présentée. La renommée de cette approche est faite sur trois concepts :

A. L'image intégrale

L'algorithme se base sur les caractéristiques de Haar (Haar features) pour localiser les visages présents sur une image d'entrée. Dans le but d'extraire rapidement ces caractéristiques, l'image est représentée sous forme intégrale. En effet, sous cette forme, l'extraction d'une caractéristique à n'importe quel endroit et à n'importe quelle échelle est effectuée en un temps constant tandis que le temps de conversion vers la représentation intégrale ne remet pas en cause ce gain de temps offert par l'utilisation de la représentation en image intégrale. La définition des caractéristiques de Haar et la manière dont la représentation intégrale accélère considérablement leur extraction sont présentés ci-après pour une image en niveaux de gris.

Dans toute image, une zone rectangulaire peut être délimitée et la somme des valeurs de ses pixels calculée. Une caractéristique de Haar est une simple combinaison linéaire de sommes ainsi obtenues.

Plusieurs caractéristiques de Haar peuvent être définies selon le nombre, les échelles, les positions et les dimensions des zones rectangulaires considérées. 4 exemples sont présentés à la figure 5.3.

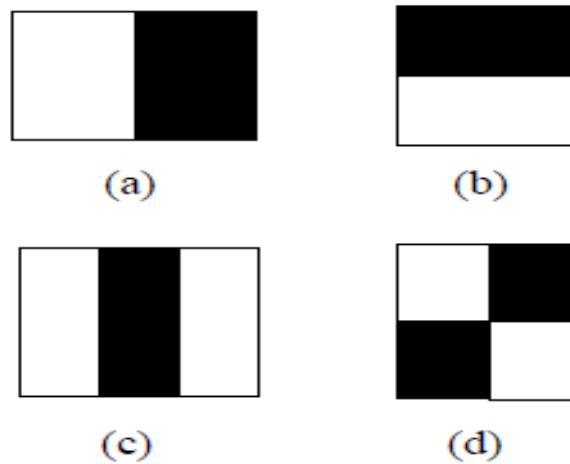


Figure 5.3 – Exemple de 4 caractéristiques de Haar. La somme des valeurs des pixels appartenant aux zones encadrées claires est soustraite à la somme des valeurs des pixels appartenant aux zones encadrées sombres pour obtenir la caractéristique de Haar. Chacune des quatre caractéristiques de Haar est représentée avec son cadre de détection respectif.

L'image intégrale est représentée mathématiquement par :

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (5.1)$$

$$\forall 0 < x \leq width, 0 < y \leq height. \quad (5.2)$$

ou $i(x, y)$ est l'image d'origine et $i(x0, y0)$ l'image sous sa nouvelle représentation. Ainsi chaque pixel a pour valeur la somme des valeurs des pixels compris dans le rectangle défini par le coin supérieur gauche de l'image et lui-même.

Le calcul de la somme des valeurs des pixels appartenant à une zone rectangulaire s'effectue donc en accédant seulement à quatre pixels de l'image intégrale : Soit un rectangle ABCD dont les sommets sont nommés dans le sens des aiguilles d'une montre en commençant par le sommet supérieur gauche et soit x la valeur sous la représentation intégrale d'un sommet X du rectangle ($X \in \{A, B, C, D\}$). La somme des valeurs des pixels appartenant à ABCD est, quelle que soit sa taille, donnée par $c - b - d + a$. Une caractéristique de Haar étant une combinaison linéaire de tels rectangles ABCD, son calcul se fait alors en un temps indépendant de sa taille.

B. Algorithme d'apprentissage basé sur Adaboost

Pour localiser les visages sur l'image d'entrée, cette dernière est scannée par une fenêtre de dimension déterminée. La fenêtre parcourt l'image et son contenu est analysé pour savoir

s'il s'agit d'un visage ou non. Comme dit plus haut, les caractéristiques de Haar sont extraites pour effectuer la classification et de ce fait la représentation intégrale de l'image accélère l'analyse. Mais, pour une fenêtre de 24x24 pixels il y a 45396 caractéristiques de Haar, les traiter toutes prendrait beaucoup trop de temps pour une application en temps réel. Pour surmonter ce problème, une variante de la méthode de boosting Adaboost est utilisée. Ci-dessous Adaboost est brièvement présenté suivi de sa variante qui constitue le deuxième apport du travail de Viola & Jones.

Adaboost est une méthode d'apprentissage permettant de "booster" les performances d'un classifieur quelconque nommé "classifieur faible". L'idée est de faire passer les candidats à classifier à travers plusieurs classifieurs faibles, chacun étant entraîné en portant plus d'attention sur les candidats mal classifiés par le classifieur précédent.

Pour arriver à ce résultat des poids sont associés aux échantillons du set d'entraînement $((x_i, y_i) \ i = 1, \dots, m)$, tout d'abord de manière équilibrée :

$$w_i^0 = \frac{1}{m} \quad (5.3)$$

pour $i = 1, \dots, m$. Le 0 en exposant indique qu'il s'agit des poids initiaux.

Adaboost sert donc à booster un classifieur déjà existant et à priori chaque classifieur faible possède le même espace d'entrée. Dans la variante d'Adaboost de Viola & Jones, les classifieurs faibles $h_j \in H$ ont pour entrée une caractéristique de Haar différente. Adaboost s'apparente alors à une sélection de caractéristiques (feature selection).

Cette variante d'Adaboost est utilisée lors de l'apprentissage pour sélectionner les caractéristiques de Haar les plus à même de détecter un visage et permet ainsi de surmonter le problème du nombre élevé de caractéristiques de Haar existant pour une fenêtre de recherche.

C. Cascade

L'idée de base derrière le concept de Cascade est que parmi l'ensemble des candidats, c'est-à-dire l'ensemble des états de la fenêtre de recherche, une partie peut être éliminée sur base de l'évaluation de seulement quelques caractéristiques de Haar. Une fois cette élimination effectuée, les candidats restants sont analysés par des classifieurs forts plus complexes (utilisant plus de caractéristiques de Haar) demandant un plus grand temps de traitement. En utilisant plusieurs « étages » de ce type, le processeur évite d'effectuer des analyses lourdes en temps de calcul sur des échantillons pour lesquels il est rapidement

possible de se rendre compte qu'ils sont négatifs. Le processus de classification apparait alors comme une cascade de classifieurs forts de plus en plus complexes ou à chaque étage les échantillons classifiés négatifs sont sortis tandis que les échantillons classifiés positifs sont envoyés aux classifieurs suivants. Ceci est représenté à la figure 5.4.

Si le premier étage rejette un faux négatif, c'est un gros problème car il ne sera jamais récupéré par la cascade. Autrement dit c'est un visage qui ne sera pas détecté. Par contre, si le premier étage transmet un faux positif, il pourra toujours être éliminé aux étages suivants de la cascade. Ce petit raisonnement permet de mettre en évidence que les premiers nœuds constitutifs de la cascade peuvent se permettre d'avoir un taux de faux positifs élevés (de l'ordre de 40-50%) mais doivent absolument assurer un taux de détection maximum.

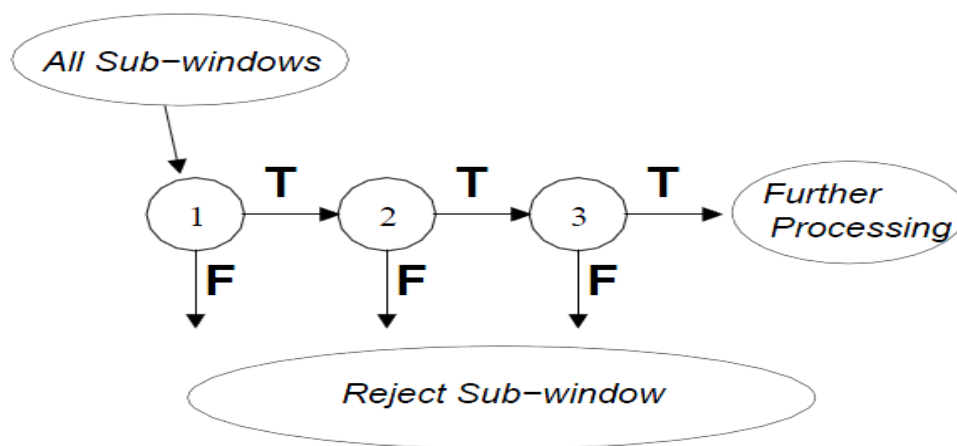


Figure 5.4 – Cascade de classifieurs forts. A chaque étage, uniquement les candidats classifiés positifs sont transmis à l'étage suivant.

Ce concept permet donc à l'algorithme de consacrer son temps à de longues analyses complexes uniquement lorsque cela en vaut la peine. Il s'agit à nouveau d'un mécanisme qui accélère la vitesse d'exécution de la méthode proposée par Viola & Jones.

5.1.1.2 Localisation de la bouche

Après la détection de visage avec l'utilisation de l'algorithme de Viola-Jones, il est possible d'extraire des zones à partir de la géométrie du visage trouvé, où les points devraient être. Ces zones sont les entrées relatives à l'extraction de la ROI.

Au moment où il est exécuté en utilisant la teinte distincte des lèvres. La lumière se reflète sur les lèvres et ce point est récupéré par une valeur de teinte définie. Contrairement

aux autres méthodes, cette méthode n'est pas indépendante de lumière, ainsi l'intensité et la direction de la lumière peut influencer les résultats (Pai et al. 2006).

Un visage humain typique suit un ensemble de normes anthropométriques, qui ont été utilisés pour affiner la recherche d'une caractéristique faciale particulière pour des régions plus petites de visage. Nous utilisons les étapes génériques suivantes pour la détection des caractéristiques faciales et l'extraction à partir de l'image du visage localisée (Khandait et al. 2009):

- 1) Pour une image couleur, la convertir en image en niveaux de gris. Réglez l'intensité des deux types d'images.
- 2) Appliquer projection horizontale pour trouver frontière gauche et droite de visage. Appliquer projection verticale pour trouver la frontière supérieure et inférieure de visage où trouver région d'intérêt d'une image.
- 3) Trouvez le gradient de la ROI de l'image détectée en utilisant Sobel / Prewitt opérateur de détection des frontières et ensuite prenez la partie inférieure du visage et prenez sa projection verticale pour obtenir la bouche.
- 4) Dessiner zone rectangulaire sur la composante caractéristique détectée.

5.1.1.3 Extraction des paramètres visuels

Dans cette étude l'extraction des caractéristiques vidéo est effectuée avec le DCT (Rodomagoulakis 2008). Il existe plusieurs types de caractéristiques qui peuvent être utilisées pour chiffrer les informations présentes dans une image. Nous avons appliqué une version modifiée de la DCT qui utilise les données contenues dans une image pour la compresser. Par exemple, la compression de l'image en format JPEG utilise cette méthode. La compression des données disponible dans l'image permet de rendre le travail de l'algorithme d'apprentissage plus facile. En plus la DCT est utilisée dans le domaine d'authentification et vérification du locuteur (Sanderson and Paliwal 2002). Cette étape se déroule en deux phases : La première est la phase de découpage de l'image, résultant de la phase de prétraitement, en sous-images. Ensuite, la seconde phase qui est l'extraction de vecteurs de caractéristiques consiste à appliquer la DCT. Ces étapes seront détaillées dans les paragraphes suivants.

5.1.1.3.1 Découpage de l'image

Le découpage de l'image consiste à subdiviser l'image en entrée en sous-images de dimension fixe qui se chevauchent dans les deux directions, l'axe des y et l'axe des x. Le découpage de l'image se passe de la manière suivante : la première sous-image de dimension $N \times N$ pixel se trouve aux coordonnées $(0, 0)$, (N, N) de l'image d'entrée. La seconde sous-image chevauche la première sous-image d'une superposition de c pixels en direction de l'axe des x. Donc, la seconde correspond à la sous-image de coordonnées $(N - c, 0)$, $(2N - c, N)$. La troisième sous-image a une superposition dans la direction de l'axe des y avec la première sous-image. La troisième correspond à la sous-image de coordonnées $(0, N - c)$, $(N, 2N - c)$. Cette procédure se répète récursivement jusqu'à ce que toute l'image en entrée soit traitée. Le résultat d'un tel découpage est montré par la figure 5.4. Dans le cadre de ce projet, la dimension des sous-images a été fixée à 16×16 pixels avec une superposition de 8 pixels.

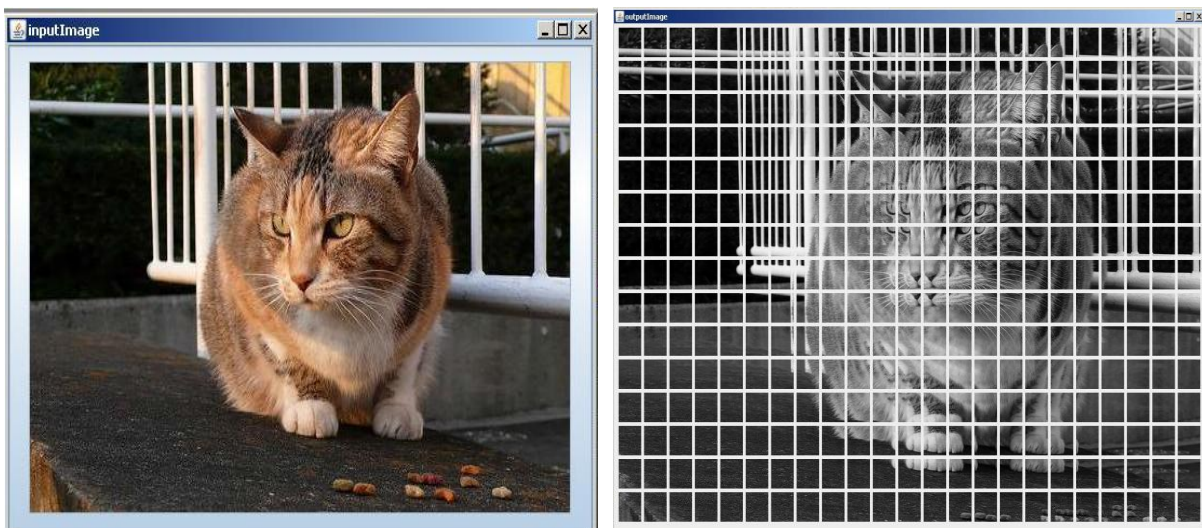


Figure 5.4 – Découpage de l'image de l'histogramme.

5.1.1.3.2 Extraction de caractéristiques

La phase d'extraction de caractéristiques présente un passage de l'image du domaine spatial au domaine fréquentiel. Comme mentionné au début de ce chapitre, nous avons choisi d'utiliser la DCT. Cette méthode consiste à présenter chaque image comme une matrice de vecteurs ou chaque vecteur correspond à une sous-image résultant d'un découpage régulier de l'image. Ces vecteurs sont des coefficients qui correspondent à des combinaisons linéaires de fonctions cosinusoïdales, ces fonctions sont la base du domaine fréquentiel.

Plus formellement, étant donnée une image qui est présentée par une matrice de sous-images de dimension $N \times N$, ces sous-images sont le résultat du découpage précédemment expliqué. Pour chaque image I un vecteur de DCT est extrait. DCT transforme chaque composante de couleur en coefficients DCT en utilisant l'équation suivante (Gupta and Garg 2012):

$$F(u, v) = \frac{1}{\sqrt{MN}} \alpha(u) \alpha(v) \sum_{x=1}^M \sum_{y=1}^N f(x, y) \cos \left[\frac{(2x+1)u\pi}{2M} \right] \cos \left[\frac{(2y+1)v\pi}{2N} \right] \quad (5.4)$$

avec,

- u est la fréquence spatiale horizontale,

- v est la fréquence spatiale verticale,

- $f(x, y)$ est la valeur de pixel aux coordonnées (x, y) ,

- $F(u, v)$ est le coefficient de DCT au point de coordonnées (u, v) , elle est dimensionnée de $M \times N$, et $\alpha(\bullet)$ est définis comme suit:

$$\alpha(w) = \begin{cases} \frac{1}{\sqrt{2}}, & w=1 \\ 1, & \text{otherwise;} \end{cases} \quad (5.5)$$

Cette matrice DCT(I) est une matrice des coefficients qui est définie à l'aide de fonctions cosinusoidales. Ces fonctions constituent la base du domaine fréquentiel. La figure 5.5 présente ces fonctions de base à deux variables $v, u = 0, 1, 2, \dots, 7$.

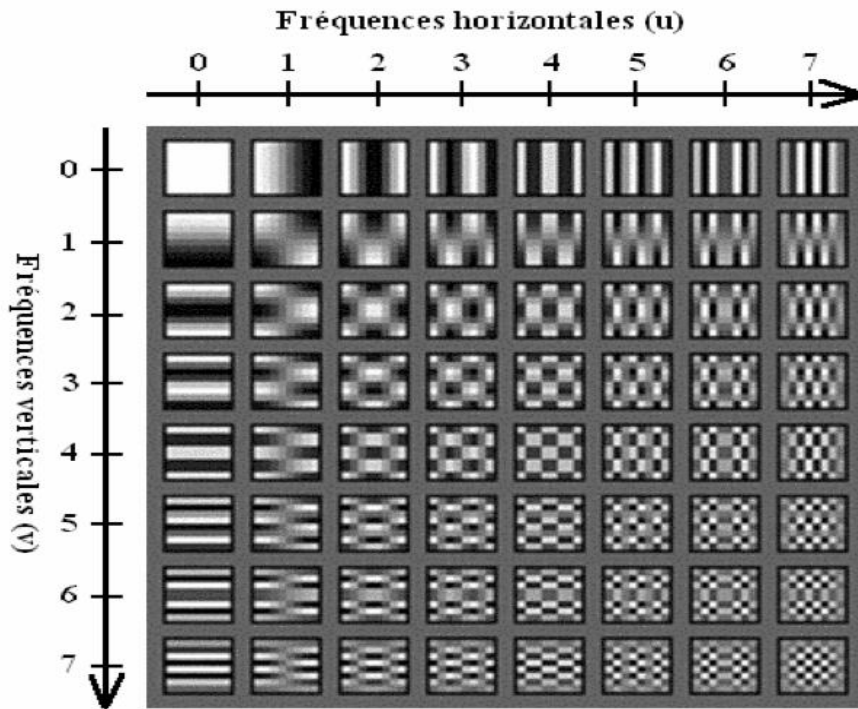


Figure 5.5 – Exemple de fonctions de base de DCT qui forme le domaine fréquentiel.

Afin d'obtenir un vecteur DCT qui est la transformée d'une sous-image I donnée, le parcours en zigzag est appliqué à la matrice $DCT(I)$. La figure 5.6 montre l'ordre dans laquelle la matrice $DCT(I)$ est parcourue selon le parcours en zigzag.

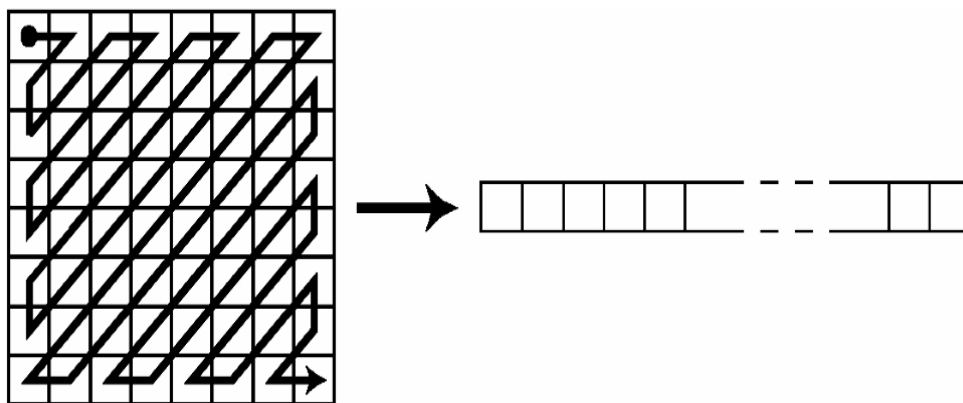


Figure 5.6 – Parcours en zigzag d'une matrice de dimension 8×8 .

Les informations les plus importantes pour représenter l'image se trouvent dans les premiers coefficients. En utilisant uniquement les premiers coefficients et la transformation DCT inverse, il est possible de régénérer une image ayant presque le même rendu visuel que

la sous-image I d'origine. Généralement, la différence entre les deux images est totalement imperceptible. Afin de compresser les informations à l'aide de la DCT, une sous-image est présentée à l'aide des M premiers coefficients de vecteur DCT.

5.1.2 Traitement acoustique

Afin de pouvoir reconnaître le contenu d'un signal de parole correctement, il est nécessaire d'en extraire des paramètres caractéristiques et pertinents pour la reconnaissance. Le signal de parole n'est pas directement utilisable à cause de sa grande complexité (grande diversité d'information) et de son caractère redondant. Le but de la paramétrisation est d'extraire l'information pertinente pour la tâche proposée.

La première étape de la paramétrisation acoustique consiste à découper le signal de parole en fenêtres de taille fixe (variable de 20 ms à environ 40 ms) réparties de façon uniforme le long du signal (toutes les 10 ms). La taille des fenêtres est choisie en considérant que les propriétés du conduit vocal peuvent être considérées comme invariables sur une petite durée égale à la taille de la fenêtre. Le signal audio est donc considéré comme stationnaire sur la durée de la fenêtre. Pour ce faire, plusieurs techniques d'analyse du signal et d'extraction de paramètres peuvent être utilisées, mais dans le cadre de cette étude, seuls les paramètres acoustiques issus de systèmes de RAP de type énergie en sous-bande et de type RASTA-PLP seront utilisés. L'espace de représentation du signal de parole ainsi obtenu est muni d'une mesure de distance euclidienne adaptée à ces paramètres acoustiques. Cette mesure de distance est utilisée comme critère de similarité au sein de l'algorithme de comparaison du système de RAP considéré. L'extraction de paramètres acoustiques différents est un élément essentiel de cette thèse.

5.1.2.1 Analyse RASTA-PLP

Afin d'augmenter la robustesse des paramètres PLP, on peut envisager l'analyse spectrale relative RASTA (RelAtive SpecTrAl), présentée par (Hermansky and Morgan 1994) comme une façon de simuler l'insensibilité de l'appareil auditif humain aux stimuli à variation temporelle lente. Cette technique traite les composantes de parole non linguistiques, qui varient lentement dans le temps, dues au bruit convolutif (log-RASTA) et au bruit additif (J-RASTA). En pratique, RASTA effectue un filtrage passe-bande sur le spectre logarithmique ou sur le spectre compressé par une fonction non linéaire. L'idée principale est de supprimer les facteurs constants dans chaque composante du spectre à court-terme avant l'estimation du

modèle tout-pôle. L'analyse RASTA est souvent utilisée en combinaison avec les paramètres PLP (Hermansky and Morgan 1994). Les étapes d'une analyse RASTA-PLP sont décrites dans la figure 5.7. Les étapes grisées sont celles qui font la spécificité du traitement RASTA. La différence entre RASTA et J-RASTA se situe au niveau du logarithme (4ème étape) : $\ln(x)$ pour RASTA et $\ln(1 + Jx)$ pour J-RASTA.

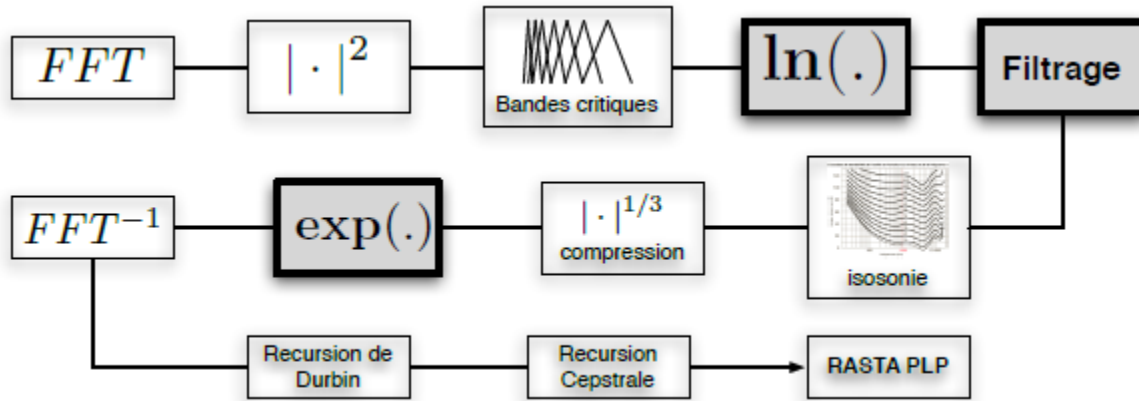


Figure 5.7 – Analyse RASTA PLP.

▪ Utilisation des dérivées premières et secondes

Dans les systèmes de reconnaissance actuels, il est très courant de compléter un jeu de paramètres par les dérivées premières (Δ) et secondes ($\Delta\Delta$) de ces paramètres. Les dérivées permettent d'inclure des caractéristiques dynamiques des paramètres acoustiques (vitesse et accélération). Le calcul des dérivées se fait sur des fenêtres centrées sur la trame analysée, ce qui assure la cohérence des informations présentes dans le vecteur. L'utilisation de ces Δ et $\Delta\Delta$ est précisément un cas de concaténation de paramètres acoustiques. Une méthode de combinaison complète de modèles utilisant un jeu de paramètres (PLP), les Δ et les $\Delta\Delta$ de ces paramètres est présentée dans (Misra et al. 2003). Chaque type de paramètres (statiques, Δ et $\Delta\Delta$) sont combinés de toutes les manières possibles pour former 7 jeux de paramètres acoustiques utilisés pour apprendre 7 modèles acoustiques différents, dont les probabilités sont ensuite combinées.

5.1.2.2 La quantification vectorielle

La quantification scalaire consiste à représenter une valeur d'un échantillon de signal pas forçement audio avec une précision réduite, par exemple la représenter avec une valeur

appartenant à un ensemble plus petit que l'ensemble original. C'est le cas typique de la conversion analogique/digitale.

Lorsque ce principe est appliqué par bloc d'échantillons (*vecteurs*), on peut parler de quantification vectorielle. La quantification vectorielle est alors une généralisation de la quantification scalaire. Mais, pendant que la quantification scalaire est dans sa forme la plus simple juste une conversion analogique/digitale, la quantification vectorielle est une méthode de codage/compression puissante. Elle est souvent utilisée dans les télécommunications pour le codage de la source, ou dans la compression des données notamment dans la compression des images. Elle est aussi un puissant outil de classification. La quantification vectorielle est définie par un doublet : un ensemble de vecteurs représentatifs appelés mots $C = c_1 c_2 \dots c_M$ qui forme un dictionnaire (*codebook* en anglais) et un critère de distorsion $d(.,.)$ (Voir la figure 5.8).

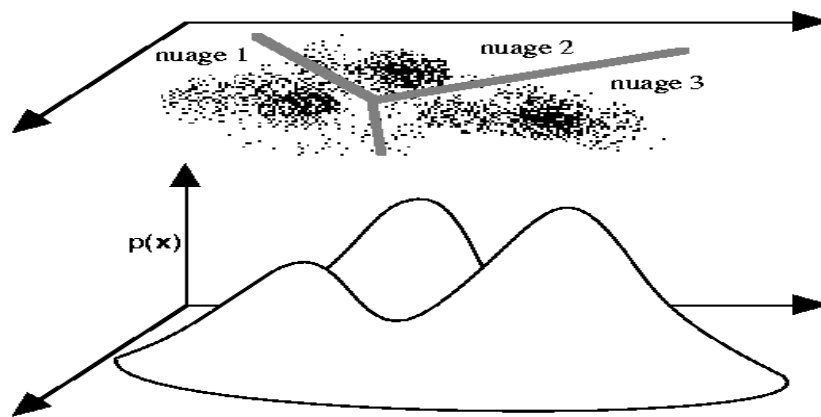


Figure 5.8 – Distribution de probabilités, un échantillon de points associés, et un découpage en nuages (clusters).

L'algorithme k-means est l'algorithme de clustering le plus connu et le plus utilisé, du fait de sa simplicité de mise en œuvre. Il partitionne les données d'un signal en K clusters. Contrairement à d'autres méthodes dites hiérarchiques, qui créent une structure en « arbre de clusters » pour décrire les groupements, k-means ne crée qu'un seul niveau de clusters. L'algorithme renvoie une partition des données, dans laquelle les objets à l'intérieur de chaque cluster sont aussi proches que possible les uns des autres et aussi loin que possible des objets des autres clusters. Chaque cluster de la partition est défini par ses objets et son centroïde. Le k-means est un algorithme itératif qui minimise la somme des distances entre chaque objet et le centroïde de son cluster. La position initiale des centroïdes conditionne le résultat final, de sorte que les centroïdes doivent être initialement placés le plus loin possible les uns des autres

de façon à optimiser l'algorithme. Cette méthode change les objets de cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés, sous réserve qu'on ait choisi la bonne valeur K du nombre de clusters. Les principales étapes de l'algorithme k-means sont :

1. Choix aléatoire de la position initiale des K clusters.
2. (Ré-) Affecter les objets à un cluster suivant un critère de minimisation des distances (généralement selon une mesure de distance euclidienne).
3. Une fois tous les objets placés, recalculer les K centroïdes.
4. Répéter les étapes 2 et 3 jusqu'à ce que plus aucune réaffectation ne soit faite.

5.1.3 Moteur de reconnaissance GA/HMM

Apprendre un HMM c'est ajuster les paramètres du modèle de manière à maximiser un certain critère. Différents critères sont disponibles dans la littérature. Nous n'allons pas tous les recenser, mais nous allons présenter une des plus importants et les plus couramment utilisés (Makhlouf et al. 2016).

Les GA constituent une large famille d'algorithmes statistiques, développés par (Holland 1983) et approfondis par (Goldberg 1999). Nous étudions ici comment une recherche de gradient « l'algorithme de Baum-Welch (BW) » peut être combinée avec des GA afin d'apprendre les HMM. Trois coopérations possibles entre les deux algorithmes sont étudiées, le GA peut être utilisé pour trouver un meilleur point de départ pour la recherche de gradient. Finalement, dans la méthode GA/HMM, le GA recherche automatiquement les trois probabilités, pour plus de détails voir chapitre 4.

Pour mémoire, nous rappelons que l'opérateur de mutation consiste à modifier chaque coefficient de chaque modèle avec la probabilité P_m et que l'opérateur d'optimisation consiste à appliquer N_{BW} itérations de l'algorithme de Baum-Welch à chaque individu.

5.1.4 La fusion audiovisuelle

L'objectif d'un système de reconnaissance audio visuelle est de combiner au mieux les performances de deux systèmes audio et vidéo afin d'améliorer les performances de reconnaissance de la parole, en particulier en présence de bruit. Classiquement, on distingue deux types de fusion: la fusion des paramètres et la fusion des scores.

5.1.4.1 Fusion des paramètres

Cette fusion est réalisée au moment de la paramétrisation des signaux audio et vidéo. Une fois les paramètres de chaque modalité sont extraits, les vecteurs audio \mathbf{o}^A et vidéo \mathbf{o}^V de dimension d^A et d^V respectivement, sont concaténés à chaque instant t pour ne former qu'un seul vecteur de paramètres audio visuels de dimension $d^A + d^V$. Dans les étapes suivantes de la chaîne de reconnaissance de la parole (estimation des paramètres, décodage, évaluation), aucune modification n'est nécessaire.

5.1.4.2 Fusion des scores

La fusion de scores ou de décision est possible lorsque l'on dispose de systèmes séparés (ici, audio et vidéo) et que leur fusion est réalisée au moment de la décision, par combinaison de leurs scores respectifs. Des poids différents peuvent être affectés à chaque système (ou parties de ces derniers) afin de privilégier l'une ou l'autre des deux modalités. Dans le cas de système de reconnaissance ou les unités sub-lexicales (de type phone, par exemple) sont modélisées par des HMM et GA/HMM, cette fusion peut avoir lieu à différents niveaux qui sont l'état ou le phone ou le mot ou encore la phrase. Lorsque la fusion est effectuée à chaque état, elle est dite synchrone, sinon elle est asynchrone.

Plusieurs stratégies de fusion de décision ont été testés (produits, des sommes, minimum, maximum, vote ...) et tout montrer une amélioration significative des résultats par rapport à la considération d'une seule modalité, qui nous mener à se concentrer dans ce travail sur l'utilisation du le modèle de la fusion séparée, c.à.d. la fusion des scores provenant de chaque reconnaiseur GA/HMM. Leurs jeux de log-vraisemblance peuvent être combinés en utilisant les pondérations qui reflètent la fiabilité de chaque flux particulier, les scores combinés prennent alors la forme suivante (l'islam et Rahman 2010):

$$\log P(\mathbf{o}^{AV} | \lambda) = w_A \log P(\mathbf{o}^A | \lambda_A) + w_V \log P(\mathbf{o}^V | \lambda_V) \quad (5.6)$$

Où λ_A et λ_V sont les GA/HMMs acoustique et visuels respectivement et $\log P(\mathbf{o}^A | \lambda_A)$ et $\log P(\mathbf{o}^V | \lambda_V)$ sont leurs log-vraisemblance. La fiabilité de chaque modalité peut être calculée par le plus approprié et le meilleur dans la performance (l'islam et Rahman 2010), la différence moyenne entre le log-vraisemblance maximum et les autres, peut être trouvé par :

$$R_s = \frac{1}{C-1} \sum_{i=1}^C (\max \log P(\mathbf{o} | \lambda^j) + \log P(\mathbf{o} | \lambda^i)) \quad (5.7)$$

Où C est le nombre de classes étant considéré pour mesurer la fiabilité de chaque modalité et $s \in \{A, V\}$. Après cela, nous pouvons calculer le poids d'intégration de la fiabilité audio A mesuré par:

$$w_A = \frac{R_A}{R_A + R_V} \quad (5.8)$$

Où R_A et R_V sont la mesure de fiabilité des sorties des GA/HMM acoustique et visuelle respectivement, et le facteur de pondération de la modalité visuelle peut être trouvée par la relation:

$$w_A + w_V = 1 \quad \text{for} \quad 0 < w_A, w_V < 1 \quad (5.9)$$

Le poids W permet de donner plus d'importance à une modalité ou à l'autre. Pour chaque système, W peut être choisi constant ou variable. Généralement, il dépend du rapport signal à bruit. Des travaux dans (Makhlouf et al. 2013a) montrent que les performances du système de reconnaissance audio visuelle sont meilleures pour un paramètre W dynamique.

5.2 Conclusion

Nous avons décrit, dans ce chapitre notre système proposé de reconnaissance de la parole audiovisuelle. Ainsi, nous avons abordé la fusion d'informations acoustiques et visuelles pour la RAP.

Nous nous intéressons dans le chapitre suivant à la description du système de reconnaissance audiovisuelle réalisé à base des HMM, et le modèle hybride GA/HMM. Egalement, la mise en œuvre de système qui a été appliqué sur deux corpus audiovisuels différents.

Réalisation

6

Comme tout modèle qui doit être expérimenté, le présent chapitre constitue un cadre d'expérimentation et d'argumentation du chapitre précédent.

Nous allons présenter dans ce chapitre les expérimentations que nous avons menées pour aller vers une collaboration des processus de reconnaissance automatique de la parole et de reconnaissance visuelle de la parole.

Nous présentons à présent les différents tests que nous avons effectués afin d'analyser les mérites des méthodes retenues dans le chapitre précédent. Les plus performantes seront validées par comparaison avec des algorithmes d'apprentissage classiquement utilisés dans la littérature.

6.1 Architecture général du système de reconnaissance

Dans nos expérimentations nous évaluons la performance des modèles audio-visuels HMM appris en utilisant les GA par rapport à l'apprentissage standard des HMM en utilisant une estimation du maximum de vraisemblance (EM).

Comme l'a fait remarquer (Alpaydin 2004), nous devons toujours garder à l'esprit que les conclusions que nous tirons de l'analyse est conditionnée par l'ensemble de données. Ainsi, nous ne comparons pas les modèles et les algorithmes d'apprentissage d'une manière indépendante de domaine. Tout résultat nous présentons n'est valable que pour l'application particulière de AVASR et pour l'ensemble de données utilisé. Comme indiqué dans le Non déjeuner théorème de gratuit (Wolpert and Macready 1997) il n'y a pas une telle chose comme le "meilleur" algorithme d'apprentissage en général. Pour n'importe quel algorithme d'apprentissage, il y aura un ensemble de données où il est très précis et une autre où il est très faible. Ainsi, nos résultats ne sont valables que pour l'application particulière d'AVASR et en particulier pour les corpus de données que nous avons choisis. Ces corpus de données sont discutés par la suite.

Dans notre travail nous allons appliquer l'algorithme de clustering K-means sur les BDD audiovisuelles CUAVE et notre propre BDD arabe (AVARB), les résultats de cette opération seront en suite introduits au HMM pour faire l'apprentissage. Afin d'augmenter la performance du système de reconnaissance proposé, nous avons utilisé une nouvelle méthode basée sur l'hybridation des deux paradigmes HMM et GA.

Pour réaliser ce système de reconnaissance, il fallait :

- Détection de visage et Localisation des lèvres dans les scènes vidéo en utilisant la méthode Viola-Jones.
- Extraction de paramètres acoustiques avec la méthode RASTA-PLP.
- Extraction de paramètres visuels avec la méthode DCT.
- Réaliser une quantification vectorielle et dégager des classes, en utilisant l'approche suivante : K-means.
- Phase d'apprentissage en utilisant les modèles HMM, et GA/HMM.
- Comparaison des taux de reconnaissance obtenus pour tirer la méthode la plus performante de reconnaissance.

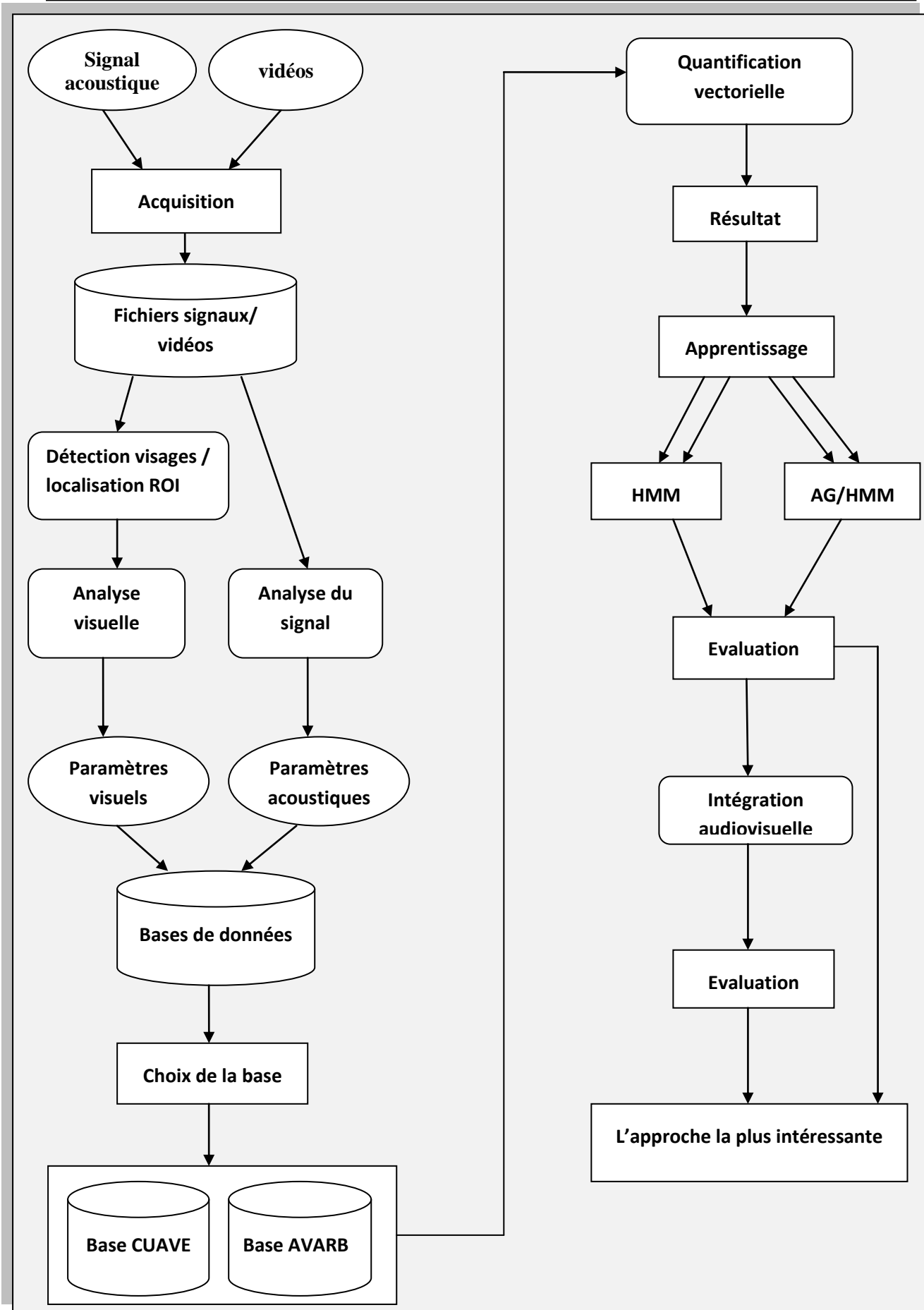


Figure 6.1 – Architecture générale du système proposé.

6.2 Base de données utilisée

6.2.1 Les bases de données audiovisuelle arabe

Dans notre travail nous avons utilisé notre propre base de données audiovisuelle de parole arabe : cette base de données multi-locuteurs a été enregistrée dans un milieu réel (une salle de cours très bruyante), Nous visons de plus la diversité des données pour un apprentissage bien amélioré, les vidéos sont capturées La à une distance moyenne égale à 16.5 cm avec une résolution de 690×450 pixel et à 30 trames/sec et avec des variations de pose (vue de profil, de face) pour un ensemble de 18 locuteurs (16 garçon et 2 filles) sauvegardées avec l'extension « .avi », alors que les fichiers audio sont sauvegardé avec l'extension « .wav », l'échantillonnage standard après des testes réalisés au sein de notre laboratoire est 16 KHz MONO (à un canal unique) car il est optimal de calculer les coefficients issus d'un signal acoustique à paramètres unique.

Notre base AVARB contient 2 corpus, le premier corpus contient des prononciations des chiffres arabes isolés (de zéro (0) à neuf (9)), alors que le deuxième corpus contient un ensemble commandes en arabe (25 mots), comme il est illustré dans le tableau 6.1 :

<i>Corpus chiffre</i>				<i>Corpus commandes</i>			
<i>code</i>	<i>Prononciation</i>	<i>Ecriture arabe</i>	<i>glossaire français</i>	<i>code</i>	<i>Prononciation</i>	<i>Ecriture arabe</i>	<i>Glossaire français</i>
1	Siffer	صفر	Zéro	1	Marhaban	مرحبا	Bienvenue
2	Wahed	واحد	Un	2	Ebdaa	ابداً	Démarrer
3	Ithnani	اثنان	Deux	3	Iqaf	إيقاف	Arrêter
4	Thalatha	ثلاثة	Trois	4	Eftah	افتح	Ouvrir
5	Arbaa	أربعة	Quatre	5	Arliq	أغلق	Fermer
6	Khamssa	خمسة	Cinq	6	Takbir	تكبير	Agrandir
7	Sitta	سنة	Six	7	Tasrir	تصغير	Réduire
8	Sabaa	سبعة	Sept	8	Tashril	تشغيل	Fonctionnement
9	Thamania	ثمانية	Huit	9	Elraa	إلغاء	Annuler
10	Tissaa	تسعة	Neuf	10	Bahth	بحث	Recherche
				11	Ekhtiyar	اختيار	Sélection
				12	Aaouda	عودة	Retour
				13	Edhar	إظهار	Affichage
				14	Qaima	قائمة	Liste
				15	Mouafiq	موافق	Accepter

				16	Doukhoul	دخول	Se connecter
				17	Khourouj	خروج	Quitter
				18	Nasskh	نسخ	Copier
				19	Qass	قص	Couper
				20	Lasq	لصق	Coller
				21	Tarjama	ترجمة	Traduire
				22	Khasaiss	خصائص	Propriétés
				23	Tatbiq	تطبيق	Application
				24	Tenfid	تنفيذ	Exécution
				25	Tahmil	تحميل	Chargement

Table 6.1 – Notre deux corpus proposés de chiffres et commandes arabes.

Les locuteurs sont de différentes régions dialectes algériennes, et chaque locuteur prononce chaque mot 9 fois avec différentes modes de prononciation (normal, lente, et rapide). Dans notre corpus basic qui contient que des mots isolés, la taille de chaque enregistrement est 2 secondes qui est un temps suffisant pour prononcer un mot lentement en arabe. La figure suivante montre quelques trames de notre base AVARB :



Figure 6.2 – quelques exemples de trames de notre base audiovisuelle AVARB.

6.2.2 La base de données CUAVE

Elle se compose de 36 locuteurs, 19 hommes et 17 femmes, poussant chiffres isolés et continue. Les vidéos des orateurs sont enregistrées en profil frontal, et pendant le mouvement. La base de données CUAVE contient environ 3 heures de parole enregistrées par une caméra Mini DV. La Vidéo a ensuite été compressée en MPEG-2 fichiers (audio stéréo à un taux d'échantillonnage 44 kHz, 16-bit). Il comprend également des fichiers audio vérifiés pour la synchronisation (taux de mono de 16 kHz, 16-bit) et des fichiers d'annotation (Patterson et al. 2002).



Figure 6.3 – Exemples de trames de la base CUAVE.

6.3 Validation du système

Une étape importante et très consommatrice en temps de développement d'un système de transcription est l'expérimentation. Il s'agit de tester les différents modules du système pour ajuster leurs paramètres. De bonnes valeurs de paramètres peuvent apporter beaucoup au niveau du taux de reconnaissance. Chaque module a ses propres paramètres et il est nécessaire de les ajuster de façon plus ou moins optimale. Ajuster les paramètres de tous les modules en même

temps est une tâche irréalisable puisque le nombre de combinaisons de paramètres à tester serait très grand et donc le temps d'expérimentation serait énorme. En général, l'expérimentation est effectuée module par module pour économiser du temps. Puis le système complet est testé également.

6.4 Traitement des données audiovisuelles

6.4.1 Séparation audiovisuelle

Une fois l'enregistrement des séquences vidéo du locuteur est réalisé à l'aide d'un appareil photo numérique Sony Cyber-Shot DSC-W530 14.1 Méga Pixel avec un zoom optique 4x grand-angle Zoom optique et 2.7 pouces moniteur LCD. La première opération consiste à la séparation des deux flux audio et vidéo. Le flux audio est extrait sous forme d'un signal à l'aide du logiciel Gold Wave de l'extension ".wav", et à partir du flux vidéo on extrait, à l'aide du logiciel BPS, des images fixes de la séquence. On passe ensuite à la construction des bases de données audio et vidéo.

6.4.2 Données visuels

Après la détection de visage avec l'utilisation de l'algorithme de Viola-Jones (voir l'exemple dans la figure 6.4), nous avons localisé la région de la bouche de chaque locuteur comme il est illustré dans les exemples dans la figure 6.5.

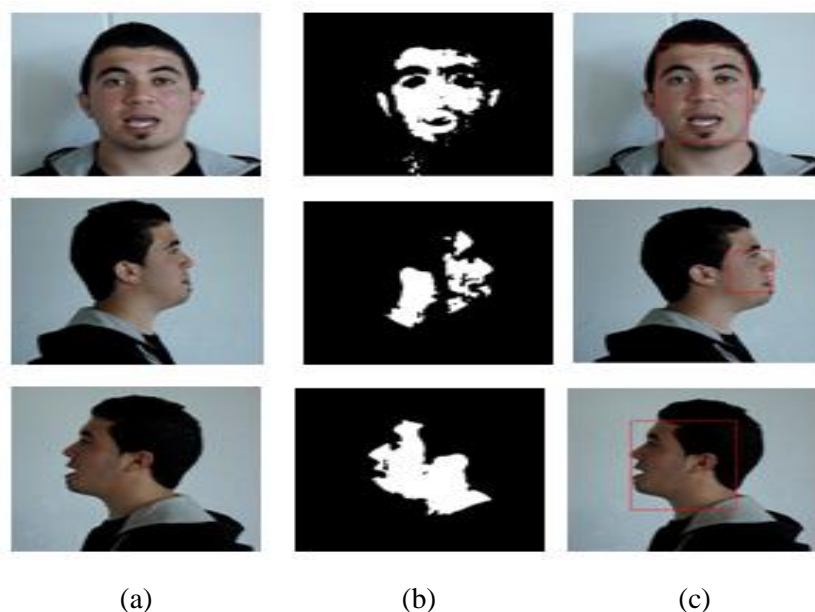


Figure 6.4 – Un exemple de détection de visage : (a) image originale (b) détection de peau avec suppression de bruit (c) résultat de détection de visage.



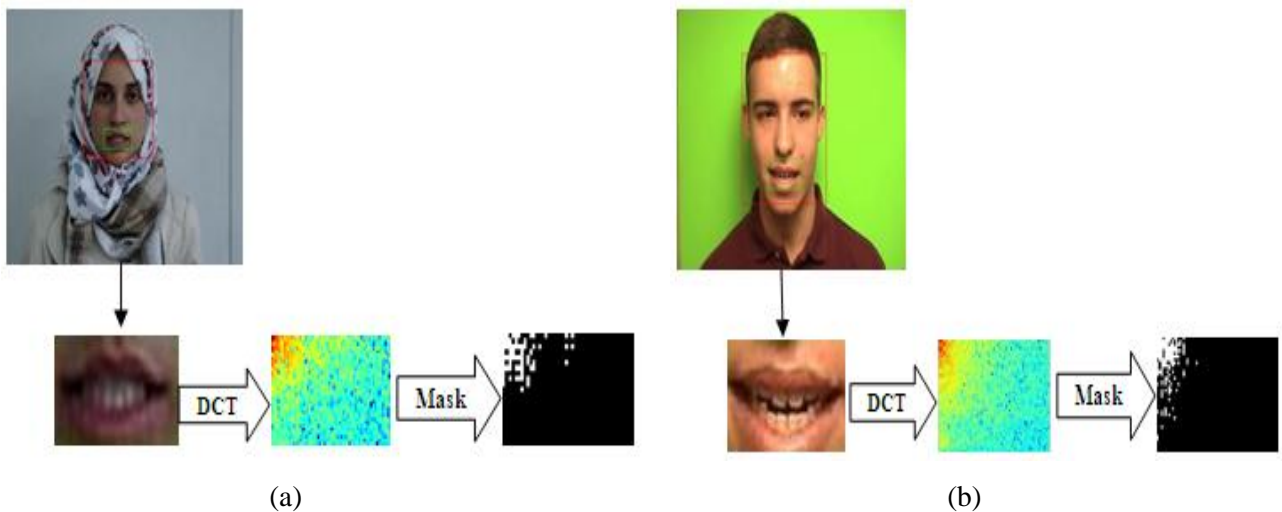
(a)



(b)

Figure 6.5 – Exemples de la région de la bouche détectée à partir de : (a) la base AVARB (b) la base CUAVE.

L'extraction des caractéristiques vidéo est effectuée avec la DCT. Les vecteurs d'entrées sont formés des coefficients basses fréquences qui se trouvent dans le coin supérieur gauche de la matrice résultante comme montré par la figure 6.6. Dans cette figure, nous avons conservé uniquement les 100 premiers coefficients de hautes amplitudes d'une image, donc le vecteur visuel dans ce cas est composé des 100 éléments. Le nombre de coefficients hautes amplitudes conservés après la transformation par la DCT est choisi de manière à conserver un maximum d'énergie totale dans les coefficients hautes amplitudes qui sera suffisant pour reconstituer les caractéristiques principales de l'image (Makhlouf et al. 2013a ; 2013b). L'énergie totale E de l'image est calculée (théorème de Parseval, à partir des coefficients de la DCT).



(a)

(b)

Figure 6.6 – Le processus de sélection des coefficients DCT avec un échantillon à partir: (a) la base AVARB (b) la base CUAVE.

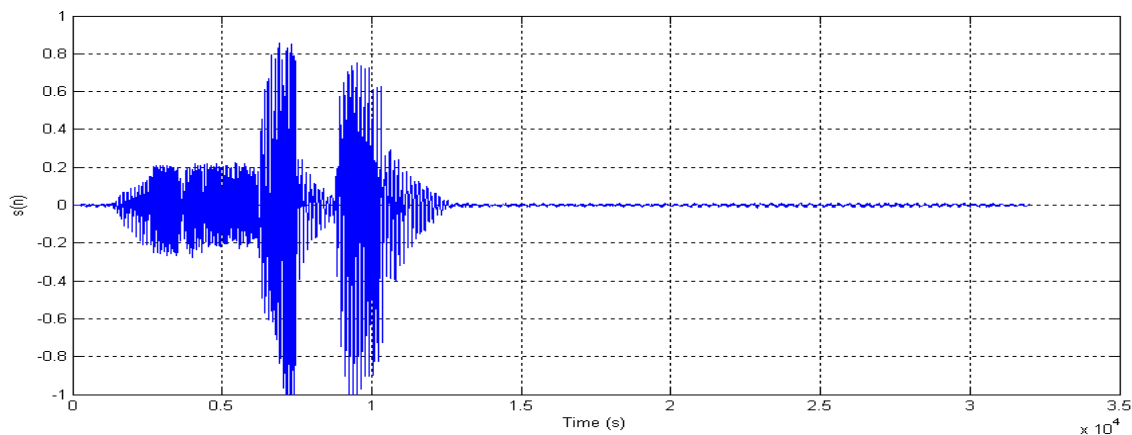
L'idée principale de l'algorithme pour encoder l'image par la DCT est de ne pas utiliser la totalité des coefficients (310500 coefficients), afin de limiter la taille mémoire et les calculs nécessaires pour l'entraînement et la reconnaissance par les modèles proposés dans notre système. Dans notre travail nous avons gardé les cent (100) premiers coefficients pour représenter l'image.

6.4.3 Données acoustiques

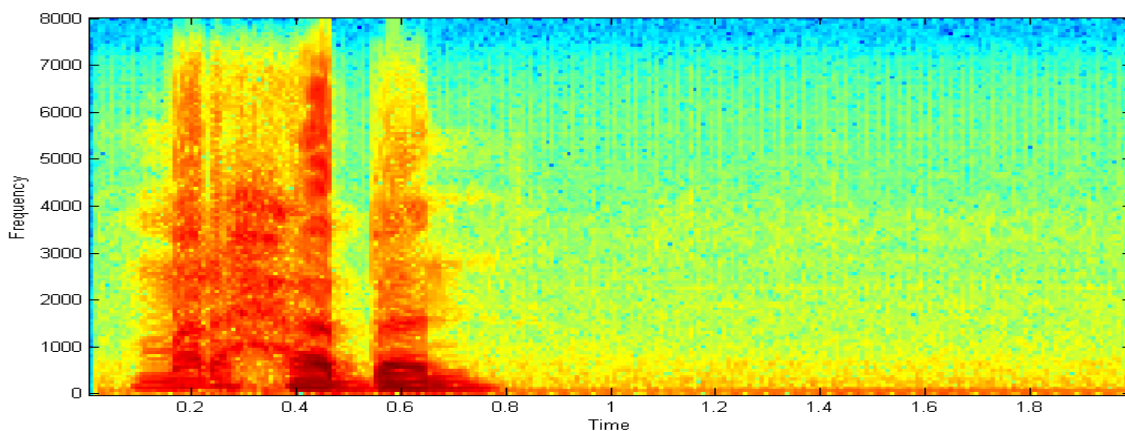
L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes. Pour la réalisation de cette phase d'extraction des paramètres, nous avons utilisé la technique RASTA-PLP (comme il est mentionné dans le chapitre 5).

Pour chaque signal vocal et avec la méthode RASTA-PLP, on extrait 9 paramètres du signal acoustique de 98 trames d'échantillonnage à 16kHz, et d'une taille de fenêtre 0.025 secondes et d'un pas de 0.010 secondes. En intégrant la première et la deuxième dérivé des paramètres, on obtient des matrices de 27 paramètres organisé comme suit : Pour chaque corpus multilocuteur, si on prend le corpus commandes par exemple, on a pour les tests 25 occurrences de commandes vocal répétés 3 fois chacune de 18 locuteurs, donc $25 \times 3 \times 18 \times 27 = 36450$ et 98 trames est la taille de la matrice, même pour l'apprentissage, sauf que l'ordre de l'occurrence entre les locuteurs sont organisés les uns après les autres.

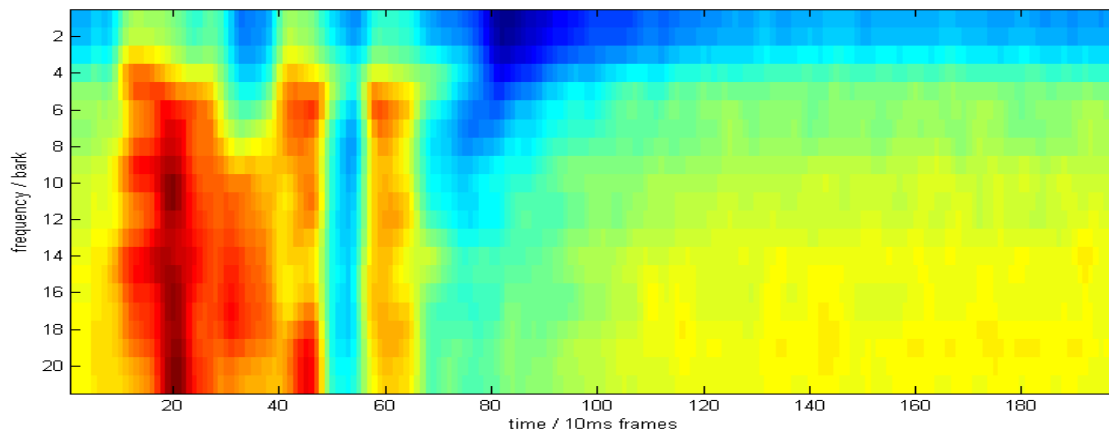
Un exemple de calcul de quelques paramètres du signal de parole utilisant cette méthode d'extraction est illustré par la figure 6.7.



(a)



(b)



(c)

Figure 6.7 – Exemple d'un signal de parole du mot arabe "/ marhaban /" (a) son spectrogramme (b) et l'ensemble des caractéristiques spectrales RASTA-PLP (c).

6.5 Modélisation par GA/HMM

Après avoir défini formellement notre approche, il est nécessaire de la tester afin de la valider.

6.5.1 Résultats obtenus et discussion

Cet algorithme optimise à la fois les paramètres (probabilités) de HMM. Il repose sur une recherche génétique d'un bon modèle parmi une population hétérogène de HMM et une optimisation par un algorithme de gradient (Baum-Welch).

Pour l'apprentissage, nous avons utilisé un nombre m des HMM de type gauche-droite avec un nombre m d'états dont m est le nombre des mots dans chaque corpus, afin de représenter les m classes.

6.5.1.1. Expérimentations avec des bruits sonore et visuel additifs

Dans cette section, nous présentons les résultats des expériences menées en utilisant des signaux audio et vidéo bruyants.

Nous avons utilisé deux types de bruit vidéo pour examiner la robustesse de notre système AVASR contrairement à audio seule ASR. Les types de bruit que nous avons implémenté sont la diminution des trames, et le bruit aléatoire gaussien. Ces types de bruit imitent des scénarios typiques où il existe une distorsion soit depuis un appareil photo défectueux ou d'un signal de transmission vidéo. De plus, La diminution de la fréquence de trames (FPS) et le bruit de bloc peut simuler la perte d'information à la suite des mouvements abrupts de la bouche et la parésie d'une partie de la bouche ou des lèvres qui peut être causée par un problème de santé. Par conséquent, ce type de bruit présente un intérêt dans des environnements d'assistance envahissants.

Le taux de reconnaissance est affecté par la qualité du signal (i.e. diminution du rapport signal sur bruit (Signal-to-Noise Ratio (SNR))). Nous examinons d'abord le cas de d'image perdue (**Frame-Dropped**). La fréquence des trames initiale était 30 fps, donc nous avons réduit à 15, 5 et 1fps puis l'interpolée de nouveau à 100fps afin de correspondre au taux de caractéristique audio. Nos mesures sont présentées dans la figure 6.8(a).

Nous présentons aussi nos résultats expérimentaux sur notre système AVASR au cours d'une gamme de niveaux de bruit. Nous avons utilisé le bruit aléatoire gaussien pour dégrader la qualité de l'image. La valeur moyenne du bruit est 0 et l'écart type était 15, 30, 50 et

100 respectivement. L'effet du bruit sur la ROI peut être vu dans la figure 6.8 et les résultats dans la figure 6.8(b).

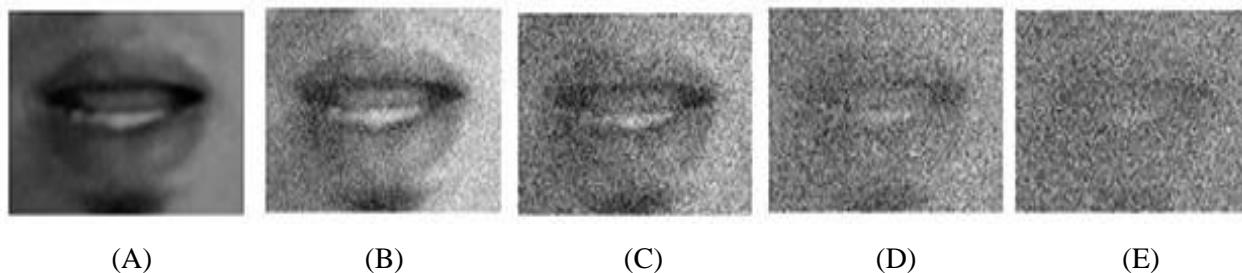
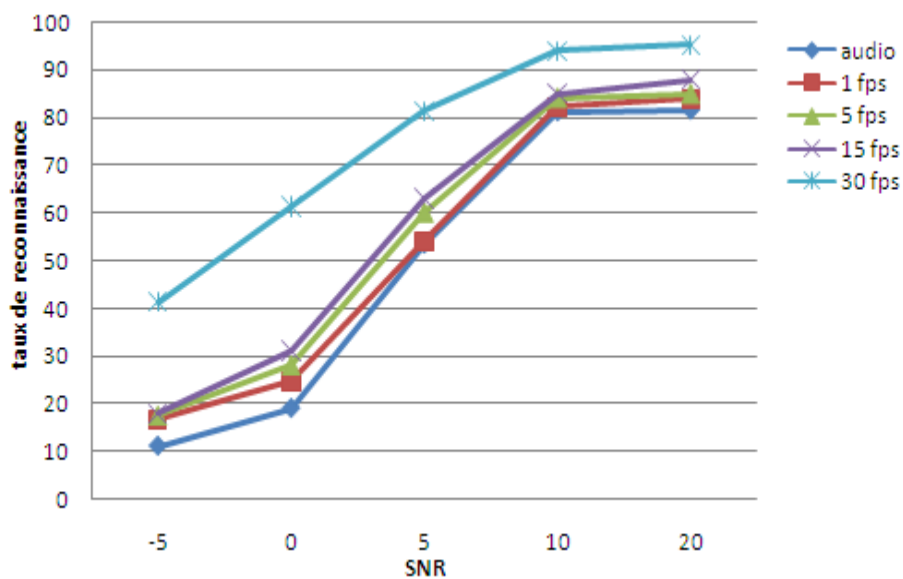
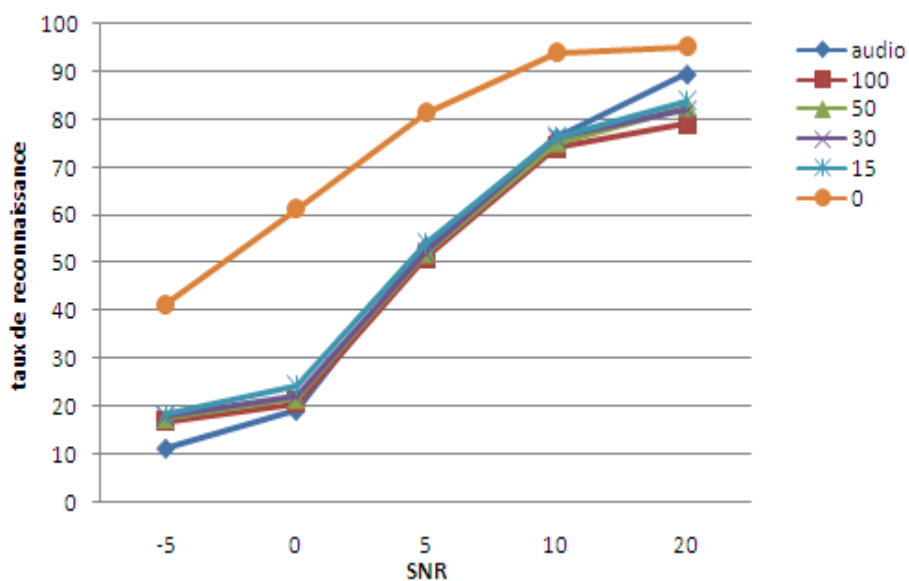


Figure 6.8 – ROI avec bruit gaussien, l'écart type =(A) 0 (B) 15 (C) 30 (D) 50 et (E) 100.



(a)



(b)

Figure 6.9 – La performance du système AVASR : (a) sous une fréquence des trames vidéo réduite (b) pour un bruit aléatoire gaussien.

Comme nous pouvons voir, les caractéristiques visuelles augmentent le taux de reconnaissance, même à 1fps. Plus précisément, la performance est supérieure pour 5 (de 56.1% à 1FPS) et 0 db (24.8% à 1FPS) à celle du reconnaiseur audio-seul (53.5% et 19.1% respectivement). Comme le montre le graphique dans 6.8(b), le taux de reconnaissance pour le système AVASR est réduit pour 10db mais pour des valeurs plus basses du SNR, le système AVASR surpasse le système de reconnaissance audio-seul. Même à un écart type de 100, le système fonctionne mieux pour 0 et 5db atteindre un taux 19.1% et 57,3% respectivement.

6.5.1.2. Expérimentations avec un bruit réel

Nous avons présenté différentes sortes d'instance avec des paramètres de contrôle différents de GA qui ont été résolus par notre algorithme pour évaluer la performance du système proposé. Nous avons exécuté chaque instance 15 fois avec un nombre différent de clusters, des valeurs de probabilité de croisement entre 0.5-0.9, et une probabilité de mutation avec la valeur 0,01. De plus, nous prenons un nombre maximum d'itérations pour l'algorithme de Baum-Welch égale à 40, les valeurs moyennes de $P(o|\lambda)$ obtenue valeurs après 150 générations (le nombre d'itérations idéale pour des meilleurs performance) sont listés dans les Tables 6.2 et 6.3 comme suit:

Nombre de clusters	P_c	P_m	Average $P(o \lambda)$	Nombre de clusters	P_c	P_m	Average $P(o \lambda)$
3	0.5	0.01	-2.3630	3	0.5	0.01	-3.7416
5	0.6	0.01	-1.5838	5	0.6	0.01	-3.2604
7	0.7	0.01	-1.1396	7	0.7	0.01	-3.4235
9	0.8	0.01	-3.3185	9	0.8	0.01	-3.9134
12	0.9	0.01	-4.0122	12	0.9	0.01	-4.3637

(a) (b)

Table 6.2 – paramètres GA pour l'entraînement du HMM pour l'audio seul: (a) base AVARB (b) base CUAVE.

Nombre de clusters	P_c	P_m	Average $P(o \lambda)$	Nombre de clusters	P_c	P_m	Average $P(o \lambda)$
3	0.5	0.01	-7.7629	3	0.5	0.01	-5.1860
4	0.5	0.01	-7.0046	5	0.6	0.01	-5.2987
7	0.8	0.01	-7.1555	7	0.7	0.01	-5.4743
9	0.8	0.01	-7.6595	9	0.8	0.01	-5.8747
12	0.9	0.01	-7.8234	12	0.9	0.01	-6.0890

(a) (b)

Table 6.3 – paramètres GA pour l'entraînement du HMM pour le vidéo seul: (a) base AVARB (b) base CUAVE.

Nous observons que les résultats varient en fonction des paramètres d'entraînement de l'AG, également au nombre de clusters obtenu par la phase de quantification vectorielle, par exemple, avec 7 clusters, $P_c = 0.7$ et $P_m = 0.01$, pour la base AVARB audio et 5 clusters, $P_c = 0.6$ et $P_m = 0.01$ pour la base AVARB visuelle sont supérieures à toutes les autres approches dans tous les cas. Par conséquent, nous les utilisons dans notre GA/HMM. Les mêmes observations pour la base de données audio CUAVE avec 4 clusters, $P_c = 0.6$ et $P_m = 0.01$, et pour la base de données visuelle CUAVE la meilleure performance est obtenue avec 3 clusters, $P_c = 0.5$ et $P_m = 0.01$.

Les figures 10 et 11 donnent le taux de reconnaissance moyennes par rapport au nombre de clusters utilisés dans l'expérience.

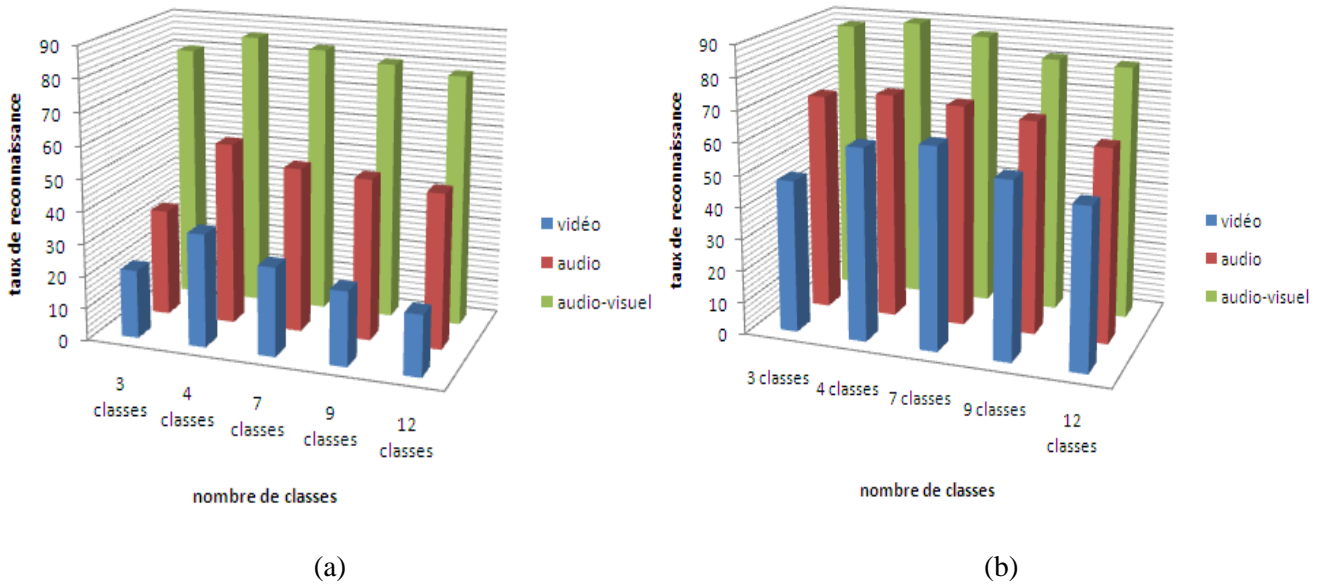


Figure 6.10 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD AVARB.

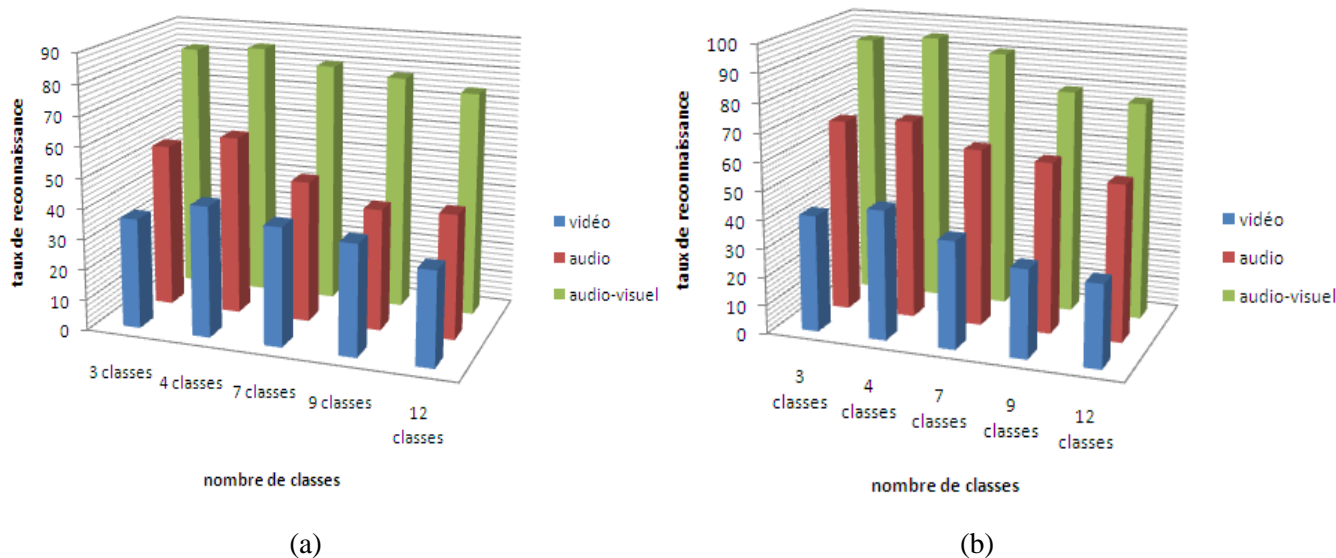


Figure 6.11 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD CUAVE.

En se basant sur les figures 6.10 et 6.11, nous constatons que les taux de reconnaissance obtenus avec notre GA/HMM sont meilleurs dans la plupart des cas par rapport à ceux obtenus avec le HMM standard (Les figures ci-dessus indiquent également que le système AVASR avec une fusion des scores dépassent significativement en atteignant des taux de reconnaissance les plus élevés. Dans la figure 6.10, nous avons noté presque les mêmes observations précédentes avec notre base de données de AVARB, c'est à dire que nous avons trouvé le meilleur taux moyen de reconnaissance égale à 93,7% et 97,6% en utilisant le HMM standard (Young et al. 2006) et le modèle hybride GA/HMM respectivement, et avec 7 classes à la fois.

Pour la base de données CUAVE les résultats montrent que le taux moyen de reconnaissance atteint un meilleur taux avec 86,8% en utilisant le modèle HMM standard avec 5 classes pour la phase de classification, et 98,1% en utilisant le modèle GA/HMM avec 3 classes.

Plus généralement, nous avons trouvé une augmentation du pourcentage variant de presque 5% à 28% des résultats de nos tests, mais cette augmentation dans les taux de reconnaissance donnés n'est pas fixe, ainsi que avec l'augmentation de la taille de la population. Il se peut donner des taux pire ou les mêmes de celle du HMM standard avant les optimisations. Cela est dû à la caractéristique de la méthode GA qui est aléatoire et aussi que ce système utilise le processus général de remplacement standard.

6.6 Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques techniques et les performances du système AVASR proposé. Les différents blocs matériels ainsi leur fonctionnement ont été détaillés.

Les résultats de l'évaluation (calcul d'erreur et les tests de reconnaissance) sont très satisfaisants et témoignent d'une grande fiabilité de mesures obtenues par ce système.

Les scores de reconnaissance obtenus ont montré que l'intégration des deux modalités acoustiques et visuelles sont supérieurs à ceux obtenus avec chaque modalité prise séparément, dans toutes les conditions expérimentales (niveau de bruit).

Conclusion et perspectives

7.1 Conclusion

Le domaine de la reconnaissance automatique de la parole est actuellement très actif. De nombreux laboratoires de recherche et des industriels effectuent des recherches dans ce domaine, avec un souci théorique et applicatif très marqué. Même si quelques problèmes de reconnaissance comme la reconnaissance de mots isolés avec un vocabulaire limité et prononcés dans des conditions calmes d'utilisation ou la reconnaissance dépendant du locuteur peuvent être considérés comme ayant atteint un niveau de performance satisfaisant, la reconnaissance automatique mérite encore de nombreux travaux de recherche pour étendre son champ d'application. Un axe important de recherche concerne l'amélioration de la robustesse d'un système de reconnaissance lorsque l'environnement de test est sensiblement différent de l'environnement d'apprentissage. Ce sujet a été le centre d'attention de ce document. Deux aspects du problème de robustesse ont été présentés : la robustesse au bruit et la robustesse au locuteur.

Nos travaux de recherche ont porté sur la fusion d'informations acoustiques et visuelles pour la RAP. Nous avons donc abordé les principaux problèmes sous-jacents à cette fusion, à savoir la paramétrisation des informations de parole et la nature des systèmes de reconnaissance dans chacune des modalités, ainsi que le lieu et la nature du processus de fusion des informations sensorielles. Nous avons choisi de résoudre ces problèmes en nous appuyant sur des études réalisées dans le domaine de la perception audiovisuelle de la parole. Nous avons développé différents systèmes pour effectuer la fusion des informations acoustiques et visuelles en prenant appui sur des modèles perceptifs. Ces systèmes ont été testés sur deux corpus audiovisuelles CUAVE.

7.2 Perspectives

Les travaux commencés au cours de cette thèse ouvrent la voie à de nombreux travaux futurs.

- La prise en compte de la parole continue ainsi spontanée est vitale pour un système de reconnaissance grand public.
- Les pauses, les répétitions, les hésitations, les phrases en suspens posent des problèmes par la suite aux autres modules de l'application visée.

- Les gens utiliseront les systèmes de reconnaissance à condition que le taux d'erreur de reconnaissance soit suffisamment faible. La reconnaissance robuste est donc nécessaire. L'utilisation d'un système de reconnaissance dans un milieu bruité et par différentes personnes devrait être habituelle.
- La prise en compte des bruits non stationnaires, dont l'importance a été soulevée à travers ce document, nécessite de continuer l'effort engagé. Nous n'en sommes qu'au début. L'étude des problèmes de détections de changement des bruits et la prise en compte de ces moments pendant la reconnaissance doit se poursuivre.
- Avec la représentation par adjacence, présentée dans le 4^{ème} chapitre, nous avons établi que le manque de compatibilité entre le GA d'une part et l'opérateur de mutation génétique défini sur la base d'approches déterministes d'autre part, nuisait à l'efficacité de l'approche. C'est donc prioritairement sur ce point que devront se focaliser de futurs développements.

Annexe A

A.1 Environnement de développement: MATLAB R2013a

MATLAB (« matrix laboratory ») est un langage de programmation de quatrième génération émulé par un environnement de développement du même nom ; il est utilisé à des fins de calcul numérique. Développé par la société américaine The MathWorks, MATLAB permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, de créer des interfaces utilisateurs, et peut s'interfacer avec d'autres langages comme le C, C++, Java, et Fortran. Les utilisateurs de MATLAB (environ un million en 20041) sont de milieux très différents comme l'ingénierie, les sciences et l'économie dans un contexte aussi bien industriel que pour la recherche. Matlab peut s'utiliser seul ou bien avec des toolbox (« boîte à outils »).

Le logiciel Matlab® et l'environnement graphique interactif Simulink® sont particulièrement performants et adaptés à la résolution de problèmes d'automatique, notamment pour la modélisation et la simulation des systèmes dynamiques.

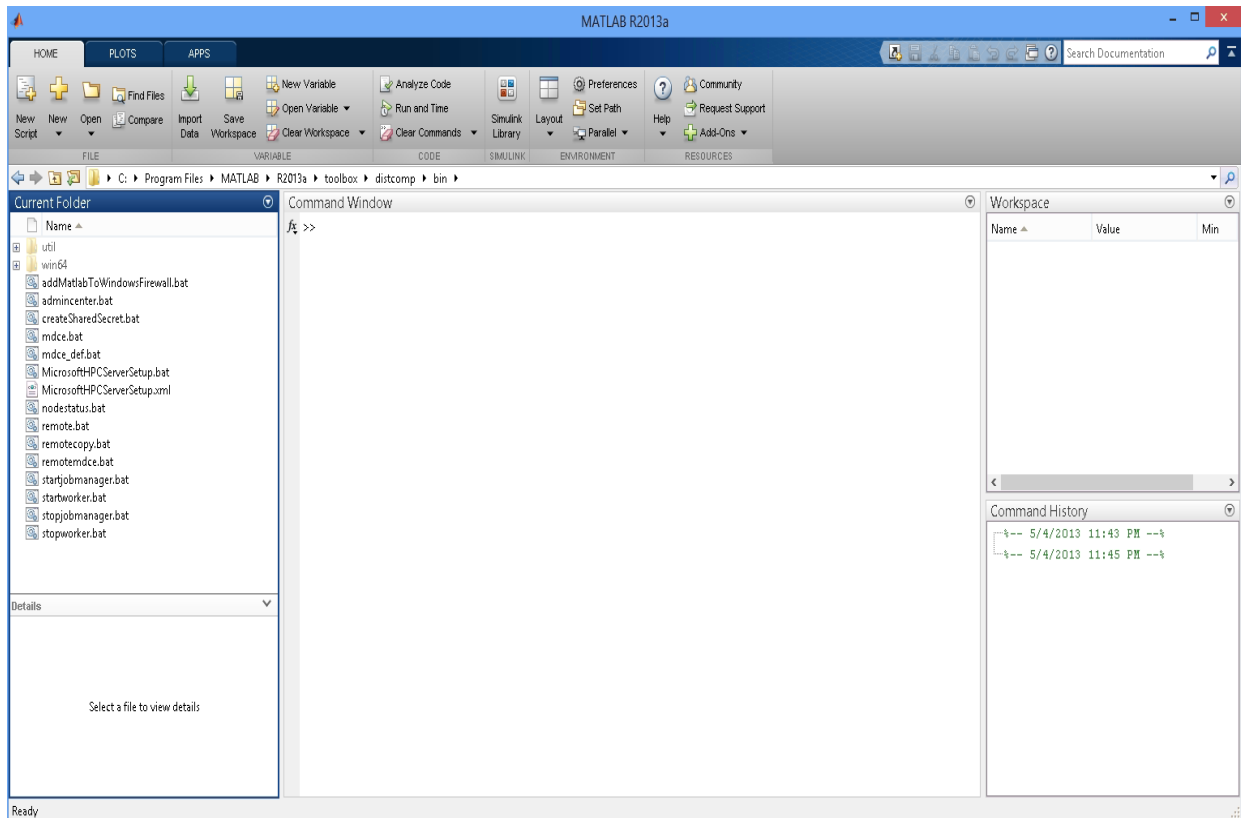


Figure A.1 – L'interface de l'environnement Matlab (R2013a).

- **Avantages :**

- collection très riche de bibliothèques avec de nombreux algorithmes, dans des domaines très variés. Exécution rapide car les bibliothèques sont souvent écrites dans un langage compilé.
- environnement de développement très agréable : aide complète et bien organisée, éditeur intégré, etc.
- support commercial disponible

- **Inconvénients :**

- langage de base assez pauvre, qui peut se révéler limitant pour des utilisations avancées.
- prix élevé

- **Pourquoi alors Matlab ?**

En effet plusieurs extensions plus « pointues » ont été conçues sous la forme de « TOOLBOXes », qui sont des paquets (payants) de fonctions supplémentaires dédiées à des domaines aussi variés que les statistiques, le traitement du signal et d'image, la logique floue, les réseaux de neurones, les ondelettes,... et qui permettent de résoudre un bon nombre de problèmes relatifs à ses domaines. Pour visualiser ces fonctions, il suffit de taper **help** suivi du nom de la famille à laquelle appartient la fonction. Pour connaître le nom de ces familles, il suffit juste de taper **help**. Il comporte plus de 1500 fonctions préprogrammées.

- **bibliothèques utilisés :**

La phase d'apprentissage est réalisée en deux étapes majeures : l'initialisation et la ré-estimation. Nous les avons conçus à partir de la plateforme HTK (Hidden Markov Model ToolKit) de l'Université de Cambridge. La boîte à outils HTK est efficace, flexible (liberté du choix des options et possibilité d'ajout d'autres modules) et complète dans le sens où elle fournit une documentation très détaillée (le livre HTK (Young et al. 2006) est une encyclopédie dans le domaine).

A.2 Structure et fonctionnement du logiciel

Ce logiciel traite une phase importante de tout type de reconnaissance de formes qui est la phase de reconnaissance. Il implémente précisément deux méthodes de prétraitement (DCT et RASTA-PLP) et l'algorithme K-means pour le clustering, ainsi 2 méthodes de reconnaissance HMM et le modèle hybride GA/HMM.

Le logiciel est implémenté sur Matlab R2013a, il est sous formes de fichier script MATLAB, ces fichiers MATLAB qui ont l'extension (.m) peuvent être considérés comme des fonctions qui peuvent être appelé à partie de l'interpréteur de commande MATALAB et qui se servent à leur tour d'un autre type de fichier des fichiers qui ont l'extension (.mat). Ces derniers fichiers représentent dans MTLAB des bases de données.

Notre application contient un fichier principale qui fait appelle aux autre fichiers .Ce fichier est nommé "interface " (voir figure A.2).

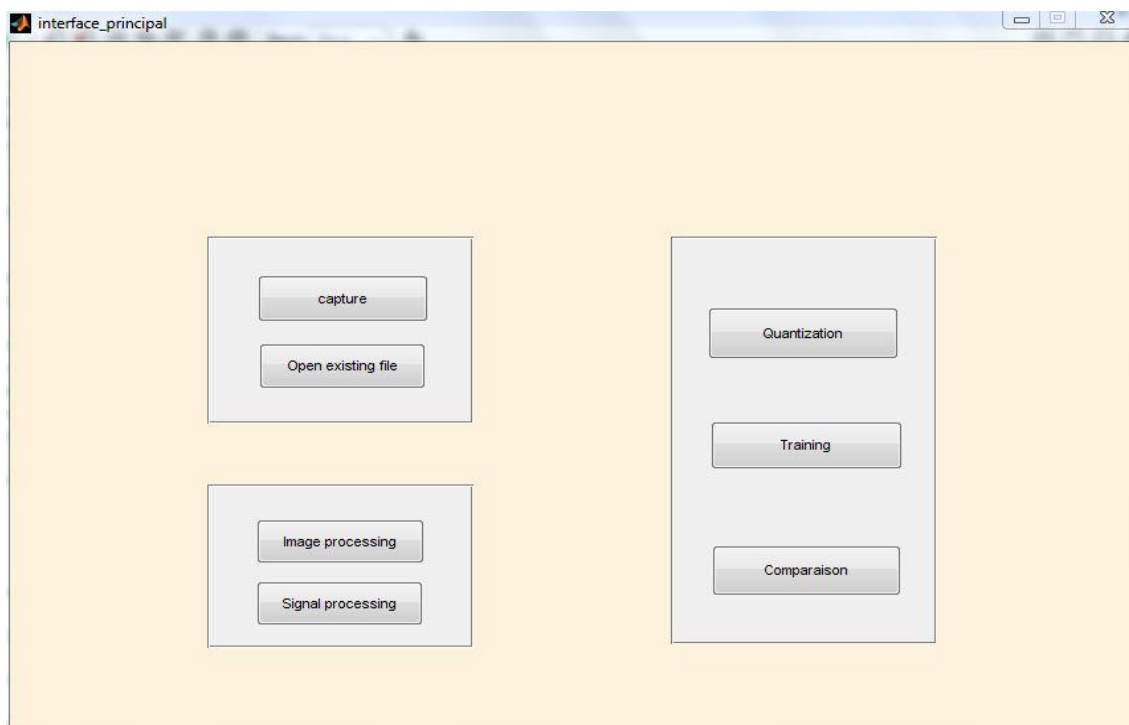


Figure A.2 – Interface principale du logiciel.

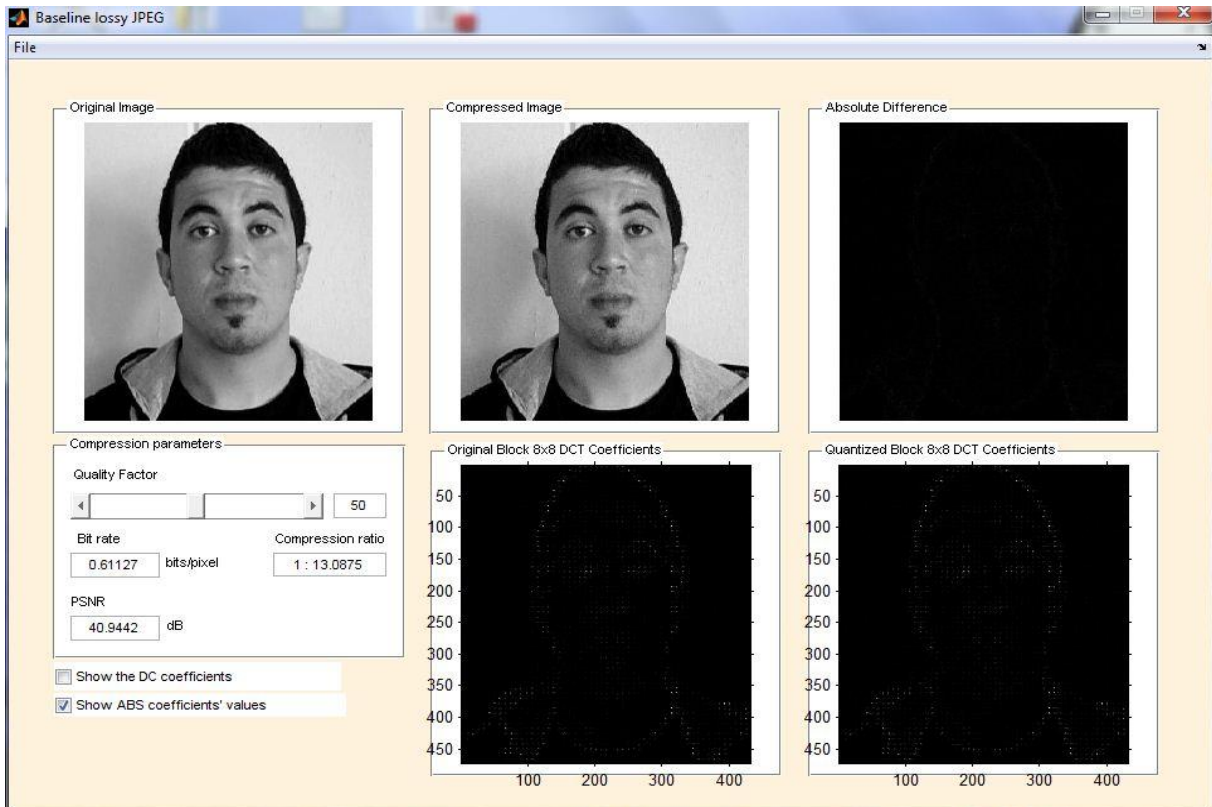


Figure A.3 – Interface d'extraction des paramètres visuels.

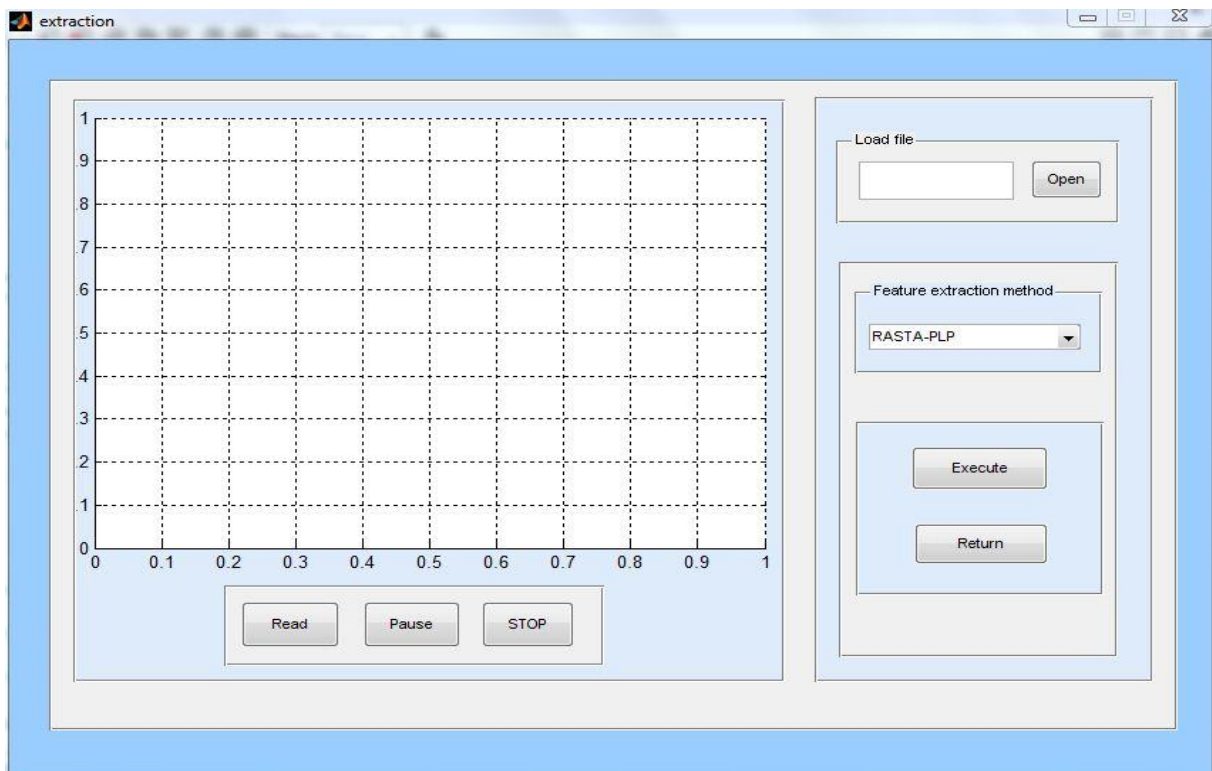


Figure A.4 – Interface d'extraction des paramètres acoustiques.

Bibliographie

- Abry C., Boë L.-J., Corsi P., Descout R., Gentil M. and Graillot P. (1980). Labialité et Phonétique, publications de l'Université des langues et lettre de Grenoble.
- Adjoudani, A., Guiard-Marigny, T., Le Goff, B. and Benoît, C. (1994). Un modèle 3d de lèvres parlantes. *In Actes des XX^e Journées d'Etude sur la Parole (JEP)*, pp. 143–146.
- Adjoudani, A. and Benoît, C. (1995). Audio-visual speech recognition compared across two architectures, *in Proc. of the 4th EUROSPEECH Conference*, Madrid, Espagne, pp. 1563-1566.
- Adjoudani, A. (1998). Reconnaissance automatique de la parole audiovisuelle. *Thèse de doctorat*, Institut National Polytechnique de Grenoble.
- Allegre, J. (2003). Approche de la reconnaissance automatique de la parole. *Rapport cycle probatoire, CNAM*.
- Alpaydin, E. (2004). Introduction to machine learning. *MIT Press*.
- Basso, A. Graf, H.P., Gibbon, D., Cosatto, E. and Liu, S. (2001). Virtual light: Digitally-generated lighting for video conferencing applications. *In Proc. ICIP*, 2: pp. 1085-1088, Thessaloniki, Greece, October 7-10.
- Benoît, C., Guiard-Marigny, T., Le Goff, B. and Adjoudani, A. (1996). Which Components of the Face Do Humans and Machines Best Speechread?, *in Speechreading by Humans and Machines*, D. Stork and M. Hennecke (eds.), Springer-Verlag, Berlin, pp. 351-372.
- Binnie C.A., Montgomery A.A. and Jackson P.L. (1974). Auditory and visual contributions to the perception of consonants, *Journal of Speech & Hearing Research*, 17, pp. 619-630.
- Berger, K. W., Garner, M., and Sudman, J. (1971) . The effect of degree of facial exposure and the vertical angle of vision on speechreading performance. *Teacher of the Deaf*, 69: pp. 322–326.
- Beyer, H.-G. (2001). The Theory of Evolution Strategies. *Natural Computing Series*. Springer, Heidelberg.
- Bregler, C., Hild, H., Manke, S. and Waibel, A. (1993). Improving connected letter recognition by lipreading, *Proc of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, IEEE*, 1, pp. 557-560.
- Bridges, C.L. and Goldberg, D.E. 1991. An analysis of multipoint crossover. *In Proceedings of the Foundation Of Genetic Algorithms. FOGA*.
- Bogert, B., Healy, M. and Tukey, J. (1963). The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Time Series Analysis*, pp. 209-243.
- Boite, R., Bourlard, H., Dutoit, T., Hancq, J. and Leich, H. (2000). *Traitement de la parole* (Presses Polytechniques et Universitaires Romandes, Lausanne).
- Bouchet, A. and Cuilleret, J. (1972). Anatomie topographique descriptive et fonctionnelle, Villeurbanne, Simep éditions.
- Broun, C.C., Zhang, X., Mersereau, R.M. and Clements, M. (2002). Automatic speechreading with application to speaker verification. *In Proc. ICASSP*, 1: pp. 685-688, Orlando, FL, USA, May 13-17.
- Brunelli, R. and Poggio, T. (1993). Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042-1052.
- Burnham, D. and Dodd, B. (1996). Auditory-visual speech perception as a direct process: the McGurk effect in infants and across languages, *Speechreading by Humans and Machines*, Stork et

- Hennecke (eds.), Springer-Verlag, Berlin, pp. 103-114.
- Cathiard, M.A. (1988). Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français. Mémoire de maîtrise, Université Grenoble II.
- Cathiard, M.A. (1989). La perception visuelle de la parole : aperçu des connaissances, *Bulletin de l'Institut de Phonétique de Grenoble*, 18: pp. 109-193.
- Cathiard, M.A. (1994). La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole. Thèse de doctorat de psychologie cognitive, UFR SHS, Université Pierre Mendès France.
- Chan, M.T., Zhang, Y. and Huang, T.S. (1998). Real-time lip tracking and bimodal continuous speech recognition. In *Proc. 2nd MMSP*, pp. 65-70, Los Angeles, CA, USA, December 7-9.
- Chiou, G.I. and Hwang, J.-N. (1996). Lipreading from color motion video. In *Proc. ICASSP*, 4: pp. 2158-2161, Atlanta, GA, USA.
- Coianiz, T., Torresani, L. and Caprile, B. (1996). 2D deformable models for visual speech analysis. In *Stork and Hennecke (1996)*, pp. 391-398.
- Collen, P., Rault, J.B. and Betser, M. (2007). Phase estimating method for a digital signal sinusoidal simulation," Software Patent PCT/FR2006/051361, 2007.
- Dai, Y. and Nakano, Y. (1996). Face-Texture Model Based on SGLD and Its Application in Face Detection in a Color Scene. *Pattern Recognition* 29(6), pp. 1007-1017.
- Dallos, P. (1973). *The Auditory Periphery: Biophysics and Physiology*. New York, USA: Academic Press.
- Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Londres, John Murray.
- Davis, S. and Melmerstein, P. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on ASSP*, 28: pp. 357-366.
- Demuynck, K., Garcia, O. and Van Compernelle, D. (2004). Synthesizing speech from speech recognition parameters. *Proc. of ICSLP*.
- Deviren, M. (2004). Systèmes de reconnaissance de la parole revisités : Réseaux Bayésiens dynamiques et nouveaux paradigmes. *Université de Nancy, Nancy, Thèse de doctorat*.
- Dodd, B. and Campbell, R. (1987) (eds.), *Hearing by Eye: The Psychology of Lipreading*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Duchnowski, P., Hunke, M. Büsching, D., Meier, U. and Waibel, A. (1995). Toward movement-invariant automatic lip-reading and speech recognition. In *Proc. ICASSP*, 1: pp.109-112, Detroit, MI, USA.
- Dupont, S. and Luetttin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141-151.
- Erber N.P. (1974). Effect of angle, distance, and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research*, 17:pp. 99-112.
- Erber N.P. (1975). Auditory-visual perception of speech, *Journal of Speech and Hearing Disorders*, 40, pp. 481-492.
- Escudier, P., Benoît, C. and Lallouache, M.T. (1990). Identification visuelle de stimuli associés à l'opposition /i/ - /y/: étude statistique, *Proceedings of the First French Conference on Acoustics*, Lyon, France, pp. 541-544.
- Eyben, F., Wöllmer, M. and Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proc. of ACM Multimedia*, pp. 1459-1462.
- Fant, G. (1973). *Speech Sounds and Features* », M.I.T. Press, Cambridge, USA.

- Fogel, L.J., Owens, A.J. and Walsh, M.J. (1966). *Artificial Intelligence through Simulated Evolution*. Wiley, New York.
- Goh, J., Tang, L. and Al turk, L. (2010). Evolving the Structure of Hidden Markov Models for Microaneurysms Detection. *UK Workshop on Computational Intelligence (UKCI)*, pp.1–6.
- Goldberg, D. and Richardson, J. (1987). Genetic algorithm with shearing for multi-model function optimization, *In J.J. Proceeding of the 2nd international conference on genetic algorithms*, pp. 41-49, Lawrence Erlbaum associates.
- Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley Reading, Massachusetts.
- Goldberg, D. (1991). Real-coded genetic algorithms, virtual alphabets and blocking. *Complex Systems*, 5: pp. 139-167.
- Gouet, V. and Montesinos, P. (2002). Normalisation des images en couleur face aux changements d'illumination. *In Proc. RFIA'02*, 2: pp. 415-424, Angers, France, January 8-10.
- Gray, M.S., Movellan, J.R. and Sejnowski, T.J. (1997a). A comparison of local versus global image decompositions for visual speechreading. *In Proc. 4th Annual Joint Symposium on Neural Computation*, pp. 92-98, Pasadena, CA, USA, May 17.
- Gray, M.S., Movellan, J.R. and Sejnowski, T.J. (1997b). Dynamic features for visual speechreading: A systematic comparison. *In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, ANIPS*, 9: pp. 751-757. The MIT Press.
- Gupta, M. and Garg, Dr.A.K. (2012). Analysis of image compression algorithm Using DCT. *International Journal of Engineering Research and Applications (IJERA)*, 2(1): pp.515–521.
- Gurbuz, S., Patterson, E.K., Tufekci, Z. and Gowdy, J.N. (2001a). Lip-reading from parametric lip contours for audio-visual speech recognition. *In Proc. 7th Eurospeech*, 2: pp.1181-1184, Aalborg, Denmark, September 3-7.
- Gurbuz, S., Patterson, E.K., Tufekci, Z. and Gowdy, J.N. (2001b). Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition. *In Proc. ICASSP*, 1: p. 177-180, Salt Lake City, UT, USA, May 7-11.
- Hlaoui, A. (1999). Reconnaissance de mots isolés arabes par hybridation de réseaux de neurones et modèles de Markov cachés. *École nationale d'ingénieurs de Tunis*.
- Hardcastle, W.J. (1976). *Physiology of Speech Production*, Academic Press, Londres.
- Harvey, R., Matthews, L., Bangham, J.A. and Cox, S. (1997). Lip reading from scale-space measurements. *In Proc. CVPR*, pp. 582-587, Puerto Rico, June.
- Haton, J.-P. (2006). *Reconnaissance automatique de la parole : Du signal à son interprétation*. Dunod Paris.
- Hermansky, H., Morgan, N., Bayya, A. and Kohn, P. (1992). RASTA-PLP Speech Analysis. *IEEE International conference on Acoustics, speech and signal processing*, 1: pp.121–124.
- Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.
- Hunke, H. M. and Waibel, A. (1994). Face locating and tracking for human-computer interaction, *Proc. Twenty-Eight Asilomar Conference on Signals, Systems & Computers*, Monterey, CA, USA.
- Hunke, H. M. (1994). Locating and tracking of human faces with neural networks. Master's thesis, University of Karlsruhe.
- Jacob, B. and Sénac, C. (1996). Un modèle maître-esclave pour la fusion de données acoustiques et articulatoires en reconnaissance. *In Actes des Journées d'Etude sur la Parole (JEP)*, pp. 363–366, Avignon, Juin.

- Jakiela, M., Chapman, C., Duda, J., Adweuya, A. and Saitou, K. (2000). Continuum structural topology design with genetic algorithm. *Comput. Methods Appl. Mech. Engrg* 186, pp. 339-356.
- Jourlin, P. (1996). Handling desynchronization phenomena with hmm in connected speech. *In Proceedings of European Signal Processing Conference*, pp. 133–136, Trieste.
- Kant, E. (1787). Critique de la Raison Pure, *Presses Universitaires de France, 11ème édition*, 1944, édition originale, 1787.
- Khandait, S.P., Khandait, P.D. and Thool, Dr.R.C. (2009). An Efficient Approach to Facial Feature Detection for Expression Recognition. *International Journal of Recent Trends in Engineering*, 2(1): pp.179–182.
- Kicinger, R., Arciszewski, T., and Jong, K. D. (2005). Evolutionary computation and structural design: A survey of the state-of-the-art. *Computers & Structures*, 83(23-24): pp. 1943-1978.
- Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal Phonétique*. 7: pp. 279–312.
- Kubrick, S. (1968). 2001 : A space odyssey (2001 : l'odyssée de l'espace). Metro-Goldwyn-Mayer (Turner Entertainment Co), April 3. <http://www.kubrick2001.com/>, <http://sfstory.free.fr/films/2001.html>.
- Kuhl, P.K. and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. *Science*, 218, pp. 1138-1141.
- Kwong, S. and Chau, C.W. (1997). Analysis of Parallel Genetic Algorithms on HMM Based Speech Recognition System. *IEEE Transactions on Consumer Electronics*. 43(4): pp. 1229 – 1233.
- Ladefoged P. (1979). Articulatory parameters, *W.P.P. 45, U.C.L.A.*, pp. 25-31.
- Lallouache M.T. (1991). Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres, PhD. dissertation, INPG, Grenoble, France.
- Laprie, Y. (2000). Analyse spectrale de la parole.
- Larr A. L. (1959). Speechreading through closed-circuit television. *Volta Review*, 61: pp.19–21.
- Lee, J. and Kim, J.Y. (2001). An efficient lipreading method using the symmetry of lip. *In Proc. 7th Eurospeech*, 2: pp. 1019-1022, Aalborg, Denmark, September 3-7.
- Le Goff, B., Guiard-Marigny, T., and Benoît, C. (1995). Read my lips ... and my jaw! how intelligible are the components of a speaker's face ? *In Eurospeech'95*, Madrid, Spain.
- Le Goff, B., Guiard-Marigny, T., and Benoît, C. (1996). Progress in Speech Synthesis, *chapitre Analysis-synthesis and intelligibility of a talking face*, pp. 235–246. Springer, New York.
- Le Huche, F. and Allali, A. (2001). *La Voix. Anatomie et physiologie des organes de la voix et de la parole* (Masson, Paris).
- Leroy, B. and Herlin, I.L. (1995). Un modèle déformable paramétrique pour la reconnaissance de visages et le suivi du mouvement des lèvres. *In 15th GRETSI Symposium Signal and Image Processing*, pp. 701-704, Juan-les-Pins, France, September 18-21.
- Leroy, B. Chouakria, A., Herlin, I.L. and Diday, E. (1996a). Approche géométrique et classification pour la reconnaissance de visages. *In Proc. RFIA*, pp. ??-??, Rennes, France.
- Lieberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech production revised. *Cognition*, 21: pp.1–36, 1985.
- Lievin, M. and Luthon, F. (1999). Lip features automatic extraction. *Proceedings of IEEE International Conference on Image Processing*, Chicago, IL, USA, 3: pp. 168–172.
- Liew, A.W.C., Sum, K. L., Leung, S.H. and Lau, W.H. (1999). Fuzzy segmentation of lip image using cluster analysis. *In Proc. 6th Eurospeech*, 1: pp. 335-338, Budapest, Hungary, September 6-9.

- Liu, L., He, J. and Palm, G. (1997). Effects of the phase on the perception of intervocalic stop consonants. *Speech Communication*, 4(22): pp. 403-417.
- Lockwood, P., Boudy, J. and Blanchet, M. (1992). Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments. *Proc. of IEEE ICASSP*, 1: pp. 265-268.
- Luettin, J. Thacker, N.A. and Beet, S. (1996a). Active shape models for visual speech feature extraction. In *Stork and Hennecke* (1996), pp. 383-390.
- Luettin, J. Thacker, N.A. and Beet, S. (1996b). Locating and tracking facial speech features. In *Proc. ICPR*, 1: pp. 652-656, Vienna, Austria, August 25-29.
- Luettin, J. Thacker, N.A. and Beet, S. (1996c). Speaker identification by lipreading. In *Proc. 4th ICSLP*, 1: pp. 62-65, Philadelphia, PA, USA, October 3-6.
- Luettin, J. Thacker, N.A. and Beet, S. (1996d). Speechreading using shape and intensity information. In *Proc. 4th ICSLP*, 1: pp. 58-61, Philadelphia, PA, USA, October 3-6.
- Luettin, J. Thacker, N.A. and Beet, S. (1996e). Statistical lip modelling for visual speech recognition. In *Proc. 8th Eusipco*, 1: pp. 137-140, Trieste, Italy, September 10-13.
- Luettin, J. Thacker, N.A. and Beet, S. (1996f). Visual speech recognition using active shape models and hidden Markov models. In *Proc. ICASSP*, 2: pp. 817-820, Atlanta, GA, USA, May 7-10.
- Luettin, J. and Thacker, N.A. (1997). Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163-178.
- Luettin, J. (1997a). Towards speaker independent continuous speechreading. In *Proc. 5th Eurospeech*, pp. 1991-1994, Rhodes, Greece, September 22-25.
- Luettin, J. (1997b). Visual Speech and Speaker Recognition, *PhD dissertation*, Université de Sheffield.
- Luettin, J. and Dupont, S. (1998). Continuous audio-visual speech recognition. *LNCS*, 1407: pp. 657-673.
- Makhlouf A., Lazli, L. and Bensaker, B. (2013a). Automatic Speechreading Using Genetic Hybridization of Hidden Markov Models. In *Proceeding of the IEEE World Congress on Computer and Information Technology (WCCIT'13)*, June 22-24, 2013, Sousse, Tunisia.
- Makhlouf A., Lazli, L. and Bensaker, B. (2013b). Hybrid Hidden Markov Models and genetic algorithm for Robust Automatic visual speech recognition. *Journal of Information Technology Review (JITR)*, 4(3): pp. 105-114.
- Makhlouf A., Lazli, L. and Bensaker, B. (2016). Structure Evolution of Hidden Markov Models for Audiovisual Arabic Speech Recognition. *International Journal of Signal and Imaging Systems Engineering, IJSISE*, 9(1).
- Malasné, N., Yang, F., Paindavoine, M. and Mitéran, J. (2002). Suivi dynamique et vérification de visages en temps réel : algorithme et architecture. In *Proc. RFIA'02*, pp.77-86, Angers, France.
- Mase, K. (1991). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6): 67-75.
- Massaro, D.W. (1987). *Categorical Perception: The Groundwork of Cognition*, chapitre Categorical partition: a fuzzy logical model of categorization behavior. *Cambridge, MA : University Press*.
- Massaro, D.W. (1989). Multiple book review of *Speech perception by ear and eye*, *Behavioral and Brain Sciences*, 12, pp.741-794.
- Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. *Cambridge, Massachusetts : MIT Press*.

- Matthews, L. Bangham, J. and Cox, S. (1996a). Audiovisual speech recognition using multiscale nonlinear image decomposition. In Proc. 4th ICSLP, 1: pp. 38-41, Philadelphia, PA, USA, October 3-6.
- Matthews, L. Bangham, J.A., Harvey, R. and Cox, S. (1998). A comparison of active shape models and scale decomposition based features for visual speech recognition. *LNCS*, 1407: pp. 514-528.
- McGurk, H. and McDonald, J. (1976). Hearing Lips and Seeing Voices, *Nature*, 264: pp. 746-748.
- Meier, U. Hürst, H. and Duchnowski, P. (1996). Adaptive bimodal sensor fusion for automatic speechreading. In Proc. ICASSP, pp. 833-836, Atlanta, GA, USA, May.
- Messer, k., Matas, J., Kittler, J., Luetin, J. and Maître, G. (1999). XM2VTSDB : The extended M2VTS database. In Proc. 2nd AVBPA, pp. 72-77, Washington, DC, USA, March 22-23.
- Michalewicz, Z. and Janikov, C.Z. (1991). Handling constraints in genetic algorithms. In *Proceedings of the Fourth International Conference on Genetic Algorithm*. ICGA.
- Milner, B. and Darch, J. (2011). Robust Acoustic Speech Feature Prediction From Noisy Mel-Frequency Cepstral Coefficients. *IEEE Trans. on ASLP*, 2(19): pp. 338-347.
- Movellan, J.R (1995). Visual speech recognition with stochastic networks. In *Gerald Tesauro, David Touretzky, and Todd Leen, editors, ANIPS*, 7: pp. 851-858, Cambridge, MA, USA. *The MIT Press*.
- Movellan, J.R and Chadderdon, G. (1996). Speechreading by Man and Machine: Models, Systems and Applications. *chapitre Channel separability in the audiovisual integration of speech : A Bayesian approach*, pp. 473-488. *Springer-Verlag, NATO ASI Series, Berlin, Germany*.
- Murty, K.S.R. and Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters*, 1(13): pp. 52-55.
- Nakano, Y. (1961). A study on the factors which influence lipreading of deaf children. *Language research in countries other than the United States, Volta Review*, 68:pp. 68-83. Cited by Quigley (1966).
- Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *Journal of Acoustic Society of America*, 28: pp.1275-1277.
- Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C. and Murphy, K. (2002). A coupled HMM for audio-visual speech recognition. In Proc. ICASSP, 2: pp. 2013-2016, Orlando, FL, USA, May 13-17.
- Neti, C. V. and Senior, A. (1999). Audio-visual speaker recognition for video broadcast news. In *DARPA HUB4 Workshop*, pp. 139-142, Washington, DC, USA.
- Neti, C., Potamianos, G., Luetin, J., Matthews, L., Glotin, H., Vergyri, D., Sison, J., Mashari, A. and Zhou, J. (2000). Audio-visual speech recognition. *Technical Report Workshop 2000, International Computer Science Institute, Center for Language and Speech Processing (CLSP)*, The Johns Hopkins University, Baltimore, MD, USA, October 12.
- O'Shaughnessy, D. (1987). *Speech Communications: Human and Machine*, Series in Electrical Engineering ed. USA: Addison-Wesley Publishing Co.
- Oudelha, M. and Aïnon, R.N. (2010). HMM parameters estimation using hybrid Baum-Welch genetic algorithm. *International Symposium in Information Technology (ITSim)*, 2: pp.542-545.
- Pai, Y., Ruan, S., Shie, M., Liu, Y. (2006). A Simple and Accurate Color Face Detection Algorithm in Complex Background. In *ICME*, pp. 1545-1548.
- Patterson, E.K., Gurbuz, S., Tufekci, Z. and Gowdy, J.N. (2002). Moving-talker speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus. *EURASIP Journal on Applied Signal Processing*, 11: pp.1189-1201.

- Pentland, A. and Mase, K. (1989). Automatic lipreading by optical-flow analysis. *Technical Report VA189-8*, ITEJ.
- Pérez, Ó, Piccardi, M. and García, J. (2007). Comparison between genetic algorithms and the Baum-Welch algorithm in learning HMMs for human activity classification, *Proceeding of EvoWorkshops '07*, pp.399–406.
- Petajan, E. (1984). Automatic lipreading to enhance speech recognition, Ph.D. dissertation, Univ. Illinois at Urbana-Champaign.
- Pigeon, S. and Vandendorpe, L. (1997). The M2VTS multimodal face database. *LNCS*, pp. 403–410.
- Potamianos, G., Cosatto, E., Graf, H.P. and Roe, D.B. (1997). Speaker independent audio-visual database for bimodal ASR. In Benoît and Campbell (1997), pp. 65-68.
- Potamianos, G., Verma, A., Neti, C. and Iyengar, G. (2000). A cascade image transform for speaker independent automatic speechreading. *In Proc. ICME*, pp. 1097-1100, New York, NY, USA.
- Potamianos, G., Luetin, J. and Neti, C. (2001a). Hierarchical discriminant features for audio-visual LVCSR. *In Proc. ICASSP*, 1: pp. 165-168, Salt Lake City, UT, USA, May 7-11.
- Potamianos, G., Neti, C., Iyengar, G. and Helmuth, E. (2001b). Large-vocabulary audio-visual speech recognition by machines and humans. *In Proc. 7th Eurospeech*, 2: pp. 1027-1030, Aalborg, Denmark, September 3-7.
- Potamianos, G., Neti, C., Iyengar, G., Senior, A.W. and Verma, A. (2001c). A cascade visual front end for speaker independent automatic speechreading. *Speech Technology*, 4: pp. 193–208.
- Rabiner, L. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Oxford University Press.
- Rao, R. and Mersereau, R. M. (1995). On merging hidden Markov models with deformable templates. *In Proc. ICIP*, 3: pp. 3556–3559, Washington, DC, USA.
- Reisberg, D., McLean, J. and Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli », in *Hearing by Eye : the psychology of lip-reading*, B. Dodd et R. Campbell (eds.), Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp.97-114.
- Revéret, L. (1999). Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole. *Thèse de doctorat*, de l'institut national polytechnique de Grenoble.
- Robert-Ribes, J., Piquemal, M., Schwartz, J. L. and Escudier, P. (1996). Speechreading by Man and Machine: Models, Systems and Applications. chapitre Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition, pp. 193–210. Springer-Verlag, NATO ASI Series, Berlin, Germany.
- Rodomagoulakis, I. (2008). Feature Extraction Optimization and Stream Weight Estimation in Audio-Visual Speech Recognition. *Phd thesis from Technical University of Crete*.
- Rogozan, A., Deléglise, P. and Alissali, M. (1996). Intégration asynchrone des informations auditives et visuelles dans un système de reconnaissance de la parole », Actes des 21èmes Journées d'Études sur la Parole, Avignon, pp. 359-362.
- Rogozan, A. (1999). Étude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audiovisuelle. *Thèse de doctorat*, Université d'Orsay - Paris XI.
- Sánchez, U.R. (2000). Aspects of facial biometrics for verification of personal identity. Ph.D. thesis, University of Surrey, Guilford, UK.
- Sanderson C. and Paliwal, K. (2002). Polynomial features for robust face authentication. *In proceedings of International Conference on Image Processing*.
- Schwartz, J.-L., Robert-Ribès, J. and Escudier, P. (1998). Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. *chapitre Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception*, pp. 85–108. Psychology Press, Hove, UK.

- Schwartz, J.-L. (2002). Traitement automatique du langage parlé 2: reconnaissance de la parole. *chapitre La parole multimodale: deux ou trois sens valent mieux qu'un*, pp. 141–178. Hermes, Paris.
- Schwartz, J.-L. (2004). La parole multisensorielle: Plaidoyer, problèmes et perspectives. *In Actes des XXVème Journées d'Etude sur la Parole (JEP)*, pp. 11–17, Fès, Maroc.
- Silsbee, P.L. and Su, Q. (1996). NATO ASI: Speechreading by Humans and Machines. *chapitre Audiovisual sensory integration using hidden Markov models*, pp. 489–495. Springer-Verlag.
- Senior, A. W., (1999). Face and feature finding for a face recognition system. *In Proc. 2nd AVBPA*, pp. 154–159, Washington, DC, USA, March 22-23.
- Shdaifat, I., Grigat, R. R. and Luetgert, S. (2001). Viseme recognition using multiple feature matching. *In Proc. 7th Eurospeech*, 4: pp. 2431–2434, Aalborg, Denmark, September 3-7.
- Shing-Tai, P., Ching-Fa, C. and Jian-Hong Z. (2010). Speech Recognition via Hidden Markov Model and Neural Network Trained by Genetic Algorithm. *Ninth International Conference on Machine Learning and Cybernetics*. Qingdao, 11-14 July.
- Sobottka, K., and Pitas, I. (1996). Segmentation and tracking of faces in color images, Automatic face and gesture recognition, pp. 236–241.
- Stevens, S.S., Volkman, J. and Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Proc. of JASA*, 3(8): pp. 185–190.
- Stork, D.G. (1997). HAL's Legacy. 2001's Computer as Dream and Reality. MIT Press, Cambridge, MA, USA.
- Sumby, W.H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, 26, pp. 212-215.
- Summerfield, Q. (1979). Use of visual information for phonetic perception, *Phonetica*, 36: pp. 314-331.
- Summerfield, Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. *Hearing Science and Hearing Disorders*, pp. 131–182.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visuel speech perception, in *Hearing by Eye: The psychology of lipreading*, B. Dodd and R. Campbell, eds.
- Summerfield, Q., MacLeod A., McGrath M. and Brooke M. (1989). Lips, teeth, and the benefits of lipreading, in *Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis (eds.), Elsevier Science Publishers, pp. 223-233.
- Taboada, J., Feijoo, S., Balsa, R. and Hernandez, C. (1994). Explicit estimation of speech boundaries. *IEEE Proc. Sci. Meas. Technol.*, 141: pp. 153-159.
- Teissier, P., Robert-Ribès, J. and Schwartz, J.-L. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6): pp. 629–642.
- Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, IEEE Computer Society Press, Jauai, Hawaii, December 8-14.
- Waibel, A. and Lee, K.-F. (1990). (eds), *Readings in Speech Recognition*, San Mateo, CA: Morgan Kaufmann.
- Walden, B. E., Prosek, A. and Montgomery (1977). Effect of training on the visual recognition of consonants, *Journal of Speech and Hearing Research*, 20: pp. 130-145.
- Wark, T. & Sridharan, S. (1998). An approach to statistical lip modelling for speaker identification via chromatic feature extraction, in *International Conference on Pattern Recognition*, pp. 123-125.

- Whalen D.H. (1990). Coarticulation is largely planned, *Journal of Phonetics*, 18(1), pp. 3-35.
- Wojdel J.C. and Rothkrantz. L.J.M. (2001a). Robust video processing for lipreading applications. *In Proc. 6th Euromedia*, pp. 195-199, Valencia, Spain, April 18-20.
- Wojdel J.C. and Rothkrantz. L.J.M. (2001b). Using aerial and geometric features in automatic lipreading. *In Proc. 7th Eurospeech*, 4: pp. 2463-2466, Aalborg, Denmark, September 3-7.
- Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), pp. 67-82.
- Wright, A.H. (1991). Genetic algorithms for real parameter optimization. In *Proceeding of the Foundation Of Genetic Algorithms*. FOGA.
- Xue-ying, Z., Yiping, W. and Zhefeng, Z. (2007). A Hybrid Speech Recognition Training Method for HMM Based on Genetic Algorithm and Baum Welch Algorithm. *IEEE 2nd International conference on Innovative Computing, Information and Control (ICICIC'07)*, pp.572.
- Yang, J. and Waibel, A., (1996). A real-time face tracker. *In: Proc. 3rd IEEE Workshop on Application of Computer Vision*. pp. 142-147.
- Yang, C., Soong, F.K. and Lee, T. (2007). Static and dynamic spectral features: their noise robustness and optimal weights for ASR. *IEEE Trans. on ASSP*, 3(15): pp. 1087-1097.
- Young, S., Evermann, G., Gale, M., Hain, s.T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2006). *The HTK Book (for HTK version 3.4)*. Cambridge University Engineering Department, Ed.
- Zemlin, W.R. (1968). *Speech and Hearing Science: Anatomy and Physiology*, New Jersey, Prentice-Hall.
- Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Proc. of JASA*, 2(33): pp. 248.

Notations

AAM	Active Appearance Model
ACP	Analyse en Composantes Principales
ASR	Automatic Speech Recognition
AVASR	Audio-Visual Automatic Speech Recognition
BW	Baum-Welch algorithm
DCT	Discrete Cosine Transform
DI	Direct Integration
DWT	Discrete Wavelet Transform
FAP	Facial Animation Parameters
FCC	Face Color Classifier
FLMP	Fuzzy-Logical Model of Perception
HMM	Hidden Markov Models
ID	Identification Directe
ICP	Institut de la Communication Parlée
IFCC	Individuel Face Color Classifier
IS	Identification Séparée
GA	Genetic Algorithm
GFCC	General Face Color Classifier
GMM	Gaussian Mixture Model
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
LUT	Look-Up Table
MFCC	Mel-scaled Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transform
MMI	Maximum Mutual Information
MSA	Multiscale Spatiale Analysis
PLP	Perceptual Linear Predictive
RAP	reconnaissance automatique de la parole
RASTA-PLP	RelAtive SpecTral Analysis-Perceptual Linear Predictive
ROI	Region Of Interest
SI	Separate Integration
SNR	Signal-to-Noise Ratio

Publications réalisées au cours de la thèse

Publications et conférences internationales :

Makhlouf A., Lazli, L. and Bensaker, B. (2012). Structure Evolution of Hidden Markov Models for an Automatic Speechreading. *Accepted paper for 7th International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, Lugano, Switzerland.

Makhlouf A., Lazli, L. and Bensaker, B. (2013a). Automatic Speechreading Using Genetic Hybridization of Hidden Markov Models. *In Proceeding of the IEEE World Congress on Computer and Information Technology (WCCIT'13)*, June 22-24, 2013, Sousse, Tunisia.

Makhlouf A., Lazli, L. and Bensaker, B. (2013b). Hybrid Hidden Markov Models and genetic algorithm for Robust Automatic visual speech recognition. *Journal of Information Technology Review (JITR)*, 4(3): pp. 105-114.

Makhlouf A., Lazli, L. and Bensaker, B. (2016). Structure Evolution of Hidden Markov Models for Audiovisual Arabic Speech Recognition. *International Journal of Signal and Imaging Systems Engineering, IJSISE*, 9(1), pp.55–66.

Co-encadrement:

Master de recherche Reconnaissance des Formes et Intelligence Artificielle (Janvier 2015- Juin 2015)

Boukhatem Chemssedine, « *extraction des paramètres vocaux à l'aide d'une nouvelle méthode d'analyse acoustique* », un master pourtant sur la mise en œuvre de la méthode J-RASTA pour faire une extraction des paramètres acoustiques.