Introduction

La reconnaissance audiovisuelle de la parole est un domaine de recherche qui a connu un intérêt grandissant durant ces dernières années. Elle se situe à l'intersection de deux grands axes de recherche : la modélisation de la perception de la parole audiovisuelle et la reconnaissance automatique des formes.

La parole est un moyen de communication audiovisuelle. Le message parlé est plus intelligible quand il est accompagné de la vision du visage du locuteur, surtout quand le milieu de transmission est dégradé. D'ailleurs, la lecture labiale permet aux malentendants de restituer une grande partie de l'intelligibilité de la parole acoustique. Chez les personnes n'ayant pas de déficience auditive, la vision du visage du locuteur permet de mieux comprendre un message complexe ou un accent étranger, même quand le signal acoustique est clair (Reisberg et al., 1987). La reconnaissance automatique de la parole 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.

Notre travail de recherche tente de vérifier ces deux conditions, c'est à dire la modélisation de la perception audiovisuelle de la parole en vue d'une implémentation logicielle, et l'extraction des informations visuelles 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 investi dans le domaine de la paramétrisation des lèvres, le traitement du signal et la reconnaissance automatique des formes, pour achever notre travail avec la réalisation d'un prototype électronique de hautes performances dédié à l'extraction des informations visuelles des lèvres.

Nous avons organisé ce manuscrit en cinq chapitres :

Le premier chapitre présente quelques notions théoriques sur la production de la parole. Cette étude est orientée vers une approche physiologique des lèvres et présente une description des muscles faciaux intervenant dans le processus de la parole, ainsi que notre travail sur l'élaboration d'un modèle de lèvres tridimensionnel.

Le chapitre 2 est consacré à la perception humaine de la parole bimodale. Dans un premier temps, nous présentons quelques expériences de perception afin de mieux comprendre les propriétés intrinsèques de la perception audiovisuelle de la parole, pour enchaîner, dans un second temps, sur les différents modèles d'intégration proposés dans la littérature.

Les différentes techniques d'extraction automatique des informations visuelles sont abordées dans le chapitre 3. Nous présentons les approches classiques en traitement d'images couleurs par l'intermédiaire de quelques expériences dans ce domaine. Notre approche est décrite ensuite.

Le chapitre 4 est centré sur la reconnaissance automatique de la parole audiovisuelle. Dans un premier temps, nous décrivons en détail des résultats de la littérature classés selon le modèle d'intégration audiovisuelle retenu. Dans un second temps, nous présentons notre démarche, les modèles d'intégration testés et les résultats obtenus pour chaque cas.

Le chapitre 5 décrit le problème de l'extraction des paramètres labiaux et présente notre prototype électronique avec ses spécifications. Ce chapitre se conclut par une évaluation des performances de notre système par rapport aux mesures de référence.

Chapitre 1

Les lèvres et la production de la parole

Le processus de production de la parole est complexe et entraîne l'activation de nombreux muscles et articulateurs dont les commandes sont parfaitement synchronisées après un long processus d'apprentissage. Dans ce chapitre, nous proposons d'étudier ce système afin de mieux comprendre le rôle de chaque articulateur intervenant dans le processus de production de la parole. Nous donnerons une description anatomique de chaque élément de ce système, avant d'enchaîner sur les caractéristiques physiques.

L'accent sera mis sur les lèvres car elles constituent une grande part de notre travail de recherche de modélisation et d'extraction des informations visibles pour la reconnaissance audiovisuelle de la parole.

1. Le système vocal

1.1 Description anatomique

Le conduit vocal est capable de générer une grande variété de sons. Il est composé de trois parties principales :

- L'ensemble subglottique constitué par les poumons et la trachée artère. Il permet de fournir l'énergie nécessaire pour générer les sons.
- Le larynx dont les principaux éléments sont les cordes vocales : il s'agit de muscles recouverts de muqueuse qui sont générateurs d'impulsions périodiques en phonation voisée.
- Le conduit vocal qui lui-même est composé de deux parties :
    1. Le conduit oral qui est composé de la cavité pharyngale et buccale. Sa longueur est de l'ordre de 17 cm mais elle varie en fonction de l'âge et du sexe. Plusieurs articulateurs comme le pharynx, le voile du palais, la mandibule et la langue peuvent modifier sa forme, donc sa fréquence de résonance.

2. Le conduit nasal qui est constitué de deux cavités nommées les fosses nasales. Le voile du palais commande l'accouplement de ce conduit au conduit oral. Quand le voile du palais est abaissé, l'air passe par les deux conduits (nasal et oral) ce qui modifie les propriétés spectrales du conduit vocal.

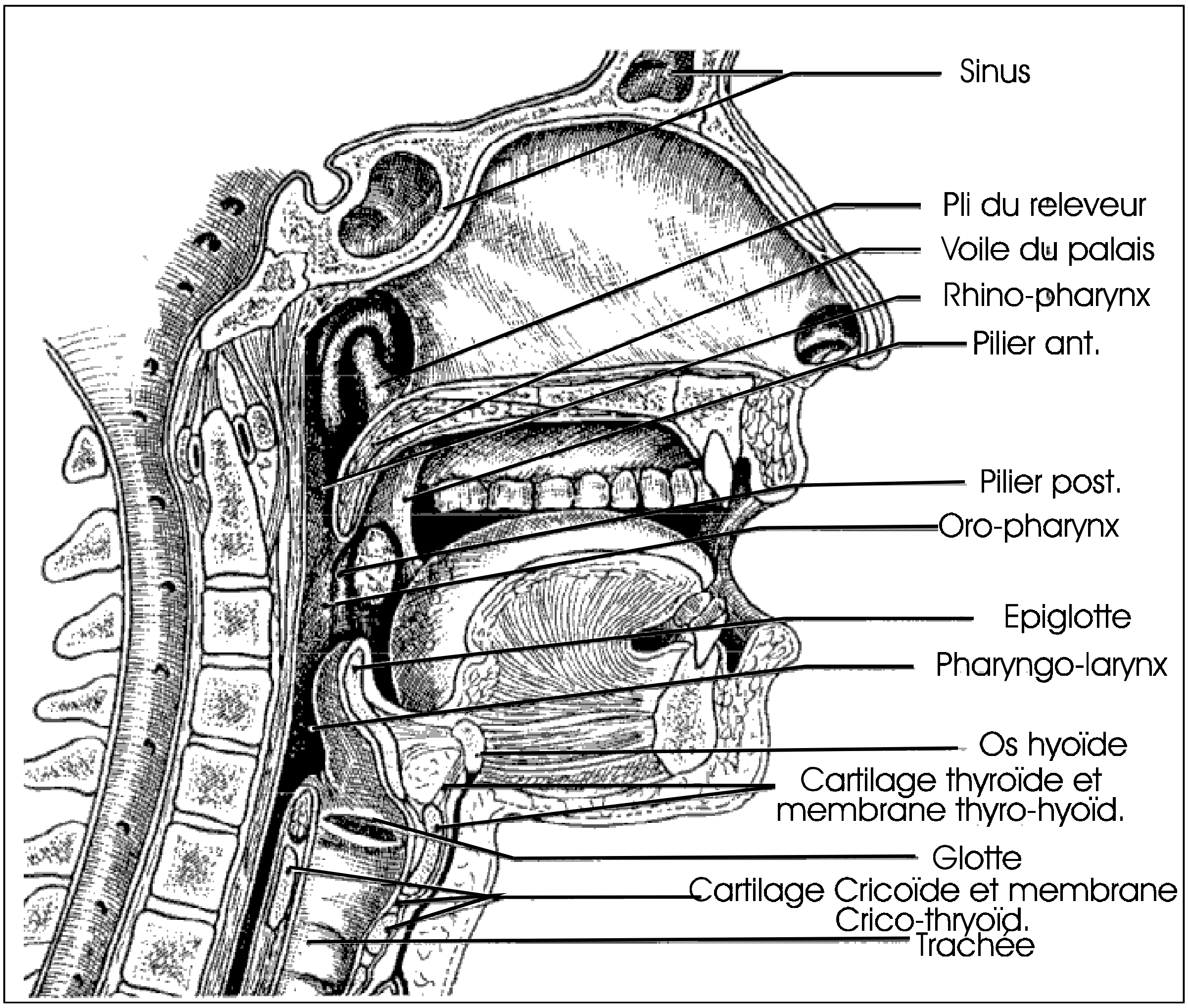


Figure 1.1 : coupe sagittale du système articulatoire d'après Cady et Kron (1971).

1.2 La production de la parole

La production de la parole repose sur le principe d'un instrument à vent et comprend un conduit (trachée, bronches, poumons) à travers lequel l'air s'écoule dans la cavité formée de la bouche et des fosses nasales en empruntant la fente limitée par les cordes vocales qui peuvent être mises en vibration. Le grand éventail des variations vocales s'explique par le fait que de nombreux muscles peuvent entrer en jeu et modifier considérablement de nombreux paramètres physiques. Par exemple en modifiant la pression du courant d'air, on peut changer la puissance sonore de la voix. La fréquence fondamentale est directement liée à la tension des cordes vocales et à l'ouverture et à la forme de la glotte. Les articulations et les muscles du larynx servent à mettre en position phonatoire les cordes vocales et la glotte.

Lorsque les cordes vocales vibrent sous l'effet du passage de l'air, la glotte ne fait pas que s'ouvrir et se refermer alternativement ; ses deux bords vibrent également de façon ondulatoire dans le sens du courant d'air. Ainsi lors de fréquences basses, autour de 100 Hz, la glotte demeure fermée plus longtemps qu'elle ne reste ouverte (dans un rapport de 5/1) ; lors des fréquences plus élevées (400 Hz), ce rapport chute à 1.4/1. Dans le chant et le sifflement, la glotte reste ouverte.

1.2.1 Les voyelles

Les voyelles orales se différencient entre elles, même si leurs fréquences fondamentales sont quasiment identiques (100 - 130 Hz), par l'adjonction de bandes de fréquences plus élevées propres à chacune d'elles. Ces zones de résonance sont appelées formants. Les formants sont produits par la déformation de la cavité oronasale. Ainsi, les voyelles /a/, /i/ et /u/ sont les extrêmes vocaliques articulatoires et acoustiques et les autres en sont les stades intermédiaires.

Les formants sont déterminés par la forme de la cavité qui est elle même modulée par la position de la langue. La Figure 1.2 montre les différentes formes de la cavité pour les voyelles /a/, /i/ et /u/.

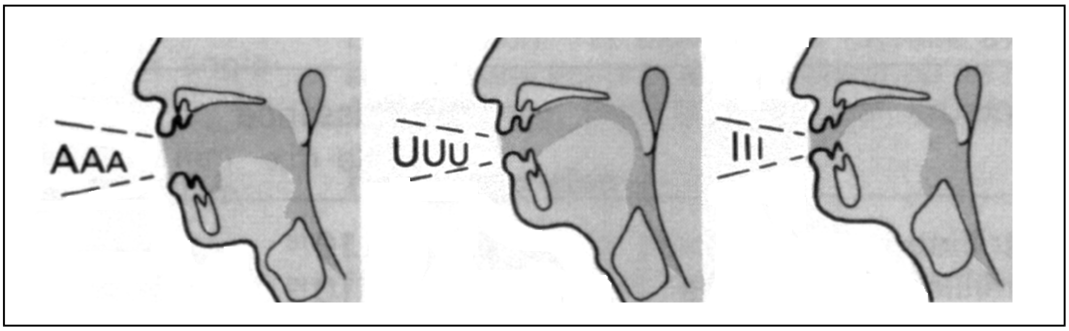


Figure 1.2 : la forme de la cavité buccale et les formants correspondant aux voyelles /a/, /i/ et /u/ (d'après Silbernagl et Despopoulos, 1992).

1.2.2 Les consonnes

Pour les consonnes en français, on distingue, selon le lieu d'articulation, les catégories suivantes :

- les bilabiales (lèvres) : /p/, /b/, /m/ ;
- les labiodentales (lèvre inférieure et incisives supérieures) : /f/, /v/ ;
- les alvéodentales (dents, alvéoles et partie antérieure de la langue) : /t/, /d/, /n/ ;
- les alvéolaires (alvéoles et partie antérieure de la langue) : /s/, /z/, /l/ ;
- les postalvéolaires (partie postérieure des alvéoles et partie antérieure de la langue) : /Σ/, /Z/ ;
- les palatales (palais dur et dos de la langue) : /ɲ/, /j/ ;
- les vélaires (palais mou et dos de la langue) : /k/, /g/, /{/.

Le mode d'articulation (vibration ou non des cordes vocales) permet de différencier /p/ et /b/, /t/ et /d/, /k/ et /g/, /f/ et /v/ et /s/ et /z/. On verra par la suite qu'en cas de lecture labiale seule, ces groupes de consonnes sont très difficiles à reconnaître car elles appartiennent à la même classe de forme de bouche.

2. Les lèvres

Les lèvres sont naturellement les organes qui transmettent la plus grande quantité d'information sur l'intelligibilité de la parole visuelle (voir chapitre 2). Nous allons donc les présenter avec plus de détails.

Dans un premier temps, nous décrivons quelques propriétés anatomiques, puis nous passerons à la géométrie et la modélisation des lèvres, plus spécifiquement le modèle 3D que nous avons développé.

2.1 Quelques éléments anatomiques

2.1.1 Plan cutané et muqueuse

Les lèvres constituent la terminaison antérieure du conduit vocal. Elles constituent deux replis musculo-membraneux mobiles, l'un supérieur, l'autre inférieur. Elles comportent une partie superficielle qui est recouverte de peau alors que la partie interne est recouverte de muqueuse. Entre la peau et la muqueuse se trouve une zone de transition appelée vermillon, qui est la partie charnue visible des lèvres.

- La lèvre supérieure part du bas de la pyramide nasale. Elle est limitée sur les côtés de la joue par le sillon naso-génien situé en arrière de l'aile du nez et qui devient ici naso-labio-génien. A la suite de la sous-cloison nasale, elle dessine une courbe dénommée arc de Cupidon.
- La lèvre inférieure est séparée de la région du menton par le sillon labio-mentonnier et latéralement par le prolongement du sillon naso-génien.

2.1.2 Plan musculaire

La production des mouvements articulatoires des lèvres est possible grâce à un tissu musculaire. Dans le fonctionnement musculaire, on distingue deux types : les muscles striés et les muscles lisses. Les muscles striés sont les seuls qui interviennent dans les phénomènes articulatoires volontaires. On peut distinguer deux catégories de muscles labiaux : les constricteurs et les dilatateurs.

La Figure 1.3 montre une vue détaillée des muscles de la face.

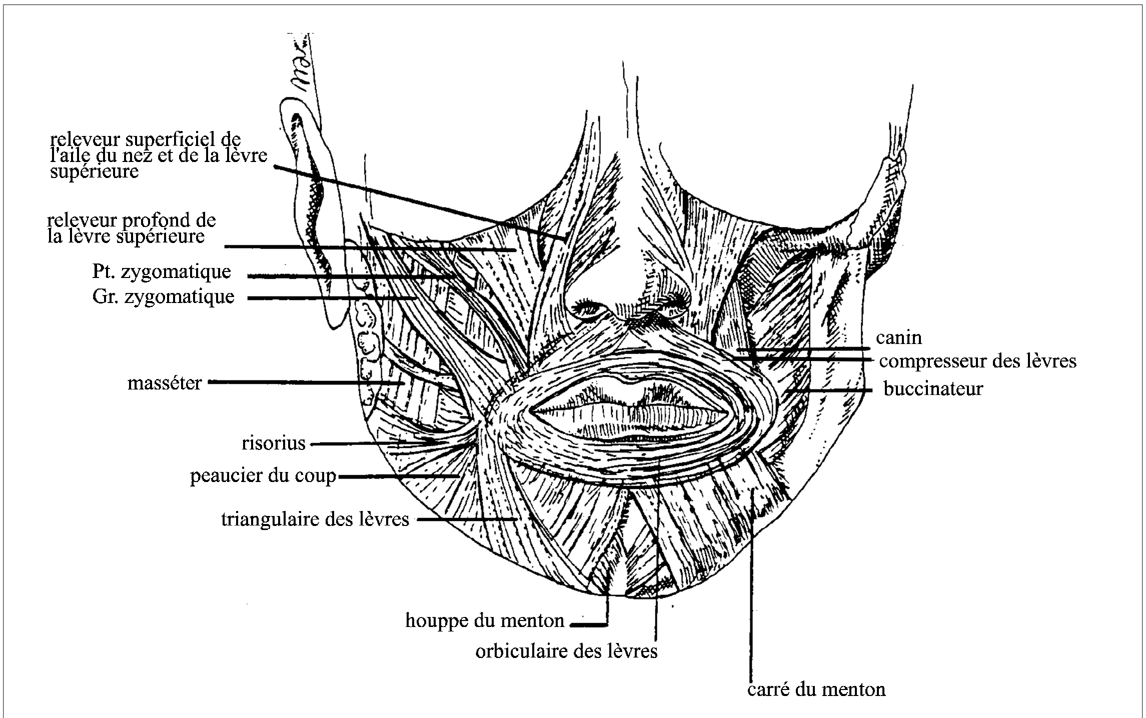


Figure 1.3: les muscles faciaux (d'après Bouchet et Guilleret, 1972).

2.1.2.1 Les muscles constricteurs

Le principal constricteur est l'orbiculaire des lèvres. Il contrôle le sphincter de l'orifice buccal. Autour de celui-ci, les fibres d'autres muscles de la face, essentiellement des muscles dilatateurs, s'imbriquent avec celles des orbiculaires. L'orbiculaire des lèvres permet l'occlusion, la protrusion et l'arrondissement des lèvres.

Le compresseur des lèvres ou le muscle de Klein est un muscle qui est particulièrement développé chez le nouveau né à cause de la succion. Il permet de comprimer les lèvres vers l'avant.

2.1.2.2 Les muscles dilatateurs

On observe deux catégories de muscles dilatateurs : les muscles dilatateurs profonds et les muscles dilatateurs superficiels.

Parmi les muscles dilatateurs profonds, on distingue :

- le canin : il est un muscle quadrilatère et aplati qui descend obliquement de la fosse canine à la commissure et à la lèvre inférieure. Il permet d'élever la commissure des lèvres et participe à l'élévation de la lèvre inférieure ;
- le buccinateur : il s'agit d'un muscle épais et aplati, de forme quadrilatère irrégulière. Son rôle est de tirer en arrière les commissures labiales ;
- le carré du menton : il est situé à la partie latérale de la région labio-mentonnière. Son rôle principal est d'attirer en bas et en dehors la moitié correspondante de la lèvre inférieure. Il participe aussi à l'ouverture labiale.
- les muscles de la houppe du menton : ce sont deux petits faisceaux musculaires situés de part et d'autre de la ligne médiane de la région mentonnière. Ces muscles sont élévateurs du menton et participent ainsi à l'occlusion labiale et à la protrusion de la lèvre inférieure.

Parmi les muscles dilatateurs superficiels, on peut nommer :

- le releveur superficiel : il s'agit d'un muscle mince situé dans le sillon naso-génien. Il participe à l'élévation de la lèvre supérieure.
- le petit zygomatique : c'est un muscle étroit et allongé qui est fixé sur la face profonde de la lèvre supérieure. Il participe à l'élévation de la lèvre supérieure.

- le grand zygomatique : c'est un muscle aplati et rubané qui se fixe à la peau et à la muqueuse et à la commissure des lèvres. Il participe à l'élévation des commissures et intervient aussi dans l'élévation de la lèvre supérieure.
- le triangulaire des lèvres : il s'agit d'un muscle large, aplati et triangulaire. Il permet l'abaissement des commissures et participe à l'abaissement de la lèvre inférieure.

2.1.3 Action des muscles labiaux dans la production de la parole

Plusieurs muscles sont activés pour assurer un mouvement labial dans le processus de la parole. On peut les classer, suivant leur activité, en trois catégories :

- les muscles protagonistes : ce sont les principaux responsables du mouvement à réaliser. Ils peuvent être aidés par d'autres muscles, mais constituent les principaux moteurs de l'action.
- les muscles antagonistes : ils ont une action opposée à celle d'autres muscles. Les muscles antagonistes sont soit inhibés pendant l'action des muscles protagonistes, soit contribuent à un mouvement contrôlé.
- les muscles synergiques : ils apportent leur contribution dans la réalisation d'une action des muscles protagonistes.

Gentil et Boë (1979) ont étudié l'activation des muscles protagonistes dans la production des sons. Ils ont classé les mouvements des lèvres en cinq catégories, et pour chaque mouvement, ils ont présenté les muscles en action.

Action	Muscles protagonistes
Abaissement/élévation de la lèvre supérieure	Orbiculaire supérieur, releveur de la lèvre supérieure, petit zygomatique
Abaissement/élévation de la lèvre inférieure	Orbiculaire de la lèvre inférieure, carré du menton
Protraction/closion des deux lèvres	Orbiculaires supérieur et inférieur
Protraction/éversion des deux lèvres	Releveur de la lèvre supérieure, la houpe du menton
Rétraction des commissures	Buccinateur

Tableau 1.1 : action des différents muscles dans les mouvements labiaux (d'après Gentil et Boë, 1979).

On peut aussi classer les muscles labiaux activés pour différents sons produits :

Sons	Muscles protagonistes
p, b, m	Orbiculaire supérieur, releveur de la lèvre supérieure, petit zygomatique, carré du menton
f, v	Orbiculaire de la lèvre inférieure, carré du menton
s, z, i, e, E, α	Buccinateur
u, o, y, O	Releveur de la lèvre supérieure et inférieure
, ↓, Σ,Z	Orbiculaires supérieur et inférieur, houpe du menton, releveur de la lèvre supérieure, carré du menton

Tableau 1.2 : activation des muscles labiaux suivant le son produit (d'après Gentil et Boë, 1979 ; adapté par Lallouache, 1991).

2.2 La géométrie des lèvres

La paramétrisation des lèvres constitue une étape importante dans notre travail de recherche. En vue de la modélisation des lèvres ou l'extraction des indices visuels pour la reconnaissance audiovisuelle de la parole, il est indispensable de connaître les éléments qui permettent de décrire les mouvements et de transmettre les informations visuelles portant sur l'intelligibilité. Après l'étude sur la physiologie des lèvres, dans cette partie, nous nous consacrons à quelques résultats de recherche sur les paramètres descripteurs des lèvres.

2.2.1 Les paramètres descripteurs

Fromkin (1964) a proposé des mesures sur la vue de face et de profil des lèvres, pour construire un ensemble de paramètres descriptifs, pour l'étude de la parole en anglais. Ces paramètres sont :

- l'étirement labial (A),
- la séparation intéro-labiale (B),
- l'aire intéro-labiale (C),
- l'aperture à l'extrémité du pavillon labial (D),
- la protrusion de la lèvre supérieure (E),
- la protrusion de la lèvre inférieure (F).

Ces mesures sont présentées sur la Figure 1.4.

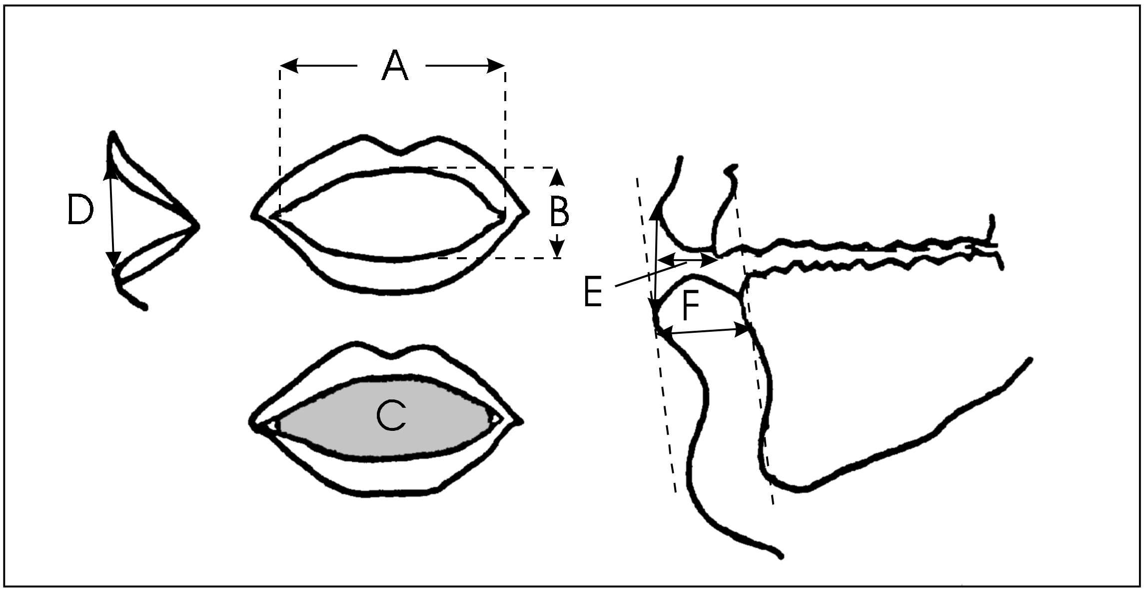


Figure 1.4 : paramètres géométriques des lèvres proposés par Fromkin (1964).

Linker (1982) a proposé une série de 24 paramètres labiaux pour l'étude des mouvements labiaux dans plusieurs langues.

Pour le français, Abry et Boë (1986) ont défini trois paramètres de face et cinq paramètres de profil pour la modélisation de la sortie du conduit vocal. Les paramètres de face étaient l'écartement, la séparation et l'aire du contour interne ; ceux de profil étaient la protrusion de la lèvre supérieure et inférieure et la protrusion du contact labial, l'aperture du pavillon et sa profondeur. Les mesures de profil étaient obtenues à partir de la position de référence des lèvres, c'est à dire l'état préphonatoire (Figure 1.5).

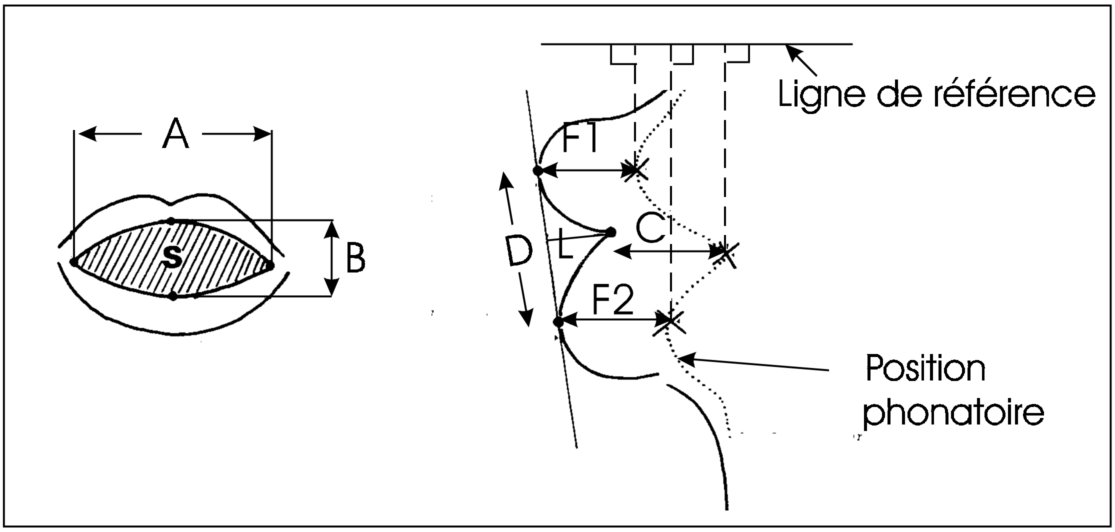


Figure 1.5 : paramètres de face et de profil d'après Abry et Boë (1986).

Lallouache (1991) a mis en œuvre un système de mesure labiale pour le calcul d'une douzaine de paramètres géométriques sur la vue de face et de profil. Pour les distances relatives, il a utilisé deux systèmes de référence : pour la vue de face, la référence a été fixée sur deux points sur les montures des lunettes et pour la vue de profil, une réglette fixée sur les lunettes. Nous reviendrons sur les détails de ce système dans le chapitre 3. Les paramètres de la vue de face de Lallouache (1991) sont :

- le paramètres B' : distance verticale entre le point le plus haut et le point le plus bas du contour externe du vermillon des lèvres,
- le paramètre A' : distance horizontale entre les deux points extrêmes du contour externe des lèvres,
- le paramètre B : distance verticale (hauteur) entre le point le plus haut et le plus bas du contour interne des lèvres,
- le paramètre A : distance horizontale entre les deux coins du contour interne des lèvres,
- le paramètre S : aire intéro-labiale,
- le paramètre S' : aire totale des lèvres,
- le paramètre M : distance verticale entre un point de référence sur le visage et un point sur le menton,
- les paramètres H et L : hauteur et longueur de l'arc de Cupidon,

Les paramètres de profil sont :

- le paramètre P1 : distance horizontale entre la référence et le point le plus avancé de la lèvre supérieure,
- le paramètre P2 : distance horizontale entre la référence et le point le plus avancé de la lèvre inférieure,
- le paramètre C : distance horizontale entre la référence et le coin du contact labial.
- les paramètres yP1 et yP2 : ce sont respectivement les distances verticales entre une ligne de référence et la protrusion de la lèvre supérieure et inférieure,
- le paramètre D : aperture à l'extrémité du pavillon labial,
- le paramètre L : profondeur du pavillon labial.

Les Figure 1.6 et Figure 1.7 montrent ces paramètres.

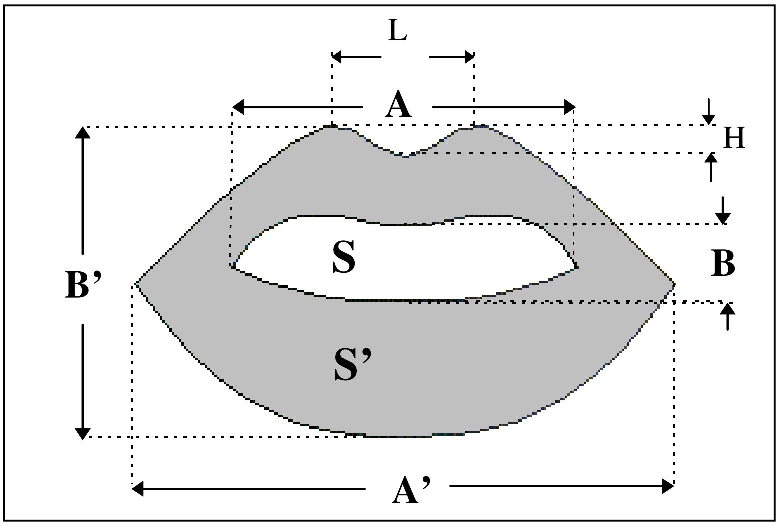


Figure 1.6 : les paramètres labiaux de face (Lallouache, 1991).

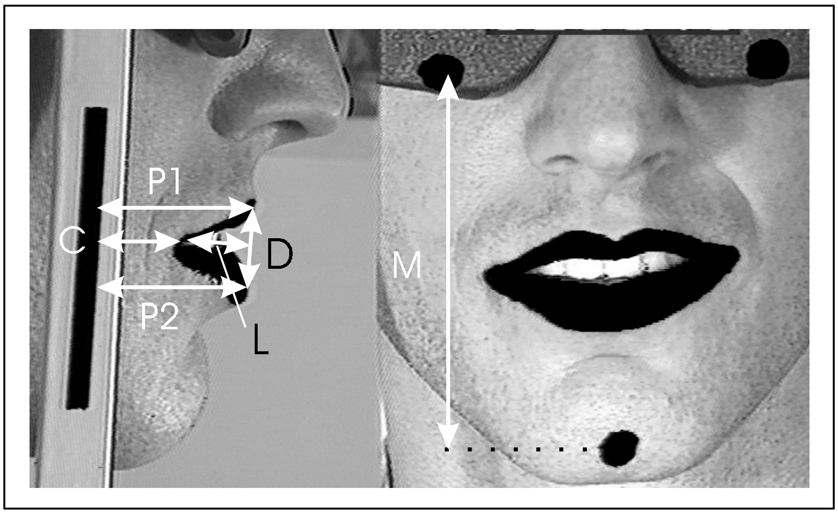


Figure 1.7 : les paramètres de profil et celui du menton. Ces distances nécessitent un point de repère fixe (Lallouache, 1991).

Nous verrons par la suite que ces mesures nous ont servi à l'élaboration du modèle paramétrique de lèvres ainsi qu'à la mise en œuvre d'un système de reconnaissance audiovisuelle. Elles constituent notre base de référence.

2.2.2 Corrélation entre les paramètres labiaux

De nombreuses études ont montré l'existence de corrélations entre certains paramètres labiaux (Lindblom et Sundberg, 1971 ; Abry et Boë, 1986). Après une analyse sur 768 images vidéo provenant des séquences de parole en français, Benoît et al. (1992) ont calculé le coefficient de corrélation entre différents paramètres de face. Le tableau suivant décrit ces résultats.

	B	B'	S	A*B
B'	0,96			
S	0,97	0,92		
A*B	0,97	0,92	0,999	
A'*B'	0,98	0,97	0,98	0,98

Tableau 1.3 : corrélations entre différents paramètres de face (d'après Benoît et al., 1992).

Ces résultats montrent aussi qu'il existe une très forte correspondance entre le produit des paramètres A (étirement interne) et B (hauteur interne) et le paramètre S (aire intéro-labiale).

A part l'existence de corrélations, certains paramètres ont une plus grande plage de variation par rapport aux autres, ce qui permet de les différencier en mode de lecture labiale seule (voir aussi chapitre 2).Les résultats de recherche de Abry et Boë (1980, 1986) ont montré que le paramètre d'étirement intéro-labial (A) est particulièrement important dans la classification des voyelles arrondies et étirées.

Benoît et al.(1992) ont étudié les variations des paramètres A et B pour les différentes formes de lèvres en français. Ils ont observé deux types de corrélations : les formes labiales non arrondies ont un étirement quasi constant et varient uniquement par rapport à la hauteur, tandis que pour les formes labiales arrondies, ces deux paramètres (étirement et hauteur) varient simultanément suivant une relation du type parabolique. La Figure 1.9 représente les résultats de cette analyse.

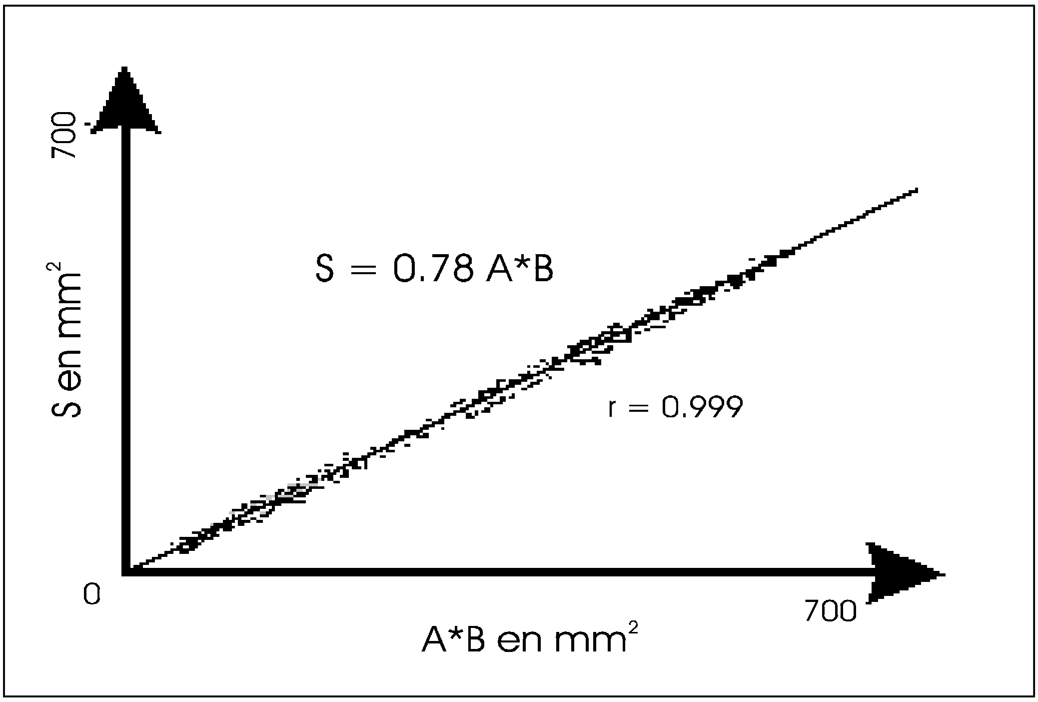


Figure 1.8 : corrélation entre le paramètre S et le produit A et B (d'après Benoît et al., 1992).

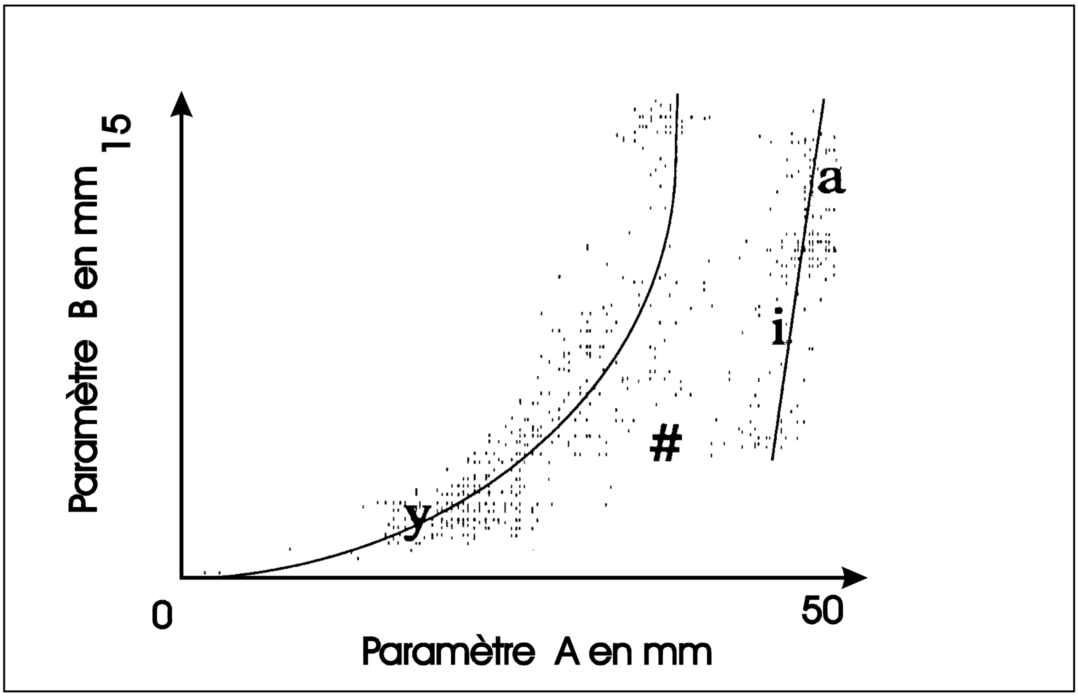


Figure 1.9 : corrélation entre les paramètres A et B pour les formes labiales dans différents contextes. Le symbole # représente la forme de bouche correspondant à la position pré-phonatoire (d'après Benoît et al., 1992).

Ces résultats mettent en évidence l'importance de ces deux paramètres dans la classification des voyelles. En d'autres termes, les paramètres A et B peuvent fournir une grande quantité d'information dans la distinction de ces voyelles et cela en mode de lecture labiale seule. Nous reviendrons sur les expériences d'intelligibilité visuelle et la pertinence des paramètres dans le chapitre 2 (page 32).

2.3 La modélisation des lèvres

Les lèvres sont les articulateurs visibles qui peuvent transmettre une grande partie de l'intelligibilité de la parole quand le canal de transmission auditif est bruité (voir chapitre 2 page 27). C'est pourquoi, une modélisation fidèle de ces organes avec un nombre réduit de paramètres de commande peut, contribuer en grande partie à la restitution de l'intelligibilité manquante. Dans cette partie, nous décrivons quelques travaux de recherche sur la modélisation des lèvres. Nous présenterons ensuite les différentes étapes à la réalisation du modèle de lèvres développé à l'Institut de la Communication Parlée.

2.3.1 Quelques modèles

Les premiers travaux de recherche sur la modélisation paramétrique des contours labiaux ont été élaborés par Lindblom et Sundberg (1971) pour l'anglais. Comme le contour interne est acoustiquement plus important dans le processus de la parole, ils ont modélisé uniquement ce contour à l'aide de deux équations polynomiales. Ces équations, commandées par les paramètres A (étirement interne) et B (hauteur interne), permettaient de reconstruire assez fidèlement le contour interne des lèvres. La Figure 1.10 montre le contour interne ainsi reconstruit et les équations descriptrices. La caractéristique importante de ce modèle est qu'il permet de vérifier la relation entre les paramètres A et B et le paramètre S (section 2.2.2 page 14).

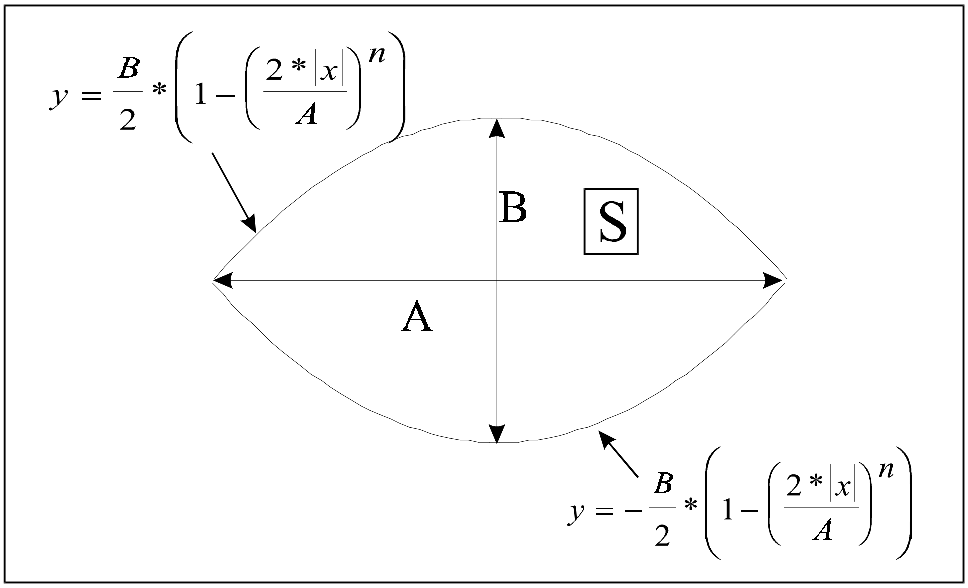


Figure 1.10 : le contour interne reconstruit à partir de deux équations polynomiales. Le centre du système de coordonnées est le point d'intersection de la séparation et de l'étirement internes (d'après Lindblom, 1971).

Abry et Boë (1986) ont modélisé le contour interne des lèvres avec deux fonctions polynomiales symétriques commandées par les paramètres A et B. Pour la vue de profil, ils ont proposé deux courbes paraboliques.

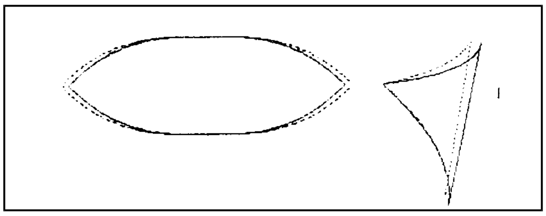


Figure 1.11 : modélisation du contour interne des lèvres vues de face et de profil (d'après Abry et Boë, 1986). Les courbes en pointillés sont les contours prédits et celles en trait plein sont les contours réels pour la voyelle /a/.

Brooke et Summerfield (1983) ont centré leur recherche sur la modélisation du visage dans lequel le contour interne des lèvres était présenté par deux arcs de cercle, le contour externe supérieur par six segments de droite et le contour externe inférieur par des arcs de cercle tangents (Figure 1.12).

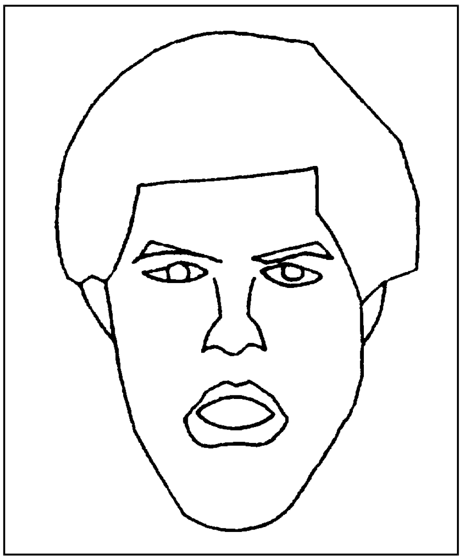


Figure 1.12 : modèle de visage de Brooke et Summerfield (1983). Les contours internes et externes des lèvres sont représentés par des segments de droite et des arcs de cercle.

Dans d'autres travaux de recherche plus récents, la modélisation des lèvres a été utilisée comme un outil pour la détection des paramètres labiaux (Hennecke et al., 1994, 1996 ; Coianiz et al., 1996 ; Vogt, 1996). Les résultats de ces recherches seront présentés dans le chapitre 3 (page 55).

2.3.2 Elaboration du modèle de lèvres de l'ICP

Dans un premier temps, Guiard-Marigny (1992) a développé un modèle de lèvres bidimensionnel. Il a centré son étude sur l'analyse des différentes formes de lèvres (Benoît et al., 1992) qui a permis de déterminer une base de données contenant les 22 formes de bouches en français (Figure 1.13).

Le modèle 2D de Guiard-Marigny (1992) est un modèle entièrement symétrique constitué de deux contours : le contour interne et le contour externe des lèvres, chacun représenté par des équations mathématiques simples. La Figure 1.14 et la Figure 1.15 montrent les contours de ce modèle et les équations mathématiques pour chaque partie.

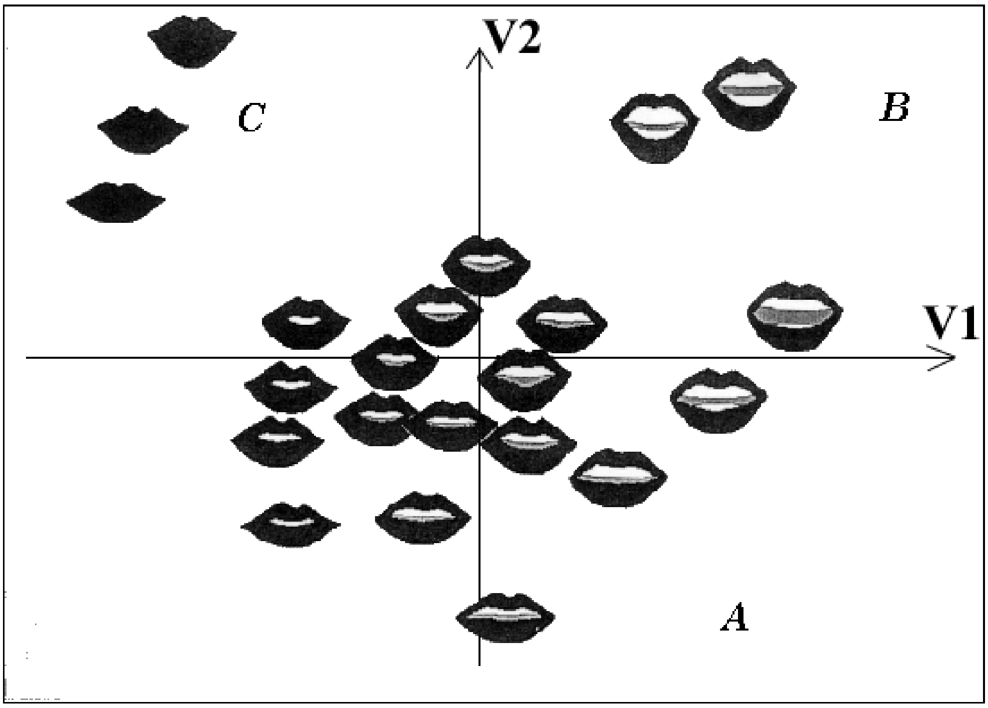


Figure 1.13 : projection simultanée des 22 formes de bouche et les trois paramètres A, B et C sur les deux premiers facteurs (d'après Benoît et al., 1992).

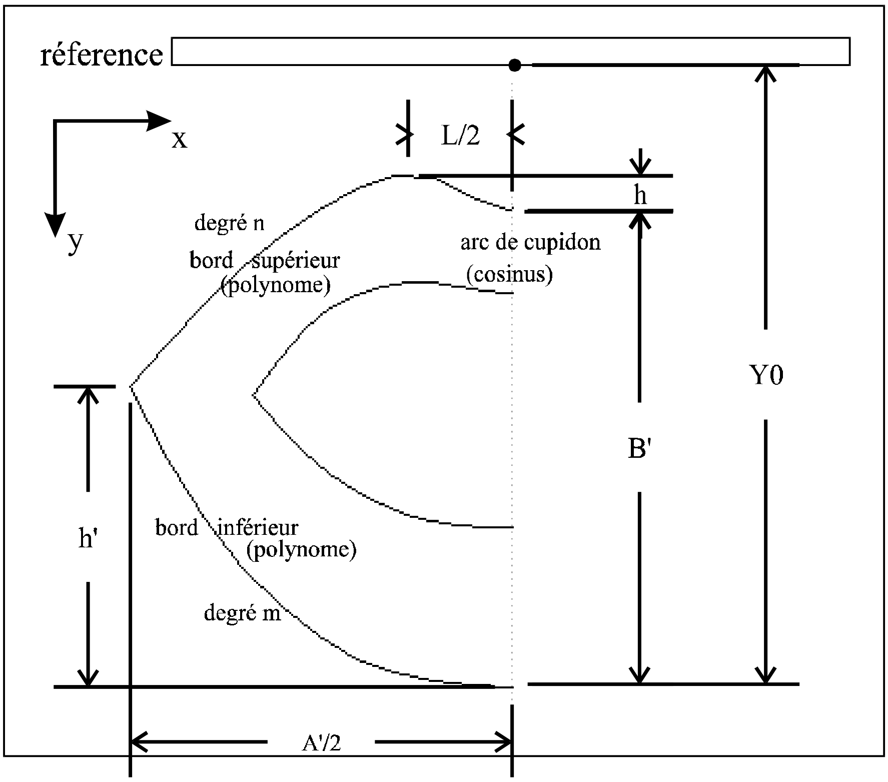


Figure 1.14 : modélisation du contour externe des lèvres utilisant des équations mathématiques simples. L'arc de Cupidon est représenté par une fonction sinusoïdale alors que les autres parties du contour sont modélisées grâce à des fonctions polynomiales (d'après Guiard-Marigny, 1996).

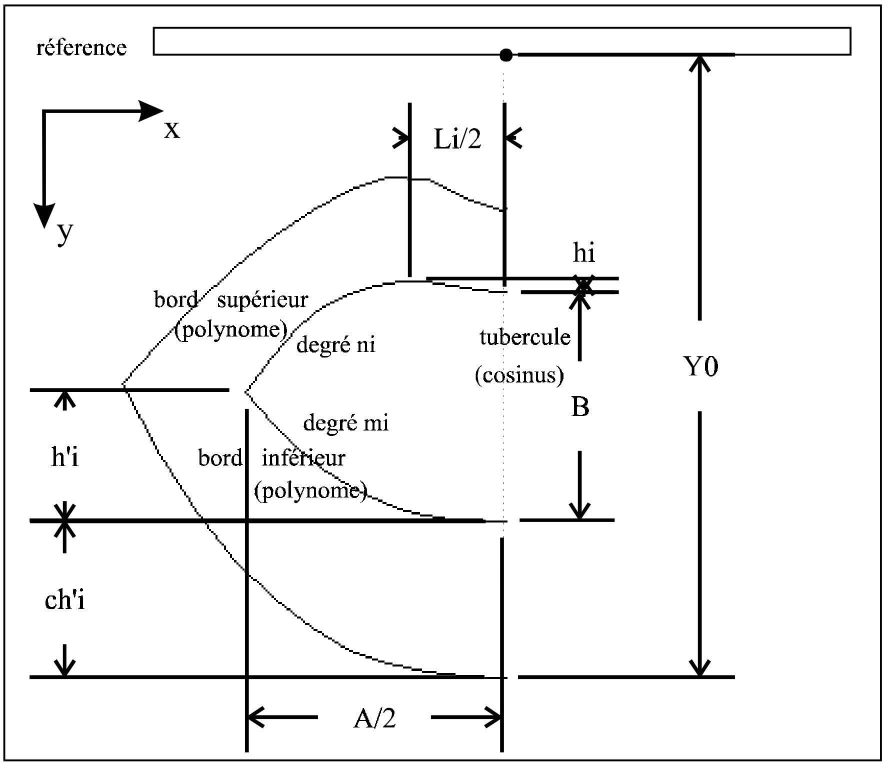


Figure 1.15 : le contour interne du modèle 2D représenté par des fonctions polynomiales et sinusoïdales (d'après Guiard-Marigny, 1996).

Ces fonctions nécessitaient de nombreux paramètres de commande (Figure 1.14 et Figure 1.15). Il a donc fallu réduire le nombre de paramètres du modèle, pour qu'il puisse être animé par un minimum de paramètres facilement exploitables. Pour cela, Guiard-Marigny (1992) a procédé à une analyse de données sur les paramètres de commande de son modèle et a obtenu des régressions qui lui ont permis de réduire les dimensions de l'espace de commande. Il n'a finalement gardé que trois paramètres pour l'animation du modèle : l'étirement interne (A), la séparation interne (B) et la protrusion du contact labial (C). Ces trois paramètres permettent de reconstruire assez fidèlement les 23 formes de contours de lèvres originales.

En partant de ce modèle 2D, nous avons élaboré un modèle tridimensionnel (Adjoudani, 1993). Nous avons repris les équations mathématiques des contours interne et externe 2D (en X et Y) pour ajouter la troisième composante (en Z). En observant la forme des lèvres dans le plan perpendiculaire au plan frontal, nous avons constaté qu'une relation mathématique ne dépendant que de la composante X pouvait reconstruire assez fidèlement la composante en Z dans le plan sagittal. En effet, la forme de la dentition et les contraintes physiologiques justifient ce choix. Nous avons donc modélisé la composante en Z grâce à une fonction polynomiale de la composante en X des contours ($z = f(x)$). Comme dans le cas de la modélisation 2D, notre base de données était constituée de l'ensemble des 23 visèmes. Etant donné que cette base contenait uniquement les contours de face, nous avons échantillonné, pour chaque visème, les contours des lèvres sur des images vidéo de face, et calculé la composante en Z sur l'image de profil. Le pas d'échantillonnage retenu offrait un nombre de points suffisants pour la modélisation : environ une quarantaine de points pour le contour externe et une vingtaine pour le contour interne. Le calcul de correspondance était réalisé à l'aide d'une interpolation linéaire entre l'image de face et celle de profil. Connaissant la taille d'un pixel sur chaque image et ayant un point de référence sur les deux images (par exemple le point le plus haut de l'arc de Cupidon), nous avons obtenu pour chaque point, ses coordonnées en X, Y et Z. La Figure 1.16 montre cette opération.

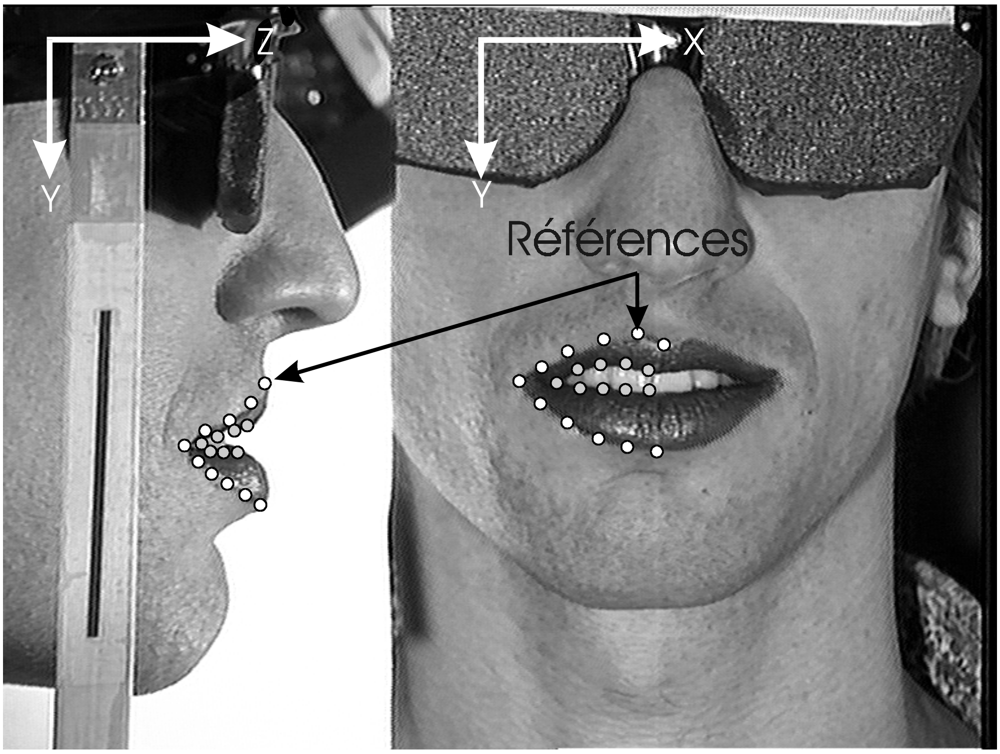


Figure 1.16 : obtention de la composante en Z de chaque point sur l'image de face. Le point de référence et une interpolation linéaire permettent le passage entre les deux systèmes de coordonnées.

Avec cette nouvelle base de données contenant les trois composantes des points de contours pour les 23 visèmes, nous avons pu valider nos fonctions polynomiales dans le plan sagittal. Les coefficients de ces fonctions ont été ajustés de façon à minimiser l'erreur entre les contours prédits et ceux des visèmes. Comme dans le cas de la modélisation 2D, nous avons réduit l'espace de commande de ces fonctions afin d'utiliser un nombre minimal de paramètres facilement exploitables sur les images vidéo. Nous avons ainsi ajouté uniquement deux nouveaux paramètres de commande pour prédire les contours des lèvres dans le plan sagittal : la protrusion de la lèvre supérieure (P1) et la protrusion de la lèvre inférieure (P2). La Figure 1.17 montre le principe de la modélisation 3D.

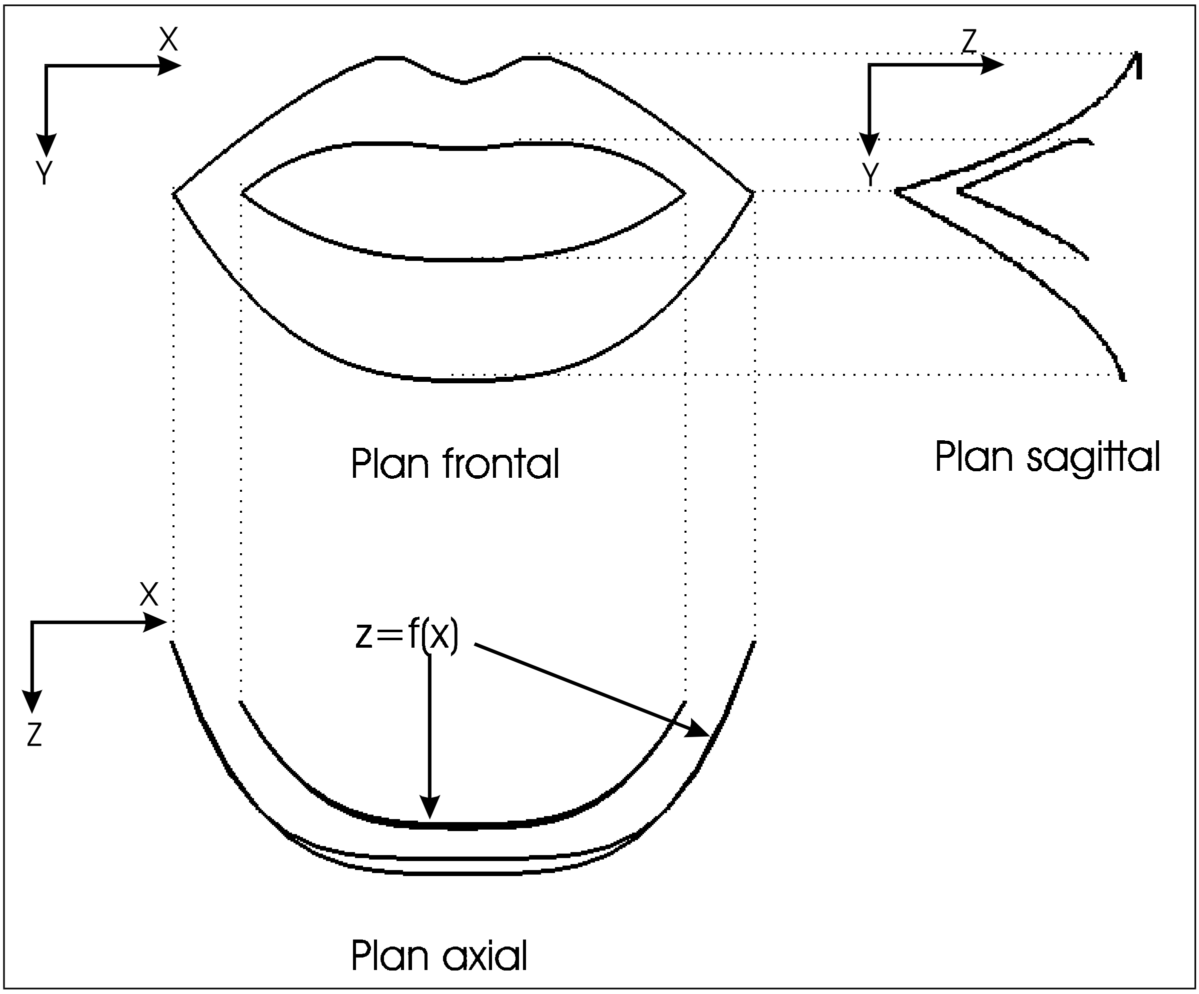


Figure 1.17 : la modélisation 3D à partir des contours 2D. Notre choix consiste à calculer la composante en Z de chaque point uniquement à partir de la composante en X.

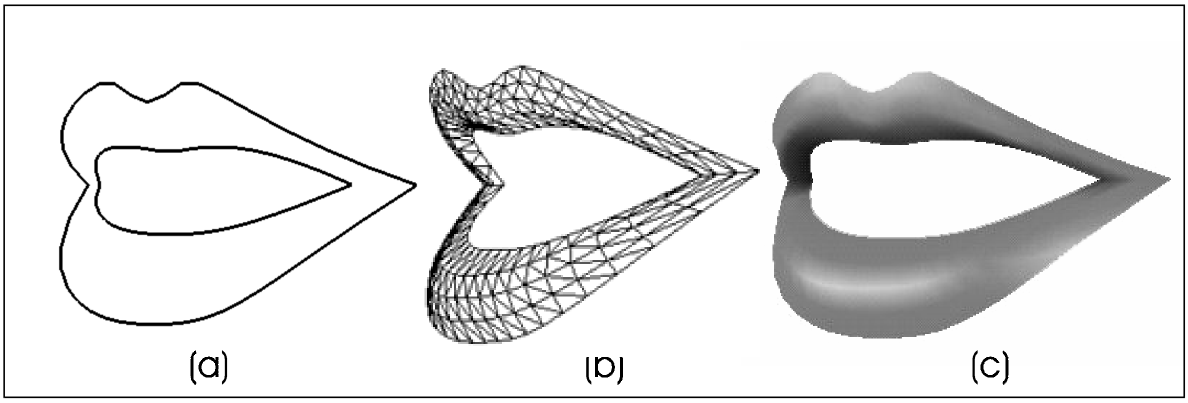


Figure 1.18 : le modèle de lèvres sous différents aspects :

a - les deux contours interne et externe,

b - l'ensemble des contours et le réseau de polygones,

c - le modèle lissé grâce à la technique de Gouraud.

Pour donner du volume aux lèvres, nous avons ajouté trois autres contours intermédiaires, obtenus par une simple interpolation linéaire entre les contours externe et interne. En reliant les points de deux contours adjacents entre eux, nous avons créé un réseau de polygones qui est ensuite lissé grâce à la technique de Gouraud (1971). Ces traitements permettent de donner un aspect naturel aux lèvres.

Le modèle est finalement commandé par seulement cinq paramètres géométriques, qui sont facilement exploitables par traitement vidéo (Figure 1.19).

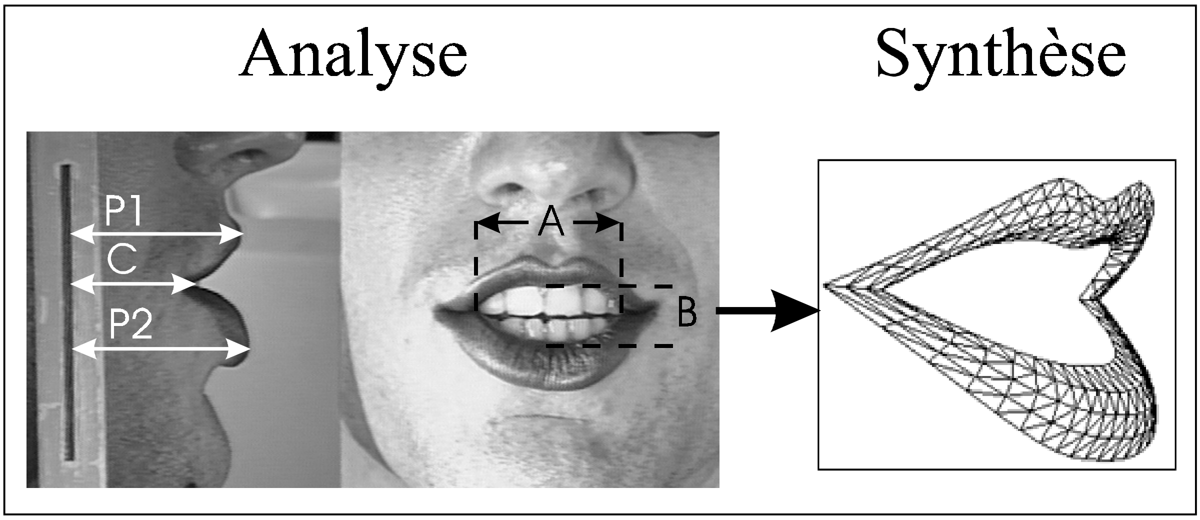


Figure 1.19 : principe de reconstruction des lèvres à partir de cinq paramètres géométriques des lèvres.

Dans une récente étude, Guiard-Marigny (1996) a obtenu des corrélations entre les paramètres de face et ceux de profil. Il a ainsi trouvé des équations qui permettent de prédire les paramètres de profil à partir de ceux de face. Ces résultats sont d'une grande importance car ils permettent d'alléger le système d'analyse en le limitant au calcul des paramètres de face.

2.3.3 Validation du modèle

La meilleure méthode pour valider un modèle de lèvres est la quantification de ses capacités de transmission d'information (voir chapitre 2). En d'autres termes, il suffit de calculer le gain d'intelligibilité du modèle par rapport au gain obtenu par des lèvres naturelles. Bien que les résultats des expériences de perception de la parole et l'évaluation de l'apport d'intelligibilité des indices visuels soient présentés en détail dans le chapitre 2 de ce manuscrit, nous décrivons une partie de l'expérience de perception de Le Goff et al. (1994) sur l'apport d'intelligibilité du modèle de lèvres par rapport aux lèvres naturelles.

Le Goff et al.(1994) ont calculé le taux de reconnaissance de 18 stimuli du type VCVCV en présence de différents niveaux de bruit additif et sous deux conditions visuelles :

- l'image binaire de face des lèvres naturelles,
- l'image du modèle 3D lissé avec la technique de Gouraud présentée de trois quarts.

Les résultats de cette expérience sont présentés sur la Figure 1.20.

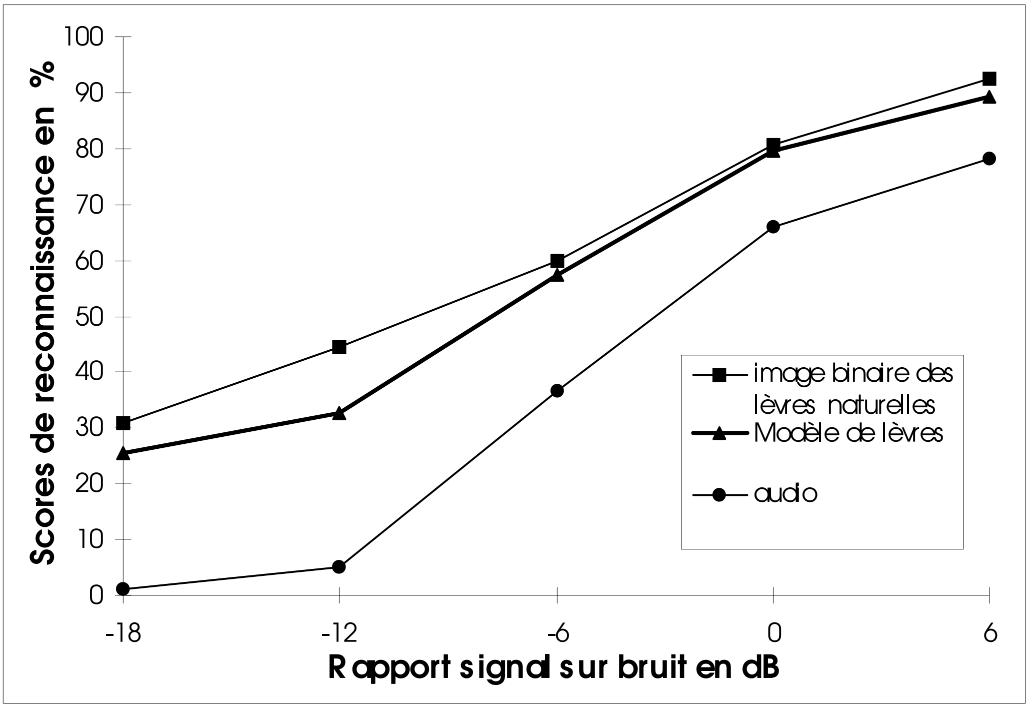


Figure 1.20 : apport d'intelligibilité du modèle de lèvres par rapport aux lèvres naturelles (d'après Le Goff et al., 1994).

Ces résultats mettent en évidence la restitution d'une grande partie de l'information visuelle par le modèle de lèvres même si dans cette expérience, la présentation du modèle était différente de celle des lèvres naturelles. Ils constituent une bonne référence à la validation du modèle 3D.

La Figure 1.21 montre les formes extrêmes des lèvres naturelles (Benoît et al., 1992) correspondant à /izi/, /yby/ et /aZa/ et celles reconstruites par le modèle.

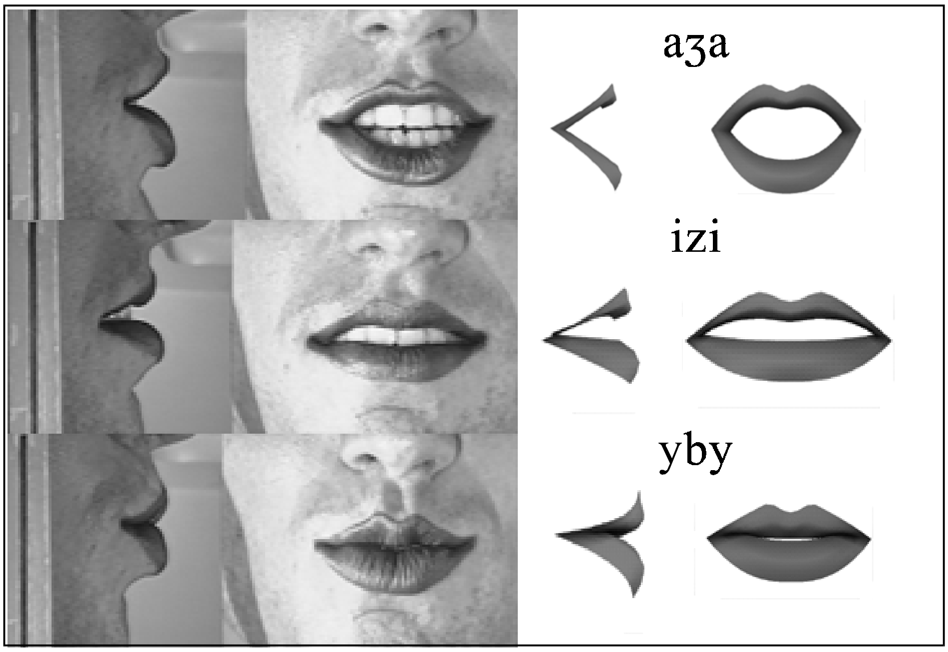


Figure 1.21 : reconstruction des formes extrêmes des lèvres et comparaison avec les images réelles.

2.3.4 Applications

La modélisation des lèvres présente de nombreux intérêts. Cette partie est consacrée à l'énumération de quelques applications potentielles du modèle de lèvres.

2.3.4.1 Transmission d'intelligibilité à bas débit

Grâce à un modèle de lèvres ou de visage commandé par des paramètres géométriques, on peut transmettre une grande quantité d'information nécessaire à l'intelligibilité de la parole visuelle à un débit très faible (c'est le but recherché dans les projets actuellement en cours comme MPEG 4, VIDAS, SNHC ...). Un simple calcul permet de vérifier cette hypothèse : si on envisage une application d'animation du modèle à distance, par exemple par l'intermédiaire du canal téléphonique, il est nécessaire de transmettre au moins 25 séries de paramètres géométriques par seconde (avec une cadence moins élevée, les mouvements rapides des lèvres ne pourront pas être analysés). Pour chaque image, cinq paramètres labiaux doivent être envoyés sur le canal. Si chaque paramètre est codé sur 8 bits, il faut donc transmettre 40 bits par image ce qui donne un débit de 1 Kbits/sec. Ce taux de transmission est largement inférieur à celui d'une transmission téléphonique normale qui est de l'ordre de 32 Kbits/sec. On peut donc envisager de transmettre les paramètres de commande du modèle en même temps que le signal acoustique codé. Un simple système de multiplexage permettra, ensuite à la réception, de séparer ces deux informations et de les traiter dans deux chaînes différentes.

2.3.4.2 Analyse des lèvres

Les techniques de détection de contours des lèvres, pour un calcul précis des paramètres géométriques, nécessitent des pré-traitements qui ne sont pas souvent pratiques dans une interface homme/machine. L'utilisation de maquillage ou de pastilles sur les lèvres sont des exemples de ce type de contraintes. Pour pallier ce problème, de nombreux chercheurs se sont intéressés à l'utilisation de modèles de lèvres pour le calcul des paramètres (voir chapitre 3, page 64). Cette technique consiste à plaquer le modèle sur l'image des lèvres de façon à minimiser l'erreur ou la fonction d'énergie. Une fois que les contours du modèle et ceux de l'image des lèvres sont le plus proche possible, on peut directement obtenir les paramètres labiaux à l'aide des équations mathématiques du modèle.

2.3.4.3 Synthèse audiovisuelle à partir du texte

La synthèse audiovisuelle à partir du texte est un outil qui permet de prédire les paramètres labiaux uniquement à partir d'une chaîne phonémique marquée en durée (Le Goff et Benoît, 1996a, 1996b). Ces paramètres servent ensuite à animer un modèle paramétrique de lèvres ou de visage.

3. Conclusion

Nous avons décrit brièvement le principe de production de la parole et quelques organes intervenant dans ce processus. Une importance particulière a été donnée aux lèvres car elles se situent au centre de notre travail de recherche. Le chapitre suivant est consacré à l'étude de l'intelligibilité transmise par les lèvres dans le cas de la lecture labiale ou un environnement auditif fortement perturbé. Nous présenterons quelques résultats de recherche sur la perception audiovisuelle de la parole, ainsi que le processus d'intégration des informations auditives et visuelles.

Chapitre 2

La perception auditive et visuelle de la parole

La perception de la parole est un processus complexe, dans lequel l'ouïe et la vision sont impliquées. Le but de ce chapitre est de montrer l'importance de la vision dans ce processus. Nous présentons quelques résultats sur la perception de la parole visuelle et audiovisuelle, et décrivons ensuite les différentes architectures d'intégration des informations auditives et visuelles qui tentent de modéliser le phénomène de fusion de données chez l'homme.

1. Quelques expériences sur la lecture labiale

De nombreuses expériences (Jackson et al., 1976; Montgomery et Jackson, 1983) ont montré l'importance des données physiques des mouvements des lèvres en lecture labiale. Ces études ont montré que les performances de la perception visuelle des voyelles en anglais chez les sujets malentendants étaient directement liées à la nature de l'entrée visuelle. Ils ont ainsi observé que les paramètres les plus informatifs étaient l'étirement (paramètre A), la hauteur (paramètre B) et l'aire intéro-labial (paramètre S).

Dans une étude similaire, Plant (1980) a mis en évidence l'importance du contour interne dans la perception visuelle pure des voyelles en anglais. Ces études ont montré que la quantité d'information transmise par la vision du contour interne était plus importante que celle du contour externe. En effet, les caractéristiques acoustiques de la parole dépendent de la forme de l'orifice buccal (contour interne) et non pas de la forme du contour externe des lèvres. Plant (1980) a aussi montré que l'étirement et la hauteur des lèvres sont importants pour expliquer les confusions perceptives.

D'autres études (Erber, 1974 ; Summerfield, 1987) ont mis en évidence l'importance de la vision de face par rapport à celle de profil dans la perception visuelle de la parole. Les résultats montrent que les performances de la lecture labiale sont plus importantes si les sujets disposent de la vue de face du locuteur. La vue de profil est porteuse d'information mais dans une échelle moins importante.

Cathiard (1988) a étudié l'importance de la protrusion sur l'identification visuelle de /si, sy, Σi, Σy/. L'expérience consistait à reconnaître la syllabe prononcée à partir des images statiques correspondant à la réalisation de la consonne ou de la voyelle sous deux conditions visuelles : la vue de face et la vue de profil. Les résultats montrent que les scores d'identification sont équivalents dans les deux cas. À l'aide d'une analyse factorielle des correspondances sur les résultats des tests de perception et les données articulatoires, elle a mis en évidence les paramètres prépondérants dans l'identification de chaque classe :

- le paramètre A pour le /i/ et le /s/ de /si/,
- les paramètres de protrusion pour le /y/ de /sy/ et de /Σy/,
- le paramètre B pour le /i/ de /Σi/,
- le paramètre A et les paramètres protrusion pour le /s/ de /sy/,
- les paramètres de protrusion et le paramètre B pour le /Σ/ de /Σi/ et de /Σy/.

Une autre expérience de Cathiard (1994) consistait à étudier la perception visuelle du mouvement d'arrondissement entre les voyelles /i/ et /y/. Les sujets devaient identifier ces deux voyelles dans les transitions /i/ /i/ ou /i/ /y/ sous plusieurs conditions visuelles : l'image statique de profil, l'image statique de face, la séquence d'images de face, la séquence d'images de profil. L'image statique correspondait à une image $I(t_0)$ choisie à divers instants t_0 et la séquence d'images était constituée de la suite $I(0), I(1), ..., I(t_0)$. Cathiard (1994) a montré que la vue de profil en dynamique et statique donnait les mêmes scores d'identification alors que les scores correspondant à la vue de face en dynamique étaient meilleurs que ceux obtenus en statique. Elle a aussi observé que les taux d'identification correcte à partir des images dynamiques de face étaient quasiment identiques à ceux de profil (statique et dynamique).

Compte tenu de l'ensemble de ces résultats, on peut affirmer que la vision de face apportent plus d'intelligibilité. Sauf dans des cas spécifiques qui concernent la classification de formes labiales étirées et protrues (Cathiard, 1994) où la vue de profil permet de fournir une quantité importante d'information, les paramètres de face sont généralement plus pertinents que ceux de profil.

2. La parole bimodale

La vision du visage du locuteur apporte une grande partie de l'intelligibilité manquante à la parole auditive quand celle-ci est fortement dégradée par la présence de bruit. Cette section est consacrée à l'étude de l'intelligibilité de la parole bimodale chez l'homme et à l'influence de la vision sur la perception de la parole. Nous allons décrire les résultats de quelques expériences de perception audiovisuelle pour mieux comprendre le mécanisme complexe de la reconnaissance bimodale. Ces résultats sont la base de notre travail de recherche pour l'implémentation d'un système de reconnaissance automatique de la parole audiovisuelle.

2.1 Quelques résultats sur l'intelligibilité de la parole bimodale

2.1.1 Apport d'intelligibilité de la vision en présence du bruit

De nombreuses recherches ont mis en évidence l'apport d'intelligibilité de la vision du locuteur quand le signal de parole est dégradé par le bruit ambiant (Sumby et Pollack, 1954 ; Neely, 1956 ; Erber, 1969, 1975 ; Binnie et al., 1974 ; Summerfield, 1979 ; Mac Leod et Summefield, 1987 ; Mohammadi et Benoît, 1992 ; Mohammadi, 1993 ; Benoît et al., 1994a, 1994b). Toutes ces études confirment le gain d'intelligibilité de la modalité audiovisuelle par rapport à la modalité auditive seule. Dans cette partie, nous présentons les résultats de quelques unes de ces études qui déterminent notre axe de recherche.

Sumby et Pollack (1954) ont étudié la perception audiovisuelle de la parole en présence de bruit. L'expérience consistait à reconnaître des mots dans deux conditions de test : audio seul et audio avec l'image du locuteur. Pour chaque cas, différents niveaux de bruit ont été ajoutés au signal audio. Le vocabulaire de test était constitué de deux séries de mots (la première de 32 mots et la seconde de 8 mots). En absence de bruit, la vision du locuteur n'apporte pas d'intelligibilité et la modalité auditive est suffisante pour décoder le message parlé. En revanche, plus le niveau de bruit devient important, plus le gain d'intelligibilité audiovisuelle augmente. D'autre part, on note que le pourcentage d'identifications correctes est directement lié à la taille du corpus de test. Nous avons reporté les résultats de cette étude sur la Figure 2.22, et avons ajouté sur cette figure les droites (V), qui correspondent à la condition visuelle seule, ou au niveau aléatoire du score audio si cette condition n'avait pas été testée.

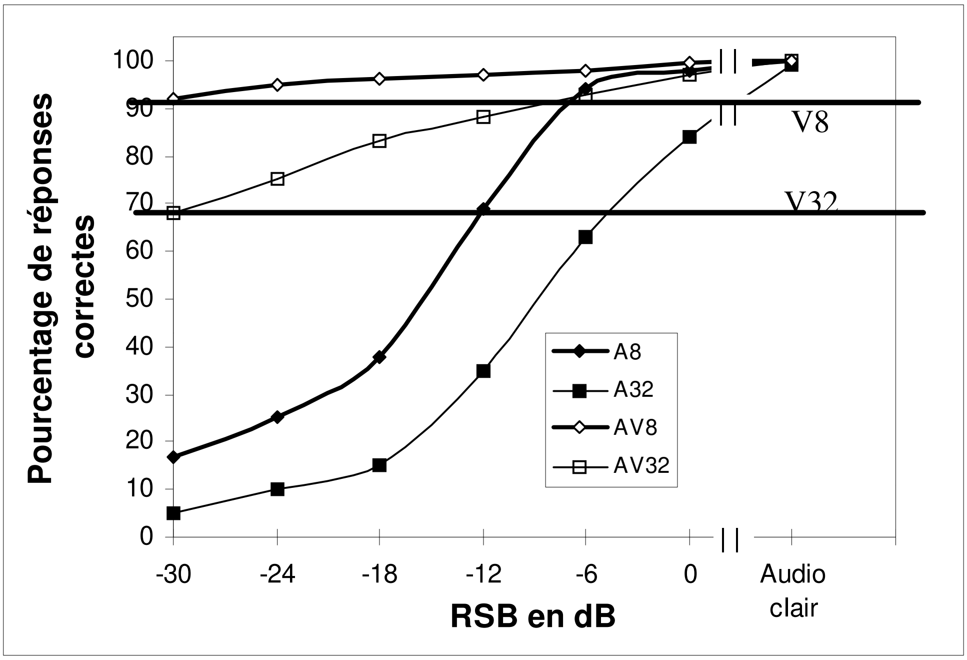


Figure 2.22 : pourcentage de réponses correctes en fonction du niveau de dégradation du signal audio. Les courbes A8 et A32 correspondent à la modalité auditive seule obtenue avec un vocabulaire de 8 et 32 mots respectivement. Les courbes AV8 et AV32 sont obtenues à partir des scores de reconnaissance audiovisuelle sur 8 et 32 mots respectivement (d'après Sumby et Pollack, 1954).

Dans une autre expérience semblable, Erber (1969) a étudié la perception audiovisuelle dans une grande plage de dégradation auditive. Les résultats de cette étude sont présentés sur la Figure 2.23.

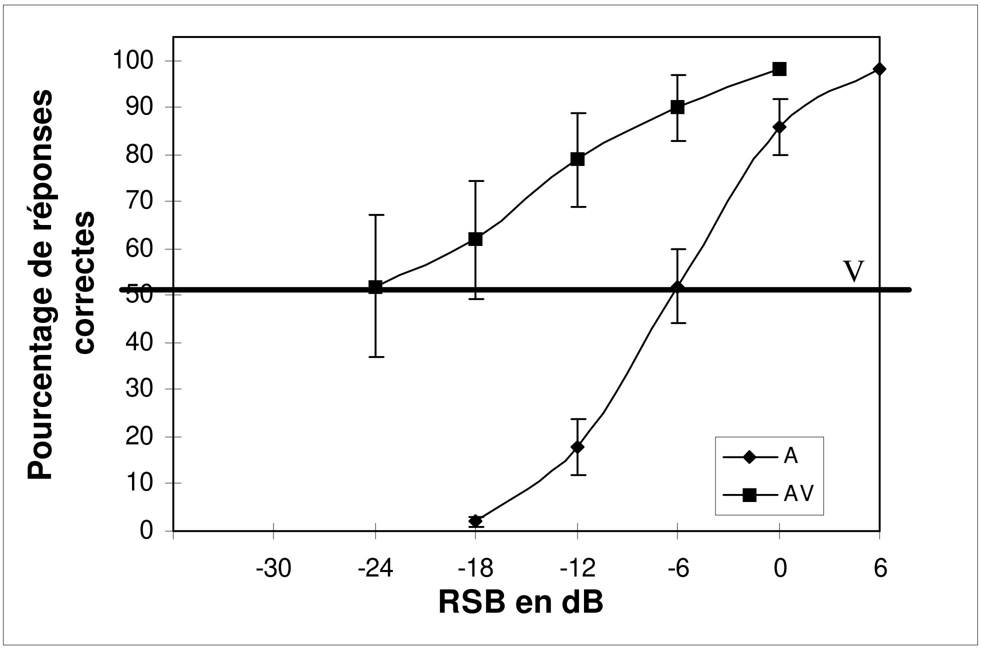


Figure 2.23 : les scores audio (A) et audiovisuels (AV) sur un vocabulaire de 250 mots (d'après Erber, 1969). La droite V indique le score visuel seul.

Cette expérience a été réalisée sur des sujets n'ayant pas de déficiences auditives en utilisant un vocabulaire de 250 mots en anglais. Les résultats prouvent encore une fois que la vision du locuteur peut nettement améliorer les scores de reconnaissance, surtout quand le milieu est fortement bruité.

Benoît et al. (1992) ont quantifié l'apport d'intelligibilité de la parole audiovisuelle en français. Les sujets sans déficience auditive et visuelle devaient reconnaître 18 mots sans signification. Comme dans les expériences citées précédemment, les sujets étaient soumis à deux conditions de test : audio seul et audio + image du locuteur. La Figure 2.24 illustre les résultats de cette expérience.

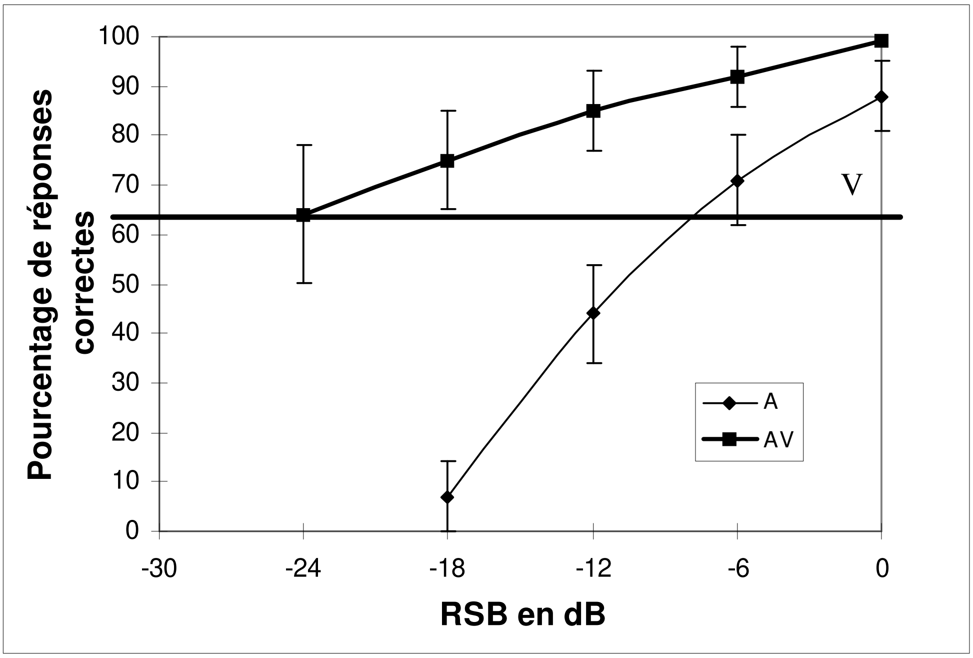


Figure 2.24 : pourcentage d'identifications correctes en fonction du niveau de bruit sur un vocabulaire de 18 mots (d'après Benoît et al., 1992).

Robert-Ribes (1995) (voir aussi Robert-Ribes et al., 1996) a élaboré une expérience de perception audiovisuelle des voyelles /i/, /e/, /y/, /O/, /u/, /o/, /a/ du français. La tâche de reconnaissance consistait à identifier ces voyelles en présence de six niveaux de bruit avec la vision de l'image du locuteur. La Figure 2.25 illustre les résultats de cette expérience.

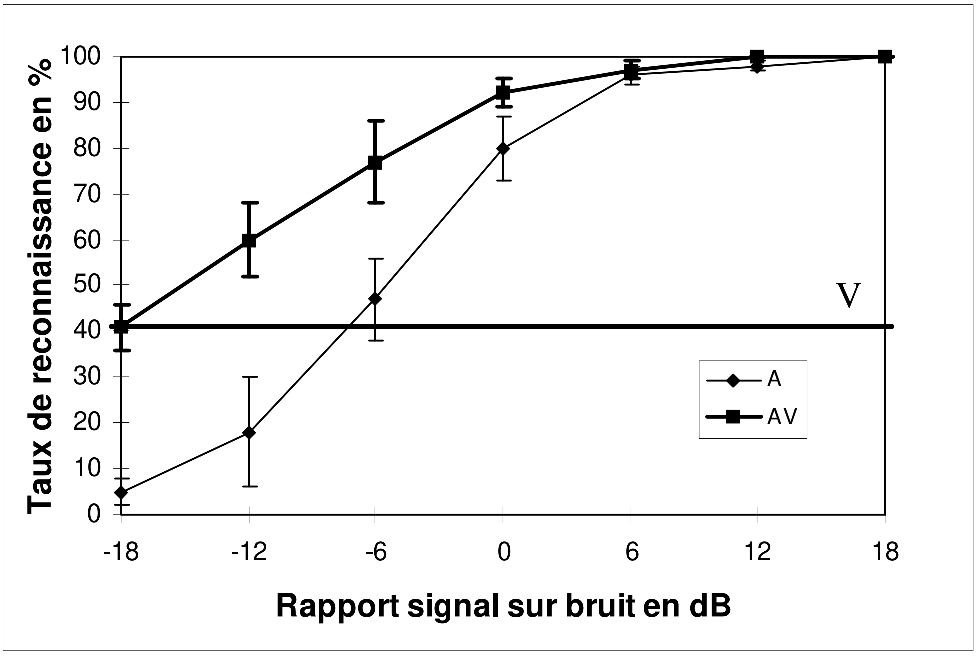


Figure 2.25 : résultats de tests de perception de Robert-Ribes (1995) sur un vocabulaire composé de sept voyelles orales.

En observant les résultats de ces expériences (parmi d'autres), on remarque une caractéristique importante de la perception audiovisuelle de la parole : pour tous les niveaux de dégradation du signal acoustique, la reconnaissance audiovisuelle (AV) est toujours supérieure à la reconnaissance auditive (A) et à la reconnaissance visuelle (V). Cette propriété constitue la condition principale qu'un modèle de fusion audiovisuelle doit satisfaire. Nous la considérons comme fondamentale dans la conception d'un système d'intégration audiovisuelle.

2.1.2 Pertinence des informations visuelles

Les résultats cités précédemment ont montré l'influence de la vision du visage du locuteur sur l'intelligibilité de la parole. Mais quel serait le gain d'intelligibilité si on remplaçait l'image du visage du locuteur par d'autres indices visuels ? Dans cette section, nous présentons les résultats de quelques études pour tenter de répondre à cette question.

Summerfield (1979) a étudié le gain d'intelligibilité de la parole avec différents types d'information visuelle. L'expérience consistait à calculer le pourcentage de mots correctement compris dans des phrases, en présence d'autres sources de parole, et cela sous cinq conditions de test :

1. signal acoustique seul,

2. signal acoustique + vision complète de la face,
3. signal acoustique + vision des lèvres seule,
4. signal acoustique + vision de quatre points lumineux placés sur les coins des lèvres et sur les intersections de l'axe de symétrie vertical avec la lèvre supérieure et inférieure,
5. signal acoustique + vision d'un cercle dont le diamètre représentait l'amplitude du signal audio non bruité.

Les résultats sont présentés sur la Figure 2.26.

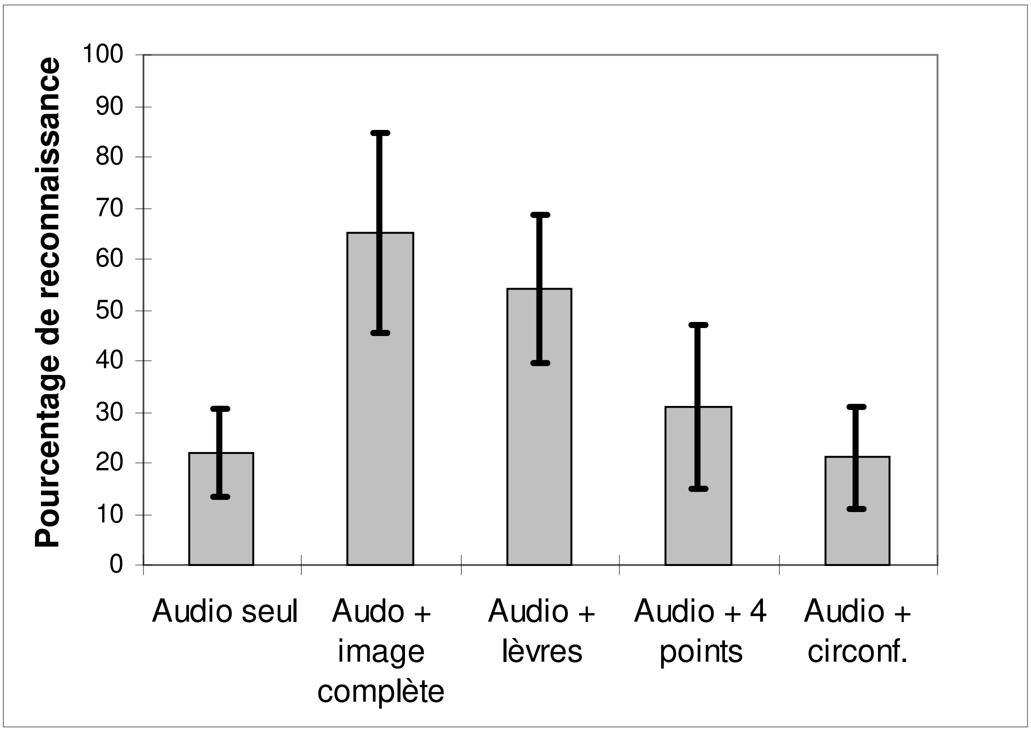


Figure 2.26 : résultats d'expérience de Summerfield (1979) sur 10 sujets. Les traits verticaux représentent les écart-types.

On peut noter que les conditions visuelles correspondant aux quatre points et au cercle à rayon modulé n'apportent pas d'intelligibilité de façon significative. En revanche, l'image complète du locuteur peut restituer une grande partie de l'intelligibilité manquante du signal acoustique. Cependant, Rosenblum et Saldaña (1995) ont montré qu'en augmentant le nombre de points lumineux sur le visage du locuteur, on peut observer une influence visuelle à condition d'avoir une pertinence articulatoire entre les points.

Le Goff et al. (1995, 1996) ont étudié l'apport d'intelligibilité dans différentes conditions de vision, et ceci en présence de différents niveaux de bruit additif. L'expérience consistait à reconnaître un mot sans signification parmi 18 sous différents niveaux de dégradation du signal acoustique et plusieurs conditions visuelles :

- audio seul,
- audio avec l'image des lèvres binarisées,
- audio avec l'image du visage complet.

Les résultats sont présentés sur la Figure 2.27.

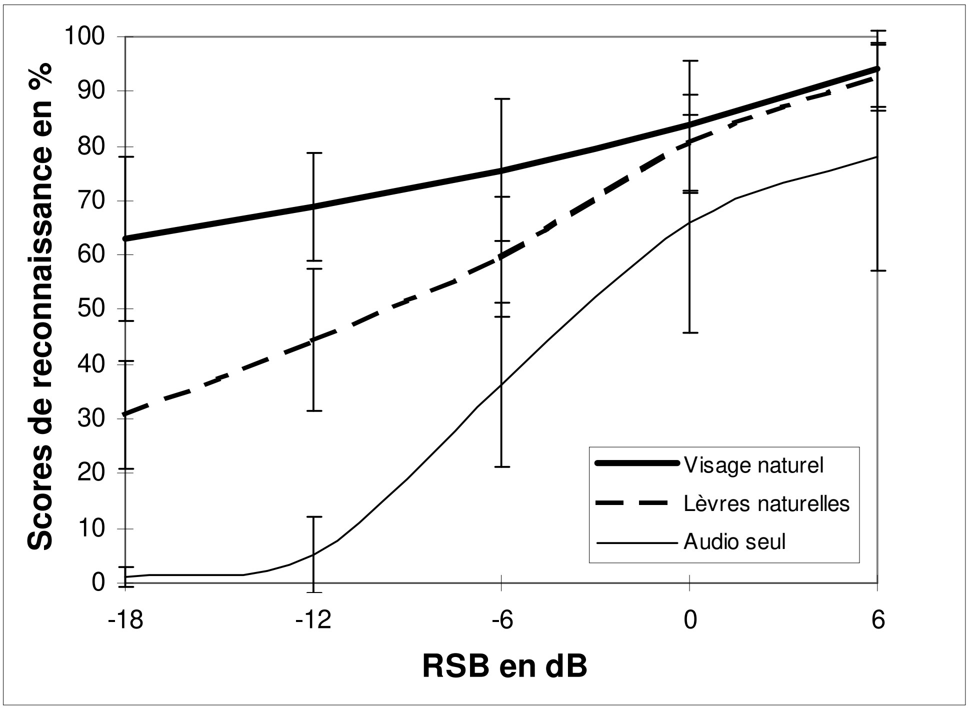


Figure 2.27 : résultats d'expérience d'intelligibilité en audio seul, audio avec lèvres naturelles et audio avec visage naturel (Le Goff et al., 1995). Les traits verticaux représentent les écart-types sur le nombre de sujets participant à l'expérience.

Ces résultats sont en accord avec ceux obtenus par Summerfield (1979). On peut remarquer que les lèvres seules peuvent transmettre une grande partie de l'information visuelle (à peu près deux tiers de l'intelligibilité manquante par rapport au visage complet). Le Goff et al. (1995) ont aussi évalué l'apport d'intelligibilité de modèles paramétriques commandés par un nombre réduit de mesures géométriques. Dans cette expérience, ils ont comparé le gain d'intelligibilité du modèle de lèvres, modèle de lèvres avec mâchoire et le modèle de visage synthétique. Le modèle de mâchoire utilisé était celui de Guiard-Marigny et Ostry (1995) et le modèle de visage, celui de Parke (1974), implémenté sur station de travail par Cohen et Massaro (1993, 1994) et modifié par Le Goff (1993) pour les commandes paramétriques et intégration du modèle de lèvres ICP (page 18). Les conditions de dégradation acoustique et le corpus de test étaient identiques à l'expérience précédente. Les résultats sont présentés sur la Figure 2.28.

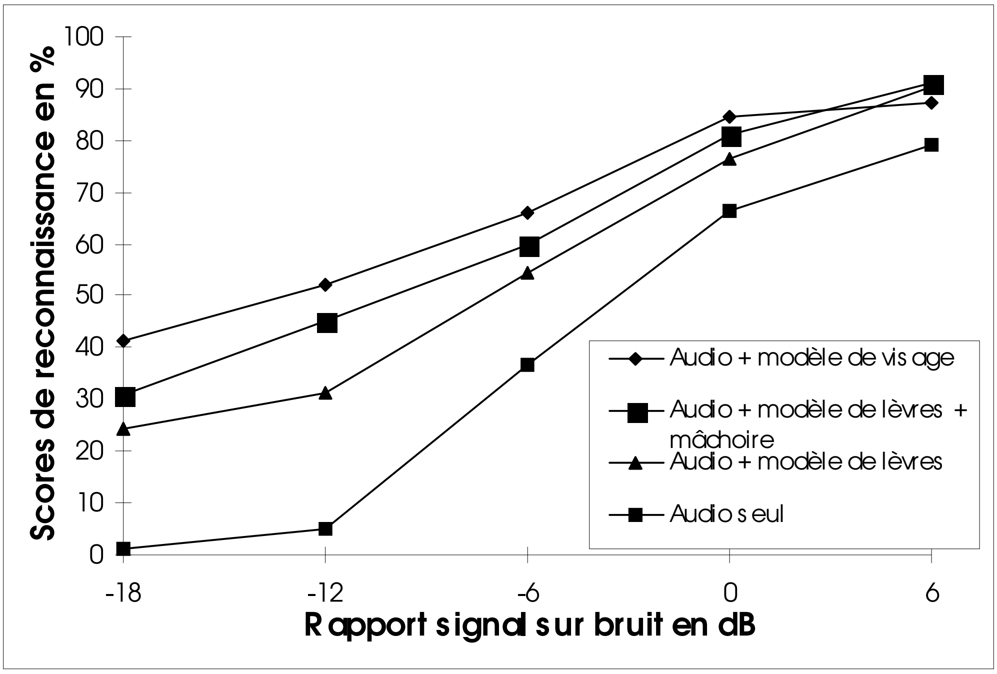


Figure 2.28 : apport d'intelligibilité du modèle de visage, modèle de mâchoire et lèvres, et modèle de lèvres (Le Goff et al., 1995).

Le Goff et al. (1996) ont quantifié le gain d'intelligibilité des différents types de présentations grâce à l'indice de Sumby et Pollack (1954). Cet indice représente la contribution à la restitution de l'information acoustique manquante. Il est obtenu par la relation suivante :

Equation 2.1 $$I_{AV} = \frac{S_{AV} - S_A}{1 - S_A}$$

dans laquelle S_{AV} et S_A sont les scores audiovisuels et auditifs respectivement.

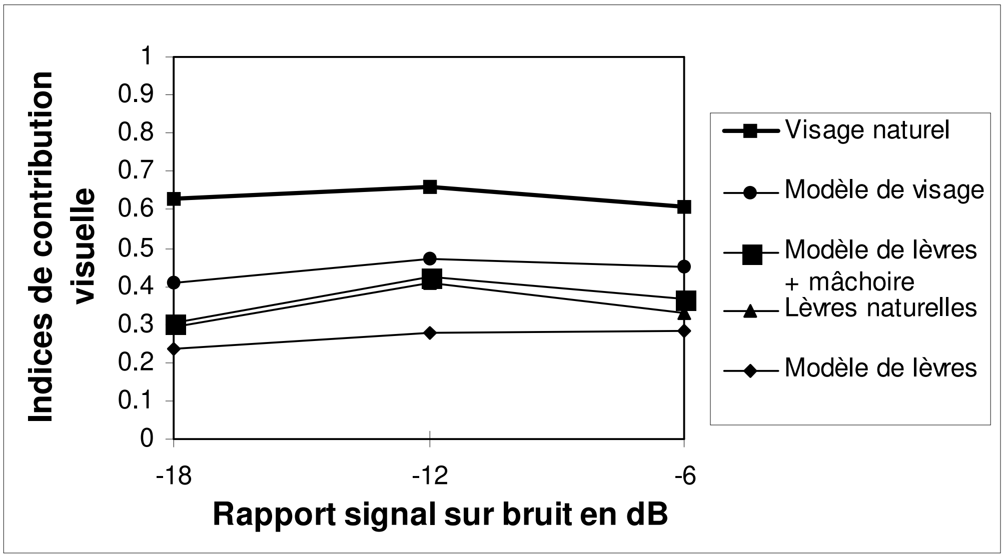


Figure 2.29 : indice de contribution visuelle pour les différentes conditions de vision (d'après Le Goff, et al., 1996).

Les résultats montrent que le modèle de visage commandé uniquement par six paramètres géométriques (A, B, C, P1, P2 et M) peut restituer environ la moitié de l'intelligibilité manquante du signal acoustique. Le modèle de lèvres commandé par cinq paramètres (les mêmes qui commandent le visage sans le paramètre du menton) apporte le quart de cette intelligibilité. Ceci prouve que ces paramètres sont porteurs d'une grande quantité d'information et peuvent être utilisés dans un système de reconnaissance automatique de la parole bimodale.

2.1.3 Présentation des indices visuels

En matière de présentation des informations visuelles, Summerfield (1983) a montré que la reconnaissance maximale est obtenue si les images respectent les conditions suivantes :

- elles doivent être présentées à une distance de 1.5 m avec une luminance suffisante,
- le locuteur ne doit pas avoir une barbe ou une moustache,
- le locuteur doit avoir des lèvres maquillées pour augmenter le contraste chromatique.

Bertelson et al. (1994) ont étudié l'influence de l'écologie des informations visuelles sur la perception de la parole. Ils ont montré qu'en inversant l'image du locuteur (la tête vers le bas), les informations visuelles étaient moins prises en compte par rapport à l'image droite. D'autres études (Reisberg et al., 1987; Green, 1987) confirment ces résultats et montrent aussi que l'écologie de la présentation des images joue un rôle très important dans la perception audiovisuelle de la parole. Les résultats de Le Goff et al. (1996) sur les scores de reconnaissance des modèles synthétiques sont en accord avec cette hypothèse.

2.1.4 Complémentarité audition/vision

Si la modalité visuelle permet d'améliorer l'intelligibilité de la parole, elle est aussi complémentaire à la modalité acoustique (Dodd, 1977 ; Binnie et al., 1974). Summerfield (1987) a étudié cette propriété de la parole audiovisuelle pour les consonnes en anglais. Il a montré que les consonnes facilement identifiables visuellement ont des caractéristiques spectrales plus sensibles au bruit et à la bande passante du signal. En revanche, les consonnes facilement reconnaissables dans la modalité acoustique sont plus difficilement identifiables dans la modalité visuelle. Par exemple les consonnes /th/ et /dh/ appartiennent à la même classe de visème, tandis que leurs caractéristiques spectrales dues au voisement sont très différentes.

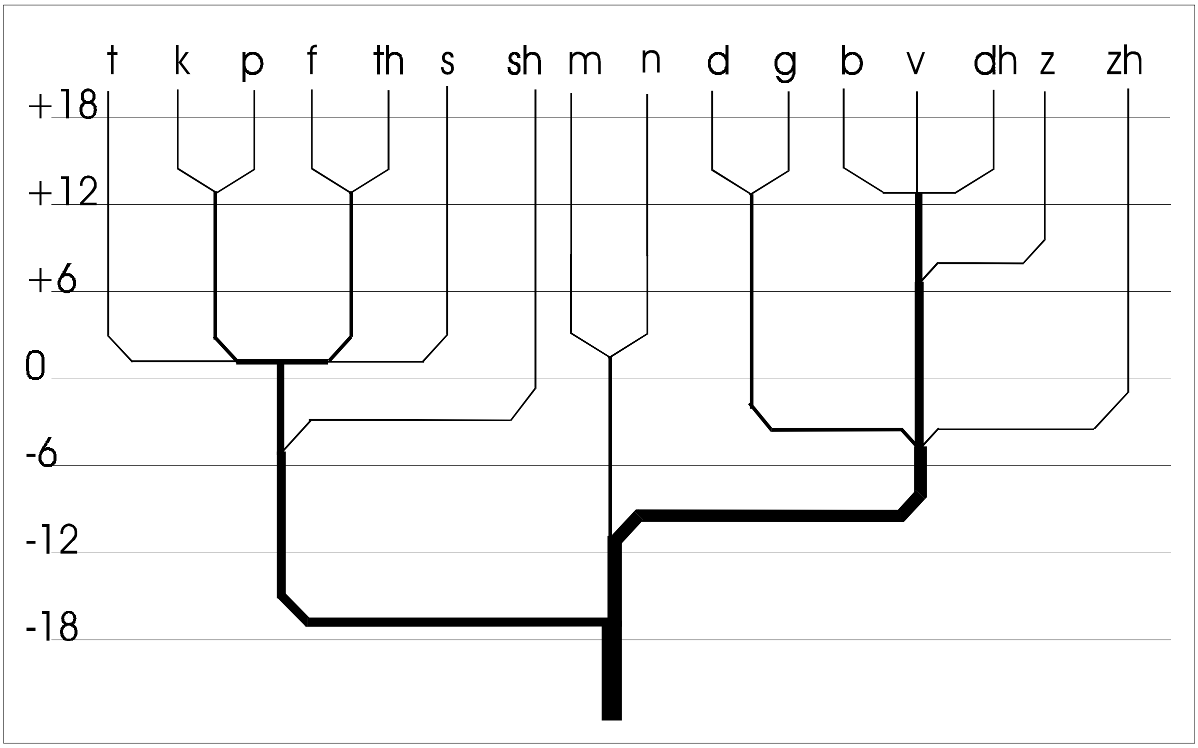


Figure 2.30 : arbre des confusions auditives des consonnes en anglais. L'échelle verticale représente les rapports signal sur bruit pour lesquels des groupes de consonnes ont été confondus (d'après Summerfield, 1987).

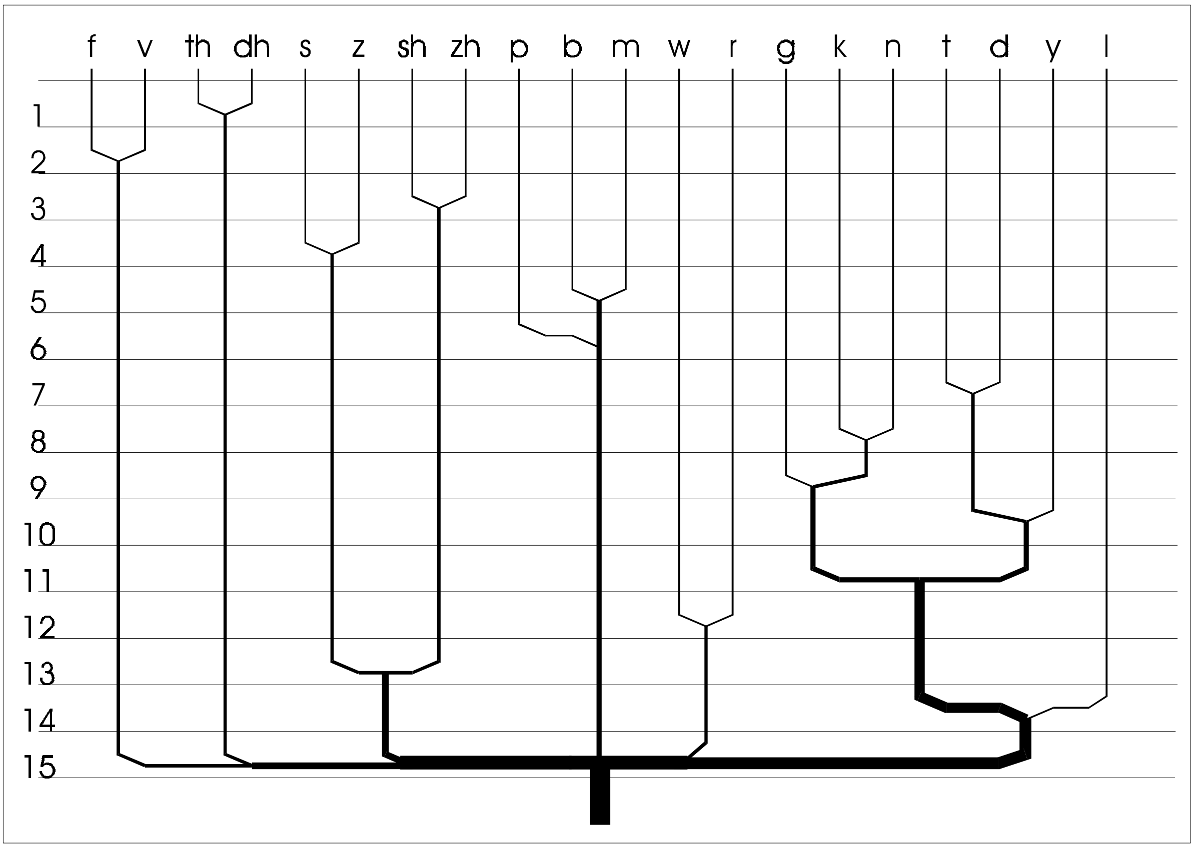


Figure 2.31 : arbre des confusions des consonnes en anglais, en condition de lecture labiale, obtenu par la méthode de classification hiérarchique (d'après Summerfield, 1987).

Les résultats de Summerfied (1987) sont présentés sur la Figure 2.30 et la Figure 2.31.

Benoît et al. (1994b) ont étudié les propriétés de la reconnaissance auditive et visuelle des voyelles /a/, /i/ et /y/ en français. Ils ont comparé l'intelligibilité de ces trois voyelles indépendamment du contexte consonantique pour les stimuli auditifs et visuels. Ils ont noté que dans la modalité auditive, /a/ était plus facilement reconnaissable que /i/ et /y/. Par contre, ils ont observé que dans la modalité visuelle seule, les voyelles les plus facilement identifiables étaient dans l'ordre /y/, /a/ et /i/. Les tests en audio seul ont été élaborés sur des stimuli dégradés (RSB = -12 dB). Les résultats de cette expérience sont présentés sur la Figure 2.32.

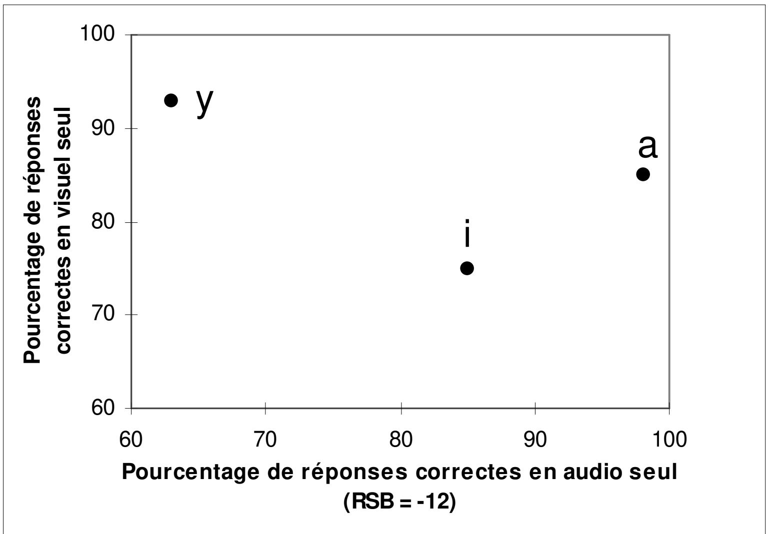


Figure 2.32 : pourcentage d'identifications correctes des trois voyelles /a/, /i/ et /u/ en français (d'après Benoît et al., 1994b).

On peut utiliser cette propriété importante de la parole audiovisuelle pour l'aide à la lecture labiale des malentendants. En fournissant des indices visuels sur les caractéristiques spectrales du signal de parole (voisé, non-voisé, nasal ...), on peut faciliter l'identification des sons qui appartiennent à la même classe de visème.

2.1.5 Incohérence entre les informations auditives et visuelles

L'expérience de McGurk et MacDonald (1976) est celle qui est la plus connue dans l'étude de la perception audiovisuelle de la parole, en présence de sources acoustiques et visuelles incohérentes. Les résultats de cette expérience sont assez surprenants : le mélange de /ga/

visuel et /ba/ auditif donne la perception de /da/ audiovisuel sur 98% des sujets (effet MacGurk). Ce résultat montre que la vision du locuteur peut influencer la modalité auditive, de telle sorte que la syllabe perçue comporte un phonème consonantique qui n'est commun à aucune des deux sources auditives et visuelles.

Le schéma de la Figure 2.33 montre le principe de cette expérience.

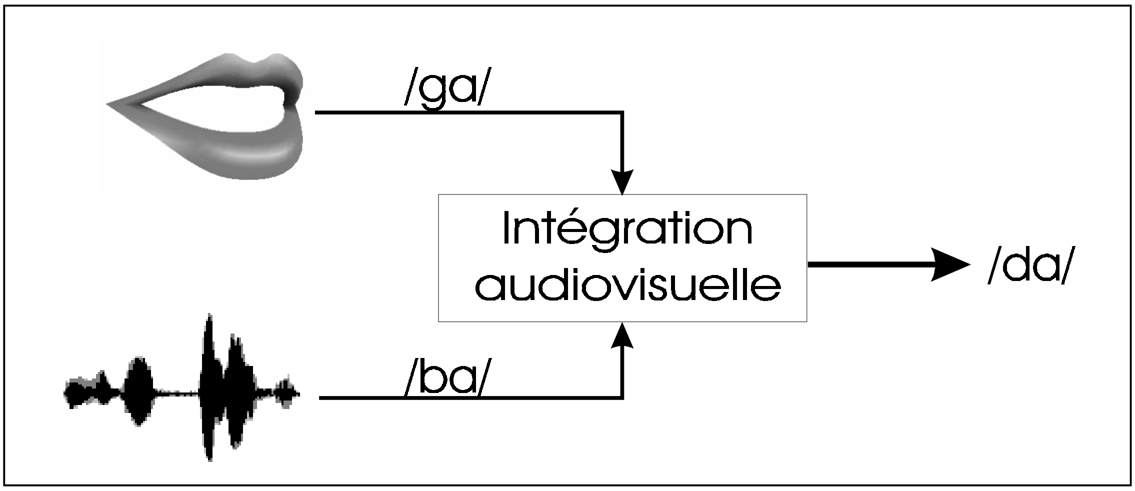


Figure 2.33 : expérience de McGurk et MacDonald (1976). Le /ga/ visuel + /ba/ auditif donne /da/ audiovisuel.

De nombreux chercheurs se sont attachés à mieux comprendre l'effet McGurk et le phénomène de fusion de données conflictuelles. Tillman et al. (1984) ont montré que la perception d'un /ba/ visuel et d'un /la/ auditif donne un /bla/ audiovisuel. Green et al. (1990) ont vérifié l'amplitude de l'effet McGurk selon que le visage et la voix présentés appartenaient ou non à la même personne. Ils n'ont observé aucune variation dans le nombre d'illusions McGurk et en ont conclu que le mécanisme d'intégration des informations auditives et visuelles est insensible à certaines incompatibilités. Ces résultats sont à interpréter avec précaution car Walker et al. (1995) ont montré que l'effet McGurk est moins marqué si l'image et la voix ne correspondent pas à une personne du même sexe, selon que les sujets sont familiers ou non avec le visage. Toujours dans l'étude de l'amplitude de l'effet McGurk, Green et al. (1991) ont montré que la voyelle qui accompagne la consonne dans l'effet McGurk joue un rôle important : l'illusion est plus fréquente si les syllabes visuelles et auditives contiennent la voyelle /i/, elle est quasiment inexistante avec la voyelle /u/.

Fuster-Duran (1996) a étudié l'influence de la langue étrangère dans l'effet McGurk. Dans son expérience, des sujets allemands et espagnols étaient soumis aux tests de perception avec des stimuli audiovisuels allemands et espagnols. Elle a noté une différence dans l'amplitude de l’illusion McGurk selon que le sujet était étranger ou non au langage du test et a conclu que les différences dans les systèmes phonétiques de ces deux langues étaient en partie responsables de cet effet. Elle a aussi montré que l’influence visuelle était plus importante dans la langue étrangère avec des mots réels. Dans le même contexte, Sekiyama et Tochura (1993) ont étudié l'effet McGurk sur des sujets japonais. Ils ont conclu que les sujets ne sont pas ou peu sensibles à cette illusion quand on leur présente des stimuli japonais. Par contre, ils ont renouvelé leurs expériences avec des stimuli anglais et ont noté l'existence de l'illusion McGurk chez les sujets. Dans une autre étude, Sekiyama et al. (1996) ont montré que la langue natale, la qualité acoustique, la maîtrise de la langue et les facteurs culturels sont des éléments qui peuvent modifier l’amplitude de l’effet McGurk.

D’autres chercheurs ont étudié l’effet McGurk avec des stimuli vocaliques. C’est le cas de Massaro et Cohen (1993) qui ont créé un continuum auditif variant de /i/ à /u/ et l’ont apparié avec l’image du locuteur qui prononce un /i/ ou un /u/. Ils n’ont pas observé l’effet équivalent à l’illusion McGurk dans leur expérience. Lisker et Rossi, (1992) ont étudié la perception de l’arrondissement pour les voyelles du français et ont présenté des stimuli auditifs, visuels et audiovisuels conflictuels à des experts phonéticiens. Les résultats montrent que l'illusion McGurk est inexistante avec des stimuli vocaliques, cependant la vision influence la détection auditive du trait d'arrondissement.

On peut finalement citer les travaux de recherche de Bernstein et Benoît (1996) qui présentent une vue d'ensemble sur les résultats d'expériences sur l'effet McGurk et la perception audiovisuelle de la parole en général.

2.1.6 Désynchronisation temporelle entre l’audition et la vision

En matière de désynchronisation audition/vision, on peut citer l’expérience de Smeele et al. (1992). Elles ont étudié l’influence de l’avance ou du retard du son par rapport à l’image sur la perception audiovisuelle de la parole. Les sujets devaient reconnaître des stimuli du type CVC avec un signal audio bruité dans différentes conditions de décalage son/image. Les résultats montrent que quand le son est en avance, l’intégration audiovisuelle de la parole est moins robuste que lorsque le son est en retard sur l’image avec le même décalage temporel.

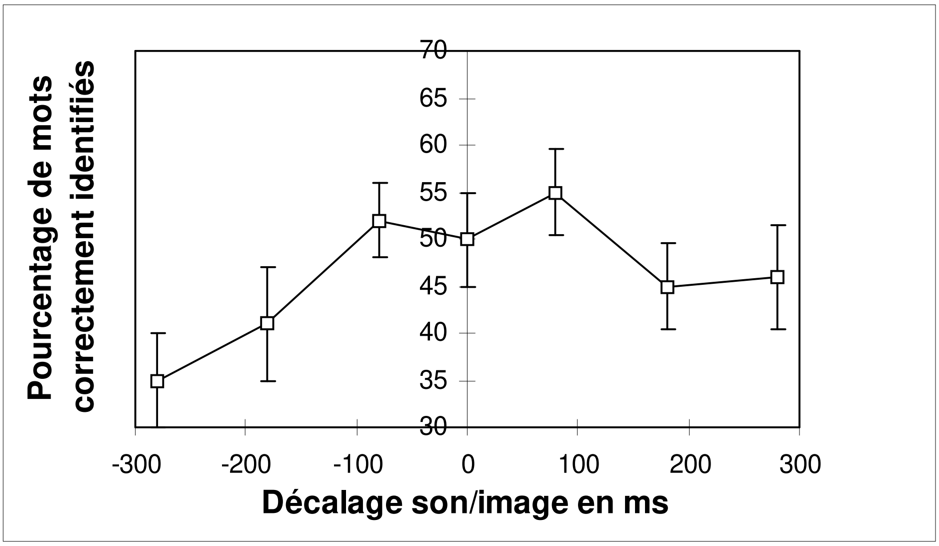


Figure 2.34 : intelligibilité audiovisuelle de la parole en fonction du décalage son/image. Les valeurs positives indiquent un retard du son par rapport à l'image (d'après Smeele et al., 1992)

Les résultats de cette étude montrent qu'au delà de 80 ms de décalage entre le son et l'image, l'intelligibilité de la parole décroît et l'auditeur est gêné par cette désynchronisation. Il est donc impératif de respecter cette limite dans des applications comme le visiophone, faute de quoi les informations visuelles ne sont pas exploitées d'une façon optimale.

Dans tous les cas, même avec un grand décalage temporel entre les deux sources (allant jusqu'à 1600 ms), l'intelligibilité audiovisuelle reste supérieure à l'intelligibilité auditive et visuelle seules (Campbell et Dodd, 1980 ; Reisberg et al., 1987).

2.2 Les modèles d'intégration audiovisuelle

Pour expliquer les phénomènes de fusion observés dans le processus d'intégration audiovisuelle de la parole, de multiples modèles ont été proposés (Campbell 1974 ; 1988, Robert-Ribes et al. 1991, Vroomen 1992, Summerfield 1987). On peut classer tous les modèles d'intégration audiovisuelle en quatre principales catégories (Summerfield, 1987 ; Robert-Ribes, 1995). Les sections qui suivent décrivent ces quatre architectures de fusion d'informations.

2.2.1 Modèle d'Identification Directe

Ce modèle est directement inspiré du modèle de Klatt (1979). Les deux sources d'information sont directement transmises vers un classifieur bimodal. Il n'y a donc aucun traitement avant la fusion des deux sources. La Figure 2.35 montre le principe de ce modèle.

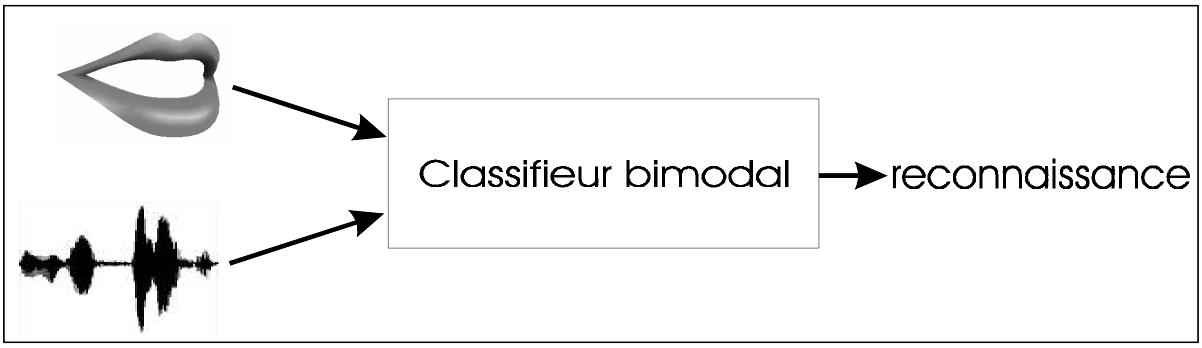


Figure 2.35 : schéma de principe du modèle d'identification directe

2.2.2 Modèle d'Identification Séparée

Suivant ce modèle, les informations visuelles et auditives sont traitées séparément et sont transmises chacune vers un classifieur. Les résultats en sortie de chacun de ces processus de reconnaissance sont fusionnés dans un module d'intégration qui donne le résultat final. On peut proposer plusieurs types d'architectures pour le module d'intégration. Il peut opérer sur des valeurs logiques où chaque modalité traite un groupe de traits phonétiques. Dans ce cas, on identifie le lieu d'articulation en fonction des paramètres spectraux de l'entrée acoustique (voisé, non voisé ...) et le résultat de classification du processus de reconnaissance visuelle (lèvres fermées, aire intéro-labial ...).

Ce module peut aussi traiter des données probabilistiques comme dans le cas du FLMP (page 46). Dans cette configuration, chacun des systèmes de reconnaissance auditive et visuelle fournit en sortie une probabilité de ressemblance à un candidat donné. Ces informations sont ensuite utilisées pour prendre la décision finale qui peut suivre une loi bayesienne. La Figure 2.36 montre l'architecture du modèle d'identification séparée.

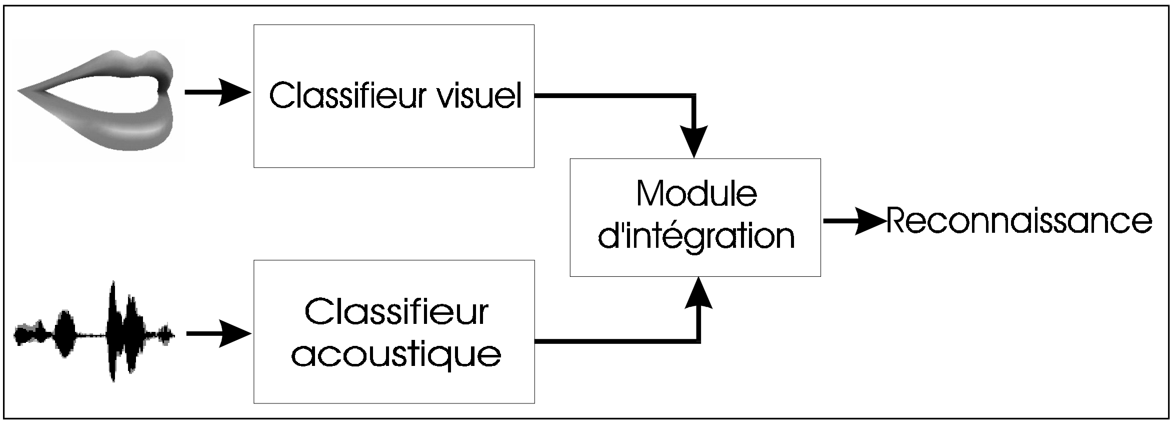


Figure 2.36 : architecture de la reconnaissance audiovisuelle la parole fonfée sur le modèle d'identification séparée.

On verra par la suite les résultats d'implémentation de ce modèle dans un système de reconnaissance automatique de la parole audiovisuelle.

2.2.3 Modèle de recodage dans la modalité dominante

Suivant ce modèle, on considère que la modalité auditive est dominante dans la perception audiovisuelle de la parole. En effet, comme les expériences mentionnées précédemment l'ont montré, quand le signal acoustique est clair, le locuteur peut identifier le message parlé sans l'aide des informations visuelles. Ce modèle propose donc de coder les informations visuelles dans une présentation de cette modalité dominante qui peut être la fonction de transfert du conduit vocal. Cette fonction de transfert est donc estimée séparément par un module de traitement du signal (coefficients cepstraux , LPC, ...) et par les indices visuels. On fusionne ensuite les deux estimations. Il faut noter que les caractéristiques de la source (voisée, non voisée ...) sont estimées uniquement à partir du traitement acoustique car le traitement visuel est incapable de les prédire. Le schéma de la Figure 2.37 montre le principe de cette architecture.

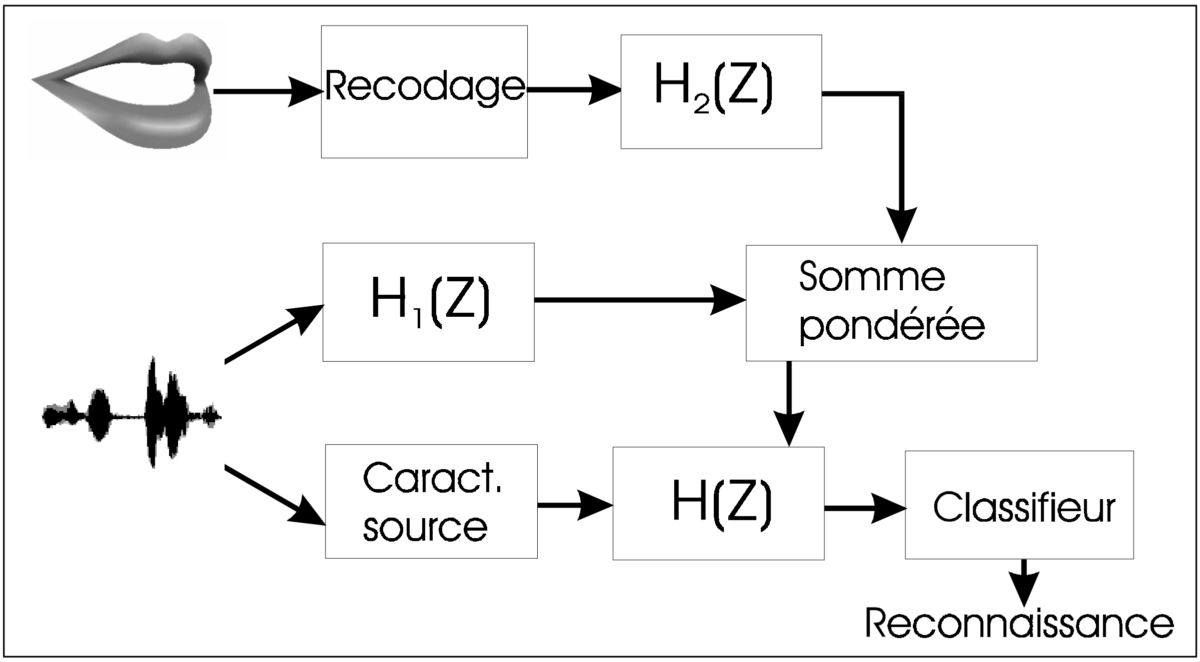


Figure 2.37 : architecture du modèle du recodage dans la modalité dominante. $H_1(z)$ est la fonction de transfert du conduit vocal prédite par les indices acoustiques et $H_2(z)$ est celle obtenue à partir des informations visuelles. H(z) est obtenue à partir des deux canaux auditif et visuel (d'après Robert-Ribes, 1995).

2.2.4 Modèle de recodage dans un espace moteur

Dans ce modèle, les deux entrées sont codées dans un espace commun qui n'est ni visuel ni acoustique. Cet espace est celui des paramètres du conduit vocal. Contrairement au modèle du recodage dans la modalité dominante dans lequel seules les informations visuelles sont codées dans l'espace des paramètres de la fonction de transfert, le modèle de recodage dans un espace moteur prévoit la projection des deux modalités dans une nouvelle représentation.

Ce modèle est l'application directe de la théorie motrice de Liberman et Mattingly (1985) qui propose que la perception de la parole est un processus qui utilise les représentations motrices. Pour percevoir la parole, il faut donc savoir la produire. Les deux sources d'information sont codées dans un autre espace pour produire les paramètres articulatoires.

Il faut noter que les informations visuelles sont insuffisantes à la définition de certains paramètres de configuration du conduit vocal, par exemple le mouvement du voile du palais qui détermine si le son est nasal ou non. On considère le protocole suivant pour la définition de l'ensemble des paramètres du conduit vocal à partir des informations auditives et visuelles : quand le canal visuel est incapable de fournir une information précise sur un paramètre, seule la source auditive est prise en compte pour obtenir cette information. Par contre, quand les deux sources sont susceptibles de fournir des données sur une des dimensions de l'espace moteur, on utilise la somme pondérée des deux projections. La figure suivante représente le principe de ce modèle.

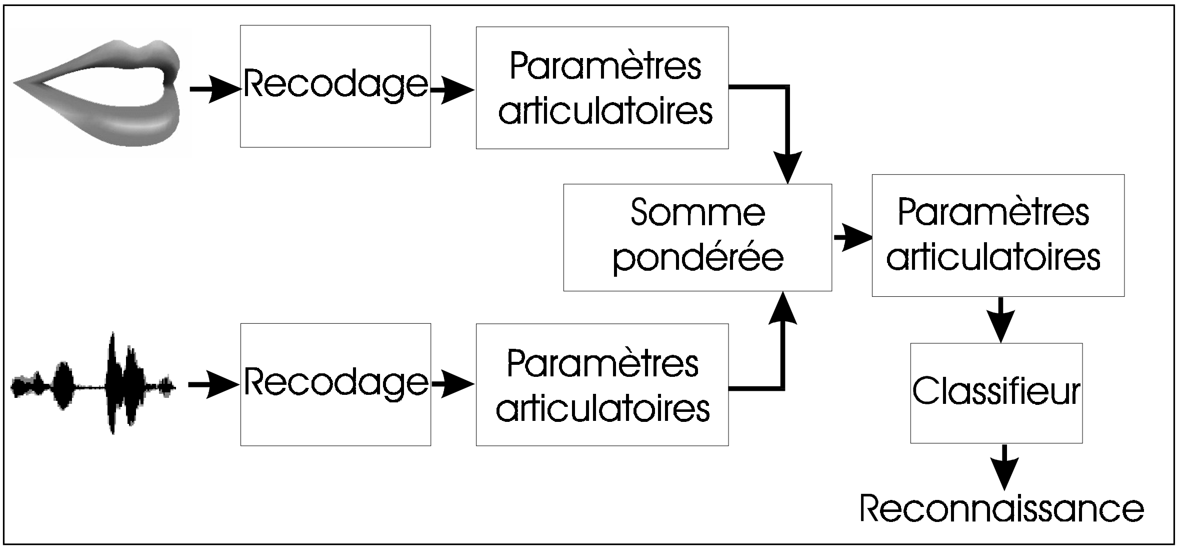


Figure 2.38 : modèle de recodage dans l'espace moteur. Chaque modalité permet de prédire les paramètres articulatoires. La décision finale est obtenue grâce à la somme pondérée de ces paramètres (Robert-Ribes, 1995).

Robert-Ribes (1995) a montré que c'est ce type de modèle qui permet d'expliquer les phénomènes observés en perception audiovisuelle de la parole par l'homme. Dans son étude, il a comparé les différents modèles d'intégration afin d'identifier celui qui permet d'expliquer les résultats expérimentaux de la perception audiovisuelle de la parole.

Il a rejeté le modèle d'identification directe car les expériences de Summerfiled et McGrath (1984) (parmi d'autres) montrent que les sujets sont capables de détecter l'incompatibilité audiovisuelle et d'estimer quantitativement une distance perceptive entre les stimuli auditifs et visuels. Il y a donc inévitablement un espace de représentation commun aux deux modalités, dans lequel les deux modalités sont comparées. Les résultats sur la fusion audiovisuelle et l'effet McGurk restent aussi inexplicables avec ce modèle.

Le modèle d'identification séparée est incompatible avec les résultats d'expériences portant sur le débit de la parole. En effet, Green et Kuhl (1989) ont montré que le débit perçu audiovisuellement est capable de modifier la frontière entre des sons voisés et des sons non-voisés. D'après Robert-Ribes (1995), "ceci est difficile à concilier avec un modèle à identification tardive (séparée), puisque la durée de la syllabe estimée visuellement est une donnée quantitative qui serait perdue dans la trace phonémique de l'entrée visuelle, et elle ne pourrait donc intervenir dans le processus de fusion tardive."

Suivant l'architecture du recodage dans la modalité dominante, la perception auditive n'est influencée par les indices visuels que si le signal acoustique est dégradé. Or, des résultats de perception audiovisuelle avec des signaux clairs montrent que la modalité visuelle est capable de modifier la perception auditive (Lisker et Rossi, 1992).

Finalement Robert-Ribes (1995) conclut que le modèle de recodage dans l'espace moteur est le seul compatible avec les résultats expérimentaux de perception de la parole bimodale et peut être considéré comme le plus plausible et le plus fonctionnel.

2.3 Quelques exemples

Nous évoquons ici quelques exemples de modèles d'intégration audiovisuelle proposés dans la littérature, et qui ont été testés sur l'homme. D'autres architectures d'intégration implémentées sur la machine seront présentées dans le chapitre 4.

2.3.1 Le modèle à identification séparée

Le modèle FLMP (Fuzzy Logical Model of Perception) développé par Massaro (1987a, 1987b, 1989) est un exemple de l'architecture à identification séparée. Fondé sur le traitement d'information provenant des deux modalités, il est constitué de trois blocs :

1. comparaison des entrées à des prototypes stockés en mémoire,
2. intégration des résultats,
3. décision finale.

La Figure 2.39 montre les étapes du modèle de perception audiovisuelle proposé par Massaro.

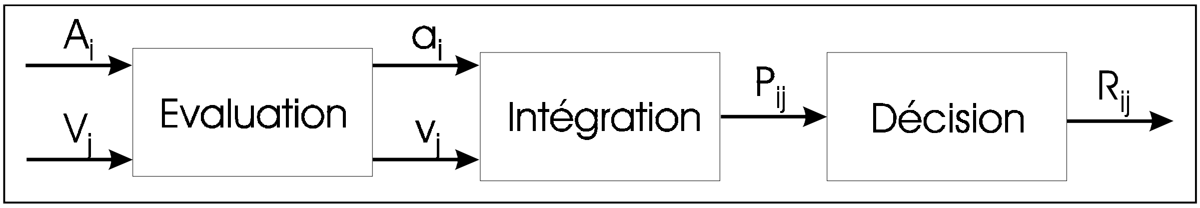


Figure 2.39 : le modèle de perception audiovisuelle proposé par Massaro (1987a, 1989).

Dans un premier temps, les informations auditives *(A_i)* et visuelles *(V_j)* sont évaluées chacune séparément en fonction de leur degré de support à un candidat R. Les indices de sortie de ce premier bloc permettent ensuite d'intégrer les deux informations et obtenir une probabilité audiovisuelle. Dans le bloc de décision, on calcule la probabilité d'avoir la classe *Rij* à la sortie selon la loi suivante :

Equation 2.2 : $$P(R_{ij} / A_i Vj) = \frac{P_{ij}}{\sum_{ij} P_{ij}} \qquad \text{avec } P_{ij} = a_i v_j$$

D'après Massaro et Friedman (1990), FLMP est équivalent à un modèle optimal à intégration bayésienne. Il est capable de prédire avec succès le résultat de la perception audiovisuelle en présence de sources conflictuelles.

Récemment, Massaro (1996) a légèrement modifié l'architecture de son modèle pour ajouter un bloc d'évaluation audiovisuelle, qui prend en compte des données mixtes provenant des deux sources. Le modèle de Massaro peut alors être considéré comme un modèle hybride selon la Figure 2.40 .

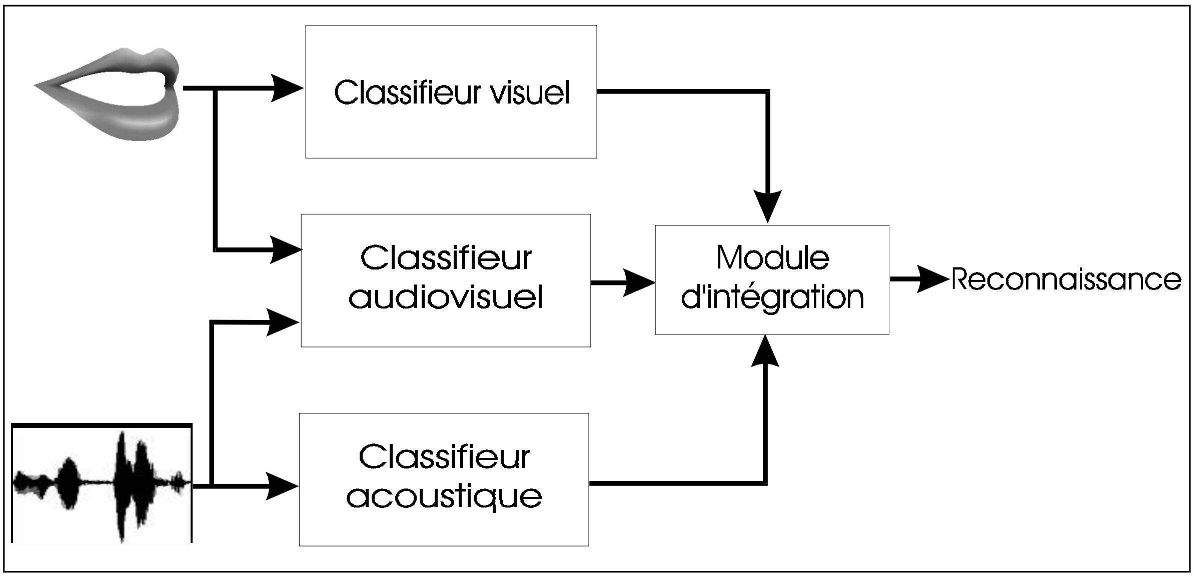


Figure 2.40 : le principe du modèle d'intégration hybride (d'après Robert-Ribes et al., 1996).

Le modèle hybride est une combinaison du modèle à identification directe et du modèle à identification séparée. L'intégration est faite après la classification dans chaque modalité (modèle à identification séparée) tout en prenant en compte la décision du bloc mixte dans lequel les données audiovisuelles sont traitées simultanément (modèle à identification directe).

Le modèle de Massaro se présente donc selon la Figure 2.41.

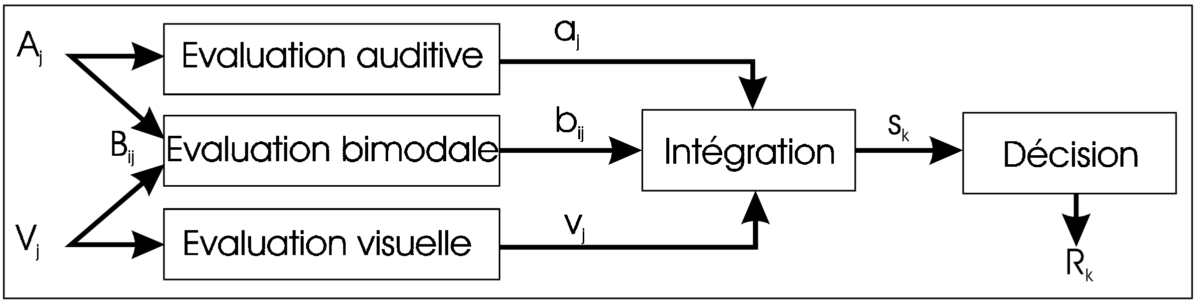


Figure 2.41 : nouvelle architecture de FLMP par Massaro (1996).

L'intégration s'effectue suivant une loi multiplicative sur les indices de sortie des blocs d'évaluation auditive, visuelle mais aussi audiovisuelle qui représentent le degré du support à un candidat. Comme dans la version antérieure de ce modèle, le bloc "décision" calcule la probabilité d'avoir la classe R_k en sortie.

2.3.2 Modèle de recodage dans l'espace acoustique

Girin et al. (1995) ont implémenté le modèle de recodage dans l'espace acoustique des données articulatoires afin d'améliorer le score d'identifications correctes des voyelles /a/, /e/, /i/, /O/, /y/, /o/ et /u/ en présence de bruit. Un filtre de Wiener ou LPC dont les coefficients sont estimés à partir des paramètres labiaux A (étirement interne), B (séparation interne) et S (aire intéro-labiale), est appliqué sur le signal auditif afin de rehausser les caractéristiques spectrales. La Figure 2.42 montre l'architecture du modèle élaboré par Girin et al. (1995).

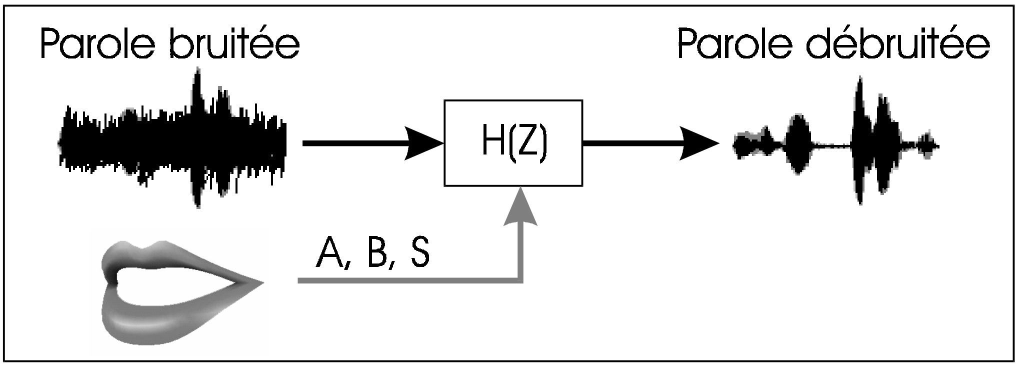


Figure 2.42 : débruitage d'un signal de parole à partir des paramètres labiaux (d'après Girin et al., 1995).

Girin et al. (1995) ont testé les performances de leur modèle sur 20 sujets sous différentes conditions de dégradation du signal acoustique. Les résultats sont présentés sur la Figure 2.43.

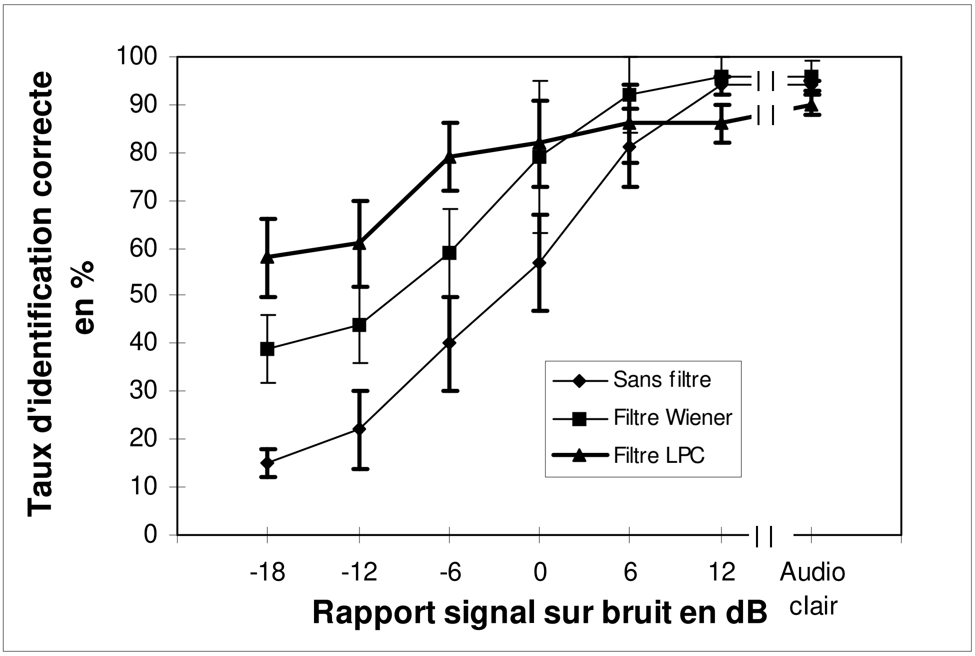


Figure 2.43 : résultats de Girin et al. (1995) sur l'identification de sept voyellesorales tenues en français.

Les paramètres A, B et S permettent de restituer une grande partie de l'intelligibilité manquante du signal acoustique quand celui-ci est fortement dégradé par le bruit additif : dans le cas d'un filtre LPC et à fort niveau de bruit, plus de la moitié de l'intelligibilité du signal acoustique est récupérée. Ces résultats prouvent encore une fois que les paramètres géométriques du contour interne des lèvres sont porteurs d'une grande quantité d'information sur l'intelligibilité.

2.4 *Récapitulatif*

Dans la section précédente, nous avons présenté les différentes architectures d'intégration. Comme ces modèles constituent une base très importante de notre travail, nous proposons par l'intermédiaire du schéma suivant, un résumé sur les caractéristiques de chaque modèle (Robert-Ribes, 1995).

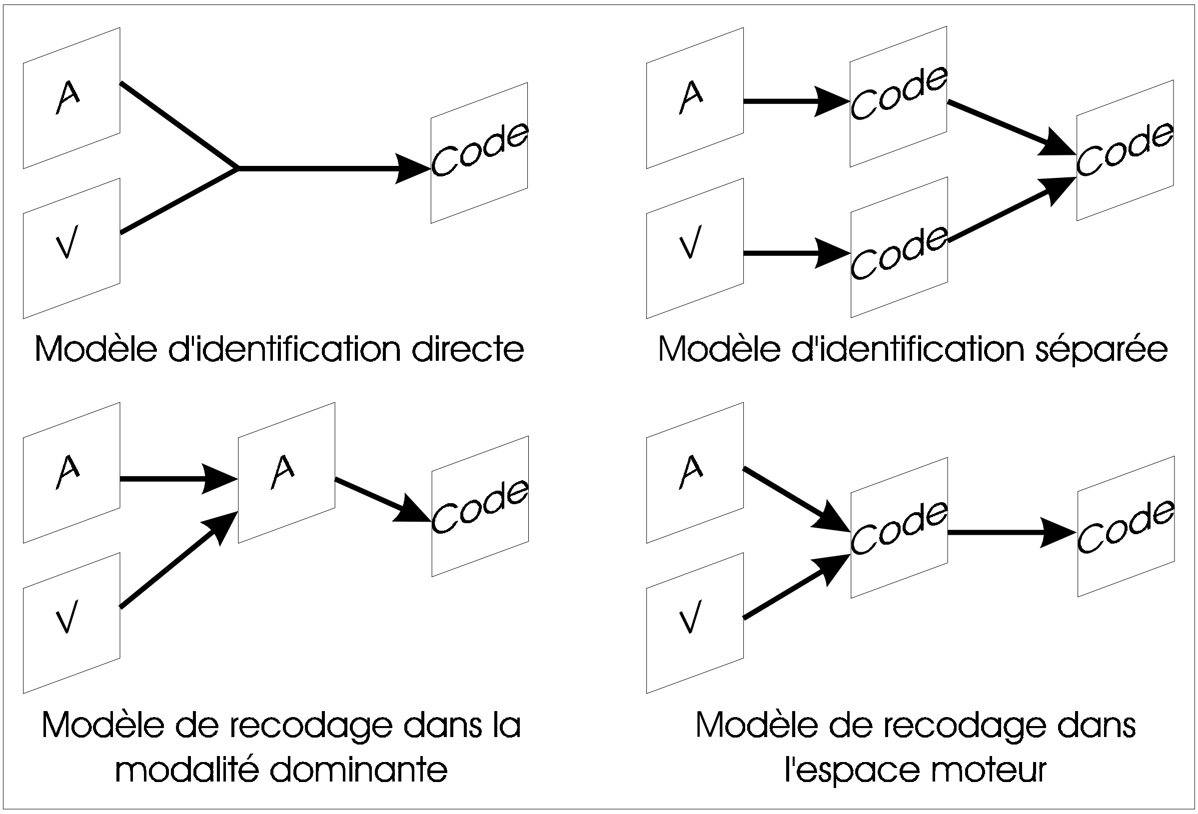


Figure 2.44 les principaux modèles d'intégration audiovisuelle (d'après Robert-Ribes, 1995).

3. Conclusion

Dans ce chapitre, nous avons souligné l'importance de la lecture labiale à la restitution de l'intelligibilité du signal acoustique, surtout quand le message parlé est noyé dans le bruit environnant. Les résultats des études de perception mettent l'accent sur le rôle fondamental des lèvres et des paramètres géométriques associés à la transmission de l'intelligibilité manquante de la parole. L'obtention de ces paramètres est donc une étape cruciale dans le processus de reconnaissance automatique de la parole audiovisuelle. Le chapitre suivant est consacré à quelques études centrées sur la définition des algorithmes de détection de la région des lèvres et le calcul des mesures géométriques.

Chapitre 3

Le traitement de la parole visuelle

Pour extraire les informations visuelles des mouvements des lèvres et de la mâchoire, il faut développer des algorithmes de traitement robustes assurant efficacité et précision sur les mesures. Ce chapitre présente les différentes techniques de traitement d'information utilisées pour l'extraction d'indices visuels. Notre démarche personnelle sera décrite dans la section 2.

1. Analyse du mouvement des lèvres

Nous avons vu que la vision des lèvres contribue en grande partie à la compréhension du message parlé surtout quand l'environnement est fortement bruité. De nombreux chercheurs se sont intéressés à l'analyse du visage du locuteur pour des applications de synthèse ou de reconnaissance. Dans le cadre de notre travail de recherche, nous nous attachons uniquement à l'analyse des organes visibles du visage intervenant dans le processus de la parole, c'est à dire les lèvres et la mâchoire.

1.1 Les systèmes mécaniques

Abbs et Gilbert (1973) ont conçu un système capable de convertir les mouvements des lèvres en variations du courant électrique..

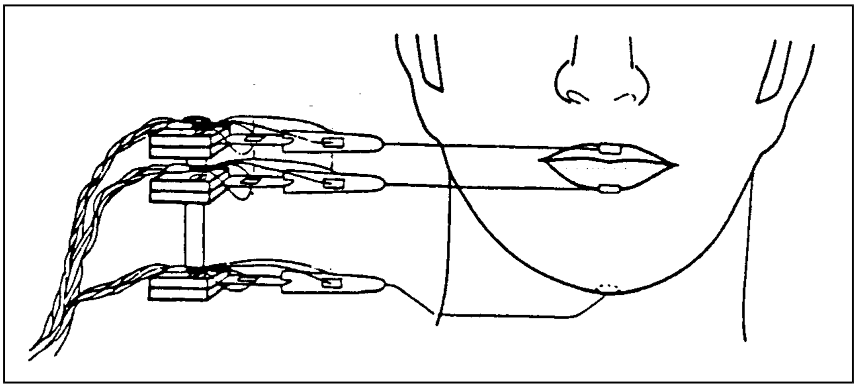


Figure 3.45 : l'analyseur mécanique d'Abbs et Gilbert (1973). Le mouvement des lèvres est transformé en variation du courant dans le circuit électrique.

En analysant ce courant, il était possible de mesurer les mouvements des lèvres et de la mâchoire.

1.2 Les systèmes magnétiques

Ces systèmes sont en général composés de plusieurs capteurs fixes et mobiles. Les capteurs fixes constituent des points de référence pour la poursuite et le repérage des capteurs mobiles qui sont placés sur les lèvres ou la mâchoire du locuteur.

Un exemple de ce type d'appareil est le kinésiographe mandibulaire conçu par Jankelson et al. (1975). Ce système est capable de suivre et d'enregistrer, dans l'espace 3D, la position de la mandibule par rapport au crâne. Il est composé de six bobines fixées sur un cadre et d'un aimant placé sur la mandibule ou sur une des lèvres. Parmi ces six capteurs, cinq sont sensibles aux variations du champ magnétique de l'aimant ; la dernière qui sert de calibrage est sensible au champ magnétique terrestre et aux champs parasites. En analysant le courant des bobines grâce à un dispositif électronique simple, on peut reconstruire en 3D l'espace du mouvement de l'aimant placé sur la mâchoire. En effet, le mouvement de cet aimant crée des variations plus ou moins importantes sur l'inductance de chaque bobine suivant sa position dans l'espace.

Un autre exemple de ce genre d'application est l'articulographe (Panagos et Strube, 1987). La particularité de ce dernier (par rapport au kinésiographe) est que dans ce système deux types de bobines sont utilisés : celles qui sont fixes (les émetteurs) et celles qui sont mobiles (les récepteurs). Les cinq bobines mobiles placées sur la langue (ou même sur les lèvres) oscillent chacune à une fréquence différente ce qui permet de les identifier dans l'espace. Les trois bobines fixes captent l'information de positionnement dans chacun des axes x, y et z. Grâce à un calculateur qui analyse en permanence les courants dans chaque spire, on reconstruit l'espace des capteurs et on peut ainsi suivre les mouvements des articulateurs en temps réel.

Ce type de système d'analyse donne de très bonnes précisions sur les mesures, mais l'utilisation des bobines ou des aimants collés sur les articulateurs peut représenter une contrainte pour le locuteur. En outre, en raison de la grande sensibilité au rayonnement électromagnétique de ces systèmes, les mesures peuvent être erronées en présence de champs électromagnétiques parasites (par exemple un moniteur vidéo).

1.3 Les systèmes optoélectroniques

Le principe de ces systèmes est fondé sur le suivi de marqueurs fixés sur les objets mobiles. En utilisant plusieurs caméras placées sous différents angles de vue et un calculateur, on peut reconstruire la trajectoire des marqueurs dans l'espace. Ces systèmes nécessitent des équipements électroniques complexes et leur mise en place est onéreuse. Cependant, ils sont largement utilisés dans l'analyse du mouvement et la reconstitution en images de synthèse. Dans le cadre de l'analyse des lèvres, cette technique permet de mesurer les distances entre plusieurs marqueurs et ainsi d'obtenir des informations pour la reconnaissance. On peut citer le système ELITE (Magno Caldogneto et al., 1989) qui utilise ce principe pour l'analyse des lèvres. Il a servi à Cosi et Magno Caldognetto (1996) dans leur système de reconnaissance audiovisuelle de la parole (Figure 3.46). Le mouvement de cinq marqueurs est suivi grâce à trois autres marqueurs qui servent de référence (les marqueurs collés sur les oreilles et sur le nez). Le système ELITE permet de calculer les paramètres A (étirement interne) et B (séparation interne) mais aussi les distances ainsi que les vitesses des marqueurs de mesure par rapport à ceux de référence.

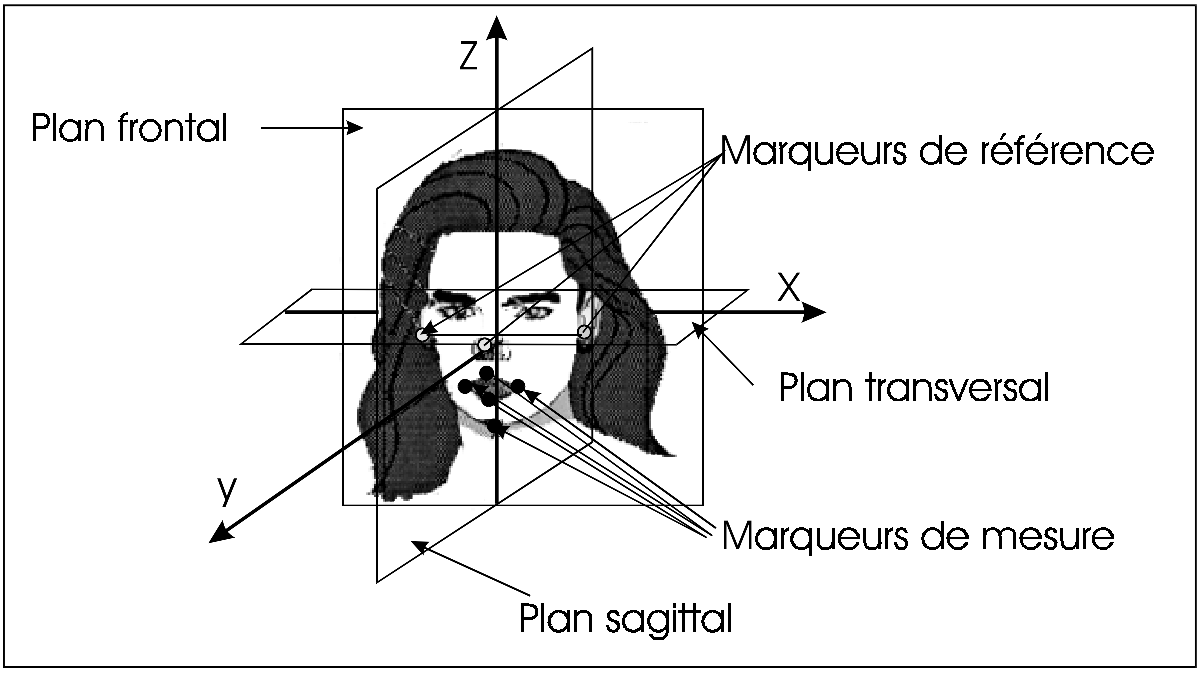


Figure 3.46 : le système ELITE de Magno-Caldognetto et al. (1989). Le suivi des marqueurs de mesure permet d'obtenir les paramètres labiaux A et B ainsi que le mouvement du menton (d'après Cosi et Magno-Caldognetto, 1996).

Cette méthode a aussi été utilisée par Brooke et Summerfield (1983) pour l'analyse des mouvements faciaux avec un nombre plus important de marqueurs. La Figure 3.47 montre cette expérience.

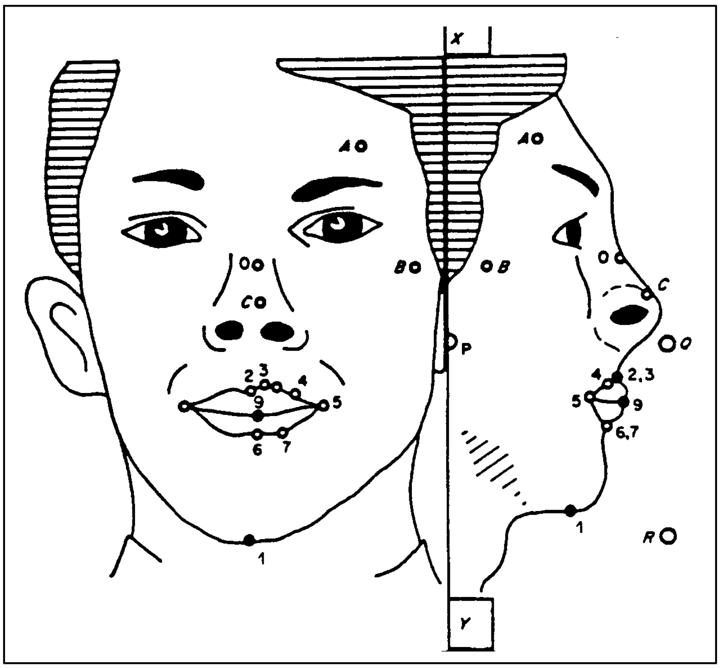


Figure 3.47 : expérience de Brooke et Summerfield pour l'analyse des mouvements faciaux (d'après Brooke et Summerfield, 1983).

Contrairement aux systèmes d'analyse magnétique, dans les systèmes optoélectroniques les capteurs doivent être absolument visibles. On ne peut donc pas analyser les mouvements des objets cachés (par exemple le mouvement de la langue dans le processus de la production de la parole).

On peut finalement citer les travaux de Finn (1986) et Finn et Montgomery (1988) dans lesuquels ils ont calculé 14 paramètres labiaux à l'aide de douze marqueurs hautement réflectifs. Certains de ces paramètres sont redondants et peuvent être facilement déduits à partir d’autres mesures. Ces distances sont présentées sur la Figure 3.48. On peut noter que certaines distances comme celles notées 10 et 11 sont pratiquement identiques (sur le schéma, ce sont les mêmes car la forme des lèvres est supposée symétrique). En effet, la première correspond à la moitié de la distance entre les deux coins des lèvres ; la deuxième représente la distance entre un coin des lèvres et l’intersection entre les segments de la séparation interne et la hauteur des lèvres.

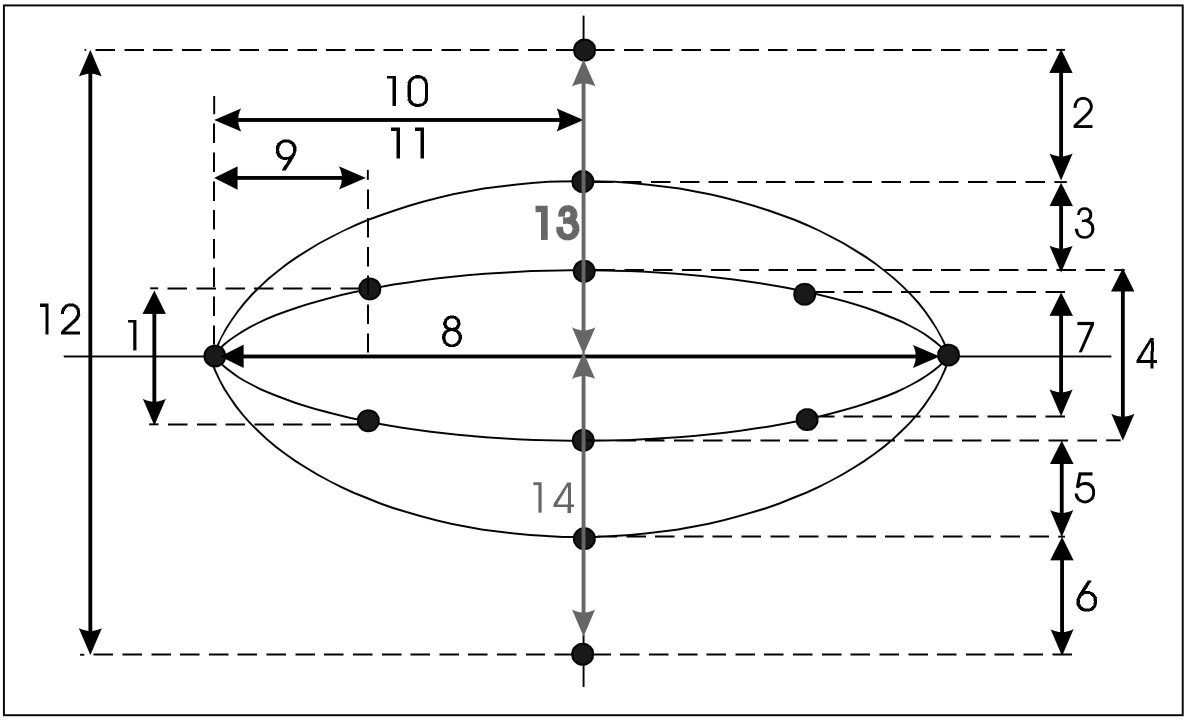


Figure 3.48 : les paramètres de mesure de Finn (1986). 12 pastilles collées sur les lèvres du locuteur permettent le suivi du mouvement des lèvres.

1.4 Analyse par traitement d'images

Cette technique consiste à numériser les séquences d'images et à extraire, grâce à des algorithmes de traitement plus ou moins sophistiqués, les informations visuelles pour la reconstitution du mouvement ou la reconnaissance. Grâce aux progrès techniques concernant la conception de processeurs hautement parallélisés, il est possible d'utiliser des algorithmes très complexes en traitement d'images pour extraire en temps réel les informations recherchées. De même, les mini-cameras couleurs (format PAL, NTSC ou Y/C) dont le poids ne dépasse pas quelques dizaines de grammes, permettent d'intégrer l'ensemble de l'appareil de prise de vue et d'éclairage sur un casque très léger. Ceci évite tout problème lié au cadrage ou à l'éclairage.

Nous allons décrire plus en détail les différents systèmes utilisant la technique du traitement d'image pour l'analyse des lèvres ou les autres organes visibles intervenant dans le processus de la parole. Tout d'abord, nous présentons quelques propriétés sur la vision humaine et les systèmes de représentation des couleurs.

1.4.1 La vision humaine

La vision humaine est possible grâce à la rétine, surface sensible à l'intérieur de l'œil. Elle est munie de deux types de capteurs photo-sensibles : les cônes et les bâtonnets. Les cônes sont individuellement reliés au cerveau et ne sont sensibles qu'à la couleur. Les bâtonnets sont reliés par grappes au nerf optique et ne sont sensibles qu'à l'intensité lumineuse. Ils sont sensibles à des énergies lumineuses très faibles pour lesquelles les cônes ne fournissent plus d'information, ce qui explique la perte de la perception des couleurs en vision crépusculaire. La densité des cônes est plus importante que celle des bâtonnets dans une zone nommée tache jaune dont le centre est entièrement démuni de bâtonnets. Les biologistes ont mis en évidence l'existence de trois types de cônes, chacun étant sensible à une couleur précise : le rouge, le vert et le bleu. C'est en combinant l'information provenant de ces trois types de capteurs que le cerveau reconstruit la couleur réelle des objets. Ce principe est utilisé pour la reconstitution des couleurs dans tous les standards de transmission de télévision couleur. En additionnant les couleurs primaires (rouge, vert et bleu) avec des coefficients multiplicatifs différents, on arrive à reconstruire toutes les couleurs de la nature.

1.4.2 La sensibilité de l'œil

La résolution chromatique de l'œil n'est pas une fonction linéaire. Les biologistes ont montré que la courbe de sensibilité visuelle atteint son maximum à la longueur d'onde de 555 nm, ce qui correspond à la couleur jaune. Comme la fréquence de la couleur verte est très proche de celle du jaune, (6×10^8 Mhz pour le vert et 5.40×10^8 Mhz pour le jaune), on peut déduire que l'œil est plus sensible à la couleur verte qu'à la couleur rouge ou bleue. Cette caractéristique de la vision humaine est très importante pour la reconstitution d'une image couleur. En outre, la résolution chromatique de l'œil est environ de deux à cinq fois plus faible que celle de la luminance. Ce qui veut dire que l'on arrive à mieux différencier deux points, l'un clair et l'autre sombre, placés à une minute d'angle de l'œil (résolution maximale en luminance) que deux points ayant des couleurs différentes placés au même angle. Dans une transmission télévisée, une bande de fréquence plus faible est allouée à la chrominance sans dégrader la qualité chromatique de l'image.

1.4.3 Quelques notions de colorimétrie

Les propriétés intrinsèques de l'œil ont permis d'établir différentes techniques de représentation des couleurs en vue de reproduire ces informations sans dégradation majeure.

Dans cette section, nous présentons ces espaces de représentation, chacun ayant des avantages et des inconvénients.

1.4.3.1 L'espace de représentation RVB

Cette représentation est la plus simple. Elle consiste à coder chaque couleur sur trois canaux différents, le rouge, le vert et le bleu. On peut donc considérer un espace cartésien dont les trois axes x, y et z représentent chacun l'intensité des trois couleurs principales. En normalisant l'amplitude de chaque composant, on obtient l'espace cubique de la Figure 3.49.

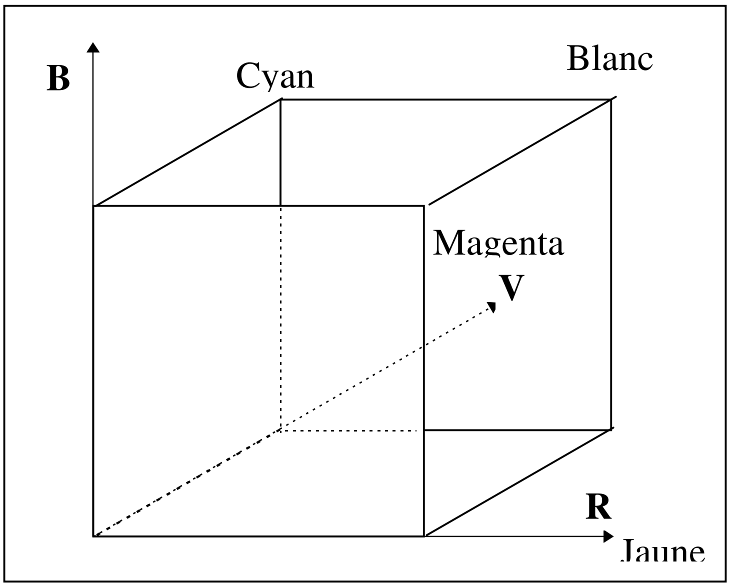


Figure 3.49 : représentation des couleurs dans l'espace RVB.

On note que dans cet espace, la droite satisfaisant la condition R=V=B représente l'ensemble des points ayant uniquement l'information de luminance, c'est pourquoi on l'appelle l'axe de luminance ou l'axe achromatique. Le noir correspond à R=V=B=0 et le blanc à R=V=B=1. Entre ces deux extrêmes se trouvent les dégradés du gris.

Pour représenter les couleurs dans cet espace, on préfère utiliser les trois coefficients normalisés r, v et b qui sont obtenus à partir des équations suivantes :

Equation 3.3 $$r = \frac{R}{R+V+B} \qquad b = \frac{B}{R+V+B} \qquad v = \frac{V}{R+V+B}$$

En ayant l'information de luminance, c'est à dire R+V+B, on peut calculer la composante couleur en connaissant deux des trois coefficients R, V et B.

On verra par la suite qu'en utilisant la représentation des couleurs dans l'espace RVB, il est très difficile d'effectuer un seuillage sur l'une des composantes seulement pour extraire l'information sur les contours des lèvres.

1.4.3.2 L'espace de représentation TLS

L'espace de représentation Teinte, Luminance, Saturation peut être utilisé pour exprimer des grandeurs colorimétriques relatives sans rattachement à des unités photométriques. La définition des trois composantes de cet espace est :

- la teinte : elle permet de distinguer les couleurs telles le rouge, le vert, le jaune, etc.
- la saturation : elle se rapporte à la notion de pureté, c'est à dire la quantité de blanc dans la couleur. En d'autres termes, elle détermine si une couleur est pastel ou soutenue.
- la luminance : c'est le terme déjà connu qui correspond à l'intensité.

Pour mieux comprendre le principe de cette transformation, on peut reprendre l'espace RVB. Le triangle grisé est le triangle de Maxwell dont les sommets représentent chacun une couleur pure.

Sur ce triangle, on voit que l'angle T représente la teinte. L'angle nul correspond au rouge, $\pi/3$ au vert et $2\pi/3$ au bleu. En reprenant les valeurs normalisées r, v et b, on obtient :

Equation 3.4 $$T = Arc\cos\left(\frac{(r-v)+(r-b)}{2\sqrt{(r-v)^2+(r-b)(v-b)}}\right) \quad \text{si b > v alors T = } 2\pi\text{-T}$$

La luminance est obtenue en faisant la somme des composantes rouge, vert et bleu. On a donc:

Equation 3.5 $$L = R + V + B$$

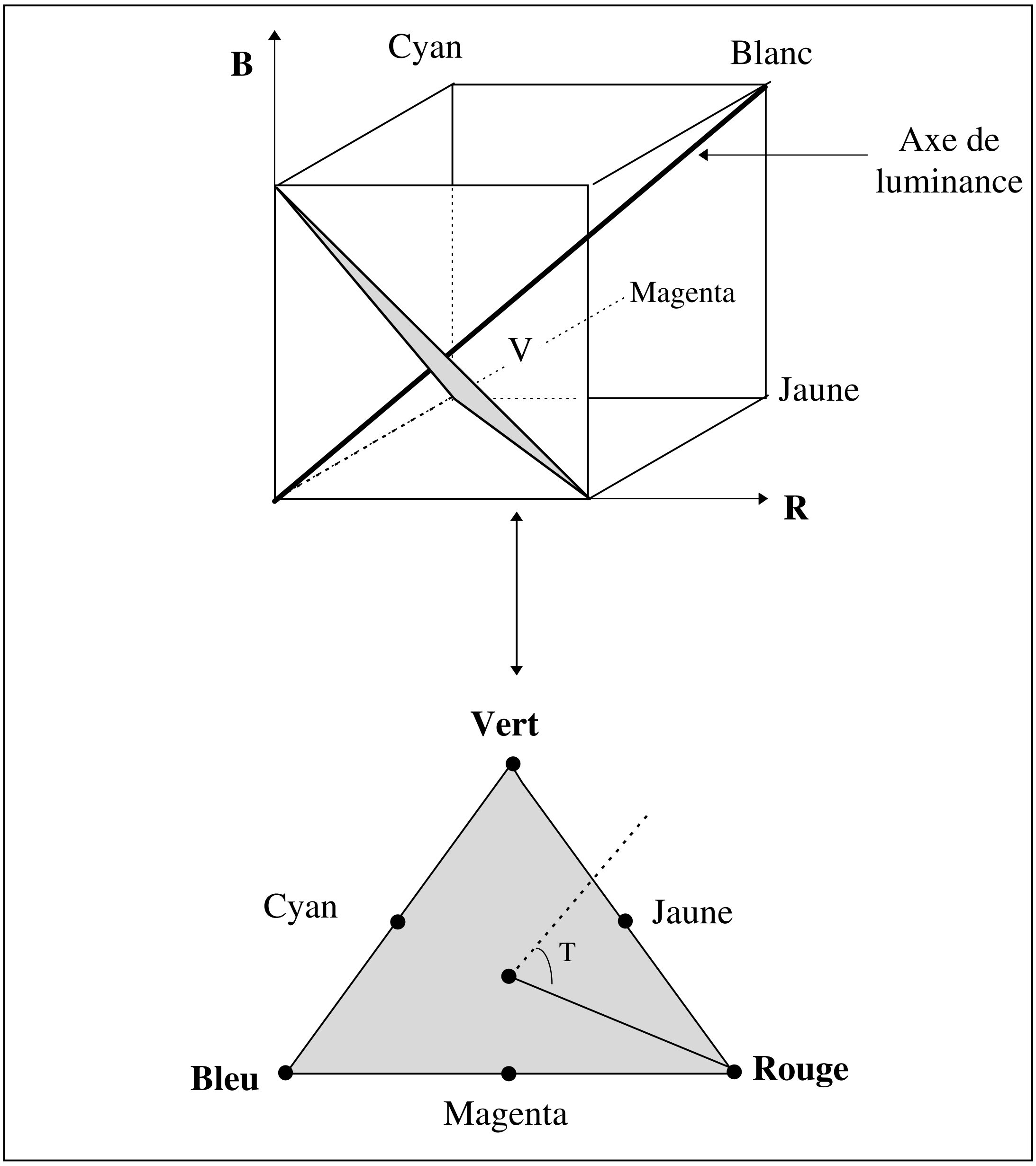


Figure 3.50 : le triangle chromatique de Maxwell.

Le troisième paramètre est celui qui concerne la pureté d'une couleur, c'est à dire la saturation. Il indique la quantité de "blanc" dans une couleur. Plus on augmente la saturation d'une couleur, plus on s'éloigne de l'axe achromatique et plus la couleur devient vive. La relation de la saturation est la suivante :

Equation 3.6 $S = 1 - \min(r,v,b)$

On peut finalement représenter l'espace de couleur TLS par la figure suivante :

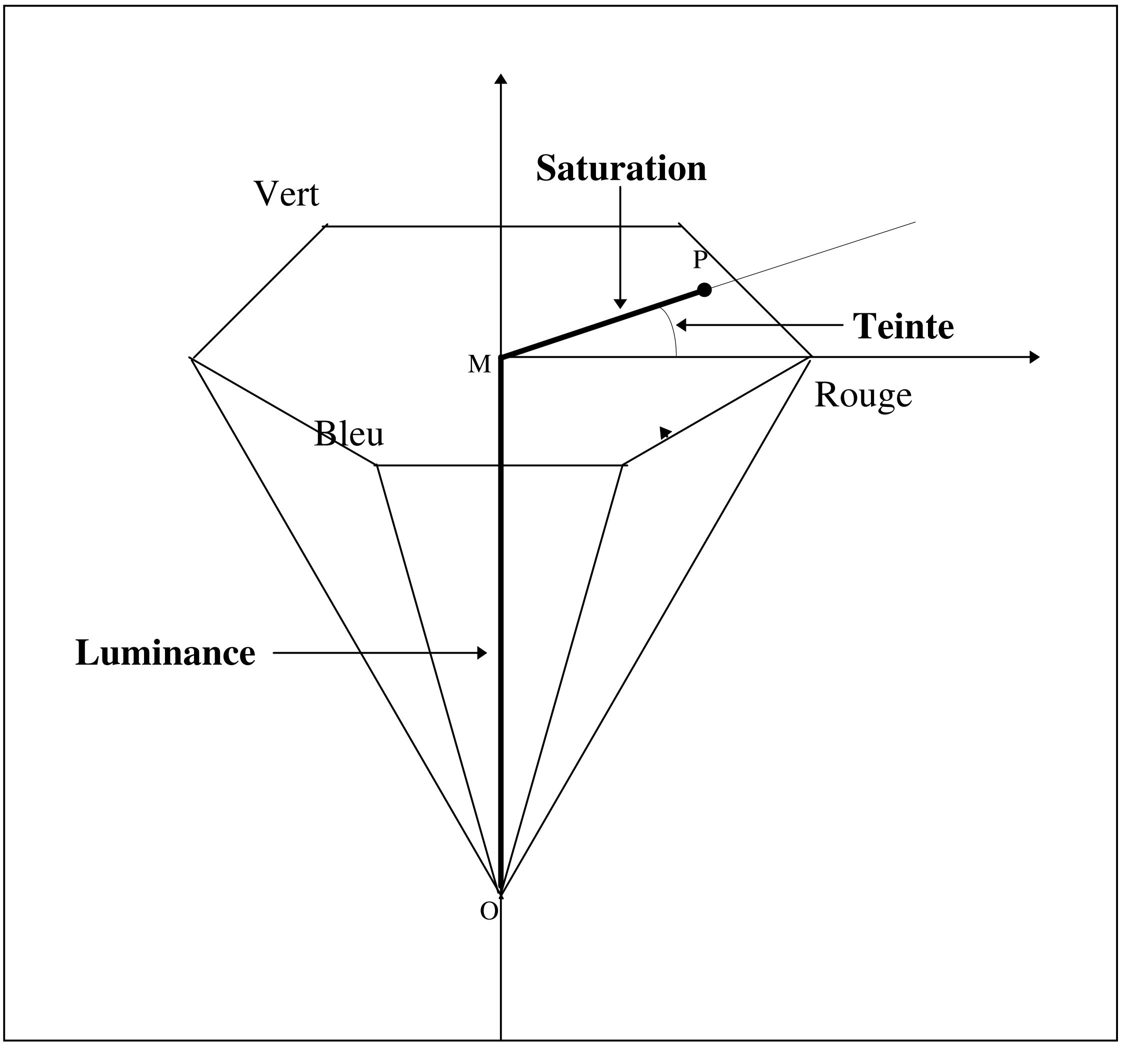


Figure 3.51 : espace de représentation TLS.

Pour une couleur quelconque (le point P), la distance OM représente la luminance, l'angle entre l'axe du rouge et MP est la teinte et finalement la distance du point de l'axe de luminance, c'est à dire MP correspond à la saturation.

1.4.4 Différentes techniques d'extraction de l'information visuelle

Les techniques en traitement d'images d'extraction des paramètres labiaux se divisent en deux grandes catégories (Revéret et Benoît, 1997) : les algorithmes orientés "image" et ceux orientés "modèle".

Dans les méthodes orientées image, les traitements s'effectuent directement sur l'ensemble des pixels de l'image. Ainsi, dans une fenêtre d'analyse qui contient les lèvres, on détecte les contours des lèvres, ce qui permet ensuite d'extraire les paramètres.

L'approche orientée modèle est fondée sur des connaissances *à priori* sur les lèvres et leur mouvement. Un modèle de lèvres qui prend en compte ces contraintes physiologiques est appliqué sur l'image. Grâce à une fonction d'optimisation, le modèle est ajusté au mieux sur l'image réelle des lèvres et les informations nécessaires sont extraites.

Cette partie est consacrée à l'étude de ces deux approches par l'intermédiaire de quelques exemples cités dans la littérature.

1.4.4.1 Approches orientées "image"

Dans une application d'interface homme/machine, il est souhaitable que les contraintes telles que le maquillage des lèvres ou l'immobilité soient supprimées. C'est le cahier des charges que Petajan et Graf (1996a) se sont fixé dans l'implémentation de leur dernier système d'analyse des lèvres. Dans d'autres recherches antérieures Petajan (1984a, 1984b, 1985) a conçu un système électronique capable d'extraire l'ombre de la cavité orale pour l'aide à la reconnaissance. Il a montré que les narines peuvent être considérées comme deux points de référence faciles à détecter sur le visage du locuteur. Ces deux points de référence ont été utilisés dans le système de Petajan et Graf (1996a). Aucune contrainte sur la mobilité ou le maquillage des lèvres n'est appliquée. Cependant, la caméra doit être placée légèrement au dessous du visage pour que les narines soient dans son champ de vision. Ce système, opérationnel dans une grande plage d'éclairage, détecte aussi la position des yeux qui permet de fournir l'angle d'orientation de la tête comme un facteur de commande. Un système de reconnaissance de visage permet de localiser les éléments du visage comme le nez, les yeux et la région des lèvres. Pour détecter les narines, un algorithme de seuillage chromatique est appliqué. Les seuils sont pré-stockés dans une LUT (*Look Up Table*). Une fois cette phase accomplie, une fenêtre d'analyse se forme autour de la région des lèvres à une distance fixe des narines. En appliquant des tests consécutifs sur les seuils du contour interne, une valeur finale pour ce seuil de détection est retenue. Chaque pixel qui est sous le seuil est considéré comme appartenant au contour interne. Il faut préciser que ce traitement ne fournit pas de paramètres discrets mais une série d'informations sur le contour interne et la langue. Pour la commande du modèle paramétrique de visage (Parke, 1974 ; Cohen et Massaro, 1993), Petajan et Graf (1996a, 1996b) ont utilisé le contour interne de ce même modèle et grâce à un algorithme de test récursif, ils obtiennent le contour sur le modèle qui est le plus proche de celui obtenu par le système d'analyse. La Figure 3.52 montre le processus d'analyse des lèvres.

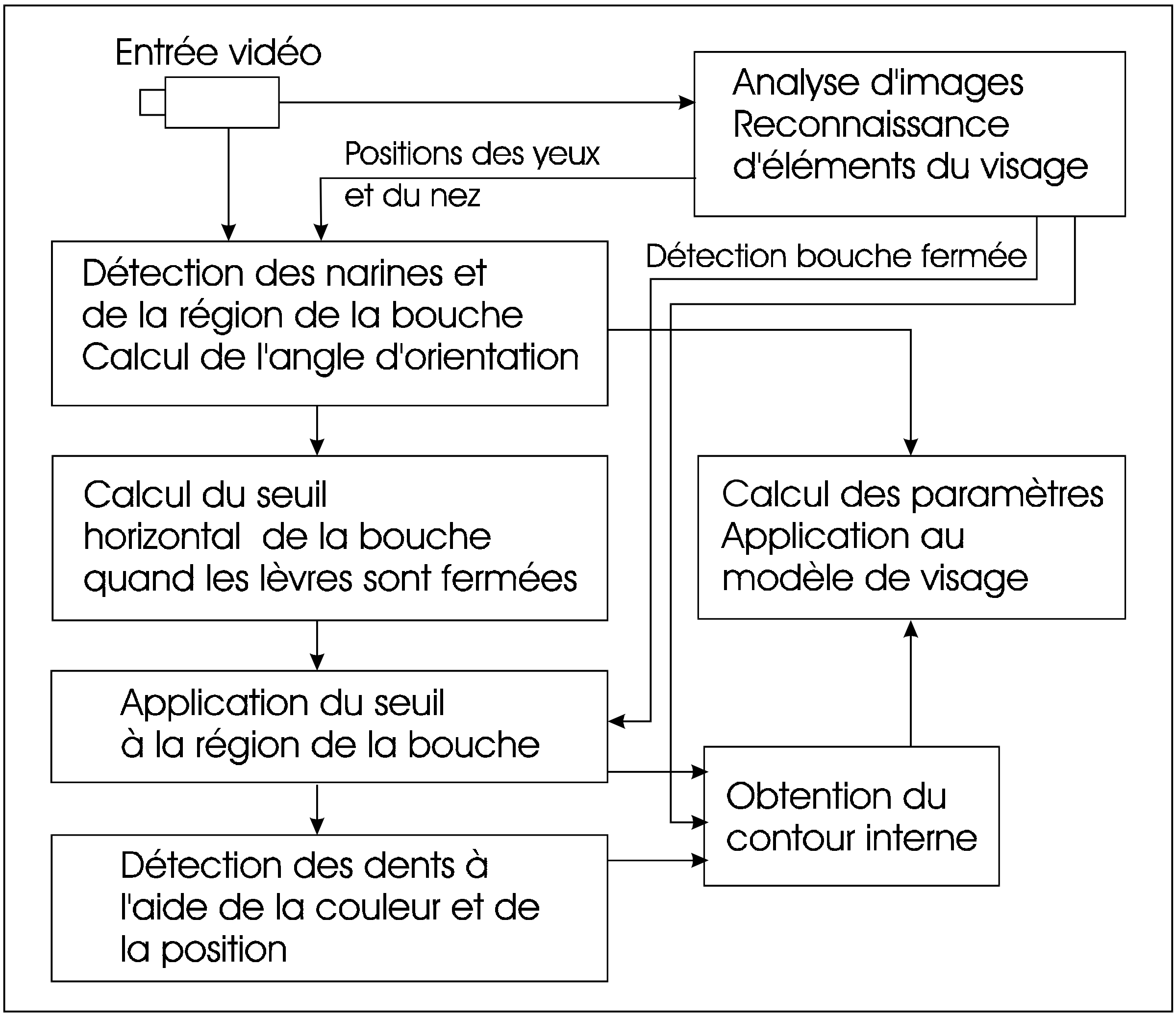


Figure 3.52 : le schéma synoptique du système d'analyse des lèvres d'après Petajan et Graf (1996a).

Figure 3.53 : résultat de la détection de la région de la bouche, des yeux et des narines (Petajan et Graf, 1996a).

Le principe d'extraction des paramètres labiaux implémenté par Bregler et Konig (1994) consiste à appliquer la méthode d'analyse en composantes principales sur un ensemble fini d'images clés. En créant un espace à N dimensions, chaque image peut être décomposée en N composantes dans cet espace. Les premières composantes représentent les basses fréquences de l'image et plus on augmente le nombre de vecteurs descripteurs, plus on s'approche de l'image de départ. En ayant une base d'apprentissage sur les paramètres des images clés de cet espace, il est possible de prédire pour chaque image, et cela en fonction du coefficient de chaque vecteur propre, les paramètres de l'image analysée. D'autres recherches dans la même direction sont en cours sur l'ensemble des 23 visèmes du français (Revéret et Benoît, 1997).

Mase (1990) ainsi que Mase et Pentland (1991) ont conçu un système d'analyse à base de flot optique. Pour chaque image, ils calculent le vecteur vitesse par rapport à l'image précédente et obtiennent ainsi une estimation sur le déplacement des points de contours des lèvres.

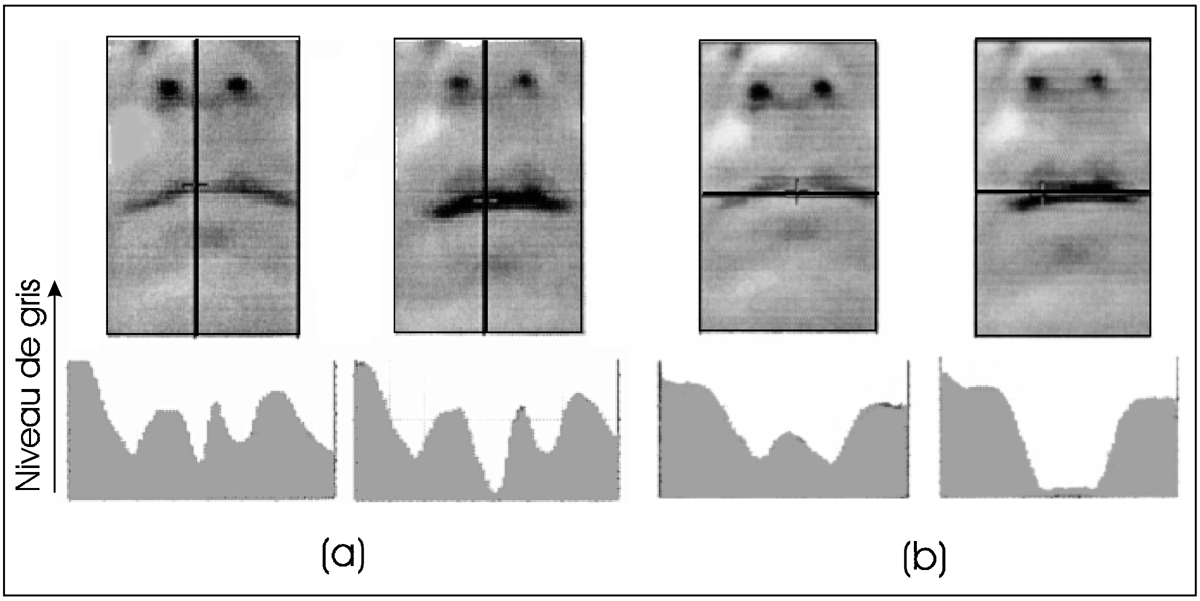


Figure 3.54 : calcul des paramètres des lèvres grâce à un seuillage de luminance sur la région des lèvres (d'après Prasad et al., 1993).

a - histogrammes de niveaux de gris sur la direction verticale de deux images correspondant à la prononciation de /ba/,

b - histogrammes de niveaux de gris sur la direction verticale pour les mêmes images.

L'extraction des paramètres labiaux peut être effectuée à partir des méthodes simplifiées de détection de contours (Stork et al., 1992 ; Petajan, 1984a), sans l'usage de l'information de chrominance. Prasad et al. (1993) ont implémenté un système de calcul de paramètres géométriques (Figure 3.54) grâce à la différence entre deux images successives filtrées passe-bas et seuillées. En localisant la région des lèvres, un simple seuillage sur les niveaux de gris de l'image permet d'obtenir les mesures requises.

1.4.4.2 Approches orientées "modèle"

Vogt (1996) a développé un système de détection de contours de lèvres fondé sur un modèle paramétrique. Sa méthode exclut toute utilisation de maquillage ou de fort éclairage. Le principe consiste à extraire quelques points du contour externe en appliquant des algorithmes de seuillage dans l'espace TLS. Grâce à un réseau de neurones, les informations sur le contour interne sont obtenues à partir de celles fournies en entrée du réseau, c'est à dire la position des dents et les coordonnées des 11 points du contour externe. Le schéma synoptique du système de traitement de Vogt est présenté sur la Figure 3.55.

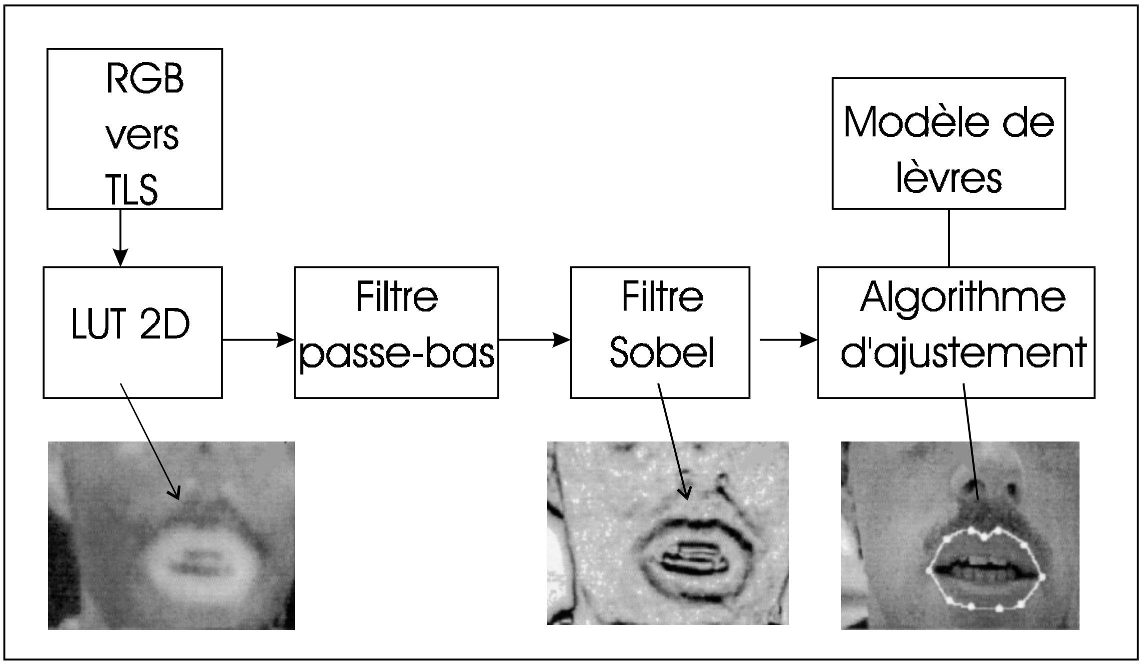


Figure 3.55 : chaînes de traitement du système de Vogt (1996).

Le réseau de neurones fournit cinq types de classifications en sortie : lèvres fermées, lèvres ouvertes, dents visibles avec la mâchoire fermée, dents visibles avec la mâchoire ouverte et finalement dents du haut visibles uniquement. Ces informations avec les points du contour externe peuvent être utilisées pour l'aide à la reconnaissance de la parole.

Coianiz et al. (1996) ont montré que l'espace de représentation chromatique TLS est bien adapté au traitement des contours des lèvres. Ils ont défini une fonction de pondération sur la valeur de la teinte de l'image, centrée autour de la couleur rouge, permettant de rehausser le vermillon. Un seuillage (la valeur du seuil est obtenue empiriquement) permet ensuite d'obtenir une image contenant juste les lèvres. Avec une analyse de concentration de pixels seuillés sur la direction verticale et horizontale de l'image, ils détectent le centre de la région des lèvres et positionnent la fenêtre d'analyse autour de cette région. Six points sur les contours des lèvres sont ensuite détectés : deux points sur le contour externe qui correspondent au point le plus haut de la lèvre supérieure et au point le plus bas de la lèvre inférieure, et quatre autres points qui correspondent aux coins des contours internes et externes des lèvres. Un calcul de gradient sur la luminance permet de détecter les deux points du haut et du bas des lèvres alors que, pour les autres, le calcul est basé sur l'information de la teinte. Ces points servent ensuite à ajuster un modèle de lèvres 2D (Henneke et al., 1994) sur l'image. La Figure 3.56 illustre deux exemples de cette analyse.

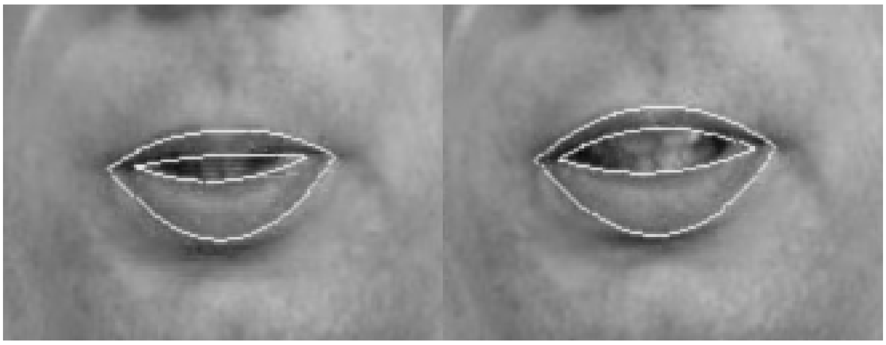


Figure 3.56 : résultats d'analyse de lèvres utilisant le modèle de lèvres 2D élaboré par Hennecke et al. (1994). (d'après Coiniaz et al., 1996).

Dalton et al.(1996) ont utilisé les contours dynamiques pour l'obtention des indices visuels. Les lèvres du locuteur sont maquillées afin de rehausser les contrastes. Les contours ainsi obtenus sont modélisés par des fonctions *spline* contrôlées par plusieurs points. Les mouvements des lèvres sont décrits par des déformations d'une forme moyenne de lèvres (bouche fermée) et sont commandés par les points de contrôle. Une analyse en composantes principales permet d’obtenir le vecteur de déplacement de chaque point. La Figure 3.57 montre la direction et l’amplitude de la première composante de cette analyse.

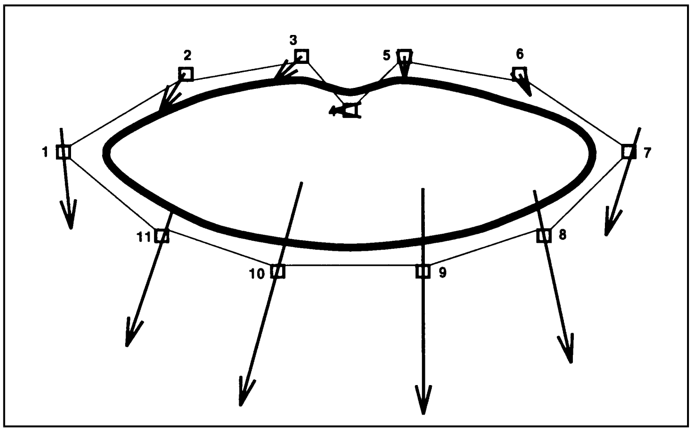


Figure 3.57 : la direction et l’amplitude de la première composante principale (d'après Dalton et al. 1996).

Figure 3.58 : une séquence d’analyse de lèvres (d’après Dalton et al., 1996).

Dans une approche semblable, Luettin et al. (1996a) ont implémenté un système d'extraction de paramètres fondé sur des modèles "actifs" de lèvres. Selon cette méthode, les frontières des lèvres sont représentées par un ensemble de points étiquetés. Pour limiter les degrés de liberté de ces points dus aux contraintes physiques des mouvements des lèvres, des connaissances *a priori* sur ces mouvements sont nécessaires. Elles sont obtenues à partir d'une analyse en composantes principales sur la base de données d'apprentissage. Pour chaque image, les contours sont réestimés en minimisant une fonction de coût qui mesure la distance de niveau de gris entre le modèle et l'image courante. La Figure 3.59 montre le résultat d'analyse de Luettin et al. (1996a) utilisant cette technique.

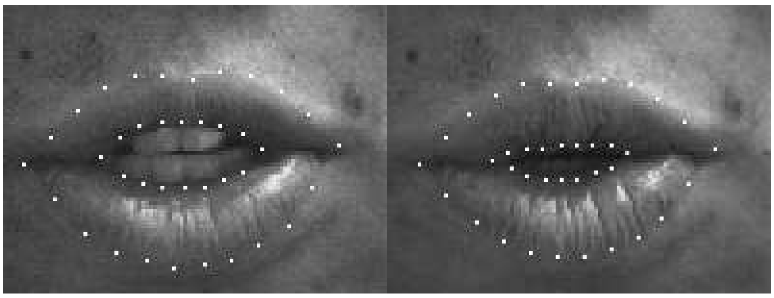


Figure 3.59 : résultats de localisation des contours des lèvres (d'après Luettin et al., 1996a).

Dans de plus récents travaux, Luettin et al. (1996b) ont amélioré les performances de ce système de détection en ajoutant une analyse en composantes principales de l'image des lèvres. Cette analyse permet d'extraire des informations supplémentaires sur le mouvement des lèvres. Ils ont montré qu'en combinant les deux approches "modèle" et "image", le taux d'identifications correctes d'un système de reconnaissance visuelle pouvait être amélioré (90.6% avec le système mixte contre 81.2% avec les contours actifs dans une reconnaissance de quatre chiffres prononcés par 12 locuteurs).

Horbelt et Dugelay (1995) ont utilisé une approche conjointes "snake" et modèle déformable pour évaluer les paramètres labiaux. En appliquant un seuillage chromatique sur la région des lèvres, une approximation de la forme des lèvres et des paramètres géométriques permet d'initialiser le modèle de lèvres. Ce modèle, dont les contours sont représentés par des paraboles, sert ensuite à initialiser les paramètres des "snakes" (courbes qui s'ajustent au mieux sur des points grâce à une fonction de calcul d'erreur). Un algorithme récursif permet d'ajuster les "snakes" et le modèle de lèvres jusqu'à ce qu'une situation stable soit obtenue. Une dizaine de paramètres labiaux ainsi que la position de la langue sont obtenus par ce processus (Figure 3.60).

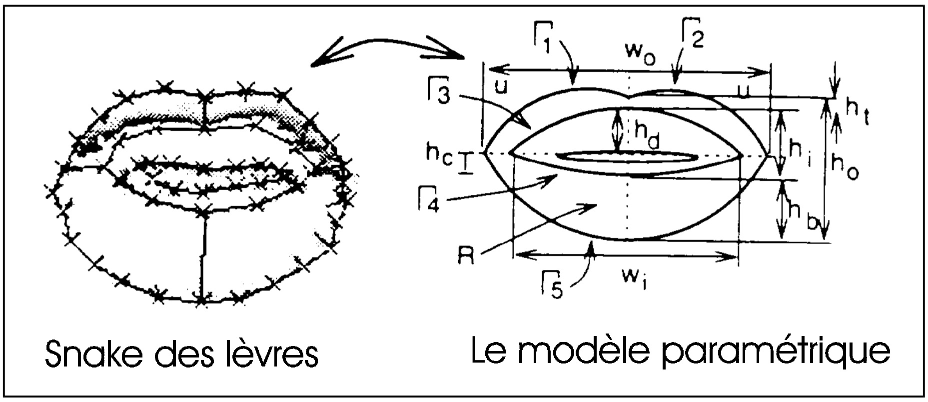


Figure 3.60 : les snakes ajustés aux lèvres et le modèle de lèvres (template) d'après Horbelt et Dugelay (1995).

Rao et Mersereau (1994) ont utilisé une méthode semblable pour extraire les informations visuelles des lèvres à l'aide d'un modèle constitué de quatre portions de paraboles. Grâce à la minimisation de la fonction d'énergie, le modèle est ajusté aux lèvres, ce qui permet de calculer les paramètres géométriques. Les paramètres d'initialisation sont obtenus à l'aide d'un filtrage linéaire qui permet de rehausser les contours.

L'implémentation d'un modèle de lèvres paramétrique a été aussi au centre des recherches de Henneke et al (1994). La Figure 3. 61 illustre ce modèle. Il est commandé par 12 paramètres qui permettent non seulement de reconstruire le mouvement mais aussi l'angle d'orientation ainsi que le centre de coordonnées.

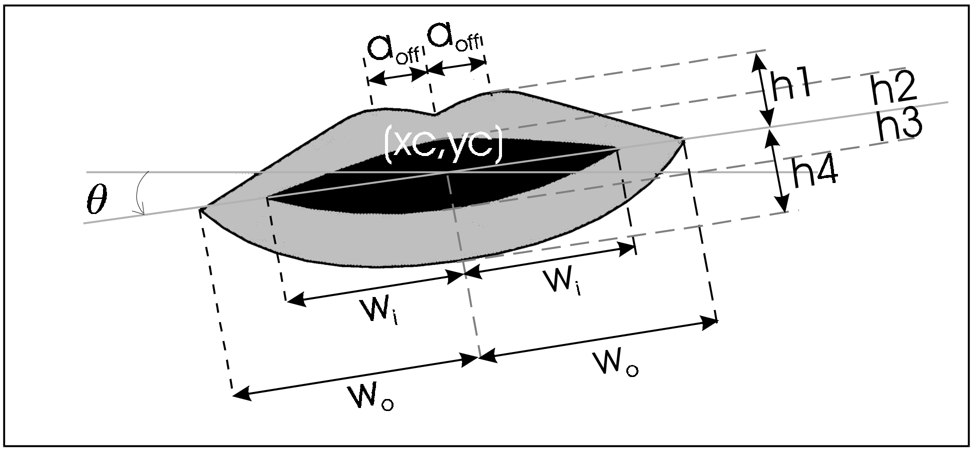


Figure 3. 61 : modèle de lèvres 2D utilisé pour le calcul des paramètres labiaux (d'après Hennecke et al., 1994).

Un filtrage gradient permet de détecter les contours sur des images en niveaux de gris. Une fonction de minimisation de coût composée d'intégrales normalisées de calcul d'aire permet ensuite de superposer le modèle sur les contours réels avec une erreur minimale.

Chandramohan et Silsbee (1996) ont proposé un système de détection de paramètres géométriques fondé sur des modèles déformables. Ils ont montré que le plaquage du modèle sur les contours des lèvres est très sensible à la variation d'éclairage et à l'initialisation des paramètres du modèle. Ils ont donc proposé une méthode qui consiste à utiliser plusieurs modèles applicables chacun à une catégorie d'images.

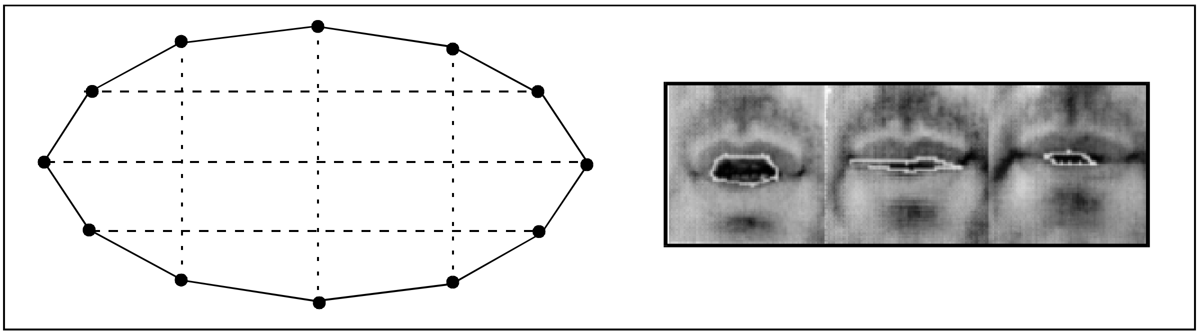


Figure 3.62 : le modèle de lèvres plaqué sur quelques images. Les ressorts sont représentés par des traits en pointillé et peuvent être activés ou désactivés suivant l'image d'entrée (d'après Chandramohan et Silsbee, 1996).

Dans un premier temps, l'image d'entrée est classée parmi des catégories d'images stockées en mémoire. Grâce à cette première phase, une série de paramètres d'initialisation ainsi que le choix du modèle sont obtenus. Le modèle est ensuite ajusté aux contours grâce à la fonction d'optimisation. Les modèles (appelés aussi templates) sont composés de plusieurs points, chacun relié à son voisin par un segment de droite. Des ressorts définissent le type d'interaction entre chaque point. Ayant une géométrie identique, les modèles sont différenciés uniquement par le nombre de ressorts en action (Figure 3.62).

Leroy et Herlin (1995) ont utilisé le principe d'un modèle déformable fondé sur des courbes elliptiques pour une application de reconnaissance de visage et d'extraction des informations labiales. Le suivi des mouvements des lèvres est obtenu en minimisant une fonction d'énergie pour chaque image.

1.4.5 Le poste Visage-Parole

Dans cette section, nous présentons en détail le poste visage-parole de l'ICP conçu par Lallouache (1991). Les mesures provenant de ce système nous ont servi de référence pour l'apprentissage et le test de notre système de reconnaissance audiovisuelle.

Il s'agit d'un système d'analyse du mouvement des lèvres et du menton entièrement automatique. Le principe de détection de la région des lèvres est basé sur le seuillage par chrominance. Pour séparer le vermillon des lèvres du reste du visage, les lèvres du locuteur sont maquillées en bleu car le contraste chromatique naturel entre le visage et les lèvres n'est pas suffisant pour extraire les contours avec précision. En représentant la couleur du visage dans l'espace RVB, on note que la composante bleue a une très faible amplitude par rapport aux deux autres. C'est pour cette raison que Lallouache (1991) utilise le maquillage bleu car cela facilite la tâche du seuillage par chrominance.

Un système de chroma-key analogique permet ensuite d'incruster du noir (niveau 0) sur les parties de l'image ayant une couleur bleue saturée. Son principe est simple : une image cache est générée à partir de l'image du locuteur avec les lèvres bleues. Cette image est binaire et seule la région des lèvres a le niveau 0. Elle commande ensuite un système de multiplexage qui sélectionne soit un signal d'amplitude vidéo nulle, soit l'image couleur du locuteur. Pour que le système soit opérationnel, il faut que ces deux signaux soient parfaitement en phase, faute de quoi la région incrustée en noir est décalée sur l'image finale.

Pour générer l'image cache, il faut connaître les coordonnées du fond de l'image du sujet dans l'espace RVB. Cependant, il est très difficile de séparer ces deux régions juste en se basant sur la composante B (s'il s'agit d'un fond bleu). En pratique on utilise un signal auxiliaire (B-V) pour générer un plan de séparation (B-V) = U_{ref}. (Macheboeuf, 1984). Ce plan permet de distinguer deux régions : la région pour laquelle (B-V) > U_{ref} (c'est à dire la région d'incrustation) et le reste de l'image.

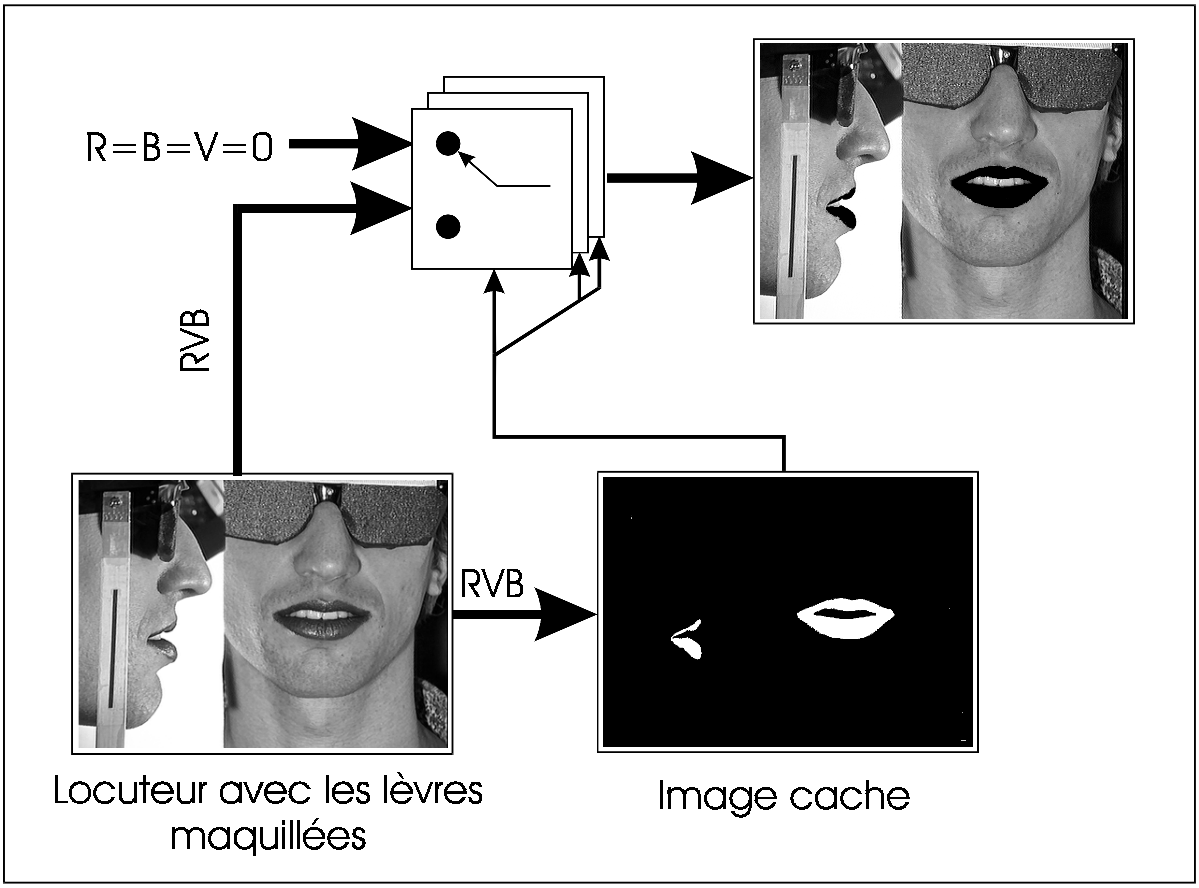


Figure 3.63 : le principe d'incrustation vidéo appliquée à la labiométrie. L'image cache commande le multiplexage entre l'image du locuteur et le niveau 0.

Pour éliminer d'autres couleurs, comme le magenta, qui entrent dans la région délimitée par le plan, on ajuste les coefficients directifs du plan de séparation de façon à avoir une sélectivité plus fine.

Le chroma-key est donc une phase de pré-traitement importante dans le système de Lallouache (1991). Ainsi, en utilisant les algorithmes très simplifiés en traitement d'images pour la détection et la poursuite de contours, on peut calculer les paramètres labiaux avec une grande précision. Afin d'extraire les paramètres de face et de profil, deux caméras sont utilisées, une étant focalisée sur la vue de face et l'autre sur la vue de profil. La Figure 3.64 montre les dispositifs d'enregistrement du poste visage-parole.

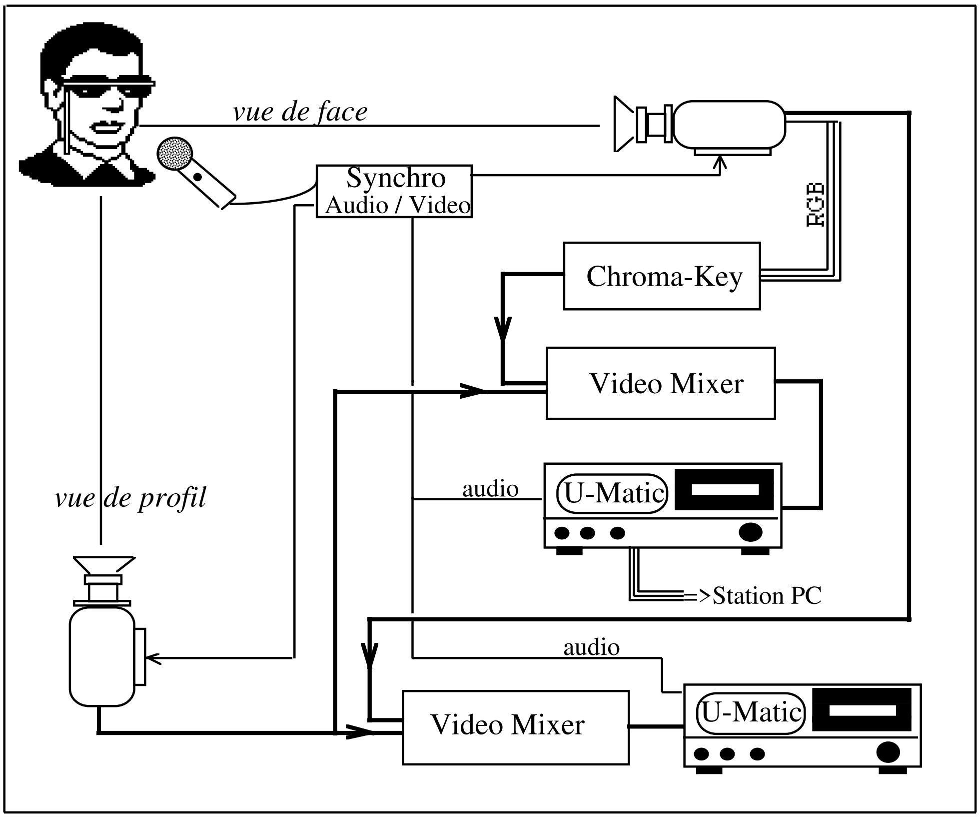


Figure 3.64 : dispositifs d'enregistrement du poste visage-parole selon Lallouache (1991).

Au début de l'enregistrement, une règle est placée devant chacune des caméras pour connaître la largeur et la hauteur d'un pixel afin de pouvoir convertir les mesures en millimètre. Après le passage par le chroma-key analogique, les signaux vidéo provenant des deux caméras sont mélangés pour constituer un seul signal, c'est à dire sur la moitié gauche de l'image on insère le signal de la caméra profil et sur l'autre moitié on prend le signal provenant de la deuxième caméra. Finalement, ce signal vidéo et le son sont enregistrés sur un support magnétique de type U-Matic. Un générateur de code image inscrit le numéro de chaque image sur ces mêmes supports. Ceci permet de référencer la totalité de l'enregistrement à la trame près. Pour retrouver la synchronisation son/image, au début de chaque stimulus, un bip sonore d'une durée de 20 ms (une trame en PAL) est ajouté au signal audio sur un canal différent. En même temps, une diode lumineuse rouge s'allume dans le champ de vision d'une des deux caméras. Il est donc possible de garder la synchronisation audio/vidéo à 20 ms près.

Afin d'obtenir une bonne qualité d'image et réduire les erreurs dues à une mauvaise incrustation par chrominance, les caméras utilisées fournissent un signal RVB sur trois canaux séparés. Ceci permet effectivement d'avoir une plus grande bande passante de chrominance par rapport à un signal composite.

La phase de l'enregistrement étant terminée, l'étape suivante consiste à traiter les images stockées sur la bande magnétique. La Figure 3.65 montre le système de traitement pour l'extraction des paramètres labiaux.

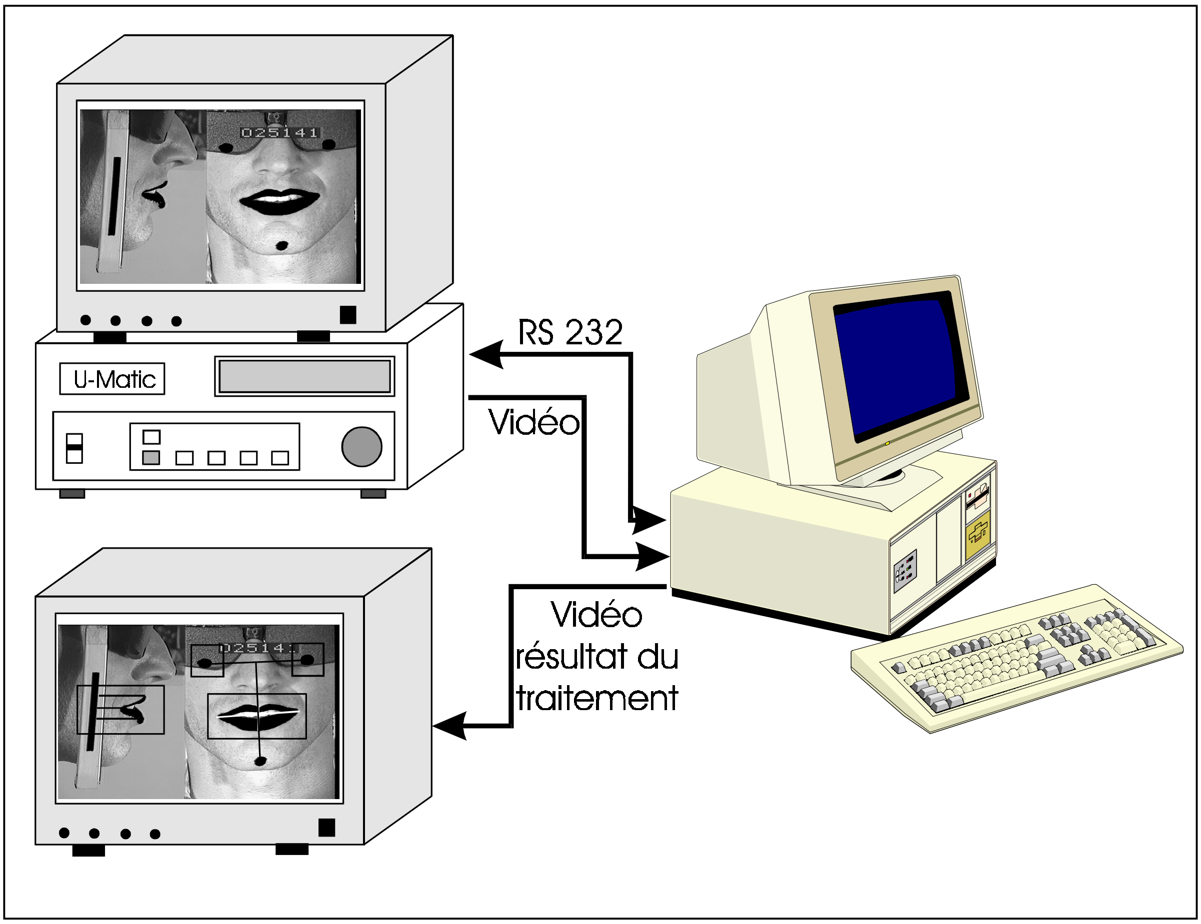


Figure 3.65 : dispositif de traitement du poste visage-parole d'après Lallouache (1991).

Sur la bande magnétique U-Matic, des séquences d’images du même type que celui présenté sur la Figure 3.66 sont enregistrées. Ces images sont préalablement converties en noir et blanc et la région des lèvres et les points de référence sont ramenés au niveau de gris correspondant au noir saturé.

Le système de traitement de Lallouache (1991) est reconnu pour être un des meilleurs au monde pour l'extraction des paramètres labiaux. Il est considéré comme un excellent outil pour les phonéticiens et les chercheurs dans le domaine de la parole audiovisuelle. Les résultats ont une précision de l'ordre du pixel (à peu près de 0.2 à 0.3 mm suivant le zoom des caméras) ce qui permet d'avoir un taux d'erreur très faible. Par contre, il présente des inconvénients dans la pratique : l'utilisation de maquillage pour les lèvres, la contrainte d'immobilité pendant l'enregistrement, la nécessité d'un fort éclairage, le temps de calcul très long.

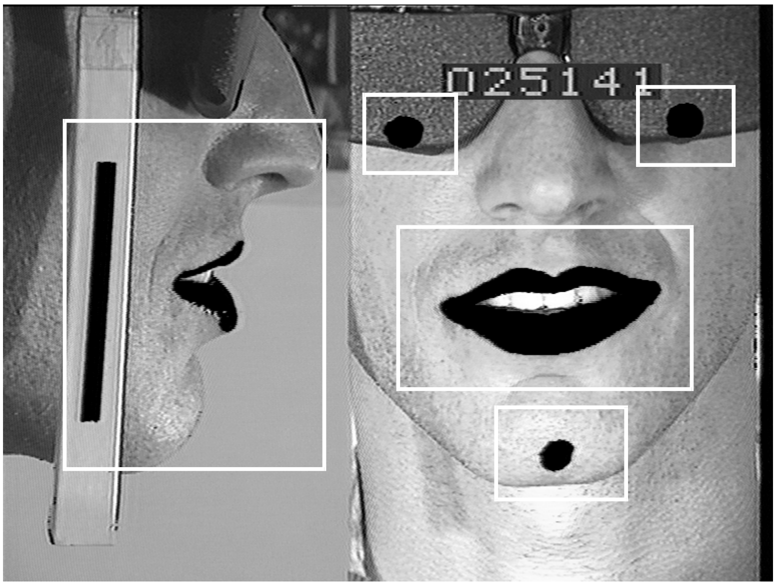


Figure 3.66 : exemple d'image utilisée pour l'extraction des paramètres labiaux. Les coordonnées des fenêtres d'analyse sont initialisées au début du traitement de la séquence et sont ensuite ajustées automatiquement.

De nombreuses recherches sont en cours, notamment à l'ICP, pour implémenter des systèmes de calcul de paramètres tout en minimisant ces contraintes. Mais il est clair que les résultats du poste visage-parole offrent une grande fiabilité ce qui n'est pas le cas des systèmes non-invasifs.

Nous avons donc utilisé ce poste pour constituer notre base de donnée audiovisuelle. Les données provenant de ce système vont ensuite servir à l'évaluation de notre carte électronique de traitement en temps réel. Celle-ci est présentée dans le chapitre 5.

2. Notre approche personnelle

Dans la section précédente, nous avons vu les différentes méthodes permettant d'obtenir les informations visuelles utiles pour la reconnaissance de la parole. Certaines méthodes n'imposent pas de contraintes physiques au locuteur et peuvent être facilement intégrées dans une interface homme/machine. D'autres nécessitent certaines conditions de prises de vue et de pré-traitements. Il paraît évident qu'en augmentant les contraintes, on diminue la complexité des algorithmes de calcul et on augmente la précision. Dans le cadre de notre travail de recherche, il est clair que pour concevoir un système de reconnaissance audiovisuelle capable d'intégrer au mieux les informations auditives et visuelles, une base de données fiable et sans bruit de mesure est indispensable. En se basant sur des données bruitées, il est impossible de détecter les points faibles du système d'intégration pour une éventuelle amélioration. Nous avons donc constitué notre base de travail avec les données du système de Lallouache (1991) dont la fiabilité de mesure a été testée par de nombreuses expériences de production (Abry et Lallouache, 1991), de perception (Escudier et al., 1990; Cathiard et al., 1991) et de synthèse (Benoît et al., 1991, 1992).

3. Conclusion

Nous avons passé en revue les différentes techniques appliquées à l'extraction des informations visuelles pour l'aide à la reconnaissance de la parole. Les méthodes fondées sur le traitement d'images sont celles qui sont le plus souvent utilisées pour des applications de reconnaissance ou de synthèse. Nous avons présenté les deux grandes catégories d'approche dans l'extraction des paramètres labiaux. Parmi ces deux groupes, nous avons sélectionné l'approche fondée sur le traitement de l'image des lèvres et l'extraction des contours, plus précisément, la méthode de Lallouache (1991). Ce système est celui qui garantit la meilleure précision malgré les contraintes qu'il impose.

Chapitre 4

La reconnaissance audiovisuelle de la parole

La reconnaissance automatique de la parole est un développement qui peut modifier l'interaction homme/machine. Malgré de nombreux progrès dans la conception des systèmes de reconnaissance, leur fiabilité reste toujours insuffisante quand les conditions de test sont différentes de celles utilisées pour l'apprentissage, par exemple en utilisation dans un environnement bruité. Les résultats montrent une chute des taux de reconnaissance quand le locuteur se trouve dans un milieu bruité. Pour pallier ce problème, des chercheurs ont commencé à utiliser les informations visuelles comme source secondaire à l'appui des informations acoustiques pour réduire les effets de bruit et améliorer les scores de reconnaissance (Petajan, 1984a, 1984b, Brooke et Patajan, 1986 ; Stork et al., 1992 ; Silsbee, 1993 ; Silsbee et Bovik, 1993 ; Bregler et al., 1993a, 1993b). En effet, on a vu dans le chapitre 2 que la vision du locuteur améliore d'une façon considérable l'intelligibilité de la parole chez l'homme. Notre travail de recherche s'appuie principalement sur les résultats expérimentaux issus de ces études et sur les modèles d'intégration audiovisuelle présentés auparavant. Notre exigence a été de concevoir un système qui intègre au mieux les informations auditives et visuelles et qui s'approche autant que possible des résultats d'intelligibilité observés chez l'homme.

1. État de l'art

Nous passons en revue quelques résultats de recherche sur la reconnaissance de la parole visuelle et audiovisuelle par la machine. Nous avons regroupé les expériences de reconnaissance audiovisuelle selon le modèle d'intégration. Une étude comparative des deux architectures est présentée par la suite.

1.1 La reconnaissance visuelle

Finn (1986) a réalisé un système de reconnaissance purement visuel dont la tâche consistait à classer les consonnes dans un contexte "a C a" dans 10 catégories de visèmes. Elle a obtenu un score de 78% de classements corrects de visèmes (si aba est reconnu ama, on considère que le classement est correcte car ces deux mots appartiennent à la même classe de visème). Elle a aussi étudié l'importance de chaque paramètre dans la reconnaissance visuelle. Les résultats ont révélé que parmi les nombreux paramètres de mesure, ceux relatifs au contour interne contenaient plus d'information que ceux du contour externe.

Goldschen (1993) (voir aussi Goldschen, et al. 1994, 1996) a conçu un système de reconnaissance visuelle de la parole à base de chaînes de Markov cachées (Rabiner, 1989). Il a centré le cadre de ses recherches autour de la reconnaissance de la parole continue en utilisant une base de données de 450 phrases (1789 mots) prononcées par un seul locuteur obtenue par Petajan (1984a, 1984b). Les tests ont donné un score de 25.3% de mots correctement identifiés pour les 150 phrases de test en utilisant uniquement les informations sur la forme de la bouche.

Brooke et Petajan (1986) ont utilisé la variation temporelle des vecteurs radiaux dans la conception d'un système de reconnaissance purement visuel. Cette étude consistait à calculer pour chaque image, la distance entre le point du milieu du segment liant les deux coins des lèvres et le point sur le contour interne pour un angle donné. L'angle zéro correspondait au coin gauche du contour interne des lèvres. La Figure 4.67 montre l'exemple d'un vecteur radial.

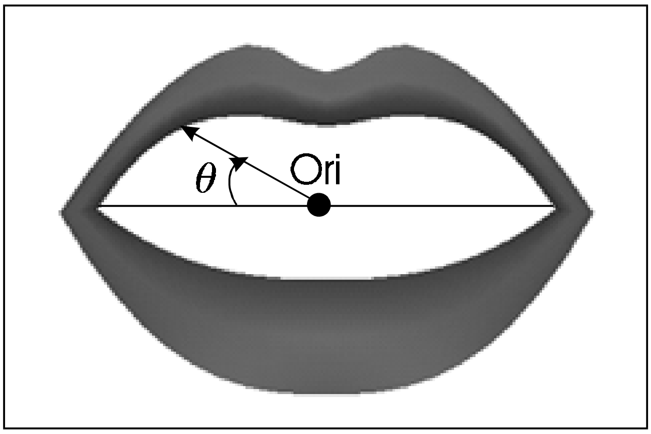


Figure 4.67 : un vecteur radial du contour interne des lèvres. Le point *Ori* est le milieu du segment joignant les deux coins des lèvres.

La variation temporelle était calculée sur deux images successives et cela pour différents angles. Brooke et Petajan (1986) ont montré que cette approche était très efficace pour la reconnaissance de voyelles. La tâche de reconnaissance était de classe les voyelles /a/, /i/ et /u/ dans des mots du type CVC. Le centre acoustique de chaque voyelle était calculé à partir du point correspondant au pic de la courbe de variation temporelle. Les résultats ont montré qu'en utilisant cette méthode, on pouvait atteindre des scores de 100 % pour le classement de ces trois voyelles.

Mase et Pentland (1991) ont développé un système de reconnaissance purement visuel en utilisant les informations sur le flot optique calculé sur quatre fenêtres d'analyse placées autour de la bouche. Les paramètres visuels étaient constitués de sommes pondérées de certaines composantes du mouvement dans les quatre fenêtres et correspondaient à la variation de la largeur et de la hauteur des lèvres. L'expérience consistait à reconnaître dix chiffres prononcés par quatre locuteurs. 70% des chiffres ont été reconnus en utilisant cette technique.

Movellan (1995) a étudié la reconnaissance visuelle pure à base de réseaux de Markov sur un corpus constitué de quatre chiffres en anglais (de un à quatre) prononcé par 12 locuteurs. Le but de cette recherche était de comparer les scores de reconnaissance obtenus à partir de trois différents types d'images comme entrée visuelle :

- images des lèvres codées sur 256 niveaux de gris,
- les mêmes images symétrisées,
- la dérivée temporelle de deux images consécutives symétrisées.

Les résultats montrent que la symétrisation de l'image des lèvres offre des performances de reconnaissance équivalentes aux images non-symétriques (55% d'identifications correctes). La dérivée temporelle apporte une information supplémentaire sur l'allure de la variation des formes des lèvres et donne un meilleur score (85% d'identifications correctes).

Rao et Mersereau (1995) ont procédé à une expérience de reconnaissance purement visuelle basée sur la programmation dynamique. Le corpus était constitué de deux classes : *wow* et *mom*. Le principe d'analyse était l'application d'un modèle et la minimisation de la fonction d'énergie (chapitre 2). Le score de reconnaissance était de 85% d'identification correcte.

Watanabe et Kohda (1990) ont conçu un système de reconnaissance visuelle des voyelles du japonais basé sur des réseaux de neurones. La caractéristique de ce système était la projection des trois paramètres visuels A (étirement interne), B (hauteur interne) et la distance entre la lèvre supérieure et la mâchoire dans l'espace des deux premiers formants. Ils ont montré que cette technique permettait d'obtenir un score de 80% d'identifications correctes des voyelles /a/, /e/, /i/, /o/ et /u/.

On peut citer aussi les études de Nishida (1986) qui a utilisé les informations visuelles pour détecter les frontières acoustiques d'un mot. Son approche consistait à calculer la dérivée temporelle des images binaires provenant de la cavité orale du locuteur. Ainsi, il a montré que les variations temporelles étaient plus importantes au début et à la fin d'un mot. Cette méthode peut être très efficace dans le cas où le signal audio est très bruité et que les frontières des mots sont difficilement détectables dans le cas de la parole continue.

1.2 Les systèmes de reconnaissance audiovisuelle

1.2.1 Intégration suivant le modèle d'identification directe

Brooke et al. (1994) ont implémenté un système de reconnaissance audiovisuelle sur un corpus composé de séries de trois chiffres prononcés par un seul locuteur (Brooke, 1996a). Le classifieur utilisé était un réseau de Markov à densité de probabilité gaussienne. Ils ont choisi une analyse par composantes principales des images provenant des lèvres du locuteur comme source d'information visuelle. Trois différents types de systèmes de reconnaissance audiovisuelle ont été implémentés : dans une première expérience, les modèles markoviens ont été entraînés avec des données acoustiques et visuelles, et ont été testés sous différents niveaux de dégradation du signal acoustique. Les résultats montrent que malgré un gain de reconnaissance dans les zones à faible bruit (RSB > 30 dB), le score de reconnaissance audiovisuelle est inférieur à celui du visuel seul pour des niveaux de bruit élevés (RSB < 25 dB).

Dans une deuxième expérience, l'algorithme de masquage a été appliqué sur les données auditives bruitées (Klatt, 1976) tout en gardant la même architecture de reconnaissance et d'intégration. Les résultats montrent une amélioration des scores de reconnaissance acoustique et audiovisuelle. Cependant, les scores de reconnaissance audiovisuelle descendent au-dessous de celui du visuel seul pour RSB < -5 dB.

Pour pallier ce problème qui est dû à une tendance vers les états à faible variance pour des signaux acoustiques dégradés, la méthode de "*grand variances*" a été implémentée dans une troisième expérience. Dans ce cas, tous les états de chaque modèle ont une même variance.

En appliquant cette technique, les performances du système dans les zones à faible bruit ont baissé. Par contre, pour les forts niveaux de dégradation acoustique, les scores de reconnaissance audiovisuelle sont à peu près équivalents à celui du visuel seul. Nous avons présenté les résultats de ces expériences sur la Figure 4.68.

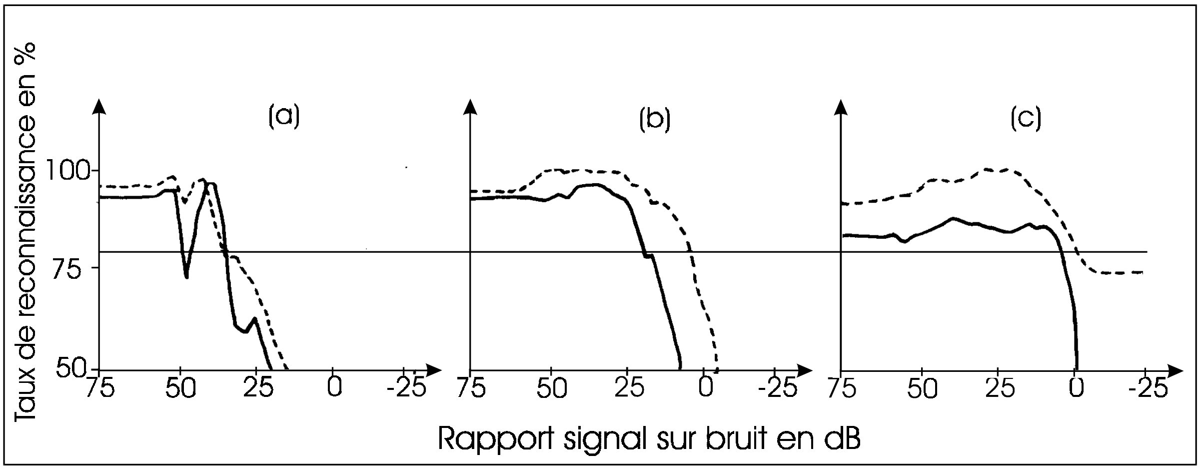


Figure 4.68 : résultats d'expériences de Brooke et al., (1994). Les courbes de la figure (a) correspondent aux résultats de reconnaissance sur les modèles markoviens à variance spécifique pour chaque état. Les résultats de la figure (b) sont obtenus avec le même système en appliquant l'algorithme de masquage. La figure (c) représente les résultats de reconnaissance avec la technique de "grand variances" et l'algorithme de masquage. La ligne horizontale correspond à la reconnaissance visuelle seule, les courbes en trait plein à la reconnaissance auditive et celles en pointillé à la reconnaissance audiovisuelle.

Le système de reconnaissance audiovisuelle de Bregler et al. (1993a) était fondé sur une combinaison de perceptrons multi-couches et réseaux de Markov. Les données mixtes auditives et visuelles sont entrées dans le perceptron qui sert à initialiser les probabilités d'émission des modèles de Markov (pendant la phase d'apprentissage). Le corpus utilisé était constitué de 26 caractères en allemand prononcés par six locuteurs. Pour évaluer les performances de ce système, les signaux acoustiques de test ont été dégradés à trois niveaux de bruit additif. Les résultats de Bregler et al (1993a) sont présentés dans le tableau ci-dessous. La condition visuelle seule n'a pas été testée dans cette expérience.

	Sans bruit	20 dB	15 dB	10 dB
Acoustique seul	89.0%	66.5%	32.7%	43.9%
Audiovisuelle	88.7%	74.0%	54.0%	52.0%

Tableau 4.4 : résultats de reconnaissance acoustique et audiovisuelle de Bregler et al. (1993a). Le rapport signal sur bruit de 15 dB correspond à un signal acoustique dégradé avec d'autres sources de parole. Les deux autres conditions test sont obtenues avec du bruit additif enregistré dans une voiture.

Silsbee et Su (1996) ont étudié la reconnaissance de 22 consonnes prononcées par un seul locuteur dans un système à base de chaînes de Markov discrètes. Les indices visuels étaient obtenus à l'aide d'une approche de traitement par modèle. Grâce à un facteur de pondération des entrées visuelles et acoustiques γ, les deux sources d'information sont intégrées dans les modèles markoviens. γ est estimé automatiquement pendant la phase d'apprentissage avec des données audiovisuelles mixtes. Silsbee et Su (1996) précisent que ce facteur est réestimé pour chaque niveau de bruit, en d'autres termes, les conditions acoustiques de test et d'apprentissage doivent être identiques pour que l'intégration soit optimale. La Figure 4.69 montre l'architecture de ce système.

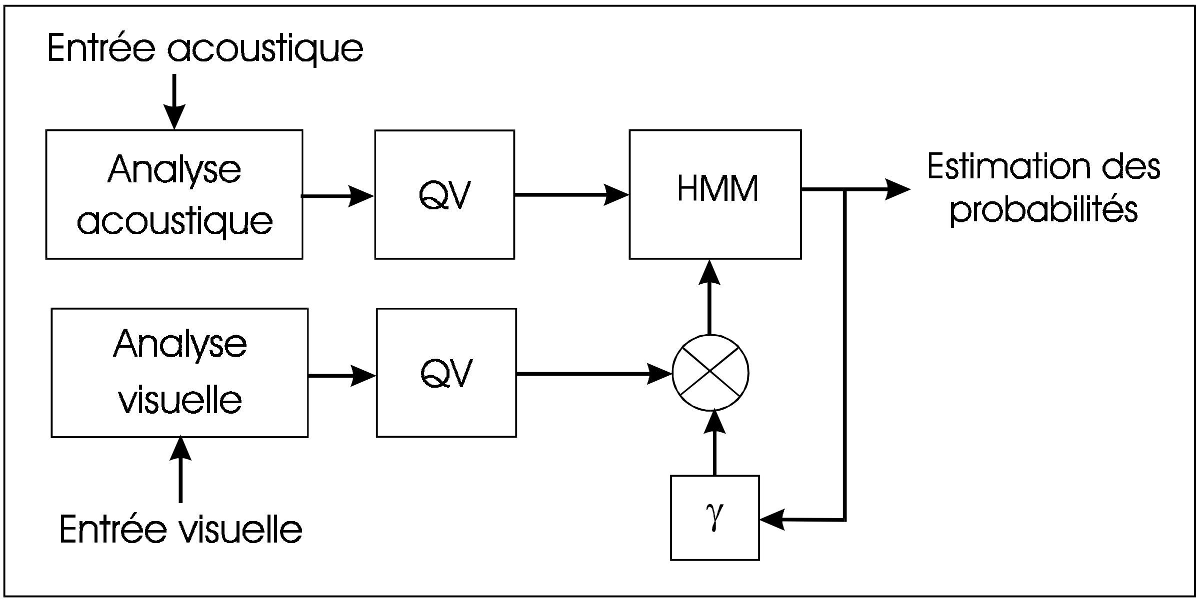


Figure 4.69 : architecture du système de reconnaissance audiovisuelle développé par Silsbee et Su (1996).

Silsbee et Su (1996) ont testé ce système sous différents niveaux de dégradation du signal acoustique. Ils ont vérifié la performance de l'estimation adaptative de γ en comparant les résultats à ceux obtenus avec des valeurs fixes pour ce facteur. Ils ont conclu que globalement, l'estimation adaptative donnait de meilleurs scores (Figure 4.70).

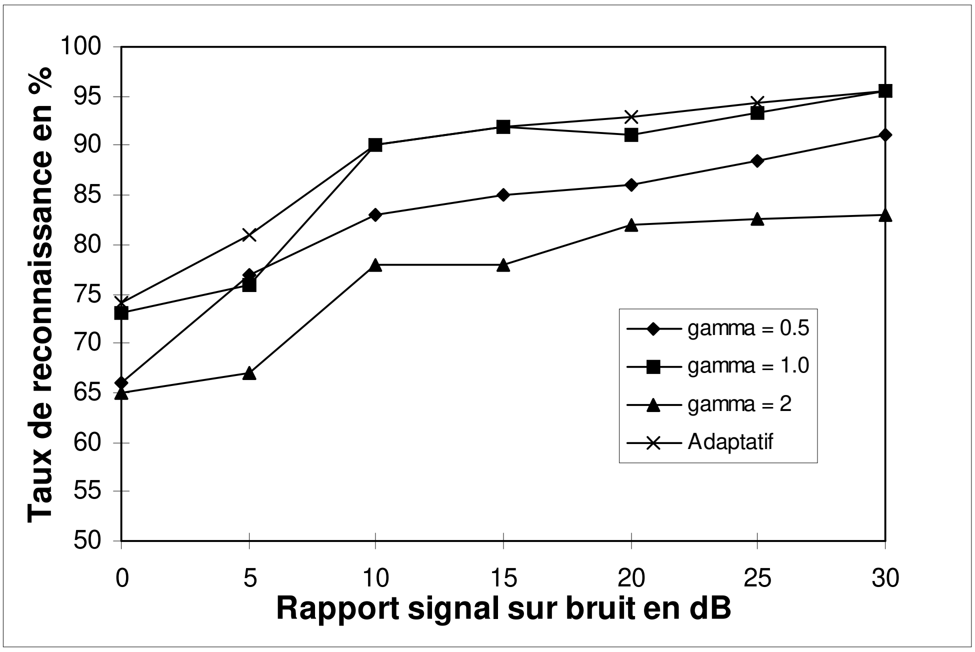


Figure 4.70 : résultats de reconnaissance audiovisuelle suivant la valeur de gamma. Les scores correspondant à la méthode adaptative sont globalement meilleurs (d'après Silsbee et Su, 1996).

Henneke et al. (1996) ont étudié l'effet d'intégration des indices visuels dans un système de reconnaissance. Ces indices sont obtenus en plaquant un modèle de lèvres (voir chapitre 3) sur les lèvres du locuteur (Hennecke et al., 1994). Grâce à ce modèle, neuf paramètres sont calculés et envoyés vers le système de reconnaissance. Ils ont implémenté un classifieur audiovisuel à base de chaîne de Markov pour la reconnaissance des syllabes /da/, /fa/, /la/ et /ma/ prononcées par dix locuteurs. Une intégration audiovisuelle fondée sur le modèle d'identification directe a permis d'améliorer les scores de reconnaissance acoustique seule (Tableau 4.5).

	Scores de reconnaissance
Acoustique seule	43%
Visuelle seule	38%
Audiovisuelle	59%

Tableau 4.5 : résultats de reconnaissance d'après Henneke et al. (1996).

Une autre expérience de reconnaissance audiovisuelle fondée sur le modèle d'intégration à identification directe est celle de Dalton et al. (1996). Ils ont implémenté un système fondé sur la méthode de *Dynamic Time Warping* pour reconnaître 40 mots prononcés par un seul locuteur. L'extraction des indices visuels était basée sur le principe de modélisation des lèvres avec des contours dynamiques (voir section 1.4.4.2). Il était possible de modifier le poids relatif de chaque source d'information dans la décision finale. Les résultats montrent qu'avec une pondération de l'ordre de 0.2, c'est à dire 20% visuel et 80% acoustique, les performances étaient meilleurs.

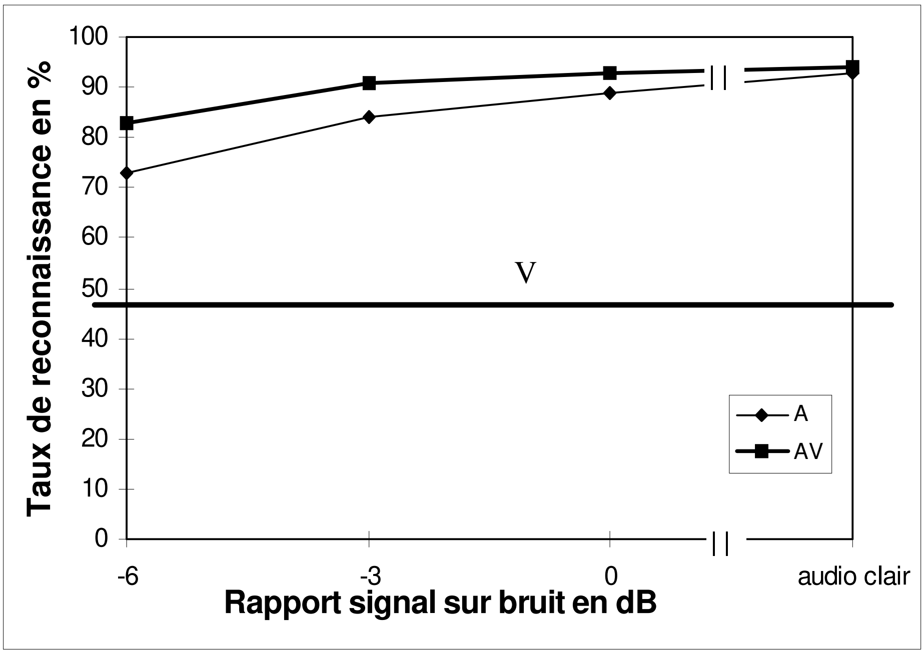


Figure 4.71 : résultats de Dalton et al. (1996) obtenus avec un facteur de pondération égal à 0.2 (20% de participation de la source visuelle dans la décision finale).

1.2.2 Intégration suivant le modèle à identification séparée

Petajan (1984a, 1984b) a développé un système électronique dédié à l'extraction de l'ombre de la cavité buccale pour la reconnaissance audiovisuelle de la parole. La Figure 4.72 reproduit un exemple de données obtenues à partir de ce système.

Petajan (1984a, 1984b) a montré que les paramètres visuels dans l'ordre décroissant de leur importance étaient :

1 - la distance verticale entre la lèvre inférieure et la lèvre supérieure,
2 - l'aire intéro-labiale,
3 - la largeur du contour interne des lèvres,
4 - le périmètre du contour interne.

L'algorithme de "*Time Warping*" a été appliqué pour l'apprentissage du noyau de reconnaissance ainsi que le modèle d'identification séparée pour la fusion des données articulatoires et acoustiques. Pour la décision finale de reconnaissance, le candidat visuel ayant le meilleur score parmi les deux premiers candidats du classifieur acoustique était sélectionné.

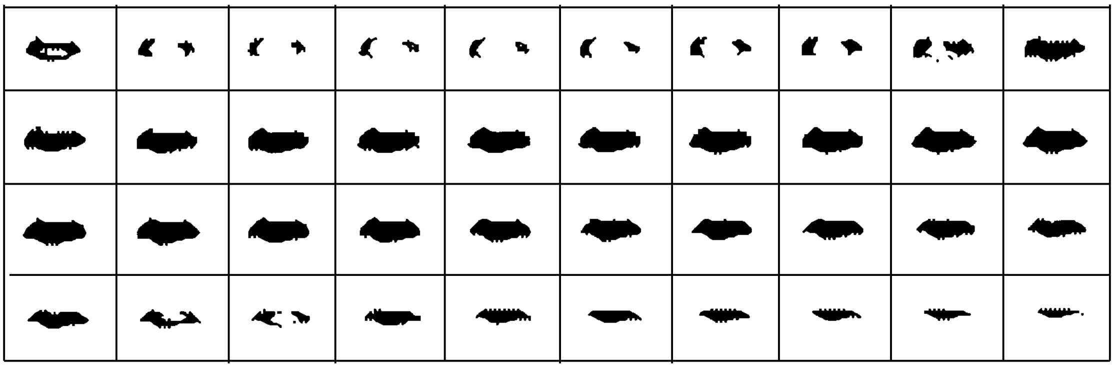


Figure 4.72 : l'ombre de la cavité orale extrait par le système électronique de Petajan (1984a, 1984b). On note la présence de deux zones distinctes sur certaines images, ce qui correspond à l'ombre de la langue.

Si les deux premiers candidats acoustiques appartenaient à la même classe de visème (ce qui est rare dû à la complémentarité audition/vision), la réponse du classifieur visuel était ignorée. Petajan (1984a, 1984b) a testé son système sur un vocabulaire composé de 100 mots, 10 chiffres et 26 lettres en anglais prononcés par un seul locuteur. Deux répétitions pour les mots, dix pour les chiffres et quatre pour les lettres ont été enregistrées. Les résultats sont présentés dans le tableau suivant :

	Vocabulaire de test		
	Chiffres	Lettres	Mots
acoustique seule	95%	64%	65%
visuelle seule	99%	65% *	60%
audiovisuelle	100%	66%	78%

Tableau 4.6 : résultats de Petajan (1984a, 1984b). Le résultat de reconnaissance en condition visuelle seule (*) sur les lettres a été obtenu sur 81 échantillons alors que les autres tests sur les lettres ont été élaborés sur un ensemble de 104 échantillons.

Petajan et al. (1987) ont élaboré un système semblable fondés sur la quantification vectorielle des images d'entrée. En utilisant le même module électronique, une séquence d'images binaires de la cavité buccale correspondant à la prononciation d'un mot est quantifiée sur 256 images de référence, ce qui réduit le débit de transmission des informations visuelles à 60 octets par seconde pour un format vidéo NTSC. Le classifieur acoustique était identique à celui de l'expérience citée précédemment. Le système a été entraîné et testé sur un vocabulaire composé des 10 chiffres et des 26 lettres de l'anglais prononcés par quatre locuteurs. Cependant, il faut noter que le système a été testé et entraîné avec le même locuteur pour chaque expérience.

Pour intégrer les deux sources d'information, Petajan et al. (1987) ont mis en œuvre 30 règles qui prévoyaient la décision finale suivant différents cas de figure. Ces règles dépendaient du corpus et ne pouvaient pas être appliquées à un autre vocabulaire. Nous avons reporté les résultats de l'expérience de reconnaissance auditive, visuelle et audiovisuelle de Petajan et al. (1987) obtenus sur le corpus des lettres prononcées par deux locuteurs.

	Reconnaissance acoustique	**Reconnaissance visuelle**	**Reconnaissance audiovisuelle**
Locuteur 2	72%	63%	86%
Locuteur 4	53%	66%	69%

Tableau 4.7 : résultats d'expérience de Petajan et al. (1987). Les tests audiovisuels ont été élaborés uniquement sur le corpus des lettres prononcées par les locuteurs 2 et 4.

On note un gain d'identification correcte de 14% pour le locuteur 2 et 16% pour le locuteur 4 par rapport à la condition acoustique seule. Ce qui correspond, pour le corpus de test contenant 300 mots, à un gain de 42 mots correctement identifiés pour le locuteur 2 et de 48 mots pour le locuteur 4.

Stork et al.. (1992) ont élaboré un système de reconnaissance audiovisuelle à base de TDNN pour la classification de dix noms en anglais commençant par les lettres b, d, f, m, n, p, s, t, v et f. L'entrée visuelle du système était constituée des paramètres sur l'étirement interne, l'ouverture des lèvres et le mouvement du menton. Ils ont comparé les résultats de reconnaissance audiovisuelle avec ceux obtenus avec le classifieur visuel seul et acoustique seul, eux aussi à base de TDNN. Les informations visuelles et auditives étaient intégrées suivant la loi bayesienne à la sortie des deux classifieurs :

Equation 4.7 :

$$p(C|A\&V) = k \cdot p(C|A) \cdot p(C|V)$$

dans laquelle k est le facteur de normalisation pour que la somme des probabilités soit égale à l'unité, $p(C|A)$ et $p(C|V)$ sont respectivement les probabilités de sortie des réseaux acoustiques et visuels quand le prototype *C* est présenté à l'entrée de chacun, et finalement *p(C|A&V)* est la probabilité audiovisuelle.

Les réseaux acoustiques et audiovisuels ont été testés sous deux conditions : condition audio clair et audio bruité. Le tableau ci-dessous résume les résultats de cette expérience.

	Audio clair	Audio bruité
Audiovisuelle	91%	74%
Acoustique	64%	43%
Visuelle	51%	

Tableau 4.8 : résultats d'expérience de Stork et al. (1992).

Prasad et al. (1993) ont mis au point un système de reconnaissance audiovisuelle de la parole fondé sur des TDNN pour la classification de cinq consonnes prononcées par quatre locuteurs dans le même contexte vocalique : /ba/, /da/, /fa/, /la/ et /ma/. Les résultats montrent qu'en utilisant une approche bayesienne dans un modèle à identification séparée, le score global est nettement supérieur aux scores visuel et acoustiques seuls (50-65% pour la reconnaissance visuelle seule, 60-80% pour la reconnaissance acoustique seule et 90-98% pour la reconnaissance audiovisuelle).

Silsbee (1993) (voir aussi Silsbee et Bovik, 1993) a conçu un système de reconnaissance audiovisuelle de la parole à base de chaînes de Markov suivant l'architecture présentée sur la (Figure 4.73).

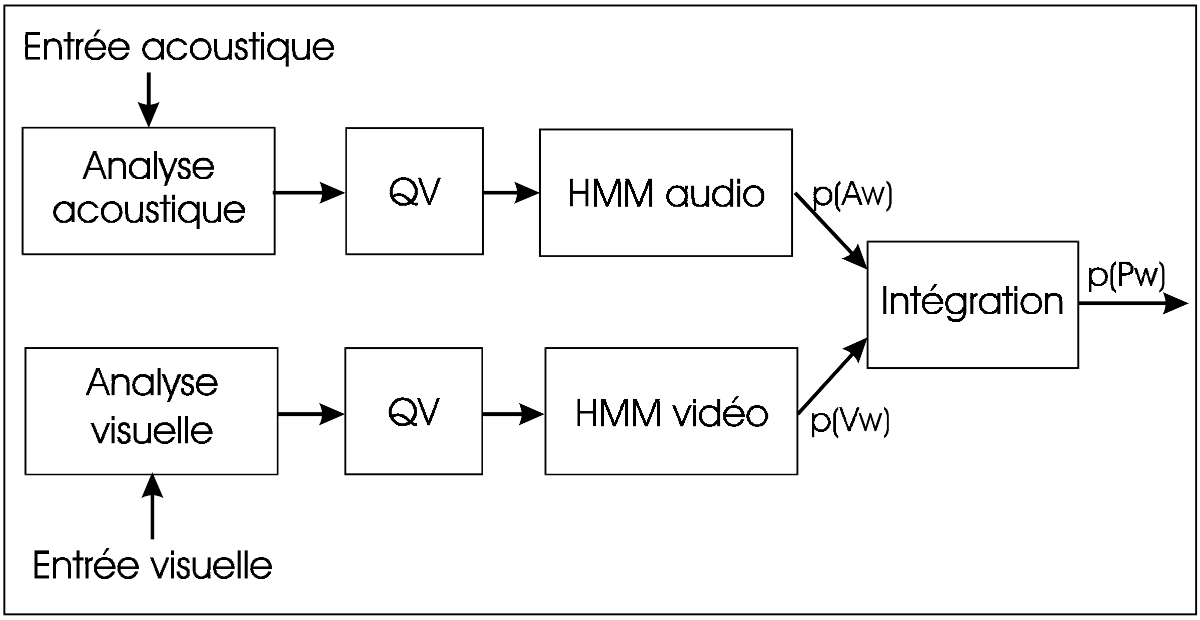
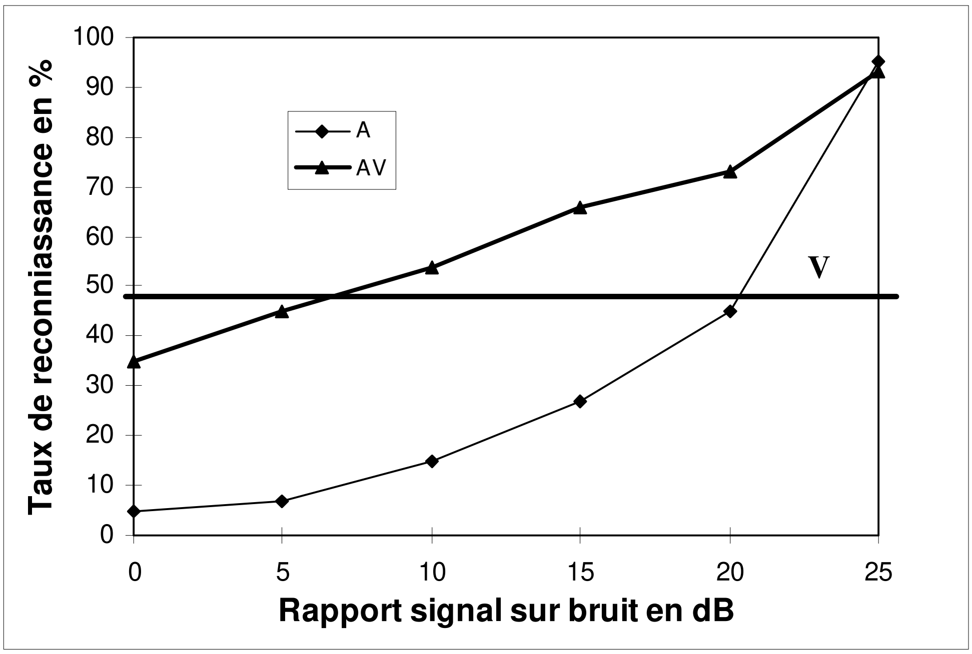


Figure 4.73 : le système de reconnaissance de Silsbee (1993).

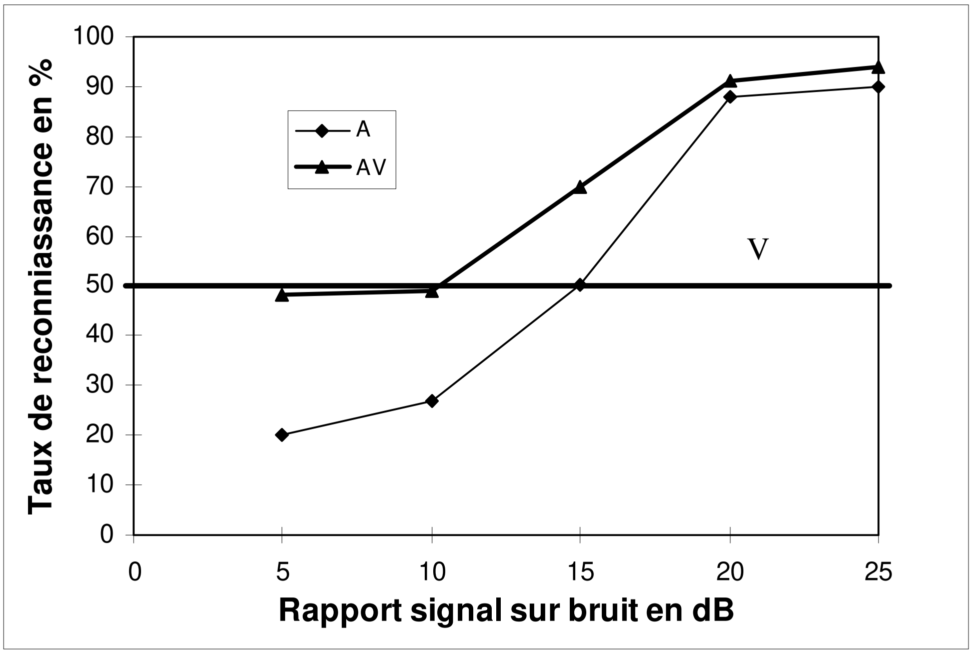
Dans chaque modalité, les données sont injectés dans deux chaînes de Markov discrètes. Les probabilités associées à chaque mot pour la modalité auditive *p(Aw)* et visuelle *p(Vw)* sont obtenues en sortie. L'intégration se fait à partir de la formule suivante :

Equation 4.8 :
$$p(P_w) = p(A_w)^{\lambda} \cdot p(V_w)^{(1-\lambda)}$$

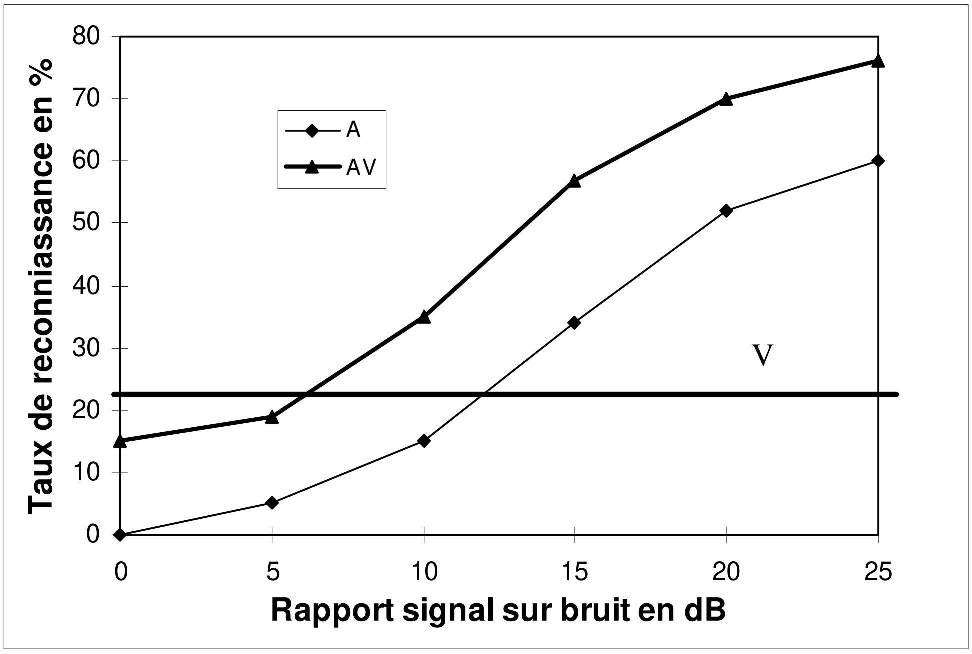
dans laquelle λ est un facteur de pondération qui permet de donner plus de poids à la décision du classifieur visuel quand l'entrée acoustique est bruitée. Silsbee (1993) prétend qu'il n'existe aucun lien permettant de calculer ce facteur en fonction du niveau de bruit du signal acoustique. Il a testé son système avec trois différents corpus audiovisuels : un premier corpus composé de 22 consonnes dans le contexte /aCa/ prononcées par un seul locuteur, un deuxième composé de 14 voyelles et diphtongues prononcées par un locuteur et un troisième composé de 500 mots prononcés par trois locuteurs. Silsbee (1993) a évalué les performances de son système à différents niveaux de dégradation du signal acoustique et en faisant varier le facteur de pondération λ. Il a conclu qu'avec la condition $0.2 < \lambda < 0.3$, les scores de reconnaissance audiovisuelle étaient meilleurs. Nous avons reporté quelques résultats de Silsbee (1993) dans la Figure 4.74.



(a)



(b)



(c)

Figure 4.74 : résultats d'expériences de reconnaissance de Silsbee (1993) avec $\lambda = 0.25$ sur le corpus de consonnes (figure a), le corpus de voyelles (figure b) et le corpus de mots (figure c). Le système a été entraîné avec des signaux acoustiques clairs et testés avec des signaux bruités. La condition 0 dB est inexistante pour le corpus de voyelles.

Waibel et al. (1994) (voir aussi Bregler et al., 1993b) ont implémenté une architecture de reconnaissance audiovisuelle fondée sur deux réseaux de TDNN pour chacune des modalités. Une quatrième couche à la sortie de ces deux réseaux intègre les informations acoustiques et visuelles. En entrée visuelle du système, des images noir et blanc, centrées sur la région des lèvres et filtrées passe-bas, étaient utilisées. Ce système a été testé sur un vocabulaire composé de 26 lettres de l'alphabet allemand prononcées par deux locuteurs. La Figure 4.75 montre l'architecture de ce classifieur audiovisuel.

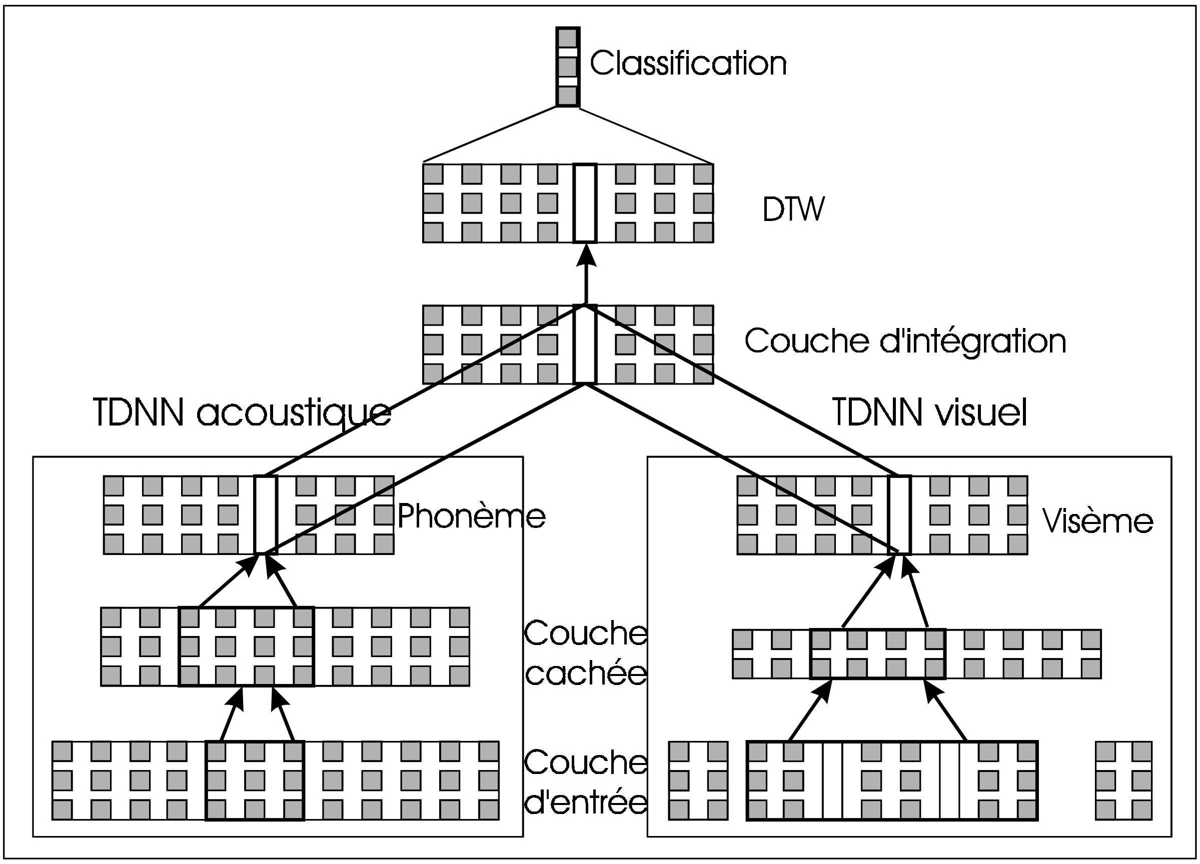


Figure 4.75 : architecture du système de reconnaissance audiovisuelle implémenté par Waibel et al. (1994).

Les performances du système ont été évaluées avec deux locuteurs sous deux conditions de test : l'entrée audio claire et bruitée. Il faut noter que l'apprentissage et le test ont été effectués avec la voix du même locuteur. Les résultats de cette expérience sont présentés dans le tableau ci-dessous.

	Audio clair		**Audio bruité**	
	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2
Acoustique	88.8%	97.0%	47.2%	59.0%
Visuelle	31.6%	46.9%	31.6%	46.9%
Audiovisuelle	93.2%	97.2%	75.6%	69.6%

Tableau 4.9 : expérience de reconnaissance de Waibel et al. (1994).

1.2.3 Système fondé sur le recodage dans la modalité acoustique

Yuhas et al. (1989, 1990) ont réalisé un système de reconnaissance audiovisuelle basé sur le modèle de recodage dans la modalité dominante. Grâce à un réseau de neurones, le système associait la forme des lèvres à un spectre acoustique. La tâche de reconnaissance était d'identifier neuf voyelles en anglais avec l'image statique des lèvres pour chaque voyelle en présence de bruit. Le spectre estimé à partir de l'image des lèvres (S_V) était combiné avec celui obtenu par l'analyse acoustique (S_A) pour constituer un spectre audiovisuel (S_{AV}) suivant la formule suivante :

Equation 4.9

$$S_{AV} = \alpha \cdot S_V + (1-\alpha) S_A$$

dans laquelle α est le facteur de pondération qui permet de donner plus de poids au spectre visuel quand l'entrée acoustique est bruitée.

Grâce à un autre réseau de neurones, le spectre audiovisuel permettait d'identifier la voyelle prononcée. Ce système en visuel seul a donné 55% d'identifications correctes. Yuhas et al. (1989, 1990) ont montré que malgré la chute sensible des performances du classifieur acoustique à forts niveaux de bruit (11% d'identification correcte à -24 dB), le système de reconnaissance audiovisuelle pouvait restituer une grande partie de l'information manquante (50% à -24 dB).

1.2.4 Quelques études comparatives

Robert-Ribes et al. (1996) (voir aussi Robert-Ribes, 1995) ont comparé les performances des deux modèles d'intégration à identification séparée et recodage dans la modalité acoustique dans une application de reconnaissance des voyelles tenues /a/, /e/, /i/, /O/, /y/, /o/ et /u/ du français prononcées par un seul locuteur. Ils ont sélectionné les paramètres A (étirement interne), B (séparation interne) et S (aire intéro-labiale) comme entrées visuelles de leur système qui était basé sur un classifieur gaussien. Les résultats sont présentés sur la Figure 4.76.

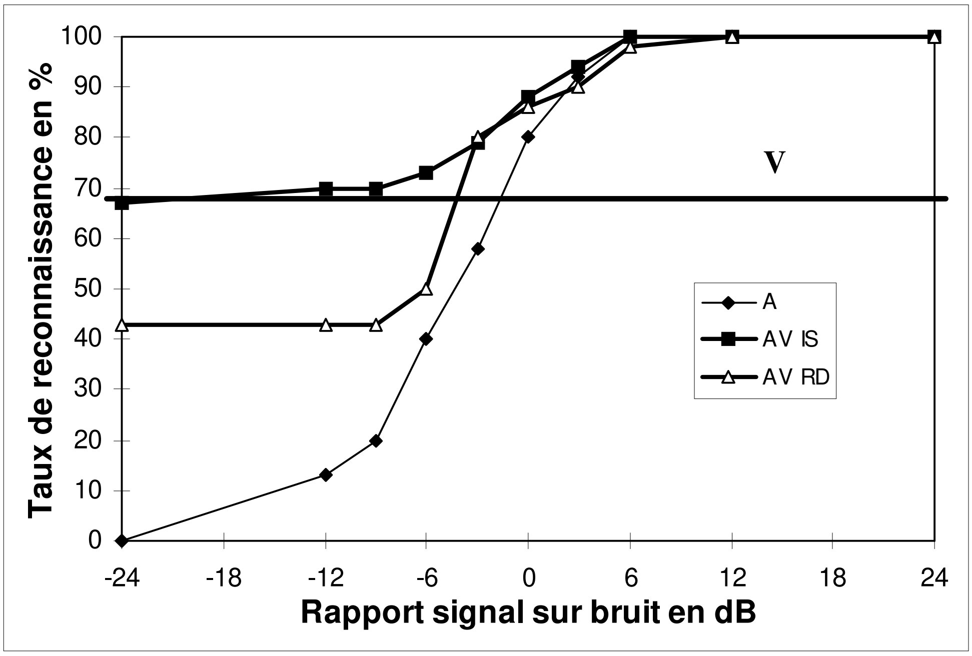


Figure 4.76 : résultats de reconnaissance visuelle, acoustique et audiovisuelle suivant deux modèles d'intégration : le modèle d'identification séparée (IS) et le modèle de recodage dans la modalité dominante (RD) (d'après Robert-Ribes et al., 1996). Le score nul représente le taux aléatoire.

Les résultats de cette expérience prouvent que le modèle de recodage dans l'espace acoustique ne peut pas exploiter les informations visuelles d'une façon optimale. En effet, pour les fortes dégradations acoustiques (RSB < -6 dB), les scores de reconnaissance audiovisuelle fondés sur ce modèle sont inférieurs à celui du visuel seul.

Dans une autre expérience, Robert-Ribes (1995) a conclu que le modèle d'intégration le plus plausible était celui du recodage dans l'espace moteur. Ce modèle intègre mieux les deux sources d'information et permet d'expliquer les propriétés de la perception audiovisuelle.

Dans une autre expérience Movellan et Chadderdon (1996) ont étudié la reconnaissance audiovisuelle en utilisant les chaînes de Markov et un corpus constitué des quatre premiers chiffres en anglais prononcés par différents locuteurs. Ils ont conçu deux systèmes de reconnaissance audiovisuelle à base de chaînes de Markov selon les modèles d'identifications directe et séparée. Les systèmes étaient entraînés avec le signal audio bruité et testés à différents niveaux de bruit additif. Les résultats montrent que généralement le système à deux réseaux indépendants offre de meilleurs performances. Cependant dans certains cas où l'audio est fortement bruité, les scores de reconnaissance audiovisuelle sont inférieurs à ceux obtenus uniquement avec les indices visuels. Nous avons reporté les résultats de ces expériences sur la Figure 4.77

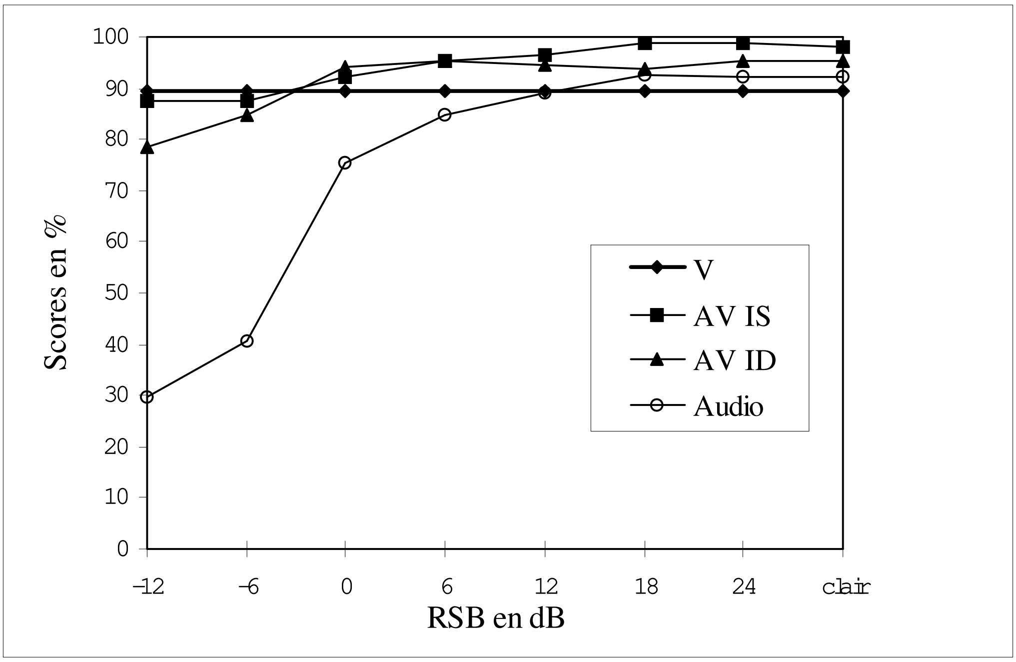


Figure 4.77 : scores de reconnaissance obtenus par Movellan et Chadderdon (1996). AV ID représente les résultats par intégration directe, AV IS est une intégration séparée selon une loi multiplicative.

Silsbee (1994) a comparé les résultats des deux modèles d'identification directe et d'identification séparée dans une application de reconnaissance de 22 consonnes en anglais. Il a implémenté une nouvelle architecture dans laquelle l'intégration s'effectuait à l'intérieur des modèles de Markov sur des vecteurs audiovisuels mixtes. Un coefficient de pondération lui a permis d'ajuster les poids relatifs de chaque vecteur de paramètres visuels et acoustiques dans le processus de reconnaissance (voir Silsbee et Su, 1996). Il a ainsi comparé les résultats de ce système à celui basé sur le modèle d'identification directe (Silsbee, 1993). Il a conclu que les performances des deux systèmes étaient pratiquement identiques. Cependant, il prétend que le modèle d'intégration directe peut donner de meilleurs scores à condition que les vecteurs acoustiques et visuels soient parfaitement synchronisés, ce qui n'était pas le cas dans son expérience.

Duchnowski et al. (1994) ont comparé les résultats de reconnaissance de deux systèmes à base de TDNN suivant que l'intégration audiovisuelle s'effectuait au niveau de la couche "phonème " ou de la couche "cachée" (voir les résultats de Waibel et al., 1994 ; section 1.2.2). La Figure 4.78 montre l'architecture de ces deux modèles.

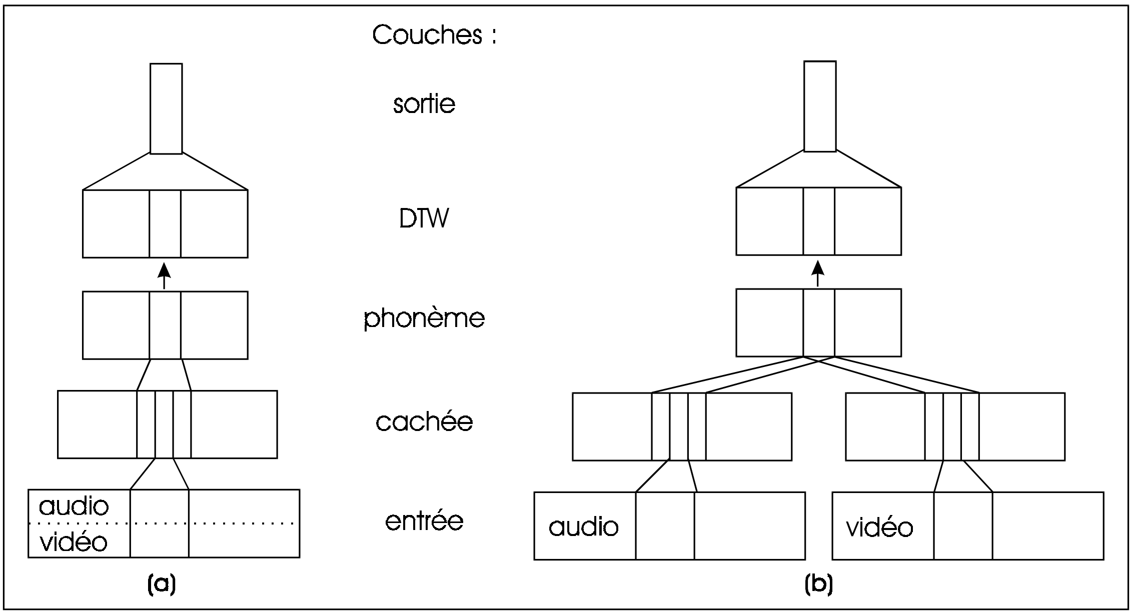


Figure 4.78 : deux architectures de reconnaissance audiovisuelle basées sur des TDNN (d'après Duchnowski et al., 1994).

Les deux systèmes devaient identifier des séquences de lettres en allemand, avec une pause entre chaque lettre, prononcées par un seul locuteur. Les performances ont été évaluées sous trois conditions acoustiques : sans bruit et avec deux niveaux de bruit blanc (les auteurs ne spécifient pas le rapport signal sur bruit). Des images brutes des lèvres étaient utilisées comme entrée visuelle dans cette expérience. Les résultats de reconnaissance de ces deux systèmes ont été comparés à ceux obtenus par Waibel et al. (1994) (voir Figure 4.75). Les taux de reconnaissance de ces trois modèles sont présentés sur la Figure 4.79.

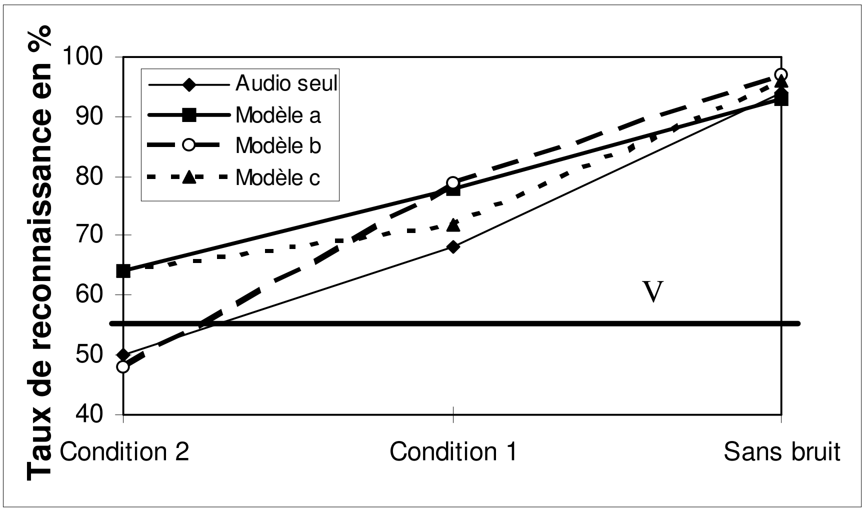


Figure 4.79 : comparaison des performances de trois modèles de TDNN. Le modèle (c) est celui implémenté par Waibel et al., (1994) décrit auparavant. Les conditions 1 et 2 correspondent à deux niveaux croissants de dégradation acoustique (d'après Duchnowski et al., 1994).

Les auteurs concluent que les deux modèles ont pratiquement des performances identiques.

Alissali et al. (1996) ont comparé les performances de trois architectures de reconnaissance audiovisuelle à base de HMM :

- le système S1 est conçu selon le modèle d'identification directe dans lequel les paramètres A (étirement interne), B (séparation interne) et S (aire intéro-labiale) sont concaténés aux paramètres acoustiques.
- le système S2 est composé de deux réseaux markoviens acoustiques et visuels selon le modèle d'identification séparée. La classification acoustique est élaborée suivant l'algorithme des N meilleures solutions. Cet algorithme permet de retenir N solutions pour chaque état et chaque instant, contrairement à l'algorithme de Viterbi qui ne sélectionne que la meilleure. Grâce à cette technique, les informations sur les limites temporelles de chaque échantillon sont fournies au classifieur visuel. La décision finale est obtenue selon une fonction linéaire.
- le système S3 était identique au système S2 à la différence que le classifieur visuel identifiait les visèmes et non pas les phonèmes.

Les tests ont été élaborés avec des signaux acoustiques dégradés à plusieurs niveaux de bruit. En comparant les résultats de reconnaissance de ces trois systèmes, on observe que les scores de S2 et S3 sont meilleurs que ceux de S1. Les auteurs concluent que le modèle d'identification séparée est plus adaptée à l'intégration des informations visuelles et acoustiques que le modèle à intégration directe. Nous avons reporté les résultats de cette expérience sur la Figure 4.80. La condition visuelle seule n'a pas été testée.

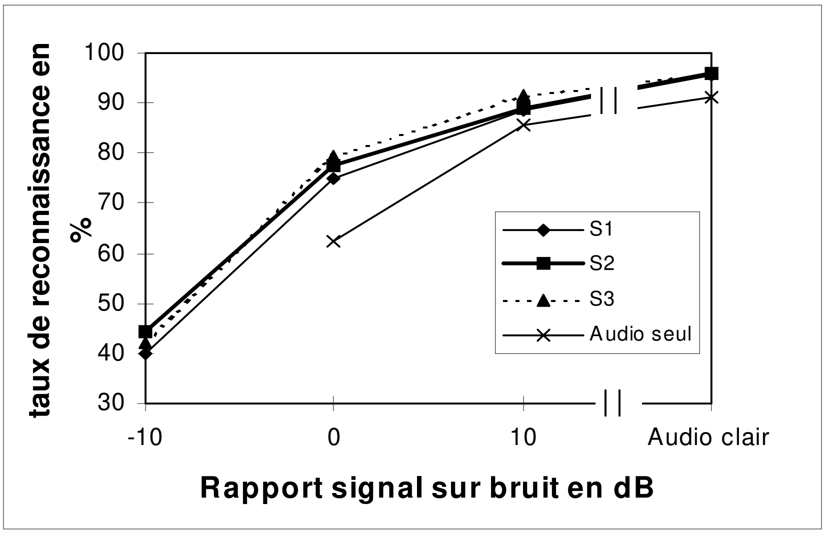


Figure 4.80 : résultats de reconnaissance obtenus par Alissali et al. (1996).

Jacob et Sénac (1996) ont étudié les performances de deux systèmes de reconnaissance utilisant des chaînes de Markov : un système classique fondé sur l'intégration directe et un autre basé sur l'architecture maître-esclave. Ce dernier était composé de deux modèles de Markov mis en parallèle. L'idée de base était de modéliser une adaptation dynamique des lois de probabilités d'un des modèles de Markov en fonction du contexte courant modélisé par l'autre. Ce contexte peut être un indice sur le voisement, la nasalité, etc. Dans l'expérience de Jacob et Sénac (1996), le modèle maître était un modèle ergodique représentant les configurations bouche fermée, bouche ouverte et bouche semi-ouverte. Le modèle esclave était un modèle dont les unités étaient des pseudo-diphones. Les vecteurs d'observation étaient composés des paramètres articulatoires ainsi que leurs dérivées concaténées aux coefficients cepstraux. Les résultats de reconnaissance sur des séquences de lettres prononcées par un seul locuteur montrent des performances quasi identiques pour les deux modèles. Les auteurs soulignent que le modèle maître-esclave peut fournir de meilleurs résultats à condition d'entraîner le système avec plus de données. Dans ce cas, le modèle maître peut être étendu pour se rapprocher des études statistiques sur l'émergence des visèmes.

Jourlin (1996) a proposé une architecture d'intégration obtenue par le produit de deux modèles markoviens acoustiques et visuels. Il a comparé les performances de cette architecture à celles obtenues avec un HMM audiovisuel fondé sur l'intégration directe. Jourlin (1996) précise que son modèle permet de gérer de façon optimale l'asynchronie à l'intérieur des unités de reconnaissance. Le corpus était composé de 26 lettres de l'alphabet français prononcées par deux locuteurs. Les résultats montrent une amélioration des performances de ce modèle par rapport au modèle classique d'identification directe pour un locuteur (92% pour le modèle classique et 93.5% pour le produit des modèles acoustiques et visuels). Jourlin (1996) n'a pas observé une nette amélioration pour le deuxième locuteur. Il précise qu'en tenant compte de la taille du corpus, il ne faut pas attacher trop d'importance aux résultats dans leur valeur absolue.

Dans une approche semblable, Brooke (1996b) a comparé les performances de ces deux types de modèles dans une application de reconnaissance de séquences de chiffres en anglais prononcées par un seul locuteur. L'entrée visuelle était constituée de 10 coefficients obtenus par analyse en composantes principales des images des lèvres. Brooke (1996b) a testé les deux systèmes sous différents niveaux de bruit additif. Les résultats sont en accord avec ceux obtenus par Jourlin (1996). Le produit des modèles markoviens acoustiques et visuels donne de meilleures performances par rapport au modèle à identification directe. Nous avons reporté les résultats de Brooke (1996b) dans le tableau suivant :

	Rapport signal sur bruit en dB					
	-22	**-13**	**4**	**5**	**14**	**23**
Audio seul	0.0%	0.0%	32.3%	84.0%	86.0%	84.7%
Modèle A	74.3%	74.3%	77.7	89.3%	96.3%	98.3%
Modèle B	79.7%	79.7%	81.0%	89.3%	95.7%	98.7%
Vidéo seul	77.0%					

Tableau 4.10 : résultats de reconnaissance de Brooke (1996b). Le modèle A représente le système à identification directe ; le modèle B est le système obtenu par le produit des HMM acoustiques et visuels.

Dans une récente étude, Su et Silsbee (1996) ont comparé les performances des modèles d'intégration directe et séparée dans une application de reconnaissance à base de HMM. Ils ont implémenté une nouvelle architecture de modèles Markoviens dont les densités de probabilités sont modifiées de façon à avoir des valeurs plus importantes dans les zones loin de la moyenne par rapport à une distribution gaussienne classique. Ceci permet de mieux représenter la densité de probabilité du système quand les conditions de test et d'apprentissage sont différentes. Su et Silsbee (1996) ont comparé les résultats obtenus grâce à cette technique avec ceux obtenus avec le modèle d'identification séparée avec pondération automatique de chaque canal (voir Silsbee et Su, 1996). Les résultats montrent que le système d'identification séparée offre de meilleures performances que le système basé sur l'identification directe. La technique de modification des densités de probabilité permet d'obtenir une amélioration dans les zones à forte dégradation par rapport au système classique basé sur l'identification directe, cependant elle dégrade les résultats quand l'entrée acoustique est faiblement bruitée (AV = 77% alors que A = 89% pour RSB = 30 dB). Dans cette expérience, l'identification séparée est la seule technique à assurer AV > A pour toutes les conditions de test.

2. Discussion

Dans la section précédente, nous avons vu que l'intégration audiovisuelle dans un système de reconnaissance audiovisuelle de la parole reste le problème majeur. De nombreux chercheurs ont adopté le modèle d'identification directe, tandis que d'autres se sont orientés vers le modèle à identification séparée. Plus rares sont ceux qui ont adopté la méthode du recodage dans l'espace acoustique ou dans l'espace moteur. L'approche fondée sur le modèle à identification directe donne une amélioration des scores de reconnaissance audiovisuelle par rapport à la condition acoustique seule. Cependant dans certains cas, la condition AV > A et V n'est pas respectée pour toutes les conditions acoustiques (Bregler et al., 1993a). Dans d'autres recherches, cette condition est satisfaite en dépit des performances globales médiocres du système (Brooke et al., 1994). Les résultats de Dalton et al. (1996) sont des exemples de cette situation : la condition AV > A et V est respectée pour tous les niveaux de bruit, en revanche, l'apport du visuel reste faible (10% de gain à -6 dB et un gain négligeable pour la condition audio clair).

L'approche fondée sur le modèle à identification séparée pose un autre type de problème : comment obtenir la décision finale suivant la décision des deux canaux visuel et acoustique ? Petajan et al. (1987) ont adopté une série de règles qui prédit le résultat final en fonction des candidats sélectionnés les deux modalités. Cette méthode est fortement dépendante du corpus et ne peut pas être applicable à un corpus à grande taille car le nombre de règles devient très vite explosif. De nombreux chercheurs se sont intéressés à une loi multiplicative sur les probabilités de sortie des deux modalités pour fusionner les données (Stork et al., 1992 ; Prasad et al., 1993 ; Movellan et Chadderdon, 1996 ; Jourlin, 1996) ou à une pondération des deux entrées (Silsbee, 1993). Certains ont adopté la solution de fusion dans un réseau stochastique (Waibel et al, 1994 ; Duchnowski et al., 1994 ; Bregler et al., 1993a, 1993b). L'utilisation d'un coefficient fixe de pondération ne permet pas d'obtenir une intégration optimale. En effet, les résultats de Silsbee (1993) montrent que la condition AV > A et V n'est pas toujours respectée.

La plupart des études comparatives entre ces deux approches donnent de meilleurs résultats avec le modèle d'identification séparée (Robert-Ribes et al., 1996 ; Duchnowski et al., 1994 ; Alissali et al., 1996 ; Su et Silsbee, 1996). D'autres ne permettent pas de conclure à l'avantage de cette méthode par rapport à l'identification directe (Silsbee, 1994 ; Silsbee et Su, 1996). Nous avons regroupé les avantages et les inconvénients de chaque modèle d'après les études comparatives sur ces deux modèles :

1. **Identification directe**

   Avantages :

   - C'est un modèle qui est facile à implémenter. Il suffit de concaténer les indices visuels à ceux du canal acoustique pour former une seule observation audiovisuelle.

- Si la taille du corpus d'apprentissage est importante, il est possible d'obtenir automatiquement le poids relatif attribué à chaque canal dans le processus de fusion (Silsbee et Su, 1996).
- La coordination temporelle entre les modalités est conservée au cours de la fusion.

Inconvénients :

- Les systèmes fondés sur ce modèle nécessitent un corpus de plus grande taille que les systèmes utilisant le modèle à identification séparée (Jacob et Sénac, 1996).
- Pour adapter le système à différents niveaux de bruit du canal acoustique, l'apprentissage doit être effectué pour chaque niveau de dégradation (Silsbee et Su, 1996 ; Hennecke et al., 1996).
- La topologie des deux sources doit être identique.
- Le problème de déphasage n'est pas géré.

2. Identification séparée

Avantages :

- Comme chaque modalité a un réseau indépendant, l'apprentissage nécessite un corpus moins important que le modèle à identification directe.
- Il est possible de gérer l'asynchronie dans le cadre d'un mot (entre son état initial et final).
- Les deux modalités n'ont pas nécessairement la même architecture de reconnaissance.
- La pondération des deux modalités peut être obtenue en fonction de la décision dans chaque modalité, sans apprentissage préalable sur des données bruitées (Adjoudani et Benoît, 1995, 1996). Cette pondération peut même être étendue aux cas où les paramètres visuels sont bruités.
- Comme les deux sources sont traitées indépendamment, ce modèle s'approche plus des hypothèses faites sur la perception audiovisuelle chez l'homme (Robert-Ribes, 1995 ; Robert-Ribes et al., 1996 ; Massaro 1996).

Inconvénient :

- Le bloc d'intégration peut être complexe et dépendant du corpus (Petajan et al., 1987).

Afin d'identifier l'architecture de reconnaissance qui satisfait notre condition fondamentale d'intégration, à savoir AV > A et V quelque soit le niveau de dégradation acoustique, nous avons procédé à des expériences sur ces deux modèles. Dans un premier temps, nous avons implémenté le modèle d'identification directe et évalué ses performances dans une grande plage de rapport signal sur bruit. Dans un second temps, nous avons conçu trois autres systèmes de reconnaissance tous basés sur l'identification séparée fusionnant chacun les informations acoustiques et visuelles suivant un algorithme différent. La partie qui suit est entièrement consacrée à notre étude expérimentale et à l'évaluation des résultats.

3. Notre approche

3.1 Objectif

Notre approche consiste à identifier l'architecture d'intégration audiovisuelle parmi les modèles présentés dans la littérature, capable de satisfaire notre condition principale (AV > A et V) pour toutes les conditions de test. Dans cet objectif, le contexte du travail est simplifié : reconnaissance de mots isolés prononcés par un seul locuteur.

3.2 Le corpus

Le corpus est constitué de neuf répétitions de 54 mots non significatifs ayant la forme $V_1CV_2CV_1z$ dans lequel V_1, $V_2 \in$ [a, i, y]; $C \in$ [b, v, z, ʒ, R, l] prononcés par un locuteur français.

Ce corpus bien qu'il soit limité en taille, représente un intérêt particulier car il a été longuement étudié à l'ICP sur la modalité visuelle à partir des analyses géométriques (Benoît, et al., 1991 ; 1992) et des tests de perception (Benoît et al., 1994b ; Le Goff et al., 1994, Robert-Ribes, 1995). Il a aussi servi pour le développement du modèle de lèvres tridimensionnel de l'ICP (Guiard-Marigny et al., 1994 ; Adjoudani et al., 1994).

3.2.1 Les paramètres visuels

Pour constituer la base de données audiovisuelles, le corpus a été analysé grâce au poste Visage Parole de l'ICP élaboré par Lallouache (1991). 16 paramètres visuels de face et de profil ont été calculés 50 fois par seconde. La Figure 4.81 est un rappel de ces paramètres. Nous avons utilisé une version logicielle légèrement modifiée du poste visage-parole afin de calculer quelques paramètres supplémentaires (les distances d1, d2 et d3 et le paramètre Mp qui est la distance verticale entre le centre d'inertie de la réglette et le point du menton vu de profil).

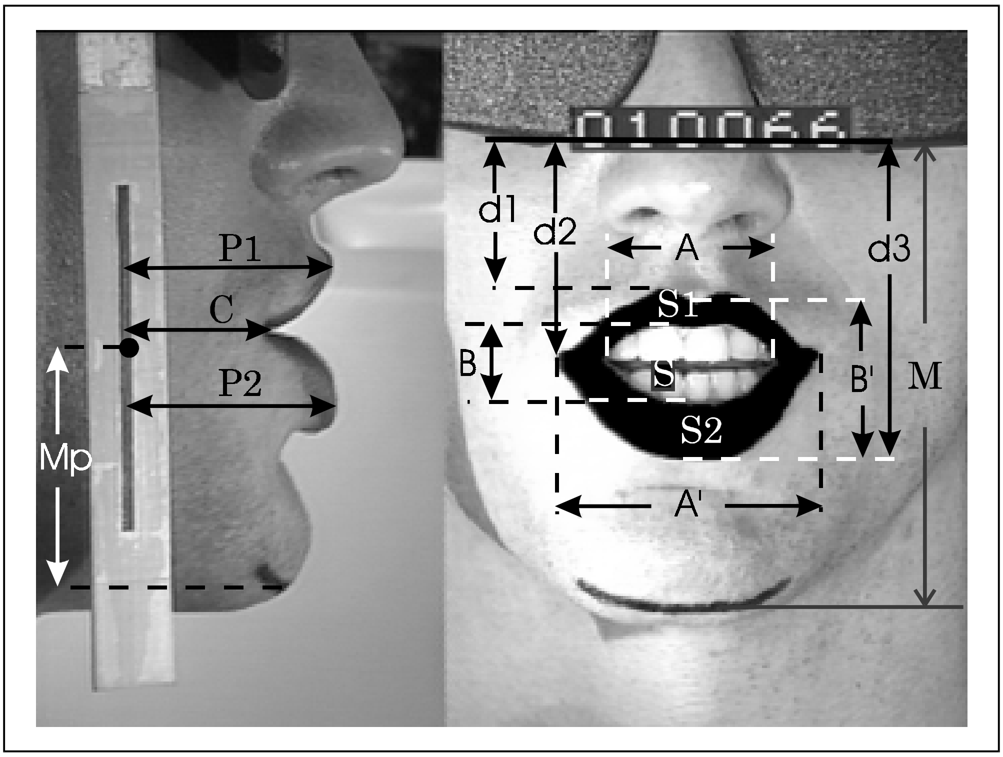


Figure 4.81 : 12 paramètres visuels obtenus avec le poste visage-parole de Lallouache (1991).

Les tableaux ci dessous montrent un exemple de résultats d'analyse sur cinq trames. Les distances sont en mm et les aires sont en mm².

Trames		A	B	S	A'	B'	S'	C	P1
8615	A	42.63	10.63	361.19	51.25	27.45	725.84	19.71	35.86
8615	B	40.13	6.08	195.69	51.26	23.11	711.80	20.43	36.21
8616	A	0.00	0.00	0.00	51.67	17.72	683.59	32.87	36.09
8616	B	0.00	0.00	0.00	52.03	17.56	626.82	34.20	36.29
8617	A	0.00	0.00	0.00	52.01	16.26	630.46	34.15	36.24

Trames		P1	P2	S1	S2	d1	d2	d3	Mp
8615	A	36.30	89.80	283.71	442.13	30.54	42.59	60.07	53.56
8615	B	37.44	88.03	286.99	424.81	31.81	42.49	57.07	53.61
8616	A	38.17	85.68	327.55	356.05	32.81	42.47	52.56	53.24
8616	B	39.18	84.55	331.92	294.90	32.67	42.59	51.12	53.07
8617	A	39.34	84.44	336.66	293.80	32.43	42.60	51.00	52.66

Tableau 4.11 : exemple de résultats d'analyse des 16 paramètres géométriques.

On note que certains de ces paramètres sont redondants, par exemple S' (l'aire totale des lèvres) = S1 (l'aire de la lèvre supérieure) + S2 (l'aire de la lèvre inférieure).

Afin d'éliminer les paramètres erronés dus à une fausse détection de contour, nous avons procédé à la visualisation de l'ensemble de ces 486 (9×54) fichiers de paramètres et de leur évolution au cours du temps. La Figure 4.82 montre un exemple d'évolution temporelle des paramètres A, B, A', B' et C pour les stimuli ABA et UBA.

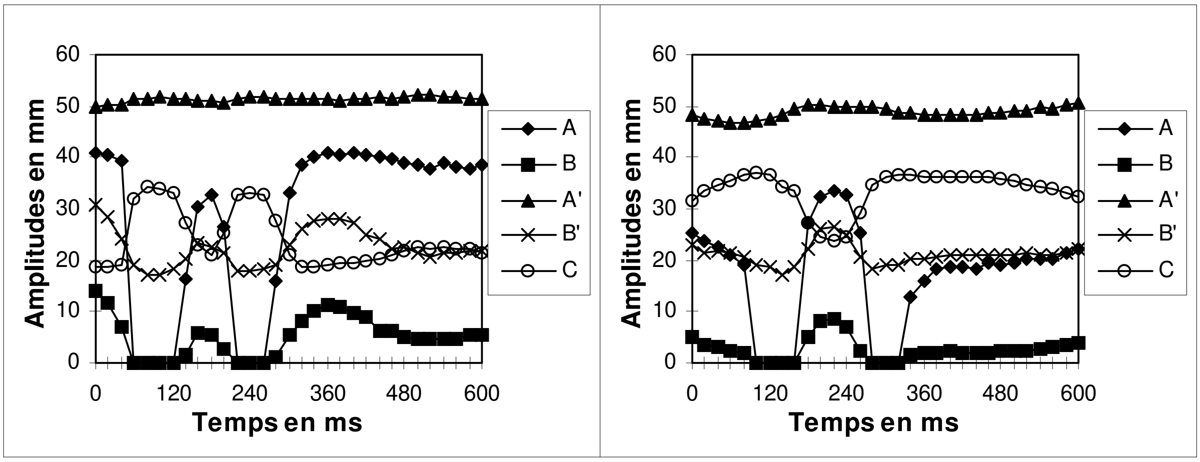


Figure 4.82 : évolution temporelle des paramètres A, B, A' , B' et C pour les stimuli ABA (à gauche) et UBA (à droite).

Les erreurs de détection ont été ainsi corrigées en utilisant le système manuel de mesure du poste visage-parole et un logiciel dédié à cette application.

3.2.2 Données acoustiques

Pour chaque fichier de paramètres visuels, le fichier audio correspondant a été numérisé sur 16 bits et sur un seul canal en utilisant une fréquence d'échantillonnage de 16 Khz. Cette configuration a permis d'avoir une bonne qualité acoustique et une bande passante assez large.

3.2.3 Synchronisation audio/vidéo

Pour la conception d'un système de reconnaissance à base du modèle d'identification directe, il est indispensable d'avoir des données acoustiques et visuelles parfaitement synchronisées. Pour ce faire, le système de traitement de Lallouache (1991) possède un dispositif qui permet de synchroniser les fichiers de paramètres avec les fichiers sonores avec une erreur maximale de 20 ms (voir chapitre 3, page 71). Ceci nous a permis de construire une base de données audiovisuelles synchronisées à la trame près.

4. Le système stochastique

Pour l'ensemble de nos expériences, nous avons choisi, dans le cadre du projet national AMIBE (Applications Multimodales pour Interface et Bornes Evoluées), l'outil de reconnaissance automatique de la parole à base de chaînes de Markov Cachées, nommé HTK, (HTK, 1992). Ce logiciel a été créé par l'équipe Parole à l'Université de Cambridge (*Cambridge University Engineering Départment, Speech group*). Ses performances dans le domaine de l'implémentation des algorithmes sophistiqués d'apprentissage et de reconnaissance et la simplicité d'utilisation de cet outil ont justifié ce choix.

En tenant compte de la tâche de reconnaissance et de la durée moyenne de chaque mot, nous avons utilisé un modèle de Markov continu contenant neuf états émetteurs pour toutes nos expériences. Chaque mot est représenté par un modèle de Markov. Une seule gaussienne représente la distribution de probabilité de chaque état ayant une matrice de covariance diagonale.

4.1 L'analyse des données

4.1.1 Analyse des données visuelles

Pour avoir une information sur l'évolution des paramètres au cours du temps, nous avons calculé la dérivée temporelle pour chaque série de paramètres. Nous avons effectué cette opération à l'aide de la relation suivante :

Equation 4.10
$$d_t = \frac{(p_{t+1} - p_{t-1}) + 2(p_{t+2} - p_{t-2})}{10}$$

dans laquelle d_t est la dérivée temporelle et p_t le paramètre visuel à l'instant t. Sur les bords, nous avons appliqué les formules suivantes :

Equation 4.11
$$d_t = p_{t+1} - p_t , t < 2$$

Equation 4.12
$$d_t = p_t - p_{t-1} , t \geq M - 2$$

M est le nombre total de trames dans le fichier (de l'ordre de 30).

4.1.2 Analyse des données acoustiques

Le vecteur audio est constitué de 12 coefficients cepstraux sur l'échelle Mel, calculés sur des fenêtres de Hamming de 40 ms, avec un espacement de 20 ms. La première dérivée ainsi que l'énergie et sa dérivée sont calculées pour chaque trame et sont ensuite concaténées pour constituer un vecteur de 26 éléments. Pour le calcul des dérivées temporelles, les mêmes algorithmes que pour les fichiers de paramètres ont été appliqués.

Pour l'analyse des bancs de filtres Mel, nous avons appliqué la formule suivante :

Equation 4.13 $$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

Des filtres triangulaires uniformément répartis sur l'échelle de Mel ont été appliqués sur la fenêtre d'analyse. La Figure 4.83 représente un exemple de ces filtres.

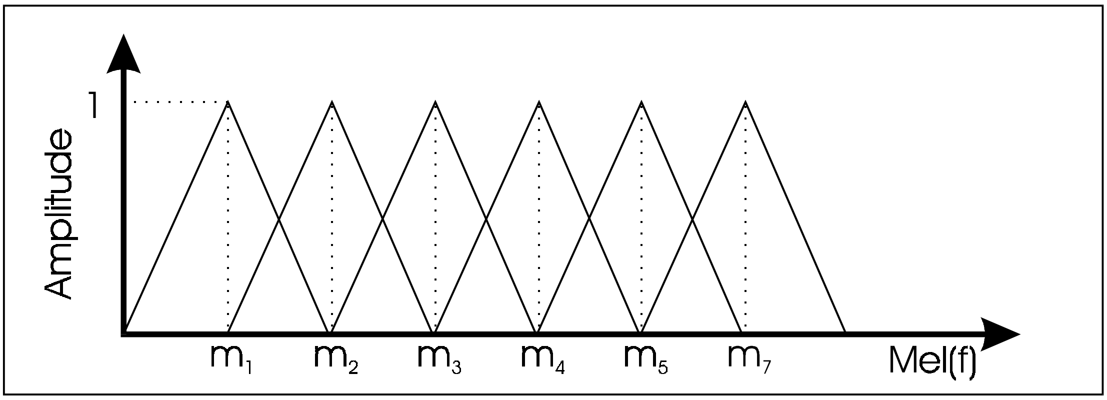


Figure 4.83 : bancs de filtres utilisés pour l'analyse des données acoustiques.

Nous avons utilisé la formule ci-dessous pour le calcul des coefficients cepstraux :

Equation 4.14 $$c_i = \sum_{j=1}^{P} m_j \cos\left(\frac{\pi i}{P}(j - 0.5)\right), 1 \leq i \leq L$$

dans laquelle m_j est la sortie du filtre j, P est l'ordre d'analyse et L est le nombre de coefficients calculés.

En tenant compte de la largeur de bande du signal acoustique et des caractéristiques spectrales de la parole, nous avons fixé l'ordre d'analyse à 24. Dans ce cas, 12 coefficients pour chaque vecteur acoustique sont suffisants pour représenter ces informations spectrales. Pour minimiser les effets de bords, une fenêtre de Hamming de 40 ms est appliquée sur les échantillons audio avant l'analyse cepstrale. La Figure 4.84 montre l'allure de cette fenêtre.

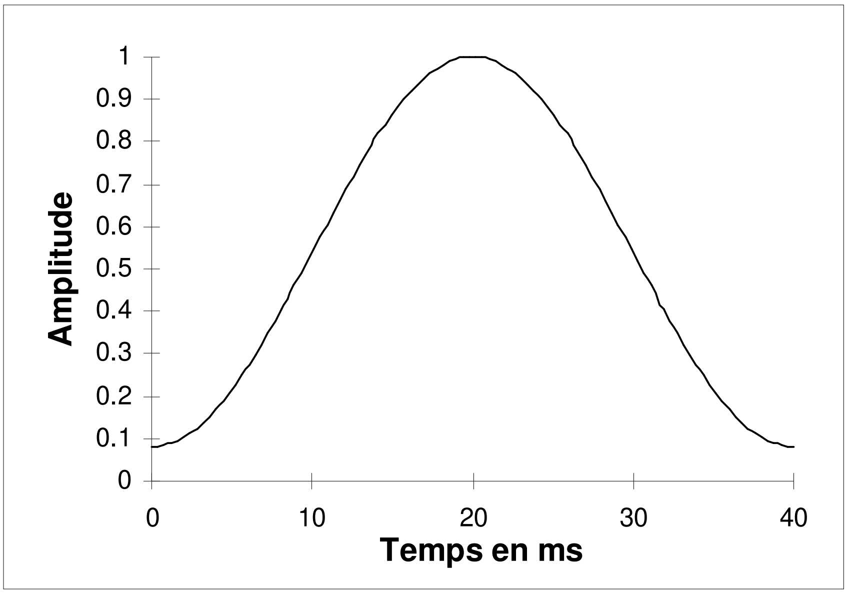


Figure 4.84 : la fenêtre de Hamming sur 40 ms.

Afin d'avoir le même nombre de vecteurs de coefficients cepstraux et de paramètres visuels pour l'implémentation du modèle d'identification directe, nous avons fixé le décalage de la fenêtre de Hamming à 20 ms, ce qui donne une analyse toutes les 20 ms. Ainsi, pour chaque seconde du signal acoustique, 50 vecteurs de coefficients cepstraux sont obtenus.

4.2 Le processus d'apprentissage

Une fois les 972 fichiers de paramètres visuels et de coefficients cepstraux préparés, l'étape suivante a été l'estimation des paramètres des modèles markoviens pour chaque mot. Pour cela, nous avons utilisé l'algorithme de Baum-Welch sur 10 itérations, dont le principe est le suivant:

le problème essentiel est d'estimer les moyennes et les variances d'un HMM dans lequel la distribution de probabilité de chaque état est représentée par une seule gaussienne, ce qui est le cas de notre étude. Cette densité de probabilité peut être calculée de la façon suivante :

Equation 4.15
$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o_t-\mu_j)' \Sigma_j^{-1}(o_t-\mu_j)}$$

dans laquelle Σ est la matrice de covariance, μ est la moyenne des vecteurs, o_t est *le* vecteur observé à l'instant *t,* et $b_j(o_t)$ est la densité de probabilité de générer le vecteur o_t à l'instant *t.*

Si un seul état existait dans la chaîne, ou si la source était unifiliaire (c'est à dire il existe une association directe entre les états et les observations), on pourrait facilement estimer la moyenne et la covariance pour chaque état *j* de la façon suivante :

Equation 4.16
$$\hat{\mu}_j = \frac{1}{T}\sum_{t=1}^{T} o_t$$

Equation 4.17
$$\hat{\Sigma}_j = \frac{1}{T}\sum_{t=1}^{T}\left(o_t - \mu_j\right)\left(o_t - \mu_j\right)'$$

où T est le nombre total d'observations.

On peut cependant utiliser ces deux équations pour calculer les valeurs initiales des paramètres du modèle. Dans un premier temps, les vecteurs d'observation sont uniformément répartis entre les différents états. Ensuite, en utilisant les équations ci-dessus, on peut calculer les valeurs initiales de la moyenne et de la variance pour chaque état. Grâce à l'algorithme de Viterbi, le chemin le plus probable, (c'est à dire la séquence d'états la plus vraisemblable) est calculé. Les vecteurs d'observation sont de nouveau attribués aux états, et la moyenne et la variance de chaque état sont calculées. Ce processus continue jusqu'à ce que les moyennes et les variances restent inchangées. En pratique, on peut fixer un seuil de tolérance au cas où l'algorithme ne converge pas.

Comme la probabilité d'avoir une séquence d'observations dépend de l'ensemble des séquences d'états possibles, chaque vecteur d'observation contribue au calcul de l'estimation des paramètres pour chaque état. Si on note *Lj(t)* la vraisemblance de se trouver à l'état *j* à l'instant *t*, on peut réécrire l'Equation 4.16 et Equation 4.17 de la façon suivante :

Equation 4.18
$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)}$$

Equation 4.19
$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t - \mu_j)(o_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)}$$

dans lesquelles la somme du dénominateur sert à normaliser ces deux valeurs.

Ces deux équations (Equation 4.18 et Equation 4.19) représentent la formule de Baum-Welch (voir Levinson et al., 1983) pour le calcul de la moyenne et de la variance du réseau de Markov. Il est nécessaire maintenant de calculer $L_j(t)$. Cela peut se faire à partir de l'algorithme de *Forward-Backward.*

Si on note $\alpha_j(t)$ la probabilité de la passe avant du modèle M ayant N états et $X = x(1), x(2), x(3) \ldots x(T)$ la séquence d'états, on a :

Equation 4.20 $$\alpha_j(t) = P(o_1, o_2, \ldots, o_t, x(t) = j | M)$$

$\alpha_j(t)$ est donc la probabilité d'observer les t premiers vecteurs en étant à l'état j à l'instant t. On peut calculer cette expression d'après la relation récursive suivante :

Equation 4.21 $$\alpha_j(t) = \left[\sum_{i=1}^{N} \alpha_i(t-1) a_{ij} \right] b_j(o_t) \quad \begin{matrix} \forall t, 1 \le t \le T \\ \forall j, 1 \le j \le N \end{matrix}$$

dans laquelle N est le nombre d'états. En effet, la probabilité d'être à l'état j à l'instant t en ayant l'observation o_t dépend de la somme des probabilités pondérées par a_{ij} pour tous les précédents états i. Les conditions initiales sont :

Equation 4.22 $$\alpha_1(1) = 1$$

Equation 4.23 $$\alpha_j(1) = a_{1j} b_j(o_1) \qquad 1 < j < N$$

La condition finale est définie par la relation suivante :

Equation 4.24 $$\alpha_N(T) = \sum_{i=1}^{N} \alpha_i(T) a_{iN}$$

La probabilité de la *passe-arrière* $\beta_j(t)$ est définie par :

Equation 4.25 $$\beta_j(t) = P(o_{t+1}, o_{t+2}, \ldots, o_T | x(t) = j, M)$$

On peut calculer $\beta_j(t)$ en utilisant la relation récursive suivante :

Equation 4.26 $$\beta_j(t) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_j(t+1)$$

avec les conditions initiales :

Equation 4.27 $$\beta_i(T) = a_{iN}$$

Equation 4.28 $$\beta_1(1) = \sum_{j=1}^{N} a_{1j} b_j(o_1) \beta_j(1)$$

On peut calculer la vraisemblance d'occupation d'un état en calculant le produit des probabilités passe avant et arrière :

Equation 4.29 $$\alpha_j(t)\beta_j(t) = P(O, x(t) = j | M)$$

La valeur de $L_j(t)$, c'est à dire la vraisemblance d'occuper l'état j à l'instant t est calculée d'après la relation suivante :

Equation 4.30 $$L_j(t) = P(x(t) = j | O, M) = \frac{P(O, x(t) = j | M)}{P(O | M)}$$

et finalement :

Equation 4.31 $$L_j(t) = \frac{\alpha_j(t)\beta_j(t)}{P(O | M)}$$

On dispose maintenant de tous les éléments nécessaires pour l'estimation des paramètres du modèle. Ces calculs sont effectués sur l'ensemble des exemples d'apprentissage.

4.3 Le processus de reconnaissance

L'algorithme le plus souvent utilisé pour la reconnaissance dans un processus Markovien est l'algorithme de Viterbi. Le modèle de tâche est défini de façon que le système décode toujours la séquence $V_1CV_2CV_1$ décrite auparavant.Cependant, la méthode de reconnaissance implémentée dans l'outil HTK (HTK, 1992) est légèrement différente par rapport à la méthode classique de Viterbi. Nous présentons ici quelques principes de cette méthode, mais avant, nous rappelons brièvement l'algorithme de calcul du chemin le plus probable de Viterbi.

L'algorithme de Viterbi ressemble beaucoup à celui du calcul passe avant à la différence près que la somme est remplacée par un calcul de maximum sur tous les chemins possibles. On peut définir la valeur $\phi_j(t)$ d'après l'Equation 4.21 :

Equation 4.32 $$\phi_j(t) = \max_i \left\{ \phi_i(t-1) a_{ij} \right\} b_j(o_t)$$

Dans laquelle $\phi_j(t)$ représente le maximum de vraisemblance d'observer les vecteurs o_1 à o_t en étant à l'état j à l'instant t. Les conditions initiales de cette relation récursive sont :

Equation 4.33 $$\phi_1(1) = 1$$

Equation 4.34 $$\phi_j(1) = a_{1j} b_j(o_1) \qquad \mathbf{1 < j < N}$$

Comme ces valeurs sont en général très faibles, il est plus pratique de travailler avec les logarithmes. Ceci évite des problèmes de précision de calcul. En prenant le logarithme des deux cotés de l'Equation 4.32, on obtient :

Equation 4.35 $$\psi_j(t) = \max_i \left\{ \psi_i(t-1) + \log(a_{ij}) \right\} + \log\left(b_j(o_t) \right)$$

dans laquelle $\psi_j(t)$ est le logarithme de $\phi_j(t)$.

Cette relation est le centre de l'algorithme de reconnaissance de Viterbi. Comme nous l'avons précisé, la méthode de reconnaissance diffère légèrement de cet algorithme. Cette méthode, livrée avec l'outil HTK, est spécialement adaptée à la reconnaissance de la parole continue. Mais nous verrons qu'elle est facilement adaptable à notre contexte de travail, c'est à dire la reconnaissance de mots isolés. Son principe est le suivant :

Imaginons que chaque état j d'un réseau de Markov possède un *jeton* à l'instant t. Ce jeton contient entre autres, une information sur le logarithme du maximum de vraisemblance d'observer les vecteurs o_1 à o_t en étant à l'état j à l'instant t. L'algorithme représenté par l'Equation 4.35 peut être ainsi transformé en *algorithme de propagation de jetons* (Young et al., 1989). On peut le résumer de la façon suivante :

1- on propage une copie de chaque jeton à l'état i vers tous les autres états j connectés en incrémentant le logarithme de probabilité de la copie par $\log(a_{ij}) + \log\left(b_j(o_t) \right)$.

2- à chaque état, on garde uniquement le jeton qui a le maximum de probabilité.

Grâce à cet algorithme, on peut concaténer plusieurs modèles les uns aux autres tout en gardant l'information sur le trajet parcouru par chaque jeton. Comme nous l'avons mentionné, cette méthode est très efficace dans le cas de reconnaissance de phonèmes où un mot est divisé en plusieurs sous-classes. Il est tout de même possible de fixer un seuil sur la probabilité de propagation des jetons pour éliminer les chemins très peu probables.

Le nombre d'états est un critère important. Une chaîne de Markov ayant un plus grand nombre d'états nécessite un ensemble d'apprentissage plus important pour une estimation acceptable des paramètres. Compte tenu de la taille de notre corpus et de la longueur moyenne des mots, nous avons choisi neuf états émetteurs, donc 11 états pour l'ensemble de nos expériences de reconnaissance. (Nous rappelons que les premiers et derniers états ne sont pas émetteurs d'observations.)

5. Expériences de reconnaissance

Dans un premier temps, nous avons procédé à l'étude des paramètres visuels les plus pertinents pour la reconnaissance audiovisuelle de la parole. Ensuite, nous avons étudié les deux principaux modèles d'intégration, c'est à dire le modèle d'Identification Directe et le modèle d'Identification Séparée. Pour le deuxième modèle, nous avons élaboré et testé différentes architectures d'intégration.

5.1 Protocole expérimental

Notre corpus est composé de 54 mots, chacun prononcé neuf fois par un seul locuteur français. Pour toutes nos expériences de reconnaissance, nous avons sélectionné sept répétitions pour l'apprentissage des modèles markoviens et les deux autres pour le test. Compte tenu de la taille limitée de ce corpus, nous avons décidé d'utiliser un principe de rotation pour les échantillons de test et d'apprentissage. Cette méthode consiste à faire une permutation circulaire entre les exemples de test et d'apprentissage pour que les résultats de reconnaissance ne soient pas biaisés par quelques échantillons du corpus.

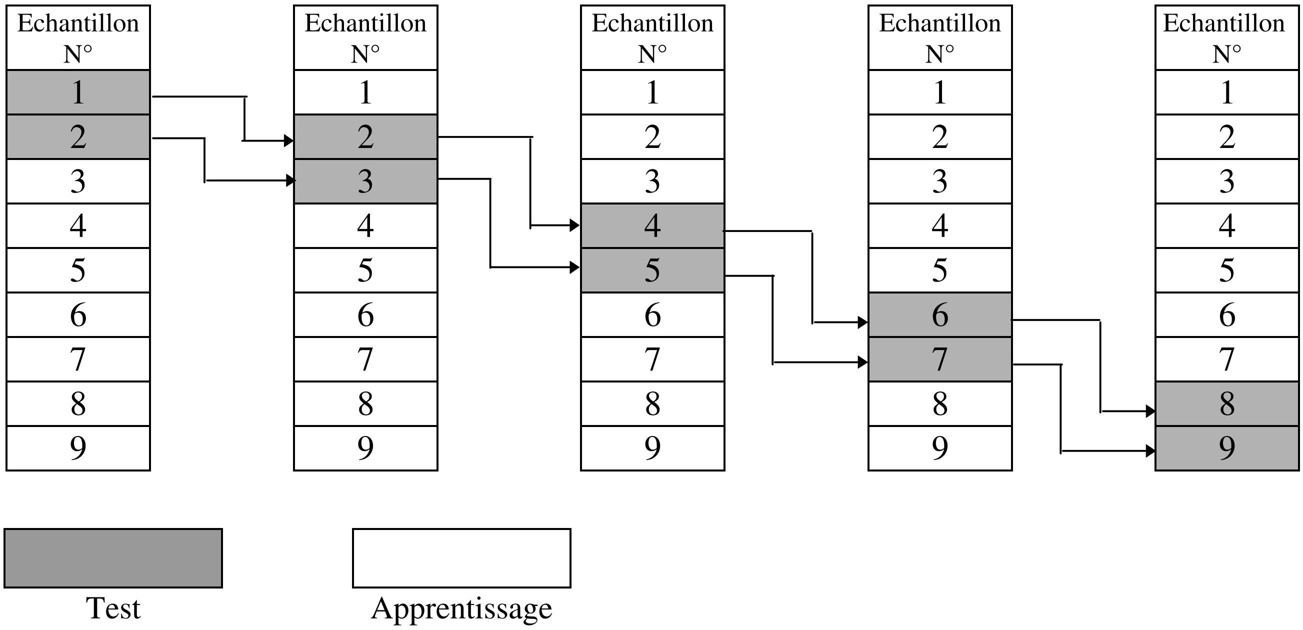


Figure 4.85 : protocole expérimental pour la sélection des échantillons de test et d'apprentissage. Les cases en fond foncé représentent les exemples de test.

Par exemple, si dans une expérience, nous avons choisi le deuxième et le troisième exemple de chaque mot dans l'ordre d'apparition et le reste des répétitions pour le processus d'apprentissage, dans la seconde expérience, le quatrième et le cinquième échantillon sont sélectionnés pour le test et le reste pour l'apprentissage et ainsi de suite. La Figure 4.85 montre le principe de la méthode de permutation des échantillons.

5.1.1 Reconnaissance auditive

Pour évaluer les performances du classifieur acoustique, nous avons procédé à l'apprentissage des modèles de Markov avec des signaux acoustiques clairs, sur 10 itérations pour chaque modèle. Nous avons ensuite testé ces modèles avec des signaux dégradés à différents niveaux de bruit blanc. Les insertions et les omissions ne sont pas autorisées. Les taux de reconnaissance chutent sensiblement en fonction de la dégradation du signal d'entrée. Nous avons reporté le résultat de cette expérience sur la Figure 4.86.

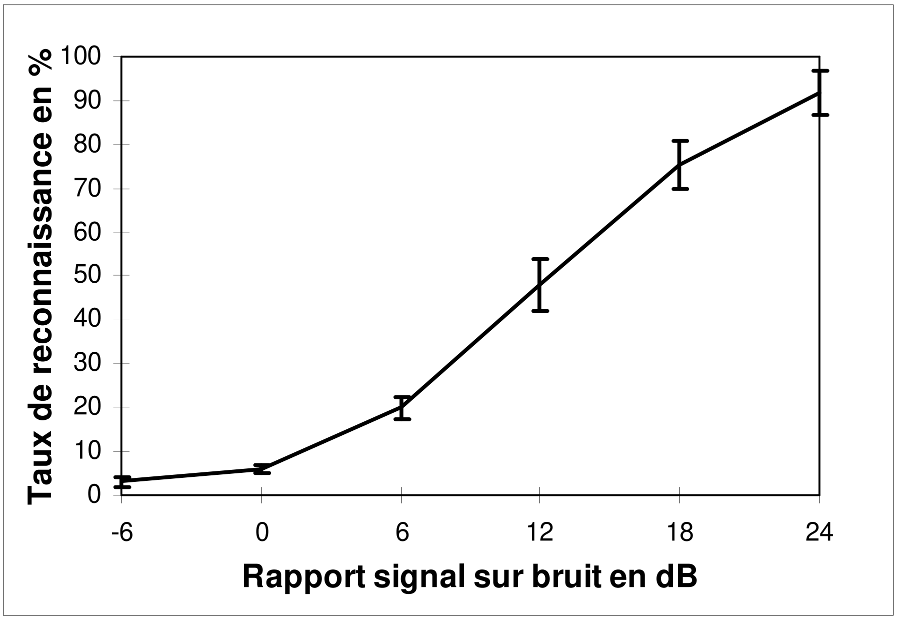


Figure 4.86 : résultats d'expérience de reconnaissance auditive seule.

Cette courbe représente la moyenne des scores obtenus à partir de la méthode de permutation des échantillons de test et d'apprentissage. Il faut noter qu'il est possible d'améliorer ces résultats en effectuant l'apprentissage sur des signaux bruités (Silsbee, 1993 ; Silsbee et Su, 1996). Mais notre objectif est de trouver une architecture de reconnaissance audiovisuelle pour avoir AV > A et V ceci quelque soit l'environnement de test. Notre système doit être capable de s'adapter automatiquement au bruit ambiant pour ajuster le poids relatif de chaque canal dans le but d'avoir une intégration optimale. En d'autres termes, quel que soit le score de chaque canal, le résultat final doit satisfaire AV > A et V. Les résultats de reconnaissance auditive nous ont donc servi uniquement de référence pour évaluer les performances du système audiovisuel.

5.2 Identification des paramètres visuels les plus pertinents

Pour concevoir un système de reconnaissance audiovisuelle performant, il est tout d'abord indispensable de connaître les paramètres visuels qui apportent le plus d'information dans la classification des mots de notre corpus. Nous avons donc procédé à de nombreuses expériences de reconnaissance utilisant uniquement les paramètres visuels avec des modèles de Markov ayant une matrice de covariance diagonale. Chaque vecteur d'observation visuel est composé de paramètres géométriques et de leur dérivée temporelle.

5.2.1 Les paramètres de face et de profil

Dans un premier temps, nous avons comparé les performances en reconnaissance visuelle seule des paramètres de face et de profil aux scores obtenus avec l'ensemble des 16 paramètres présentés auparavant. La Figure 4.87 montre les résultats des expériences utilisant la méthode de permutation des échantillons.

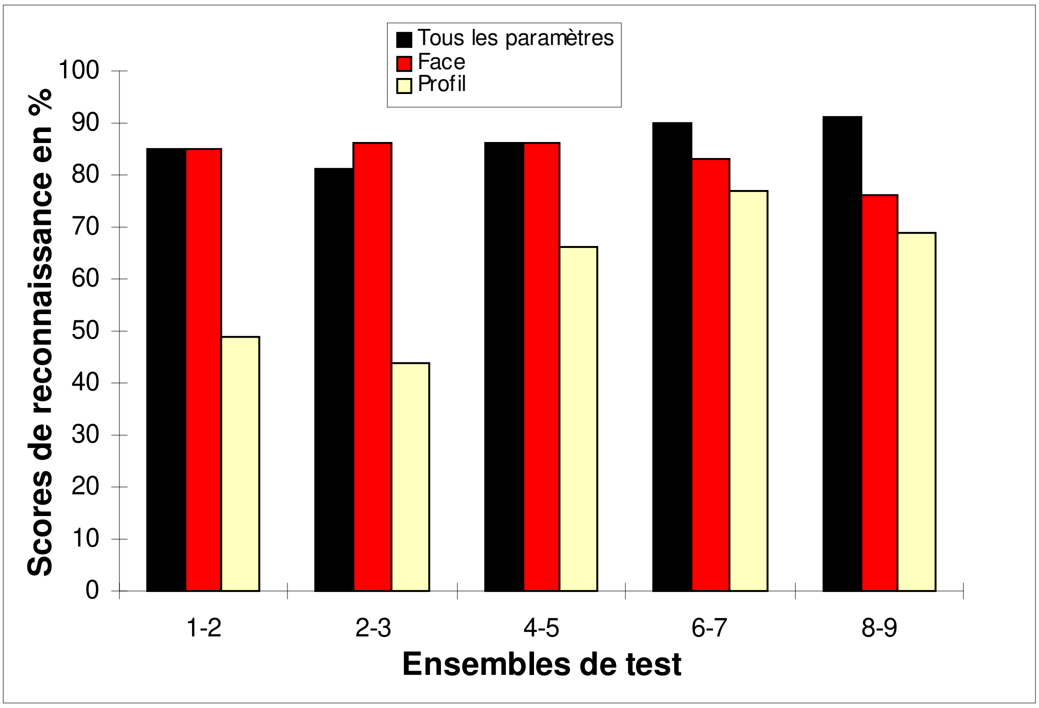


Figure 4.87 : scores de reconnaissance visuelle en utilisant différentes configurations de paramètres : les paramètres de face (83.2% en moyenne), les paramètres de profil (61% en moyenne) et ceux de face et de profil (86.6% en moyenne).

Les paramètres de face sont : A, B, S, A', B', S', S1, S2, M, d1, d2, d3.

Les paramètres de profil sont : C, P1 et P2 et Mp.

Il est évident qu'en utilisant l'ensemble des paramètres de face et de profil, on obtient globalement de meilleurs scores de reconnaissance qu'avec la configuration face seule ou profil seul. On peut noter la grande plage de variation des scores de profil (de l'ordre de 30%) suivant l'ensemble apprentissage-test. En effet, les paramètres de profil sont moins stables et peuvent présenter des variations rapides à cause du traitement séparé sur des trames non-entrelacées. Par ailleurs, le nombre de ces paramètres est nettement inférieur à ceux de face, ce qui rend le score final plus sensible au bruit de mesure.

Pour mieux analyser ces résultats, nous avons calculé la moyenne et l'écart-type sur tous les ensembles de test. Ils sont présentés dans le tableau ci-dessous.

	Tous les paramètres	Paramètres de face	Paramètres de profil
Taux moyen	86.6%	83.2%	61%
Ecart-type	4.04	4.21	13.95

Tableau 4.12 : moyennes et écart-types des scores de reconnaissance visuelle. Les écart-type ont été obtenus sur chaque ensemble de test-apprentissage.

Les résultats prouvent que les paramètres de profil transmettent nettement moins d'information (22% de moins en moyenne sur le taux de reconnaissance) que les paramètres de face. Ils sont en accord avec ceux obtenus par Erber (1974) et Summerfield (1987) sur la perception bimodale de la parole chez l'homme. D'autre part, on peut noter l'ecart-type important sur les scores de profil. Ceci montre l'existence d'une grande variabilité de ces paramètres parmi les ensembles de test. Le nombre réduit de ces paramètres est aussi responsable de cet effet.

5.2.2 Différentes combinaisons de paramètres de face

Pour connaître le poids relatif de chacun des paramètres de face, nous avons procédé à des expériences de reconnaissance utilisant différentes combinaisons de ces mesures. Pour chaque série de résultats, nous avons comparé les performances à celles obtenues avec tous les paramètres de face.

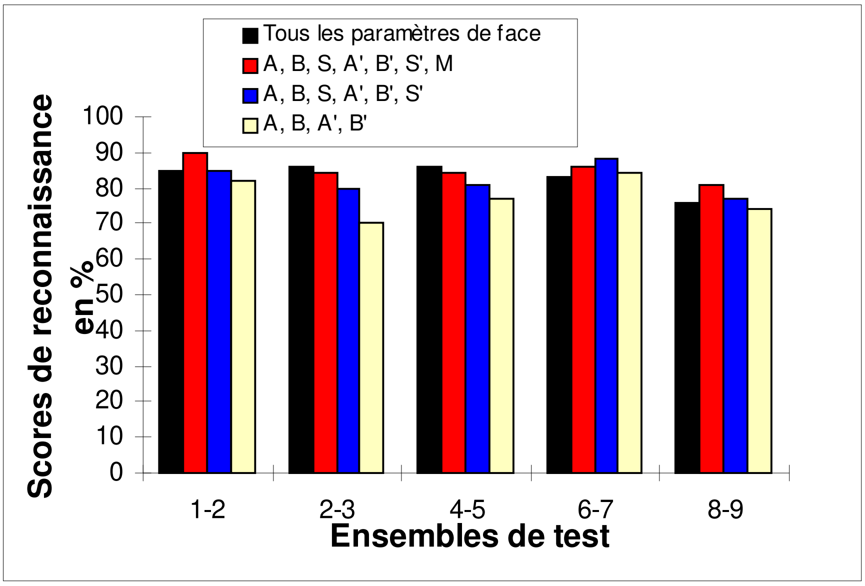


Figure 4.88 : résultats de reconnaissance visuelle avec différentes combinaisons de paramètres géométriques de face (moyenne de 83.2% pour tous les paramètres, 85% pour A, B S, A', B', S' et M', 82.2% pour A, B, S, A', B' et S'et 77.4% pour A, B, A' et B'.

La moyenne et l'écart-type de ces résultats sont reportés dans le Tableau 4.13.

	Tous les paramètres de face	A, B, S, A', B', S', M	A, B, S, A', B', S'	A, B, A', B'
Taux moyen	83.2%	85%	82.4%	77.4%
Ecart-type	4.21%	3.32%	4.04%	5.73%

Tableau 4.13 : moyennes et écart-types en pourcentage des scores de reconnaissance visuelle avec différentes combinaisons de paramètres de face.

On peut noter que le score moyen obtenu avec l'ensemble des paramètres de face est légèrement inférieur à celui obtenu avec les mêmes paramètres sans d1, d2, d3 (de l'ordre de 2%). Nous avons observé l'évolution de ces paramètres au cours du temps pour deux stimuli "ABA" et "UBU".

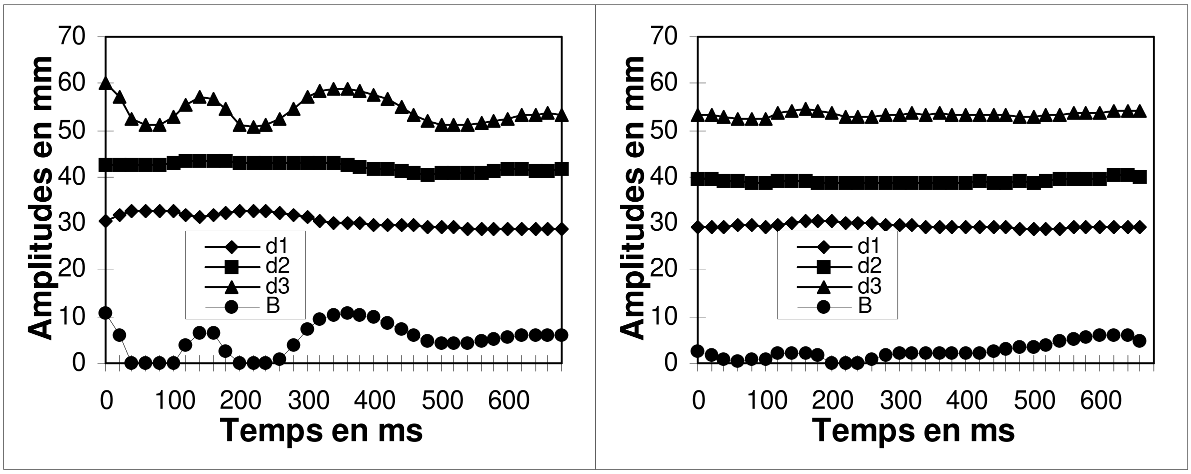


Figure 4.89 : évolution temporelle des paramètres d1, d2, d3 et B. La figure de gauche correspond au stimulus ABA, celle de droite à UBU.

Les paramètres d1 et d2 ont très peu de dynamique et ont pratiquement la même allure pour ABA et UBU alors que ces deux stimuli sont visuellement très différents. L'évolution du paramètre d3 est très proche de celle de B. En effet, comme le paramètre d1 qui mesure le mouvement du haut de l'arc de Cupidon est pratiquement constant, on peut affirmer que d3 et B sont redondants.

Il est clair que les paramètres d1 et d2 ne transmettent pas d'information de reconnaissance visuelle d'une façon significative.

5.2.3 Choix des paramètres

En tenant compte des résultats de la section précédente, nous pouvons affirmer que :

- les paramètres de face sont plus pertinents que ceux de profil,
- parmi les paramètres de face, les distances A, B, A' et B' semblent apporter une plus grande quantité d'information que les autres. Les études menées par Plant (1980) sur la perception visuelle de la parole chez l'homme ainsi que celles de Petajan (1984a, 1984b) et Finn (1986) sur la reconnaissance automatique de la parole sont en accord avec cette hypothèse.

Compte tenu de l'ensemble des résultats obtenus dans cette partie, nous avons sélectionné les paramètres A, B, S, A', B' et S'.

Le choix des paramètres S et S' peut paraître surprenant car comme on l'a vu dans le chapitre 1, ces deux valeurs sont fortement corrélés avec les produits A×B et A'×B' respectivement (Benoît et al., 1992). Cependant, les résultats du Tableau 4.13 montrent que l'ajout de ces deux paramètres permet d'obtenir un léger gain de reconnaissance (de l'ordre de 5%). D'autres chercheurs comme Robert-Ribes (1995) ou Girin et al. (1995) ont utilisé le paramètre S avec les deux paramètres A et B dans leurs expériences de perception ou de reconnaissance automatique de la parole. En outre, avec une légère modification des algorithmes de calcul des paramètres A, B, A' et B', les paramètres S et S' peuvent être facilement calculés.

Nous n'avons pas utilisé le paramètre M dans nos expériences car il nécessite un point de référence sur le menton. Il est difficilement exploitable dans une application pratique de reconnaissance audiovisuelle.

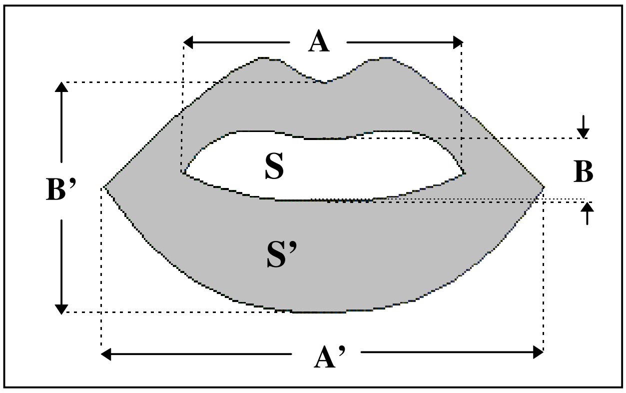


Figure 4.90 : les six paramètres visuels utilisés pour le système de reconnaissance audiovisuelle.

Tous les paramètres sélectionnés, c'est à dire A, B, S, A', B' et S' sont facilement identifiables à l'aide des algorithmes de traitement simplifiés sur les images *chroma-key*. Le schéma de la Figure 4.90 rappelle les mesures représentées par ces paramètres. Pour toutes les expériences de reconnaissance audiovisuelle, notre score de référence visuel a été celui qui a donné le plus faible résultat avec la méthode de permutation circulaire des ensembles de test et d'apprentissage, ceci afin de mieux évaluer le système et quantifier l'apport d'information des paramètres visuels. D'après les résultats de la Figure 4.88, ce score est obtenu en utilisant les échantillons 8 et 9 pour le test et les autres pour l'apprentissage. Cette répartition de fichiers de test et d'apprentissage est celle de référence dans toutes les expériences qui suivent.

5.3 Modèles d'intégration testés

Comme nous l'avons précisé, nous avons centré notre étude sur les deux principaux modèles d'intégration, c'est à dire l'identification directe et l'identification séparée. Ceci nous a permis d'évaluer nos résultats par rapport aux études similaires dans la littérature. Les sections suivantes décrivent notre démarche personnelle dans l'implémentation de ces deux modèles.

5.3.1 Modèle d'Identification Directe

En concaténant les paramètres d'analyse acoustique, c'est à dire les coefficients cepstraux et leur dérivée temporelle et l'énergie, aux six paramètres visuels des lèvres et leur dérivée, un vecteur d'observation audiovisuel est obtenu. Nous avons donc élaboré l'apprentissage de nos modèles markoviens sur ces vecteurs mixtes. La Figure 4.91 rappelle le principe de cette expérience.

Figure 4.91 : schéma de principe du modèle d'identification directe

Chaque vecteur audiovisuel est composé de 26 éléments de l'analyse acoustique (12 coefficients MFCC + dérivées + énergie) et de 12 éléments visuels (six paramètres géométriques des lèvres + leur dérivée).

5.3.1.1 Résultats

Les résultats de cette expérience sont présentés sur la Figure 4.92. La région grisée, centrée sur la moyenne, représente l'écart-type sur les scores de reconnaissance visuelle seule. Pour chaque niveau de bruit, l'écart type est représenté par des barres verticales. La courbe en pointillé est la moyenne des scores sur les combinaisons apprentissage-test. La courbe A indique les scores du classifieur acoustique seul obtenus avec la combinaison apprentissage-test de référence. La ligne V montre le score visuel obtenu avec cette combinaison.

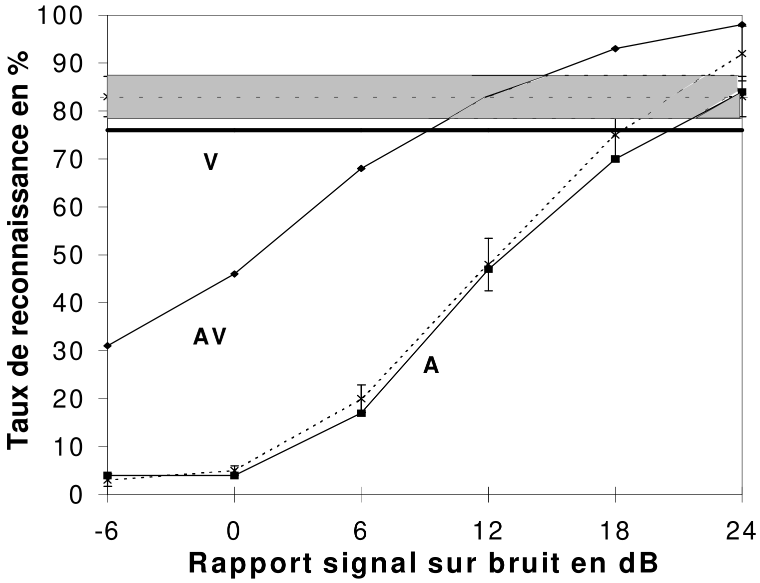


Figure 4.92 : scores de reconnaissance visuelle (V), auditive (A) et audiovisuelle (AV) pour différents niveaux de dégradation acoustique suivant l'architecture d'identification directe.

Les résultats de cette expérience montrent que les informations visuelles permettent d'améliorer les scores de reconnaissance du système audio seul pour tous les niveaux de bruit. Cependant il est clair que ce modèle ne permet pas de satisfaire notre condition majeure, c'est à dire AV > A et V pour toutes les conditions de test. On remarque que les scores AV sont inférieurs au score visuel seul quand le rapport signal sur bruit de l'entrée audio dépasse la limite inférieure de 10 dB.

5.3.2 Modèle d'Identification Séparée

Nous avons implémenté ce modèle en utilisant deux réseaux de Markov acoustique et visuel séparés. Chaque modèle de Markov a été entraîné avec des données visuelles ou acoustiques. Ainsi, les deux classifieurs sont entièrement indépendants. Pour le test, les échantillons visuels ou acoustiques sont séparément présentés à l'entrée de chaque modalité. La Figure 4.93 rappelle l'architecture de ce système.

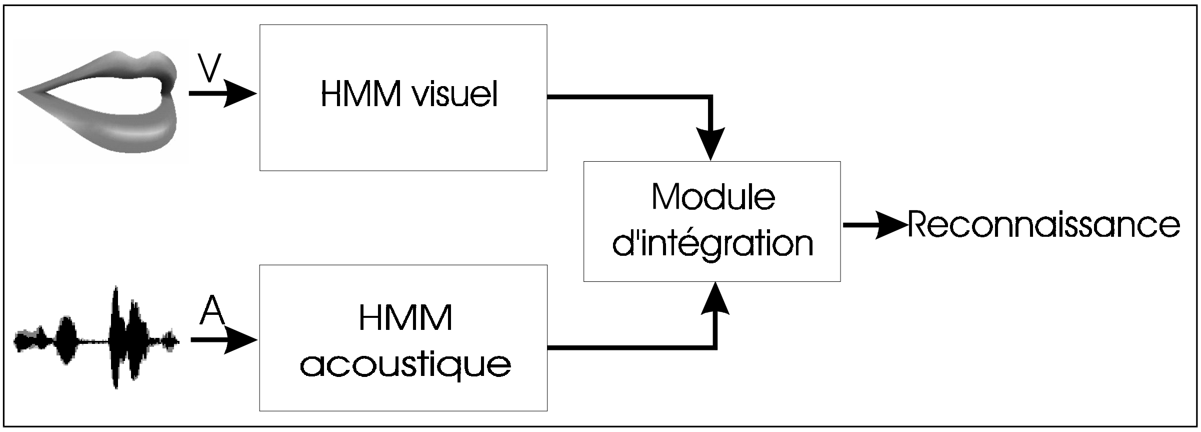


Figure 4.93 : architecture du système de reconnaissance audiovisuelle à base du modèle à identification séparée.

Le problème central de ce type d'architecture repose sur la structure du module d'intégration. Nous avons proposé trois concepts pour ce bloc. Les sections qui suivent décrivent chaque réalisation et les résultats obtenus.

5.3.2.1 Maximisation des produits croisés des probabilités de sortie

Cette première approche consiste à calculer le maximum des produits des probabilités conjointes des deux modalités. Elle a été utilisée par de nombreux chercheurs qui ont implémenté le modèle à identification séparée (Prasad et al., 1993 ; Movellan et Chadderdon, 1996 ; Stork et al., 1992). Elle consiste à sélectionner, pour chaque mot, le candidat qui maximise la vraisemblance dans les deux canaux. On peut résumer ces opérations par l'équation suivante :

Equation 4.36: $$P_{AV}(C_{best}|W) = \max_i \; P_V(C_i|W) \times P_A(C_i|W)$$

dans laquelle C_{best} représente le candidat qui maximise ce produit. $P_V(C_i|W)$ et $P_A(C_i|W)$ sont respectivement les probabilités de sortie du modèle de Markov vidéo et audio C_i , quand le mot W est l'entrée du système. Le schéma synoptique d'implantation de cette méthode d'intégration est présenté sur la Figure 4.94.

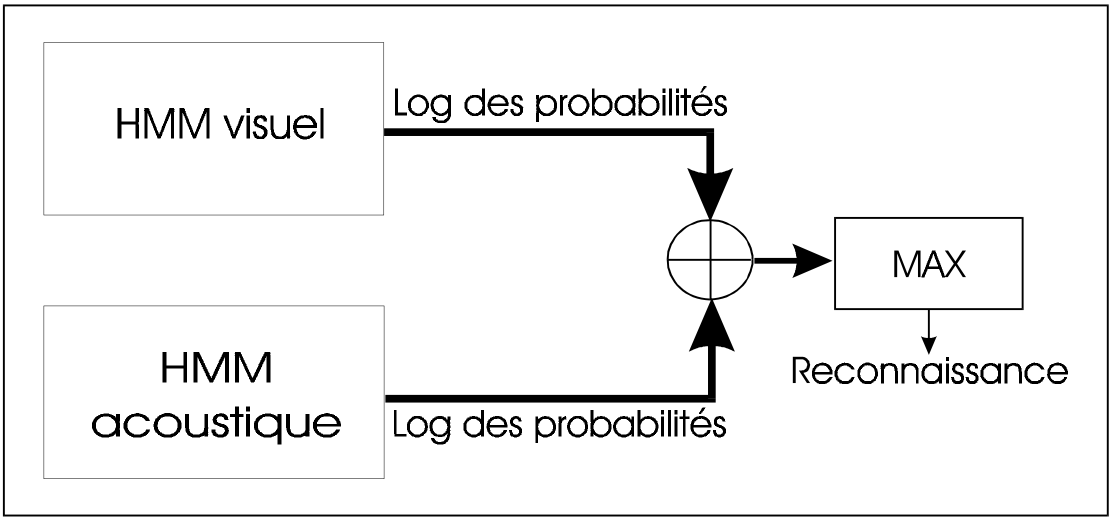


Figure 4.94 : schéma synoptique du modèle d'intégration basé sur la maximisation des produits des probabilités conjointes.

5.3.2.1.1 Résultats

Les résultats obtenus à partir de cette architecture sont présentés sur la Figure 4.95.

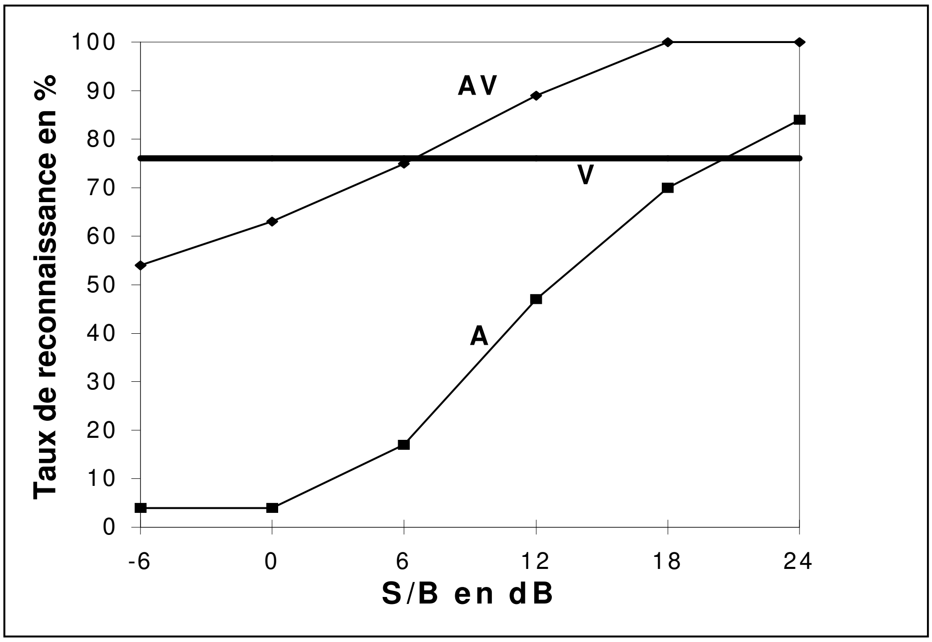


Figure 4.95 : résultats de l'intégration audiovisuelle (AV) suivant la méthode de maximisation des probabilités conjointes. Les courbes A et V représentent les scores respectivement en audio et vidéo seul.

Comme dans le cas de l'architecture de l'Identification Directe, les résultats de la reconnaissance audiovisuelle sont toujours supérieurs à ceux de la reconnaissance acoustique seule, mais cette architecture ne satisfait pas la condition principale de l'intégration audiovisuelle pour les signaux audio fortement bruités (RSB < 6 dB). Cependant, elle offre une bonne synergie quand le rapport signal sur bruit de l'entrée audio est supérieur à 12 dB.

5.3.2.2 Sélection du meilleur candidat acoustique ou visuel

Cette méthode ne peut pas être considérée comme une architecture d'intégration car elle ne permet pas une fusion des données provenant des deux canaux. Elle consiste à sélectionner une des modalités suivant son degré de certitude. Le schéma synoptique suivant résume le processus d'intégration suivant ce principe.

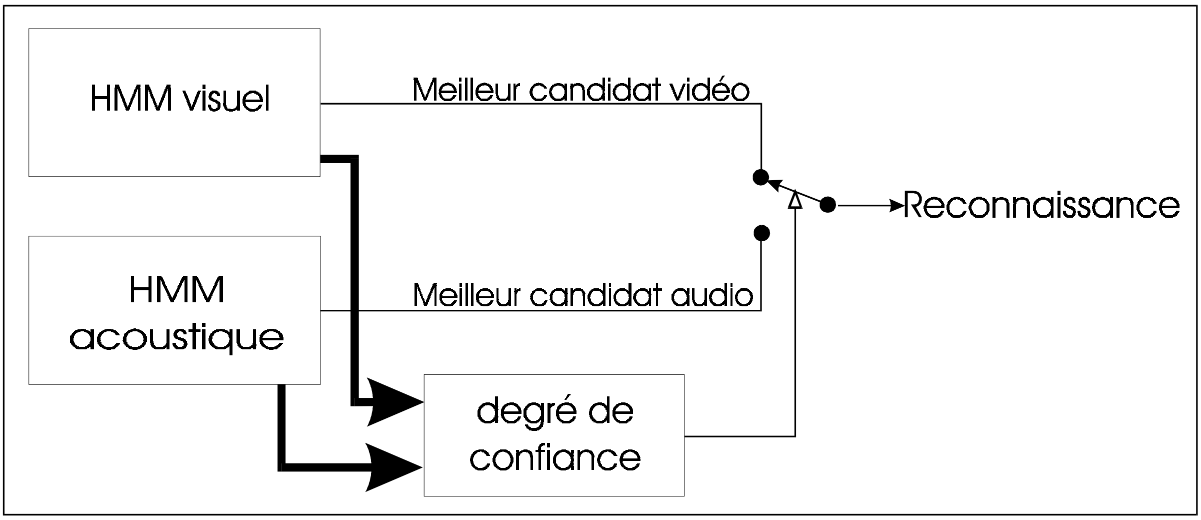
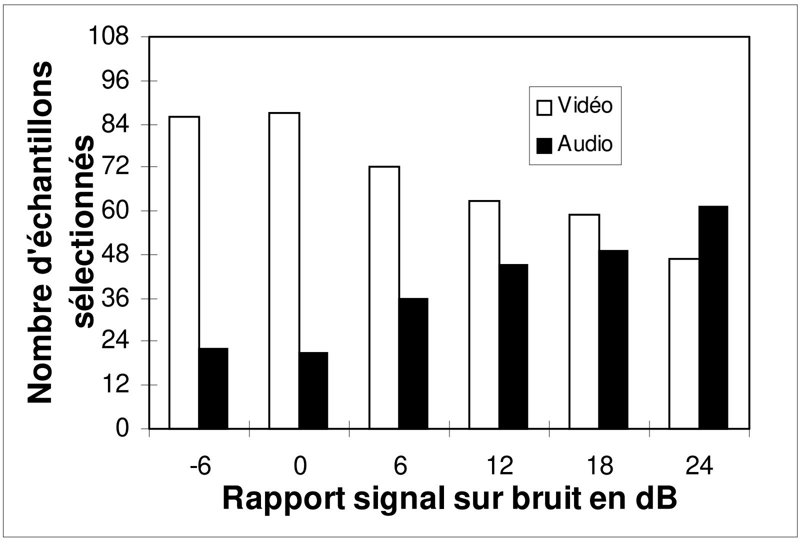


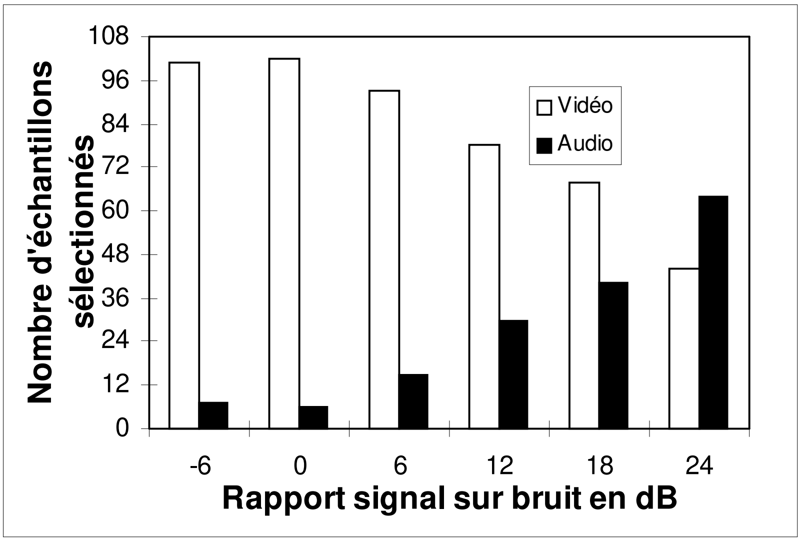
Figure 4.12 : schéma synoptique de la méthode de sélection du meilleur candidat acoustique ou visuel.

Grâce aux probabilités de sortie de chaque modèle de Markov, le degré de confiance de chaque canal est évalué. Ceci permet ensuite de commander un "interrupteur" qui sélectionne le canal ayant une plus grande certitude dans sa sélection.

Pour évaluer ce degré de confiance, on présume que la certitude de chaque modalité a un lien direct avec la différence des probabilités de sortie pour les N premiers candidats de chaque canal. Une grande différence des probabilités de sortie pour les N premiers candidats dans une modalité représente la certitude de cette modalité dans sa sélection. Pour trouver la valeur optimale de N, nous avons testé les performances de ce module pour différentes valeurs de N sur une grande plage de dégradation de l'entrée acoustique. Pour chaque condition, nous avons calculé le nombre de fois que chacun des candidats visuels ou acoustiques a été sélectionné. Nous avons reporté les résultats de ces tests pour $N = 2$ et $N = 4$.



(a)



(b)

Figure 4.96 : nombre de sélections de chacun des candidats audio et vidéo. Les résultats de la figure (a) ont été obtenus avec N=2, ceux de la figure (b) avec N=4.

La valeur N=2 n'est visiblement pas optimale pour estimer le degré de confiance du canal audio, à savoir selection des candidats audio quand signal acoustique est peu bruité, et sélection des candidats visuels en présence d'un fort niveau de bruit acoustique. En effet, pour des signaux fortement dégradés, l'estimateur de confiance attribue un poids trop important au canal audio dont les résultats sont aléatoires pour ces conditions de test (3%, voir Figure 4.86). En revanche, pour N=4, les résultats sont meilleurs et on note que le degré de confiance du canal audio est relativement bien estimé pour des signaux audio fortement bruités. Pour les signaux clairs, plus de confiance est attribuée à la décision du classifieur acoustique.

Nous avons renouvelé cette expérience avec d'autres valeurs de N. Pour notre corpus, les résultats ont montré qu'en fixant N=4 l'estimation de la certitude/incertitude était optimale.

Nous avons calculé la dispersion des probabilités de sortie du classifieur acoustique pour les quatre premiers candidats en fonction du bruit additif du signal auditif. Cette dispersion a été calculée de façon suivante :

Equation 4.37 $disp = ((prob(C_1) - prob(C_2)) + (prob(C_1) - prob(C_3)) + (prob(C_1) - prob(C_4)))$

dans laquelle $prob(C_1)$ est le logarithme de probabilité du meilleur candidat, $prob(C_2)$ est le logarithme du second candidat, etc.

La Figure 4.97 montre les résultats de ces calculs. La moyenne et l'écart-type pour chaque niveau de bruit ont été calculés sur 108 échantillons de test.

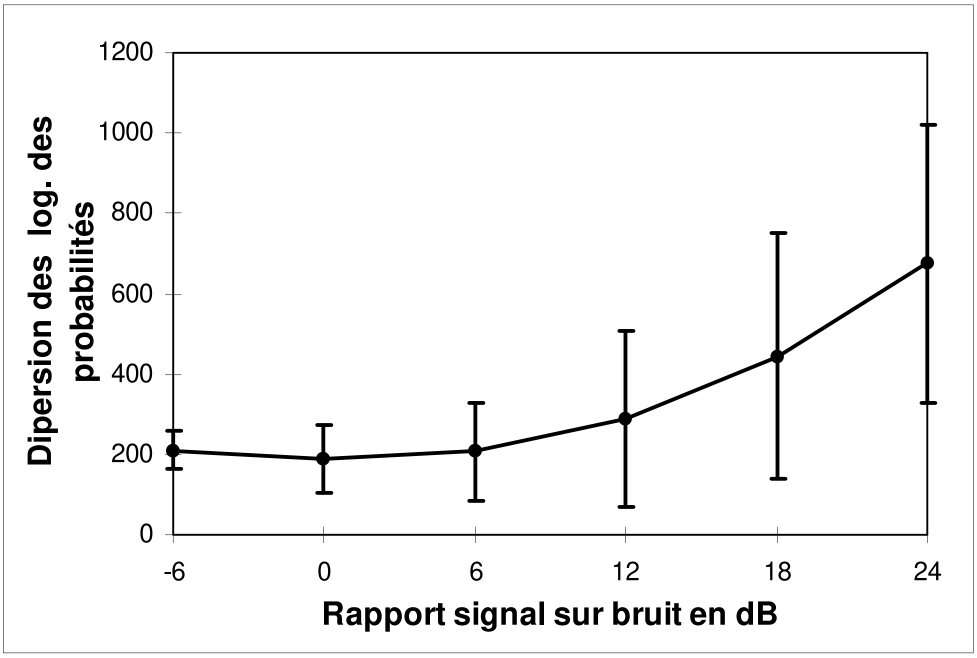
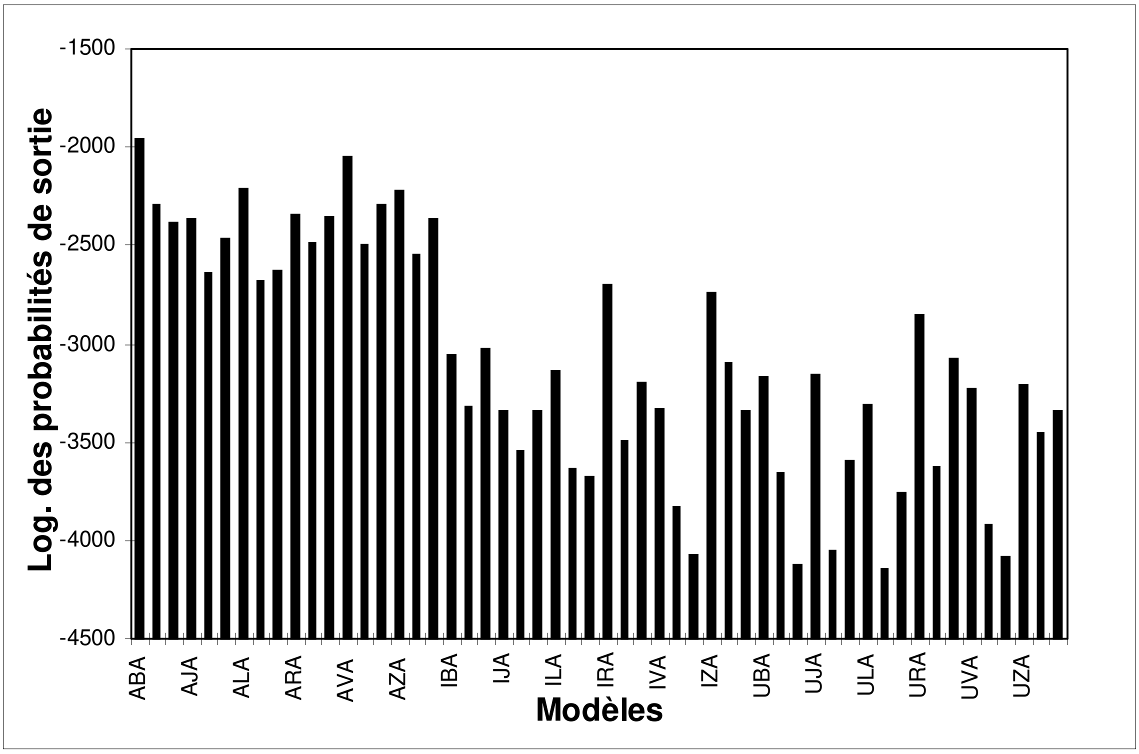
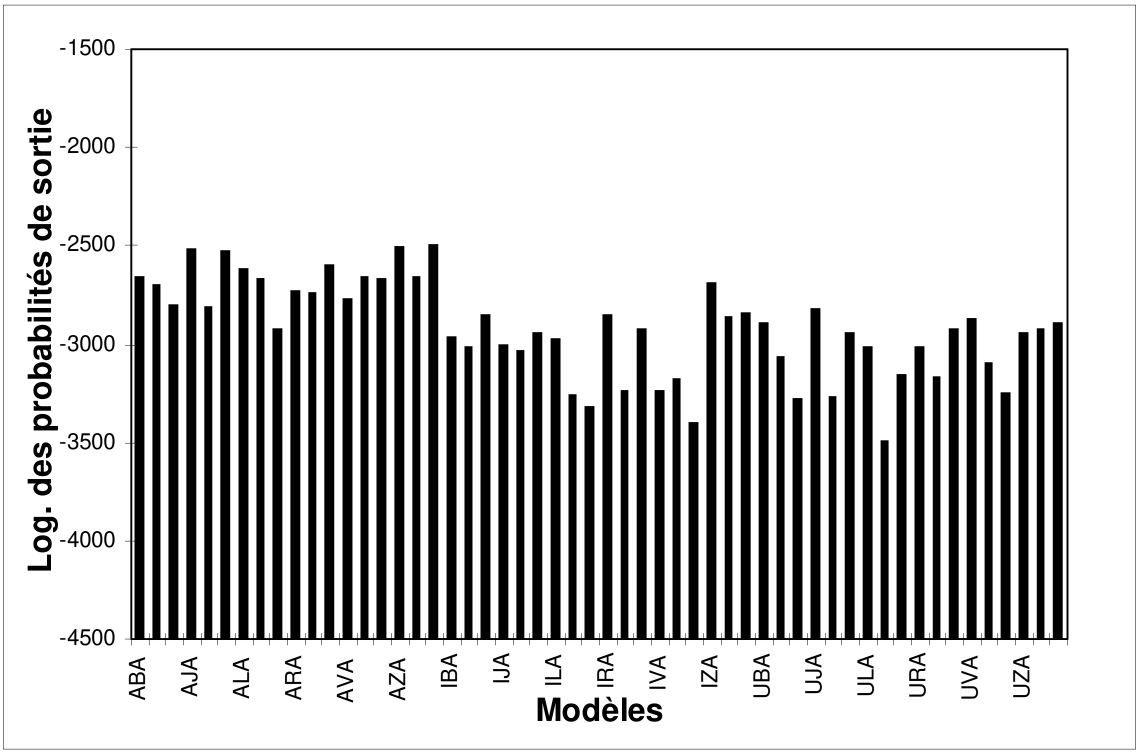


Figure 4.97 : variation de la dispersion des probabilités audio en fonction du niveau de bruit du signal acoustique.

Les résultats de la Figure 4.97 montrent que quand le rapport signal sur bruit de l'entrée audio diminue, la moyenne et l'écart type de la dispersion des quatre premiers candidats diminuent aussi. Dans de bonnes conditions acoustiques, les propriétés spectrales de chaque mot le rendent plus reconnaissable parmi les autres candidats (grande variance).



- a -



- b -

Figure 4.98 : probabilités de sortie des modèles de Markov dans deux conditions de test pour le mot ABA; pour la condition (a) l'entrée est clair et la condition (b) correspond au signal audio fortement bruité.

Par contre, dans le milieu bruité, les probabilités de sortie sont proches car le bruit élimine les caractéristiques spectrales intrinsèques à chaque mot (faible variance). Cet indice de certitude/incertitude donne le degré de similitude entre le mot d'entrée et les autres candidats.

La Figure 4.98 montre un exemple de la variation de la dispersion des probabilités de sortie pour le mot ABA dans deux conditions : audio clair et audio fortement bruité.On remarque une plus grande dispersion des probabilités de sortie quand le niveau de bruit est faible, que lorsque le signal acoustique est fortement dégradé.

5.3.2.2.1 Résultats

Les résultats de l'intégration suivant le principe de sélection du canal le plus fiable sont présentés sur la Figure 4.99.

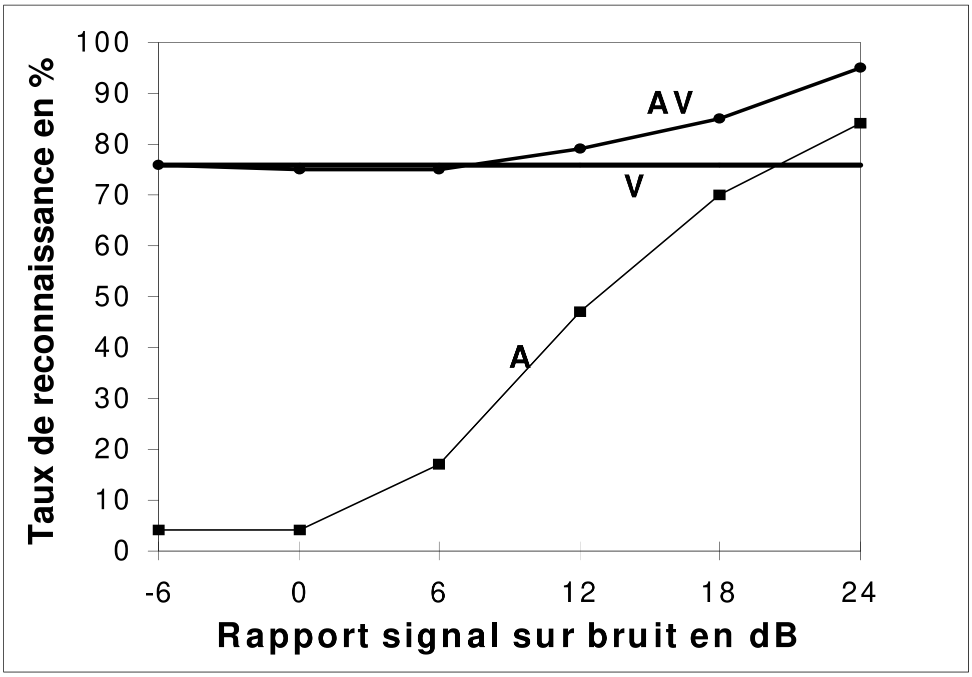


Figure 4.99 : scores de reconnaissance audiovisuelle suivant la méthode de sélection du meilleur candidat audio ou vidéo.

Ces scores montrent que notre condition principale, c'est à dire AV est supérieur ou égal à V et A, est respectée pour tous les niveaux de bruit du signal acoustique. Cependant, il est clair que cette méthode ne peut pas intégrer au mieux les informations visuelles et acoustiques puisqu'elle sélectionne le meilleur candidat de la modalité qui a un degré de certitude plus important. En revanche, ces résultats prouvent que notre démarche dans l'estimation du niveau de confiance peut être intéressante. Cependant, elle doit être implémentée dans un autre système d'intégration dans lequel les deux modalités sont fusionnées de façon à avoir une bonne synergie dans les conditions claires ou très peu dégradées. Dans ce but, nous avons développé un nouveau système d'intégration qui consiste à intégrer les informations visuelles et auditives suivant une pondération de chaque canal en fonction du niveau de confiance.

5.3.2.3 Intégration utilisant les probabilités pondérées

En utilisant l'indice de certitude/incertitude décrit dans le paragraphe précédent, nous avons développé une méthode d'intégration qui permet de pondérer chaque modalité en fonction de son degré de certitude. On peut récrire l'Equation 4.36de la façon suivante :

Equation 4.38:

$$P_{AV}\,(C_{best}|W) = \max_i \; P_A\,(C_i|W)^{\alpha} \times P_V\,(C_i|W)^{(1-\alpha)}$$

dans laquelle C_{best} est le candidat qui maximise le produit des probabilités pondérées, i le nombre de mots du vocabulaire, W l'échantillon présenté à l'entrée du système et α, le coefficient de pondération. Ce coefficient est obtenu à partir de l'Equation 4.39.

Equation 4.39:

$$\alpha = \frac{\sigma_A}{\sigma_A + \sigma_V}$$

dans laquelle σ_A et σ_V sont respectivement la dispersion des probabilités du canal audio et vidéo calculée suivant la méthode décrite précédemment.

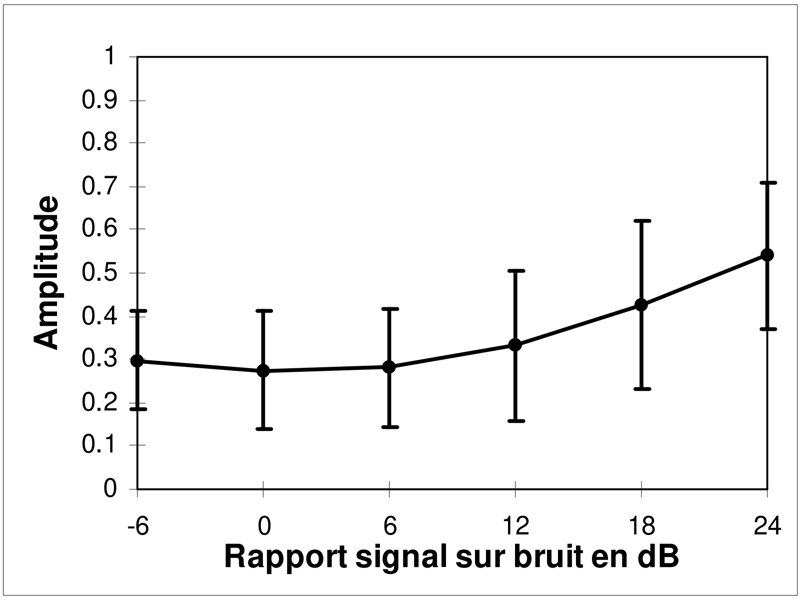


Figure 4.100 : variation du coefficient de pondération α en fonction du rapport signal sur bruit de l'entrée audio. Pour chaque condition de test, la moyenne et l'écart type ont été calculés sur les 108 échantillons de test.

La Figure 4.100 représente la variation de α en fonction du niveau de dégradation du signal d'entrée audio. La moyenne et l'écart type ont été calculés sur 108 échantillons de test. On peut noter une certaine corrélation entre α et le bruit acoustique.

Pour de grandes valeurs de RSB, la valeur moyenne de α reste proche de 0.5, ce qui signifie que le même poids est attribué aux deux canaux acoustique et visuel (Equation 4.39). Cependant, en dégradant le signal audio d'entrée, la valeur moyenne de α diminue en fonction du RSB. Cette technique est aussi opérationnelle si l'entrée des paramètres visuels est bruitée. On verra par la suite que cette propriété est très importante car les données provenant des systèmes de calcul des paramètres en temps réel sont souvent bruitées à cause des algorithmes de calcul simplifiés ou de la variation des conditions d'éclairage par exemple.

On peut représenter ce modèle d'intégration par le schéma synoptique suivant :

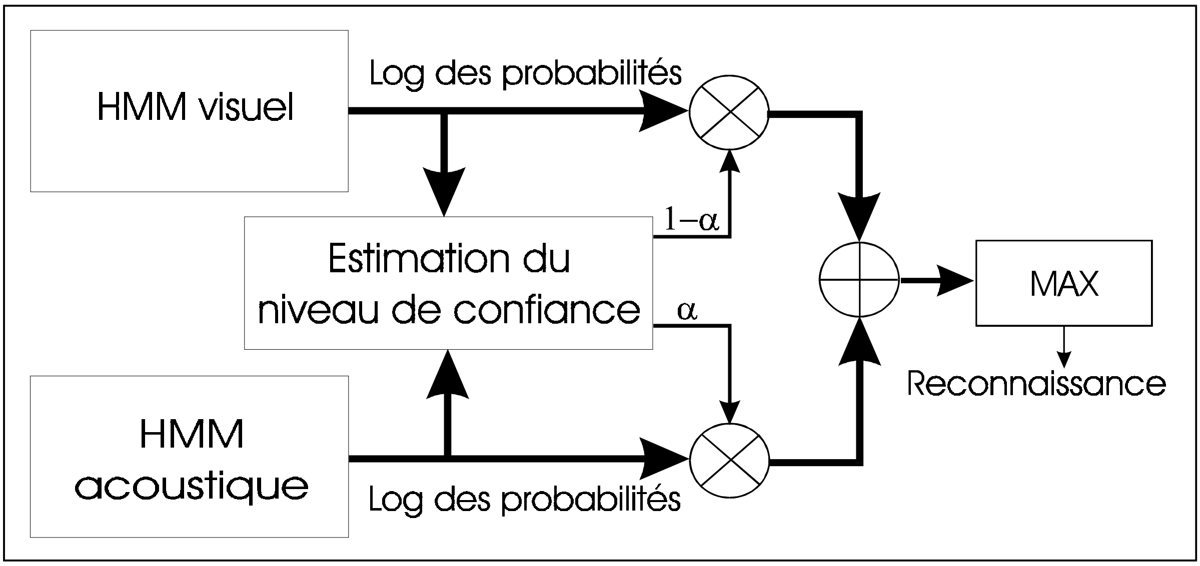


Figure 4.101 : schéma synoptique de l'architecture d'intégration audiovisuelle par pondération.

Comme dans l'architecture précédente, les probabilités de sortie de chaque canal permettent d'estimer le niveau de confiance relatif de chaque source. Le résultat de cette estimation définit le coefficient normalisé de pondération. On calcule ensuite le produit des probabiltés pondérées et le candidat qui maximise ce produit est sélectionné.

5.3.2.3.1 Résultats

La Figure 4.102 représente les résultats obtenus à partir de cette architecture utilisant la pondération automatique.

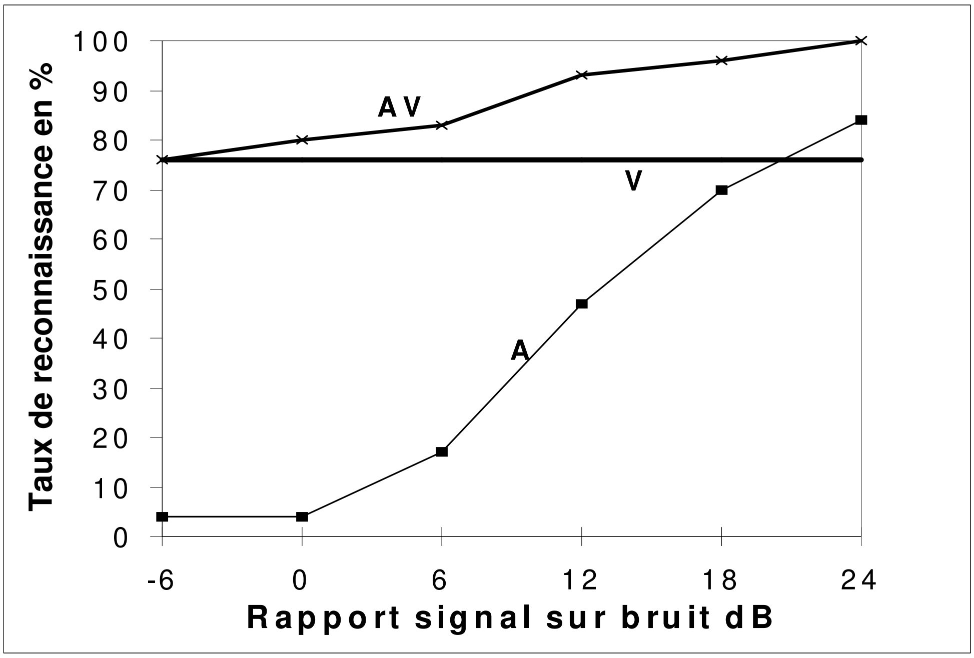


Figure 4.102 : résultats de reconnaissance audiovisuelle sur 108 échantillons de test suivant le principe de pondération des canaux suivant le niveau du bruit d'entrée.

Les résultats montrent que dans toutes les conditions de test, les scores de reconnaissance audiovisuelle sont toujours supérieurs ou égaux aux scores vidéo et aux scores audio. Cette technique permet aussi d'avoir une grande synergie quand le signal acoustique est clair ou faiblement dégradé.

5.3.3 Architectures mixtes

5.3.3.1 A + V + AV (directe)

Dans une autre expérience de reconnaissance, nous avons implémenté l'architecture d'intégration audiovisuelle mixte proposée par Robert-Ribes et al. (1996). Les modèles de Markov acoustiques et visuels permettent de traiter les deux modalités séparément, tandis que les réseaux audiovisuels les intègrent suivant l'architecture de l'identification directe. La Figure 4.103 montre le principe de ce système.

L'estimateur de confiance calcule le coefficient de pondération de chaque canal (α, β et γ) suivant la dispersion des probabilités des quatre meilleurs candidats. Le mot qui maximise le produit de ces probabilités pondérées est la solution de sortie de cette architecture. σ_A, σ_V et σ_{AV} sont respectivement les dispersions des probabilités de sortie calculées sur les quatre premiers candidats des classifieurs acoustique, visuel et audiovisuel. Comme les probabilités de sortie des réseaux markoviens dépendent du nombre de coefficients des vecteurs d'entrée, nous avons normalisé les valeurs de dispersion en fonction de ce nombre.

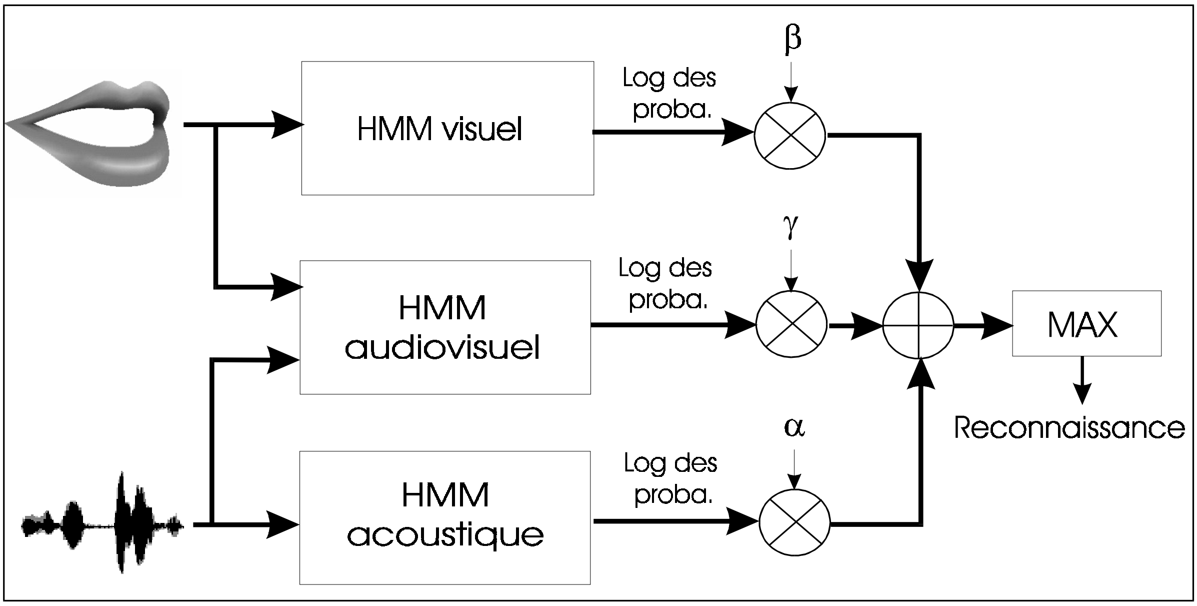


Figure 4.103 : schéma synoptique de l'implémentation de l'architecture mixte.

La variation de σ_A, σ_V et σ_{AV} en fonction du rapport signal sur bruit du signal acoustique est présentée sur la Figure 4.104.

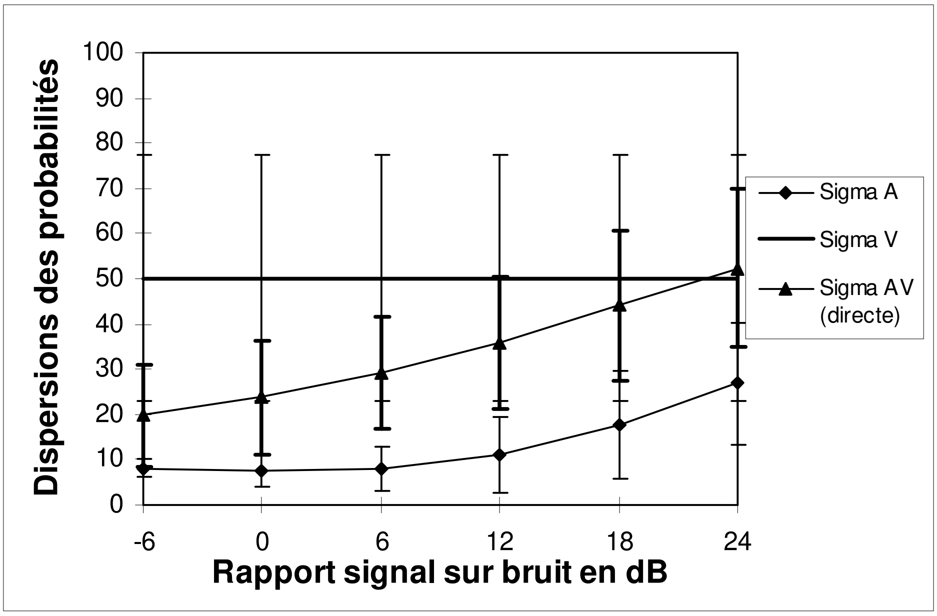


Figure 4.104 : moyennes des dispersions de probabilités calculées sur les quatre premiers candidats. Les traits verticaux représentent l'écart-type obtenu sur 108 échantillons de test.

Les équations suivantes permettent d'obtenir les coefficients de pondération de chaque canal :

Equation 4.40 $$\alpha = \frac{\sigma_A}{\sigma_{AV} + \sigma_A + \sigma_V}$$

Equation 4.41 $$\beta = \frac{\sigma_V}{\sigma_{AV} + \sigma_A + \sigma_V}$$

Equation 4.42 $$\gamma = \frac{\sigma_{AV}}{\sigma_{AV} + \sigma_A + \sigma_V}$$

La Figure 4.105 montre la variation de la moyenne de ces coeffiecients, sur 108 échantillons de test, en fonction du rapport signal sur bruit. Dans les zones à fort niveau de bruit additif sur le signal acoustique, le poids relatif du canal acoustique ainsi que celui du canal audiovisuel (à identification directe) diminuent (faible niveau de confiance de ces deux classifieurs).

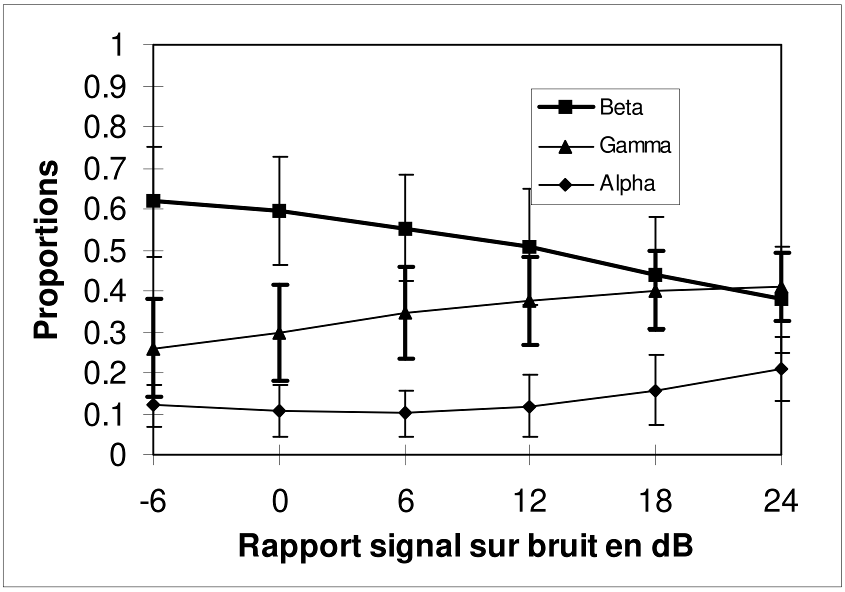


Figure 4.105 : moyennes et écart-types des coefficients de pondération des canaux acoustique, visuel et audiovisuel en fonction du bruit additif du signal audio.

Finalement, les trois canaux sont intégrés d'après la relation suivante :

Equation 4.43 $$P_{final}(C_{best}|W) = \max_i \; P_A(C_i|W)^{\alpha} \times P_V(C_i|W)^{\beta} \times P_{AV}(C_i|W)^{\gamma}$$

dans laquelle C_{best} est le candidat qui maximise le produit des probabilités pondérées, *i* le nombre de mots du vocabulaire (c'est à dire 54) et *W* l'échantillon présenté à l'entrée du système.

5.3.3.1.1 Résultats

Théoriquement, ce modèle est capable de conserver les informations sur la coordination audiovisuelle grâce au module d'intégration directe, mais aussi de traiter séparément les deux modalités avec le bloc de classification à identification séparée. Il permet ainsi d'obtenir une intégration audiovisuelle optimale (Massaro, 1996). Les résultats de cette expérience sont présentés sur la Figure 4.106.

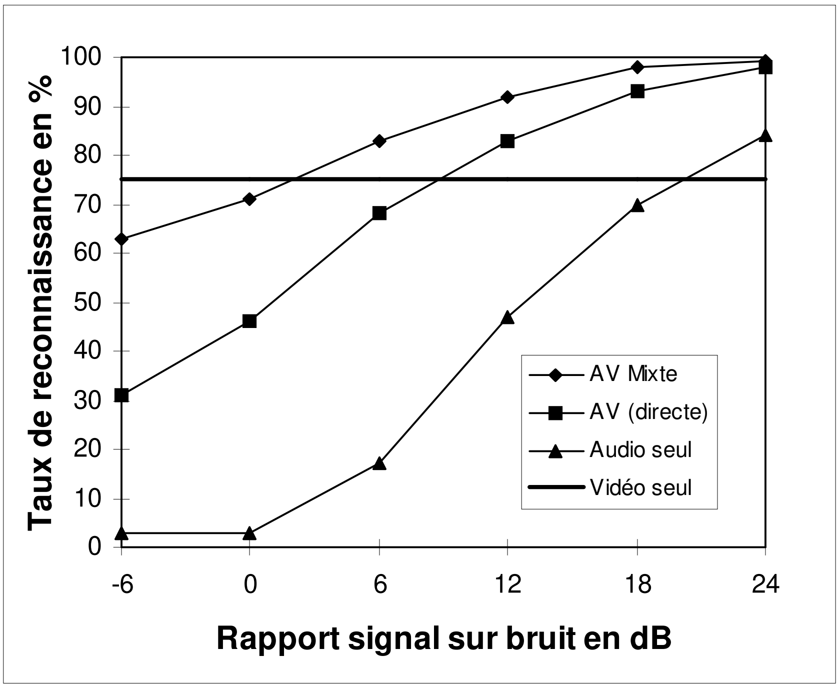


Figure 4.106 : résultats de reconnaissance obtenus à partir de l'architecture d'intégration mixte (AV mixte). La courbe AV représente les scores d'identification du modèle d'identification directe.

Les résultats obtenus montrent que ce système permet d'améliorer les scores de reconnaissance du modèle à identification directe. Cependant, la condition AV > A et V n'est pas respectée pour les zones à fort niveau de bruit acoustique (RSB < 3 dB).

Dans les zones à forte dégradation du signal acoustique, le poids attribué à la modalité visuelle seule est plus important que celui des canaux acoustique seul et audiovisuel (intégration directe). Malgré ceci, la décision finale du système hybride tel que nous l'avons implémenté est influencée par celle du canal audiovisuel dans les zones à fort niveau de bruit.

Cette expérience montre que malgré une estimation assez fiable du degré de confiance de chaque canal (acoustique, visuel et audiovisuel), notre principe de pondération des probabilités de sortie n'est pas adapté à un tel système hybride. La concaténation des informations visuelles aux informations acoustiques augmente la dispersion sur les probabilités de sortie ($\gamma > \alpha$ pour tous les niveaux de bruit). Néanmoins, dans les zones à bruit élevé, les scores finaux restent en deça des performances obtenues avec l'architecture précédente qui n'utilise que A et V.

5.3.3.2 V + AV (directe)

Nous avons implémenté une autre architecture mixte basée sur les modèles d'identification directe et séparée. Le schéma de la Figure 4.107 montre le principe de ce système.

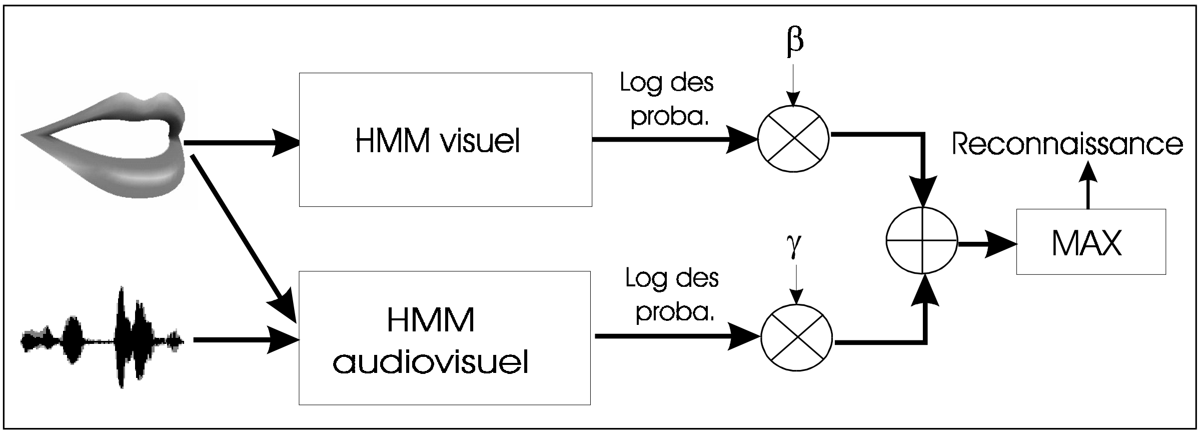


Figure 4.107 : architecture du système de reconnaissance audiovisuelle basée sur le principe de AV +V.

Dans ce système, les résultats de classification des modèles markoviens audiovisuels et visuels sont intégrés selon le modèle d'identification séparée. Comme dans l'expérience précédente, la décision finale est obtenue grâce à une pondération des sorties des deux types de classifieurs suivant notre calcul du degré de confiance de chaque canal. La dispersion des probabilités de sortie permet d'estimer le niveau de confiance de chaque canal et de pondérer la sortie de chaque classifieur :

Equation 4.44 $$\beta = \frac{\sigma_V}{\sigma_{AV} + \sigma_V}$$

Equation 4.45 $$\gamma = \frac{\sigma_{AV}}{\sigma_{AV} + \sigma_V}$$

σ_V et σ_{AV} sont respectivement les dispersions des probabilités des classifieurs visuel et audiovisuel.

La Figure 4.108 montre la variation des moyennes des coefficients β et γ, sur 108 échantillons de test, en fonction du rapport signal sur bruit. En augmentant le niveau de bruit sur le signal acoustique, le poids de pondération du canal audiovisuel diminue.

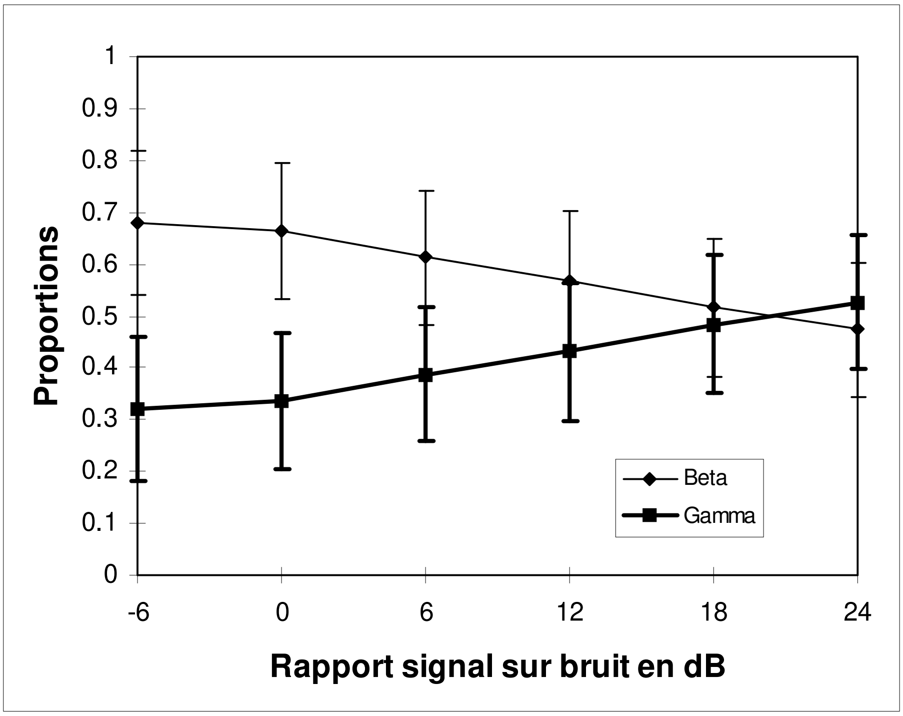


Figure 4.108 : moyennes et écart-types des coefficients de pondération des deux classifieurs audiovisuel et visuel.

L'équation suivante représente l'opération d'intégration audiovisuelle :

Equation 4.46 :
$$P_{final}(C_{best}|W) = \max_i P_V(C_i|W)^{\beta} \times P_{AV}(C_i|W)^{\gamma}$$

5.3.3.2.1 Résultats

Les résultats de cette expérience sont présentés sur la Figure 4.109.

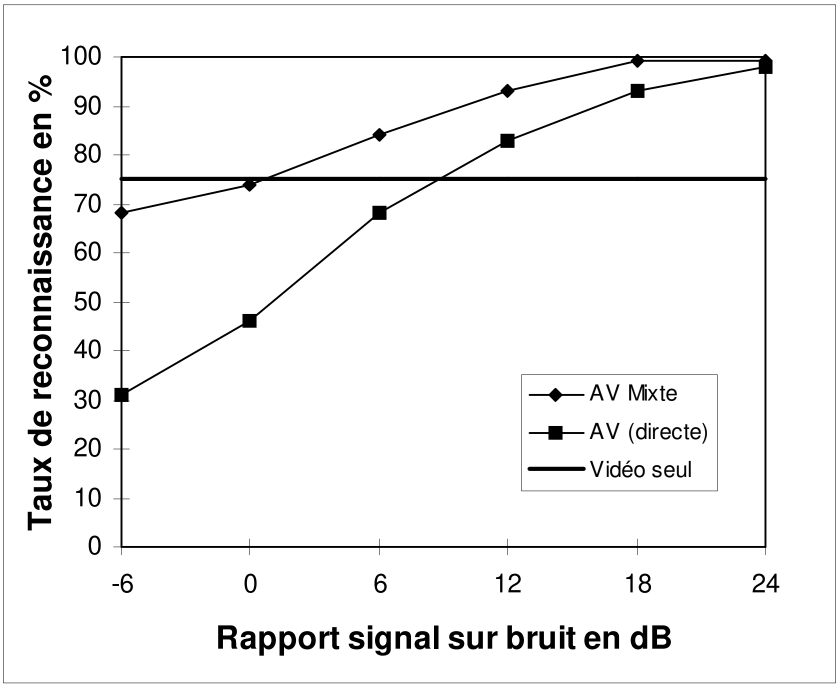


Figure 4.109 : résultats de reconnaissance du système basé sur une architecture mixte AV + V.

Ces résultats sont semblables à ceux obtenus dans l'expérience précédente. Cette architecture permet d'améliorer les scores de reconnaissance du classifieur à identification directe. Cependant, pour les forts niveaux de bruit additif sur le signal audio, la condition AV (mixte) > A et V n'est pas respectée. Comme dans le cas précédent, la pondération des probabilités de sortie suivant la loi multiplicative, telle que nous l'avons implémentée, ne donne pas de résultats satisfaisants.

5.4 Discussion

Nous avons présenté les résultats de deux principales approches d'intégration audiovisuelle. Le modèle à identification directe est capable d'améliorer d'une façon significative (33% en reconnaissance audiovisuelle contre 3% en acoustique seule pour la condition de -6 dB) les scores d'un système de reconnaissance acoustique quand le milieu est bruité. Nos résultats prouvent que malgré cette performance, l'intégration des deux sources d'information n'est pas optimale. A condition de pondérer chaque canal en fonction de son degré de confiance, le modèle à identification séparée peut donner de meilleurs résultats. Les tests montrent que ce modèle est capable de satisfaire la condition AV > A et V pour tous les niveaux de bruit du signal auditif. Par ailleurs, grâce au traitement séparé des deux sources, la complémentarité audition/vision est mieux exploitée, même si on peut regretter la non prise en compte de la coordination audiovisuelle.

Parmi les travaux que nous avons cités précédemment, le système élaboré par Silsbee (1993) est celui qui est le plus proche de notre architecture, à la différence près que Silsbee (1993) pondérait les deux canaux avec un facteur fixe qu'il avait obtenu après de nombreux tests. Dans de récents travaux (Silsbee et Su, 1996), ce facteur de pondération est estimé automatiquement pendant la phase d'apprentissage des modèles markoviens dans une architecture à identification directe. D'autres chercheurs (Waibel et al., 1994) ont utilisé un facteur de pondération dans le système de reconnaissance à base de TDNN. Ici aussi, ce facteur est obtenu pendant la phase d'apprentissage du réseau.

Notre système est capable de s'adapter aux différents niveaux de bruit acoustique pour en extraire un indice de fiabilité de chaque source afin d'intégrer celles-ci d'une façon synergétique. Notre module de pondération basé sur une méthode simple, ne nécessite pas d'apprentissage sur les données bruitées et permet d'ajuster le poids relatif de chaque source dans une architecture à identification séparée. Cependant, les résultats montrent que malgré une estimation assez pertinente du niveau de confiance de chaque canal, cette pondération n'est pas fiable dans une architecture d'intégration hybride. Afin de bénéficier des avantages du modèle hybride sur la conservation de la coordination audiovisuelle de la parole, d'autres méthodes de pondération automatique restent à être testées

6. Conclusion

Dans ce chapitre, nous avons présenté les résultats de quelques recherches sur la reconnaissance visuelle et audiovisuelle de la parole. Les études centrées sur l'identification audiovisuelle de la parole sont confrontées à un problème majeur : l'intégration des deux modalités. La plupart des chercheurs dans ce domaine tentent de résoudre ce problème par l'intermédiaire des deux modèles d'identification directe et séparée. En se basant sur les résultats de ces recherches, nous avons implémenté des architectures d'intégration basées sur ces deux modèles pour la reconnaissance de mots isolés prononcés par un seul locuteur.

Après cette étude expérimentale, nous avons procédé à la conception et la mise en oeuvre d'un système électronique d'acquisition et de traitement vidéo pour l'application pratique de ces connaissances. Ce prototype est capable de transmettre en temps réel les paramètres labiaux à un système de reconnaissance. Il est présenté dans le chapitre suivant.

Chapitre 5

Une application temps réel

1. Introduction

Le poste visage-parole développé par Lallouache (1991) permet de calculer avec une grande précision les paramètres labiaux. Cependant, il n'est pas opérationnel pour une application temps réel. Afin d'extraire les informations visuelles des mouvements des lèvres, un système de calcul des paramètres en temps réel a été développé par Mégard et Butré (1995) Nous avons encadré ces deux élèves ingénieurs, et c'est à partir des plans qu'ils ont conçus que nous avons achevé la conception et la réalisation du premier prototype de détection de contours labiaux. Une application de reconnaissance audiovisuelle ou un système d'analyse et de synthèse en temps réel, tels sont les exemples d'utilisation de ce système. Dans ce chapitre, nous présentons tout d'abord une description détaillée du système électronique LIPTRACK et ses caractéristiques techniques. Ses performances par rapport aux mesures de référence, c'est à dire celles obtenues par le poste de Lallouache (1991), seront présentées à la suite.

2. Le système électronique *LIPTRACK*

LIPTRACK est le nom de notre carte électronique. Son nom est l'acronyme du *Labiomètre, Indépendant, Programmable, Temps Réel, Avec Chroma-Key*. Nous présentons dans cette section les principales fonctionnalités techniques de ce prototype.

2.1 Objectifs

Le but de cette conception électronique est d'implémenter un système d'extraction de contours labiaux capable de calculer les paramètres labiaux (A, B, A' et B') sur des lèvres maquillées en temps réel, et cela pour une entrée vidéo standard (PAL, Y/C ou NTSC).

Dans un premier temps Mégard et Butré (1995) ont développé le schéma électronique de base du labiomètre temps réel. D'autres efforts ont ensuite été nécessaires à sa mise au point et à l'implantation des programmes en assembleur. Le cahier des charges consistait donc à réaliser une carte électronique dédiée à la labiométrie temps réel, capable de traiter des images au format PAL, Y/C ou NTSC. Les paramètres labiaux doivent être transmis en temps réel, c'est à dire 25 images entrelacées par seconde ou 50 trames par seconde pour le système PAL. La nécessité de traiter une trame à la fois (et non pas une image) est essentiellement motivée par la rapidité du mouvement des lèvres lors de la prononciation des occlusives labiales (/p/, /b/, /m/) par exemple. Dans ces cas, sur une image entrelacée, les contours sont flous et inexploitables. La Figure 5.111 montre un exemple d'image entrelacée et non-entrelacée pendant un mouvement rapide des lèvres.

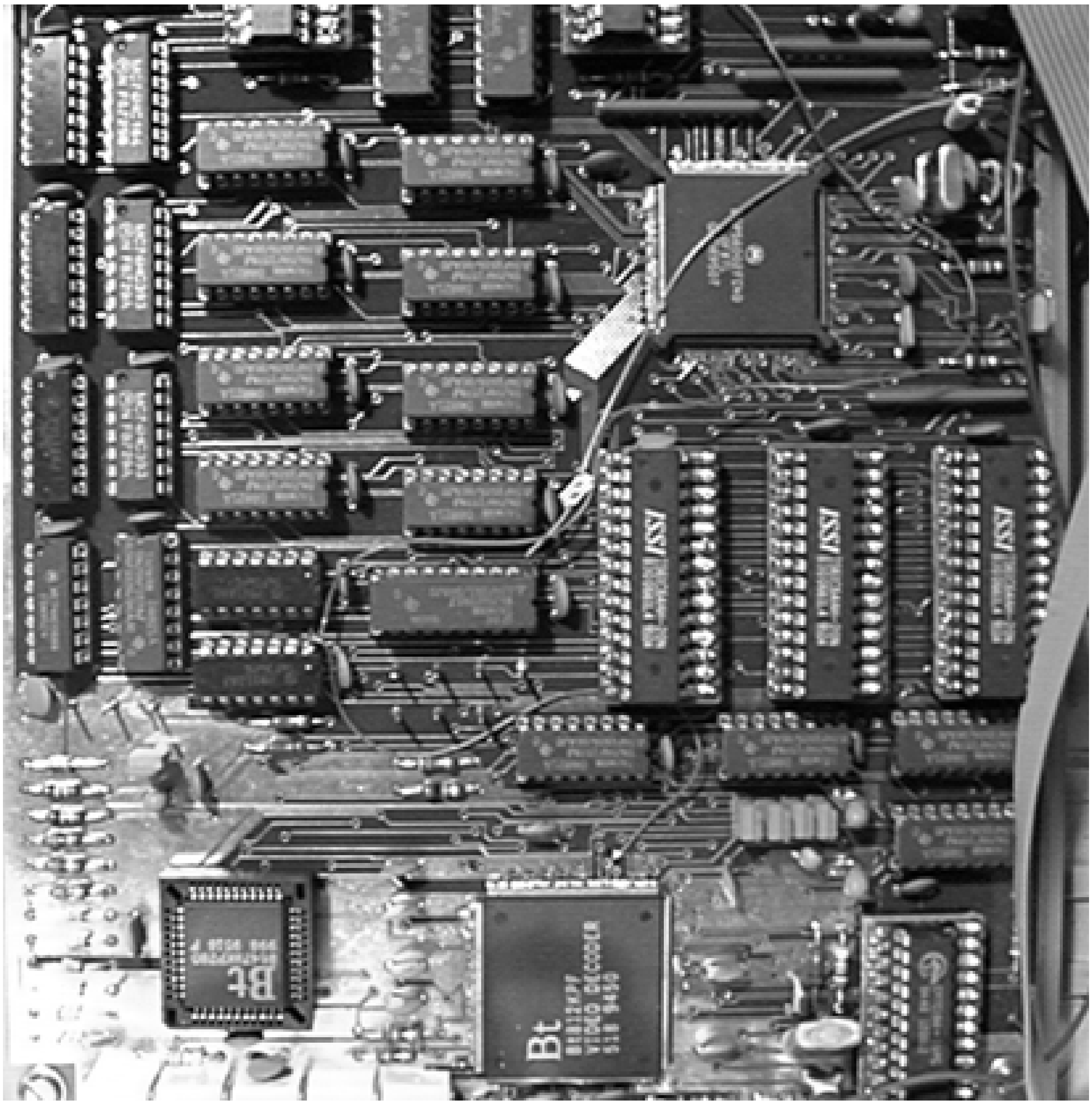


Figure 5.110 : une vue du système électronique Liptrack.

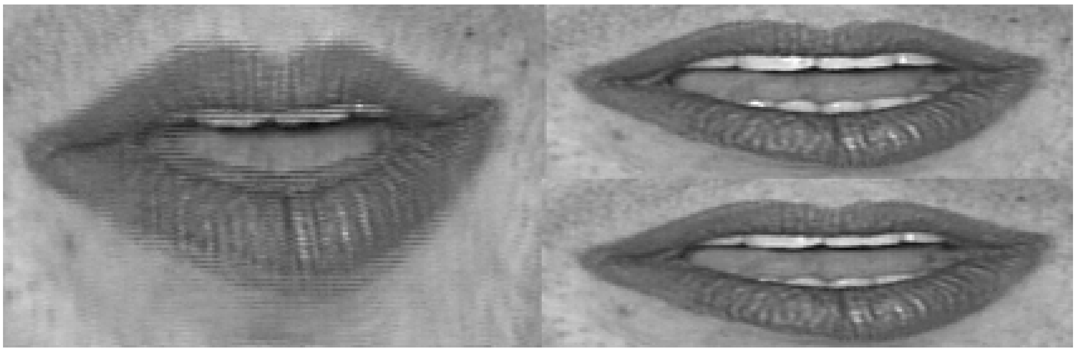


Figure 5.111 : le désentrelacement des images vidéo est nécessaire avant tout traitement sur les lèvres : à gauche une image entrelacée, à droite ses deux trames.

L'adaptabilité du système aux différents standards vidéo est un critère important car il permet de traiter des images provenant de sources variées (magnétoscope, caméra ...). Compte tenu des caractéristiques du standard Y/C, ce format répond mieux à nos exigences puisque la bande passante de luminance est plus large par rapport aux autres systèmes. En outre, les informations couleur sont codées séparément. En utilisant ce type de signal, les contours sont plus précis et l'erreur sur le calcul des paramètres est moins importante.

Comme dans le cas du poste visage-parole, les lèvres du locuteur sont maquillées en bleu afin de faciliter la détection et la poursuite des contours. Les paramètres labiaux sont ensuite calculés et envoyés vers un autre ordinateur (un PC ou une station de travail muni d'un port série RS-232). Ces paramètres servent par exemple à améliorer les scores de reconnaissance de la parole ou à animer un modèle de lèvres ou de visage en temps réel. Un exemple d'intégration de ce prototype dans une interface homme/machine est présenté sur la Figure 5.112.

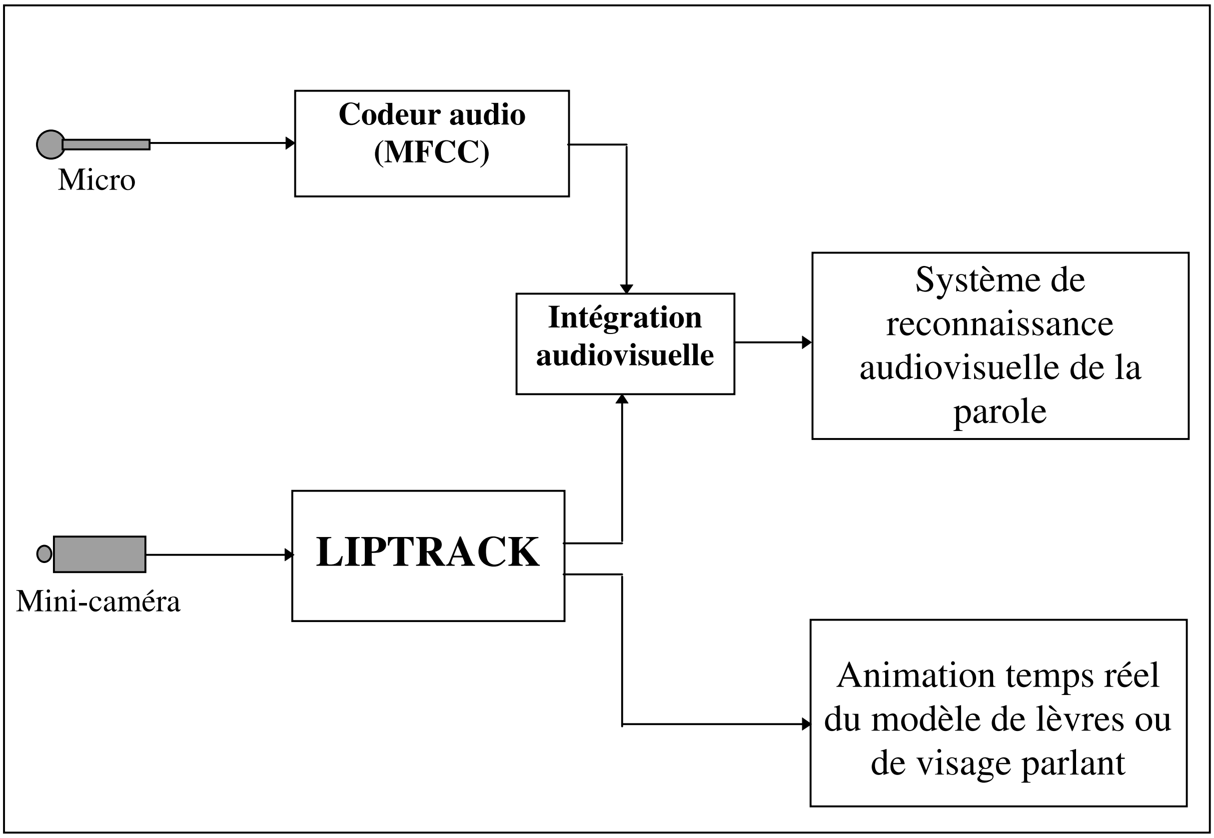


Figure 5.112 : Liptrack au centre de deux applications multimédia : la reconnaissance audiovisuelle de la parole et l'animation en temps réel d'un modèle paramétrique de lèvres ou de visage.

2.2 Principes et schéma synoptique

La Figure 5.113 représente les différents blocs du labiomètre temps réel.

Le signal d'entrée vidéo en format PAL, Y/C ou NTSC est décodé, puis numérisé sur 16 bits par le convertisseur analogique/numérique. Le signal vidéo numérique est entré dans une une table de correspondance (*LUT : Look Up Table*) qui permet de sélectionner la région bleue du maquillage (1 pour le bleu du maquillage, 0 pour le reste de l'image). Cette image binaire est ensuite stockée en mémoire ou reconvertie en signal analogique RVB par un convertisseur numérique/analogique. Ceci permet de surveiller le bon déroulement de la binarisation de la région des lèvres. Le DSP 56002 harmonise l'ensemble des opérations de transfert de données et calcule les paramètres géométriques des lèvres 50 fois par seconde pour une entrée 625 lignes.

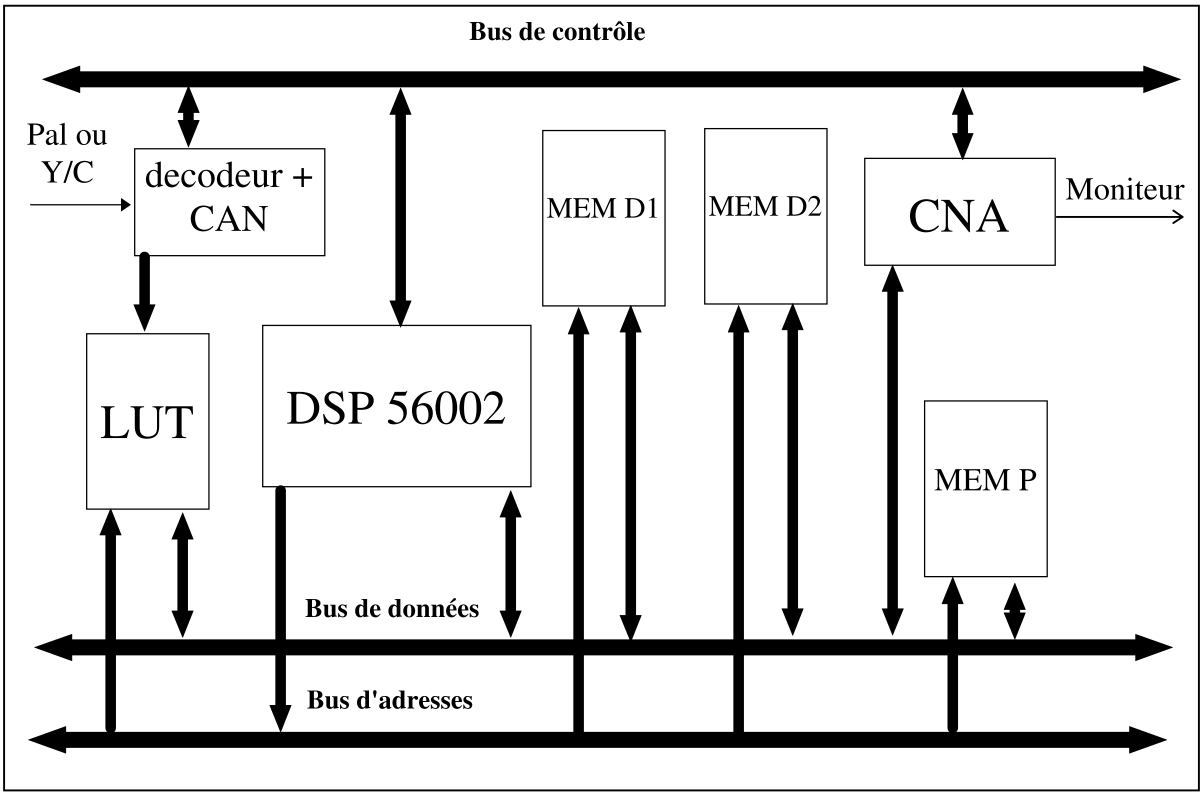


Figure 5.113 : schéma synoptique du système de labiomètrie temps réel.

2.2.1 Description des principaux composants

2.2.1.1 Le DSP

Il s'agit du DSP 56002 de Motorola. Ce composant, très utilisé en traitement du signal pour ses capacités de calcul et son nombre d'instructions, fonctionne à une fréquence interne de 40 Mhz. Ses 24 bits de données et 16 bits d'adresses permettent d'effectuer des calculs très précis et d'adresser une grande zone de mémoire vive. L'architecture interne de ce DSP est présentée sur la Figure 5.114.

Ce DSP possède une interface série RS 232 qui permet de transmettre et de recevoir des données vers et depuis le monde extérieur. Nous verrons par la suite que ce port facilite l'élaboration d'un protocole de transfert entre le PC et la carte.

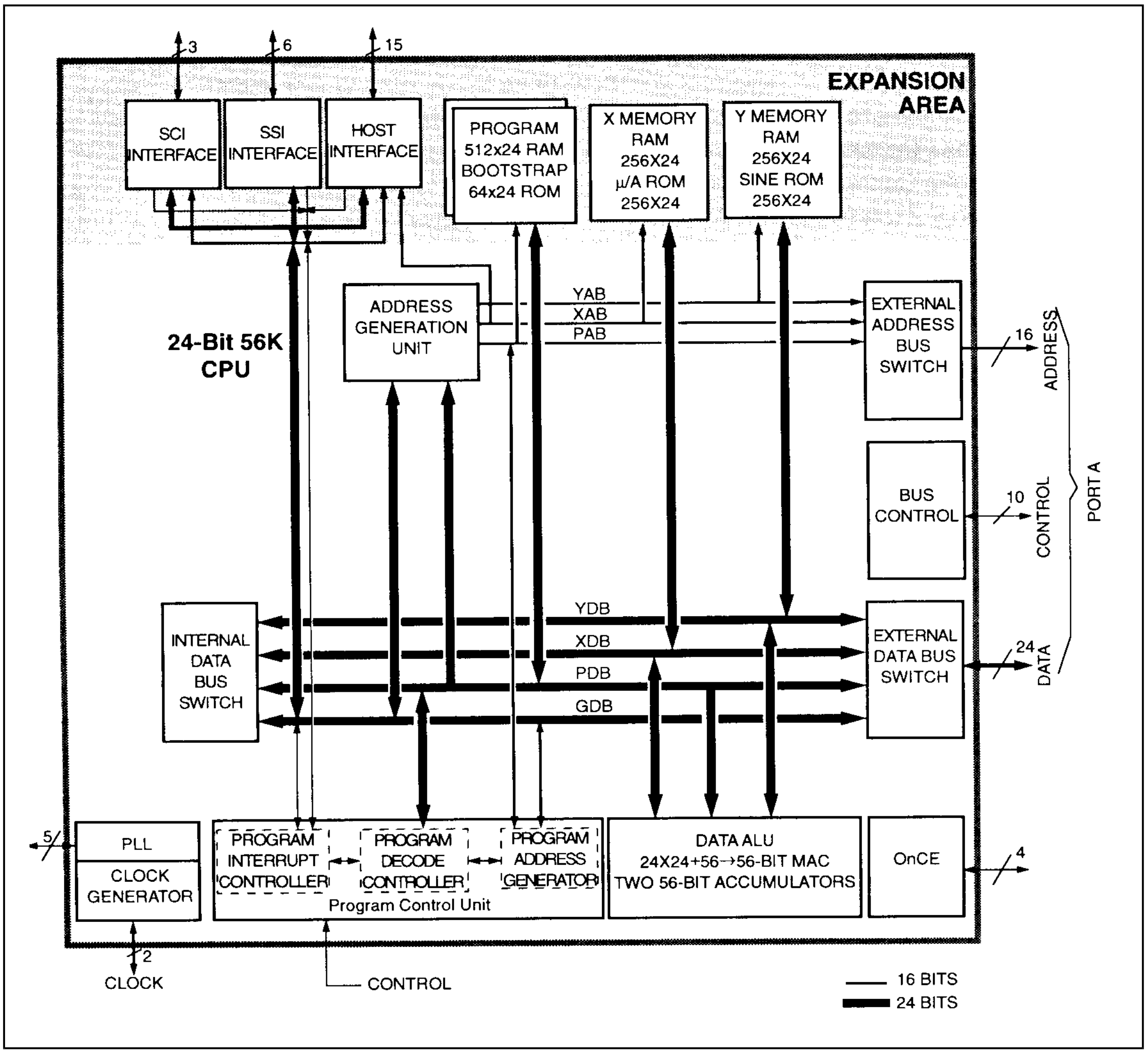


Figure 5.114 : architecture interne du DSP 56002 selon Motorola (1993).

2.2.1.2 Les convertisseurs et décodeurs vidéo

Le signal d'entrée de la carte doit être du type standard de transmission télévision afin d'intégrer un plus grand nombre d'applications. Pour cela, nous avons choisi un composant du type *Brooktree* (Bt 812) qui permet de numériser un signal PAL, NTSC ou Y/C sur 16 bits RVB ou YUV. Les registres de contrôle de ce composant qui commandent les horloges internes pour la détection des *bursts* doivent être programmés avec beaucoup de précision. Par exemple pour détecter les informations couleur d'un signal PAL, la valeur à programmer dans le registre de détection couleur est fonction de l'horloge externe à quelques Hertz près. En cas d'absence d'information couleur dans le signal, ce composant se met automatiquement en mode noir et blanc et continue à numériser les images. Il fournit aussi en sortie les différents signaux de synchronisation lignes et trames. On verra par la suite que ces signaux harmonisent l'ensemble du processus de calcul des paramètres.

Afin de surveiller le bon fonctionnement du système, le signal de sortie de la LUT est reconverti en analogique et envoyé vers un moniteur de contrôle. Pour cela, un composant du même fabricant (Bt 478) a été choisi. Il permet de reconvertir le signal binaire en un signal vidéo en utilisant les tops de synchronisation lignes et trames provenant du décodeur vidéo.

2.2.1.3 Les mémoires

Pour pouvoir effectuer des lectures et écritures rapides en mémoire image et mémoire LUT, des composants à très faible temps d'accès ont été nécessaires. Nous avons choisi des RAM 64K 4 bits fonctionnant a 15 ns pour les mémoires image ainsi que pour la LUT. En ce qui concerne les mémoires programme, des RAM 8K 8 bits à temps d'accès de l'ordre de 20 ns ont été utilisées. Ces choix ont été très stricts car des temps d'accès plus grands ne permettaient pas le fonctionnement de la carte en temps réel (stockage d'une trame en moins de 20 ms et calcul des paramètres).

2.3 Description fonctionnelle

Nous consacrons cette partie à l'étude plus approfondie des différents modules de la carte et leur principe de fonctionnement.

2.3.1 Initialisation et description logicielle

Au moment de la mise sous tension de la carte, le DSP exécute le programme pré-stocké dans sa mémoire morte et initialise le port série RS-232. Ainsi, il se met en mode attente de réception des données. Les premiers octets transmis par l'ordinateur connecté sur le port définissent la taille totale du programme et l'adresse de départ dans la mémoire programme. Les données qui suivent sont stockées à partir de l'adresse spécifiée. Quand la transmission est terminée, le DSP se met en mode exécution et initialise les convertisseurs vidéo. L'initialisation du décodeur vidéo nécessite une écriture dans 40 registres internes de ce composant. Une fois cette phase terminée, le DSP se met en mode attente de réception des données pour charger sa mémoire LUT. A la fin de la transmission de ces valeurs, le DSP autorise les interruptions provenant du décodeur-convertisseur vidéo d'entrée et exécute le programme de calcul des paramètres labiaux suivant les tops de synchronisation verticale. 50 paramètres sont ainsi transmis vers la machine hôte chaque seconde. Pour permettre de vérifier la bonne réception des données, nous avons élaboré un protocole de *Hand Shake* qui permet de garder la synchronisation et la détection d'erreurs de transmission sur la ligne. Ainsi le DSP se met en mode d'attente tant que la machine hôte ne lui a pas renvoyé l'écho de la donnée transmise. Ce protocole permet plus particulièrement d'animer des modèles de visage de synthèse très lourds en calcul (Le Goff, 1993 ; Portel, 1996) sans perdre la synchronisation. Dans ce cas, le DSP transmet les données suivant le rythme qui lui est imposé par la machine hôte. On peut résumer le processus d'initialisation de la carte sur l'organigramme suivant :

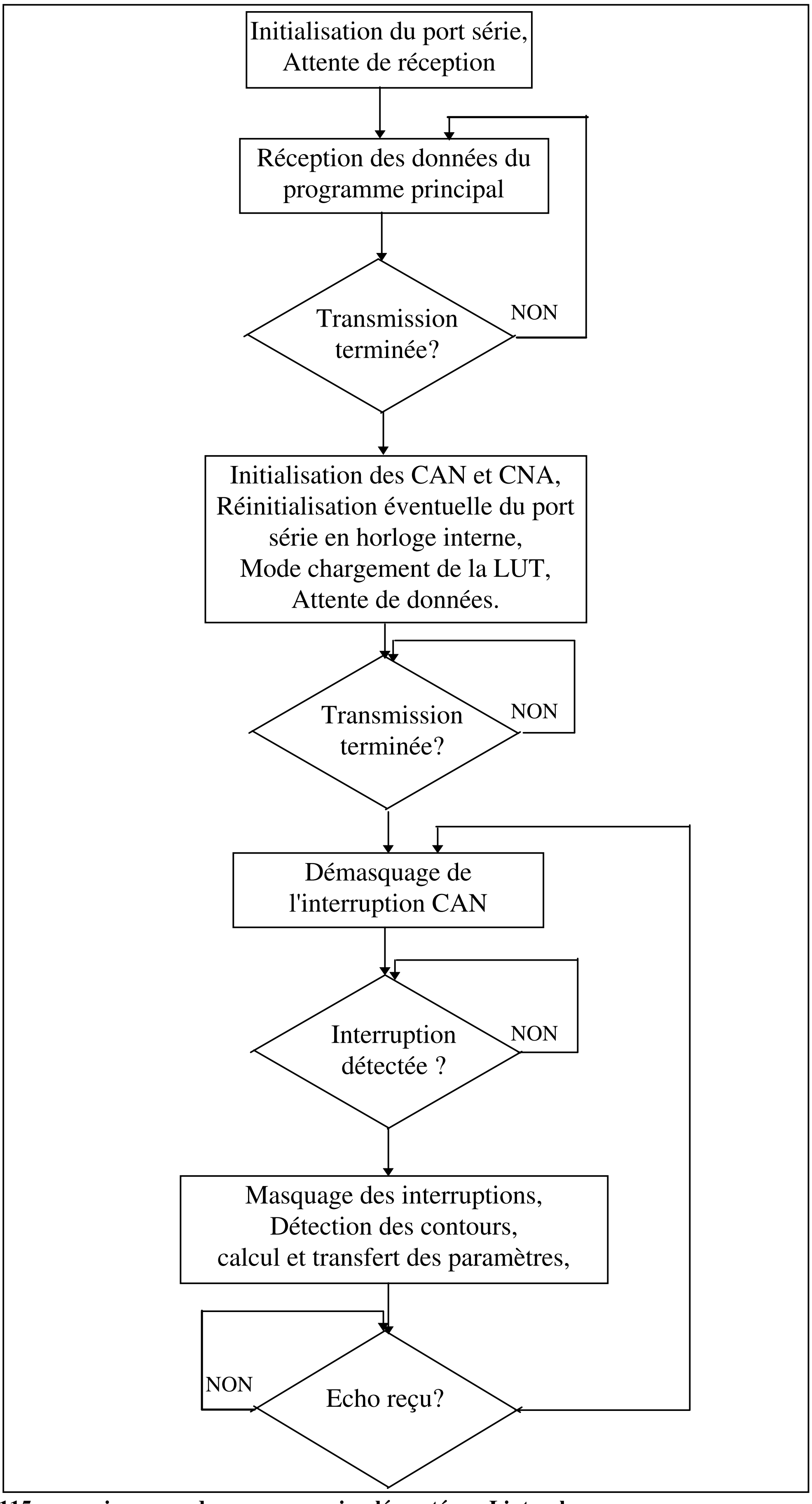


Figure 5.115 : organigramme du programme implémenté sur Liptrack.

En remplaçant les mémoires vives de programme par des EPROM, le DSP peut être directement initialisé. Dans ce cas, la première phase d'initialisation du port série ainsi que la transmission des codes de programme ne sont plus nécessaires.

2.3.2 Le Chroma-Key numérique

A la suite des travaux d'Angola (1993) pour la détection des contours labiaux sur des images *chroma-key*, Bouffartigue (1994) et Mégard (1995) ont implémenté des algorithmes de poursuite de contours des lèvres sur des images couleur. Ils ont étudié les différentes transformations pour un seuillage chromatique optimal. Mégard (1995) a centré ses recherches sur la détection du vermillon par l'intermédiaire d'une LUT. Elle a tout d'abord vérifié l'effet de seuillage chromatique dans l'espace RVB. Comme on l'a vu précédemment, cet espace est une représentation très simplifiée des couleurs. Pour appliquer un seuillage des lèvres, il suffit de vérifier, pour chaque pixel de l'image, si sa composante B est supérieure à un certain seuil et si les autres composantes sont inférieures à un seuil. Cette technique, malgré sa simplicité, ne donne pas de résultats satisfaisants. Mégard (1995) a montré qu'en utilisant cette méthode, certains pixels appartenant au visage ou aux dents sont classés parmi les pixels du vermillon. La Figure 5.116 montre la représentation des pixels des lèvres et du visage dans l'espace RVB.

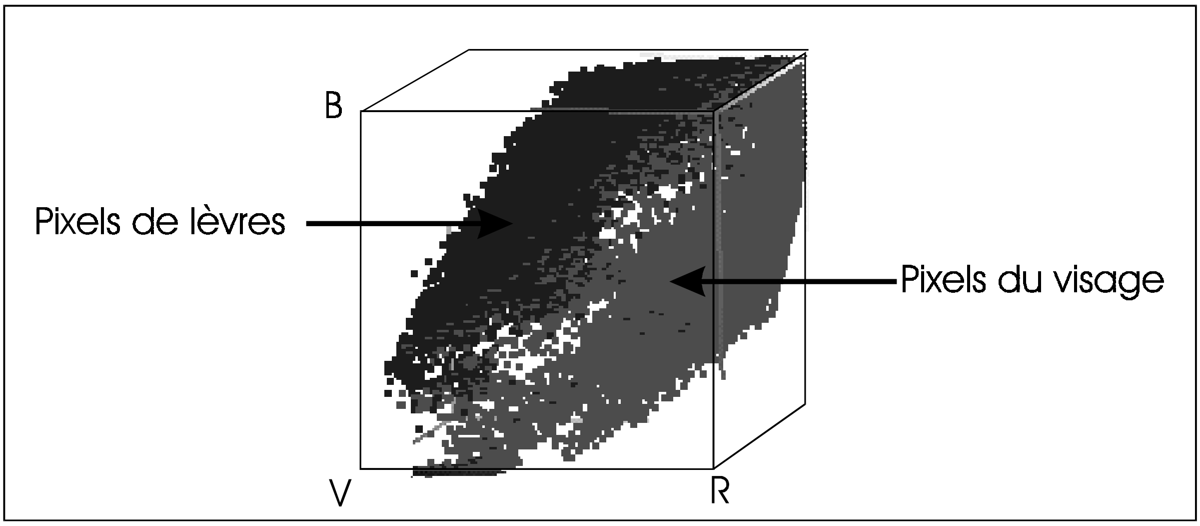


Figure 5.116 : représentation de la région des lèvres et du visage dans l'espace RVB (d'après Mégard, 1995).

Il est clair que dans ce cube, la ligne de séparation des pixels des lèvres et ceux du visage est très floue. En d'autres termes, il est très difficile de détecter les contours internes et externes sans avoir un taux d'erreur important juste en appliquant un seuillage sur une ou plusieurs composantes. Cependant, un autre axe de projection peut séparer le bleu des lèvres du visage. Cet axe est la transformation TLS. La Figure 5.117 montre une représentation dans l'espace TLS des lèvres maquillées en bleu.

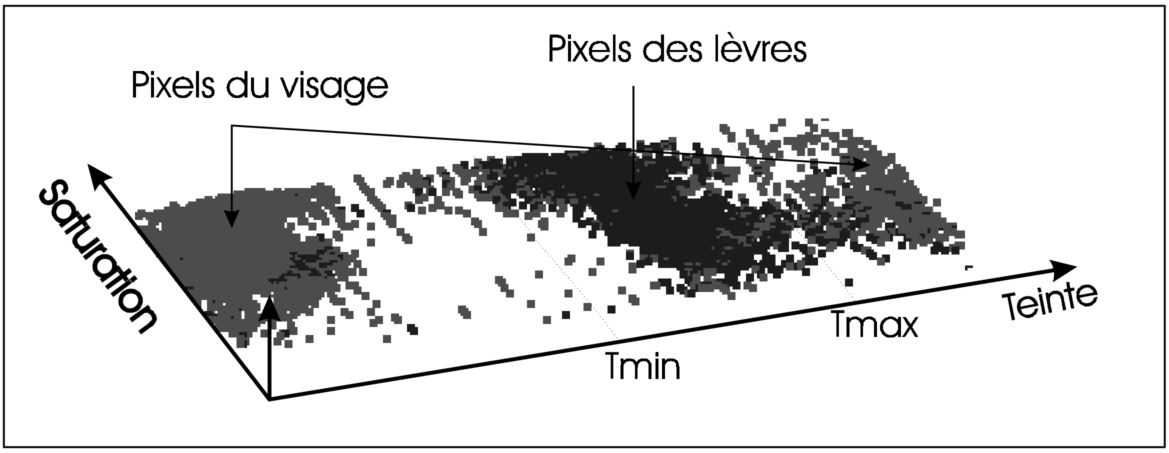


Figure 5.117 : représentation des lèvres dans l'espace TLS (d'après Mégard, 1995)

On peut remarquer que les pixels des lèvres sont concentrés dans une portion de l'espace déterminée par Tmin et Tmax. En choisissant bien ces deux valeurs, on peut facilement séparer la région des lèvres et procéder à une binarisation de l'image.

Afin de calculer les valeurs Tmin et Tmax pour chaque prise de vues, un outil logiciel a été élaboré sur une station Silicon Graphics munie d'une carte d'acquisition d'image (Mégard, 1995; Guiard-Marigny, 1996). Il permet de numériser une image quelconque du locuteur et d'appliquer, sur cette image, le seuillage TLS avec différentes valeurs de Tmin et Tmax. L'utilisateur a la possibilité d'entrer ces deux valeurs et de visualiser la qualité du seuillage. En fonction du résultat, il peut modifier ces deux valeurs jusqu'à l'obtention d'un résultat satisfaisant.

Cet outil a été utilisé pour le calcul de la LUT du Liptrack. Il faut rappeler que le décodeur vidéo numérise les données sur 16 bits, c'est à dire 5 bits pour le rouge, 6 bits pour le vert et 5 bits pour le bleu. Ceci entraîne une légère erreur sur la teinte car le canal du vert a plus de précision que les autres canaux (plus de bits alloués). C'est pourquoi il est nécessaire de prendre en compte de ce phénomène dans le calcul de la LUT.

Quand les valeurs de Tmin et Tmax sont obtenues pour la prise de vues, il ne reste plus qu'à transférer la LUT vers la mémoire de la carte.

2.3.3 Binarisation de l'image

Le signal d'entrée vidéo est décodé puis numérisé grâce au composant intégré *BrookTree.* Les 16 bits RVB (5 bits en R et B et 6 bits en V) sont envoyés vers une table de correspondance (Look Up Table). Cette LUT est préalablement chargée par le DSP de façon à détecter uniquement la région des lèvres. C'est à dire, elle ne laisse passer que la région du bleu saturé correspondant au maquillage. L'ensemble des ces opérations est présenté sur la Figure 5.118.

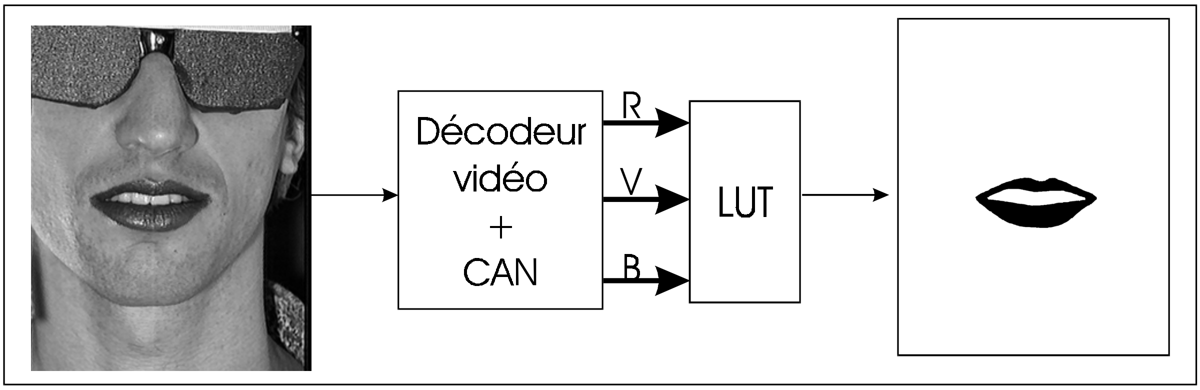


Figure 5.118 : processus de binarisation du vermillon des lèvres.

La LUT est une mémoire vive de 64K sur 4 bits. Il est évident que seul le bit de poids faible de cette mémoire est utilisé puisque l'image de sortie doit être entièrement codée sur un seul bit. La sortie du décodeur vidéo constitue l'adresse de cette mémoire. La Figure 5.119 montre un exemple de sélection de la région des lèvres.

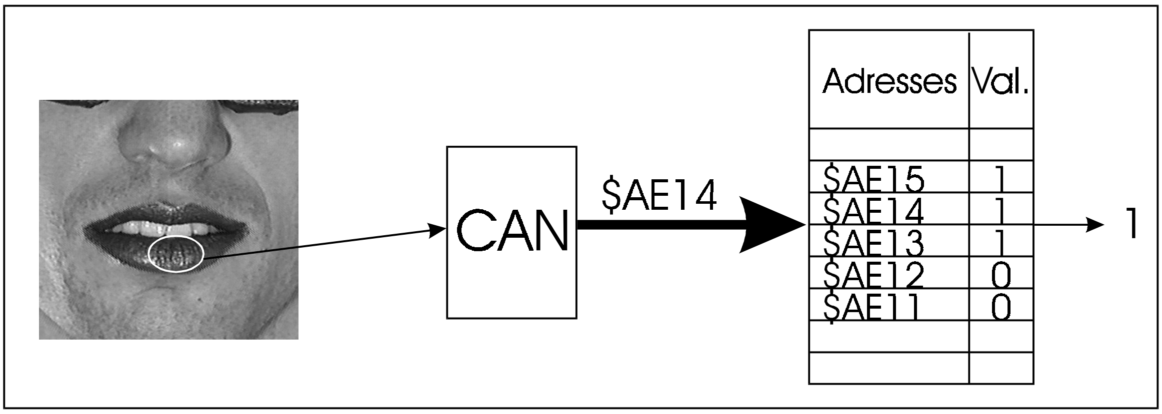


Figure 5.119 : un exemple de détection de la région des lèvres. Le résultat de la conversion numérique constitue l'adresse de la RAM de la LUT.

Il est donc très important de charger la LUT avec des valeurs qui permettent de bien séparer le vermillon des lèvres du reste de l'image.

2.3.3.1 Le transfert de la LUT

Comme nous l'avons précisé, les communications entre la machine hôte et la carte se font essentiellement par l'intermédiaire du port série RS 232. Le protocole de communication est décrit dans la section 2.3.4.3.

A chaque adresse de la mémoire LUT correspond une valeur codée sur un bit. Pour accélérer le processus de chargement, un octet est envoyé pour chaque trame de données. A partir de cette donnée sur 8 bits, le DSP démarre l'écriture en mémoire en prenant le bit de poids faible de l'octet transmis. Après écriture, l'adresse de la LUT est incrémentée et le deuxième bit de poids faible est écrit en mémoire. Ce processus continue jusqu'à ce que tous les bits de la trame soient correctement stockés. Dans ce cas le DSP envoie la trame d'écho pour signaler la fin du chargement du paquet. Ce processus continue tant que toutes les valeurs de la LUT ne sont pas transmises. Le schéma de la Figure 5.120 décrit le transfert d'un octet de la LUT vers la carte.

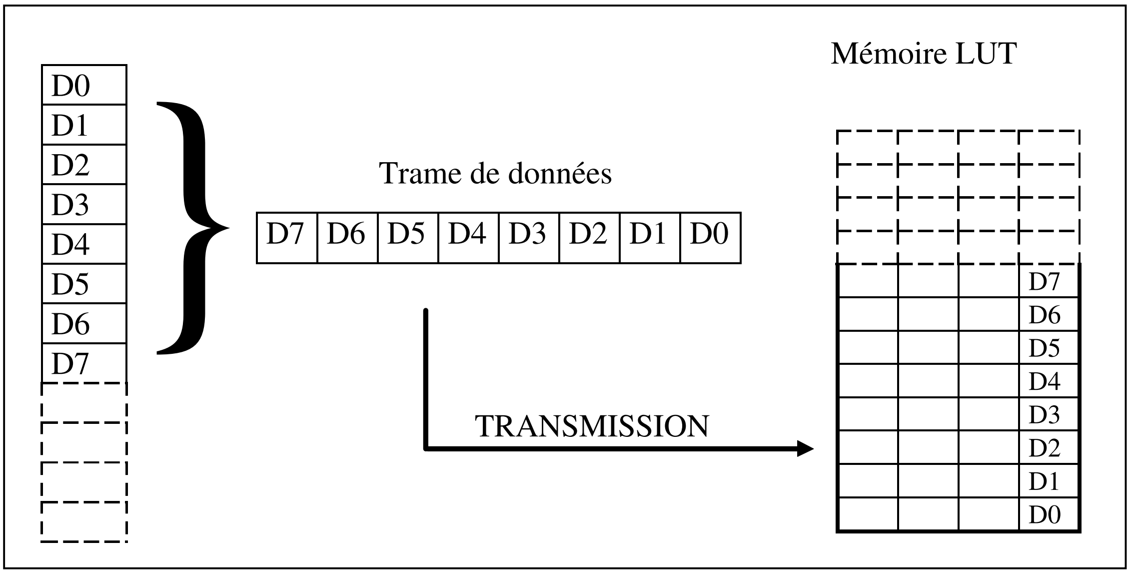


Figure 5.120 : transmission des données vers la mémoire LUT.

En utilisant cette méthode et une vitesse de transmission de l'ordre de 9600 bauds, le transfert de la LUT ne dure que quelques dizaines de secondes.

2.3.3.2 Mémorisation des données

L'image de sortie est donc une image binaire ("1" pour le maquillage bleu et "0" pour le reste). Le convertisseur numérique/analogique transforme cette image binaire en un signal vidéo pour alimenter l'entrée d'un moniteur de contrôle. Pour le maintien de la cadence de 50 paramètres par seconde, une configuration de deux mémoires *ping-pong* a été implémentée. La Figure 5.121 représente le principe de la mémorisation de l'image.

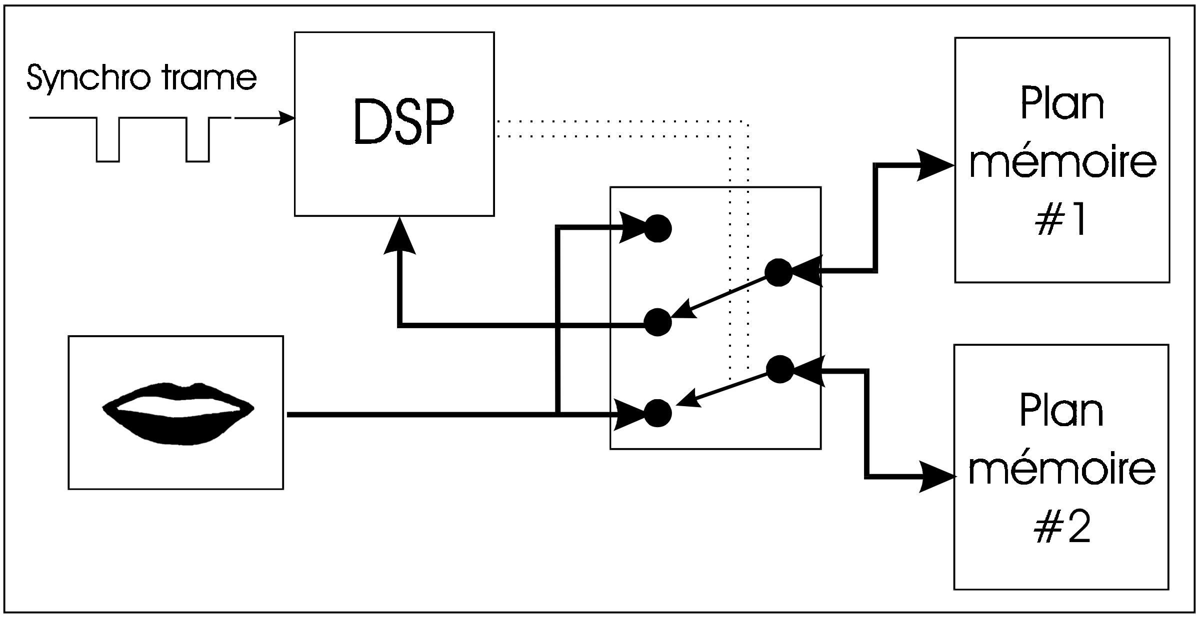


Figure 5.121 : principe de mémorisation d'une image et calcul sur l'image précédente

Sur cet exemple, alors que l'image courante est en train d'être chargée en mémoire N° 2, le DSP continue ses calculs sur l'ancienne image stockée en mémoire N° 1. Une fois les calculs terminés, le DSP démasque les interruptions qui proviennent de la synchro trame et autorise le basculement des deux mémoires.

Afin d'économiser de l'espace mémoire pour le stockage de l'image binaire, quatre pixels sont stockés à la même adresse suivant l'ordre de balayage de l'image (Figure 5.122).

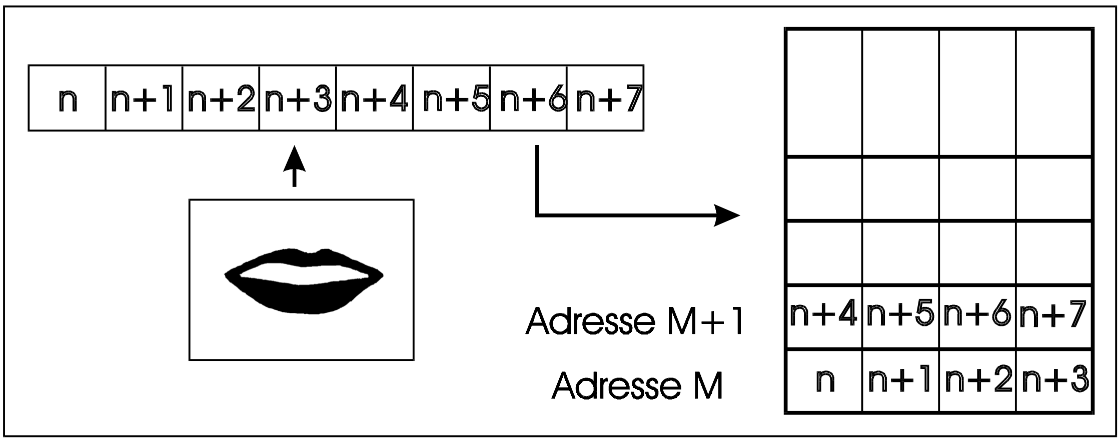


Figure 5.122 : stockage de 8 pixels en mémoire par séries de 4.

Pour retrouver une donnée dans cette mémoire, en plus de son adresse, il faut connaître sa position à cette adresse. Ceci a l'inconvénient de compliquer les programmes de traitement, mais permet de réduire le nombre d'accès en mémoire. Il permet aussi de multiplier par quatre l'espace adressable de l'image.

2.3.4 Détection des contours et calcul des paramètres

Dans cette partie, nous décrivons les algorithmes de détection et de poursuite de contour implémentés sur le système de labiométrie. Grâce à l'utilisation des images binaires, ces algorithmes sont simplifiés.

2.3.4.1 Détection du point de départ du contour

L'analyse d'images se fait à partir d'une fenêtre dont les coordonnées (la largeur, la longueur et l'adresse du pixel du point haut à gauche) sont stockées. Ces données sont paramétrisées de façon à ce que l'utilisateur puisse facilement les modifier pour différentes prises de vue.

Comme chaque case mémoire contient quatre pixels, il est nécessaire d'établir une convention pour connaître la position de chaque pixel (Figure 5.123). A chaque déplacement sur le contour, il faut calculer la nouvelle adresse ainsi que la nouvelle position en mémoire du prochain pixel par rapport au dernier pixel traité.

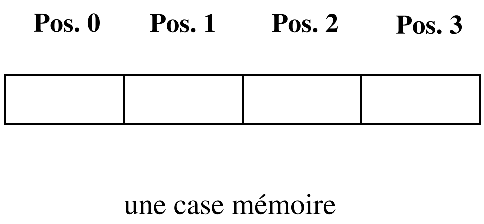


Figure 5.123 : les différentes positions pour une case mémoire

La Figure 5.124 décrit la détection du premier point du contour sur l'image binaire.

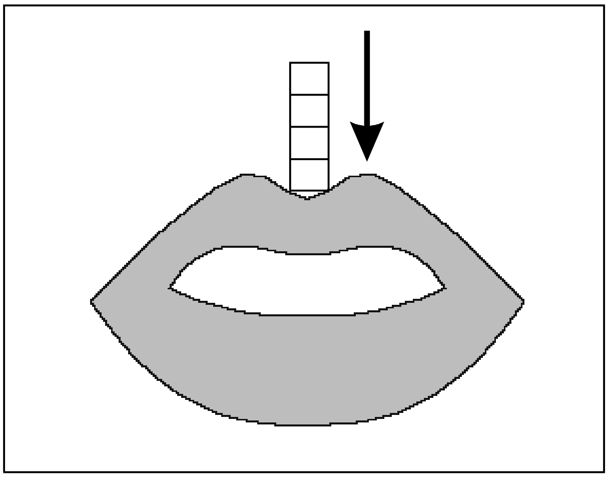


Figure 5.124 : déplacement de quatre pixels consécutifs au milieu de la fenêtre d'analyse pour la détection du premier point contour.

L'algorithme de détection du premier point du contour externe peut être représenté par l'organigramme suivant :

nous avons

X : adresse du milieu de la fenêtre d'analyse
M : nombre de pixels pour chaque ligne d'image
n : nombre de pixels noirs détectés

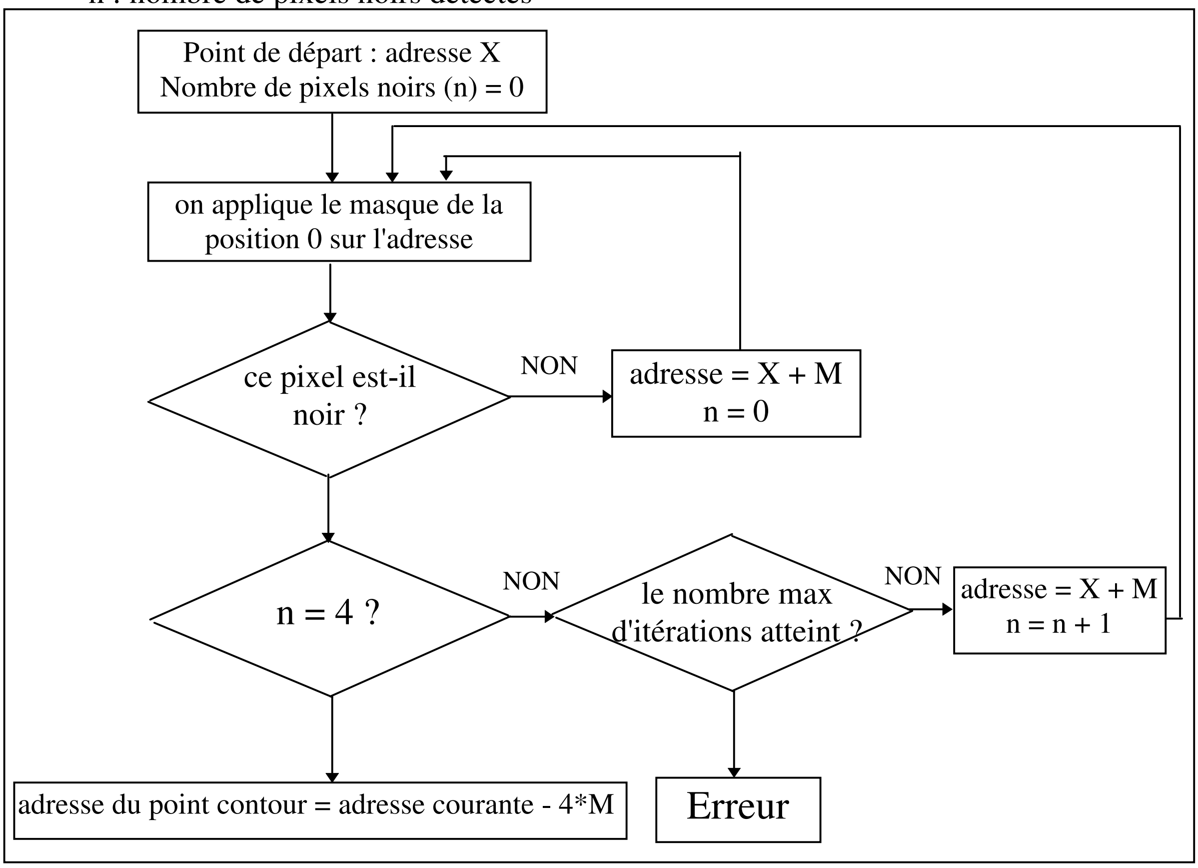


Figure 5.125 : organigramme de détection du premier point de contour externe de la lèvre supérieure

Pour minimiser les erreurs de détection, nous avons fixé le nombre de pixels noirs consécutifs à quatre. Dans ce cas les pixels noirs isolés de l'image ne sont pas pris en compte.

Un test sur le nombre d'itérations permet d'arrêter le processus et d'envoyer un message d'erreur au cas où le contour est introuvable.

En ce qui concerne le contour interne, l'algorithme est très semblable à celui du contour externe. Cependant, il doit être exécuté après le calcul des paramètres du contour externe. En voici d'autres différences :

- on teste si les pixels sont BLANCS, sinon on continue la descente à l'intérieur des lèvres.
- tant que trois pixels blancs consécutifs ne sont pas détectés, on continue la recherche, ceci afin d'éviter la poursuite du contour sur des pixels blancs isolés des lèvres.
- si on dépasse le trois quart de la hauteur totale du contour externe, on suppose que les lèvres sont fermées et on arrête la recherche.

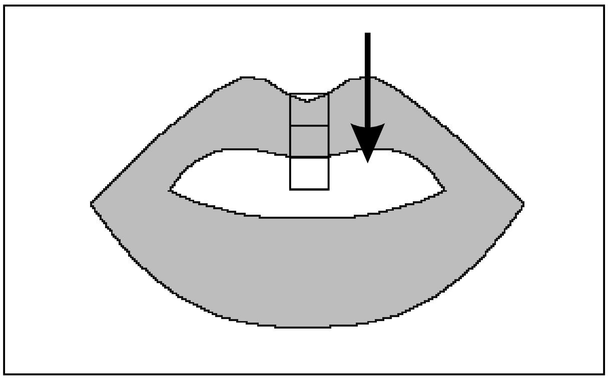


Figure 5.126 : détection du premier point du contour interne

L'organigramme suivant montre l'algorithme de détection du premier point du contour interne.

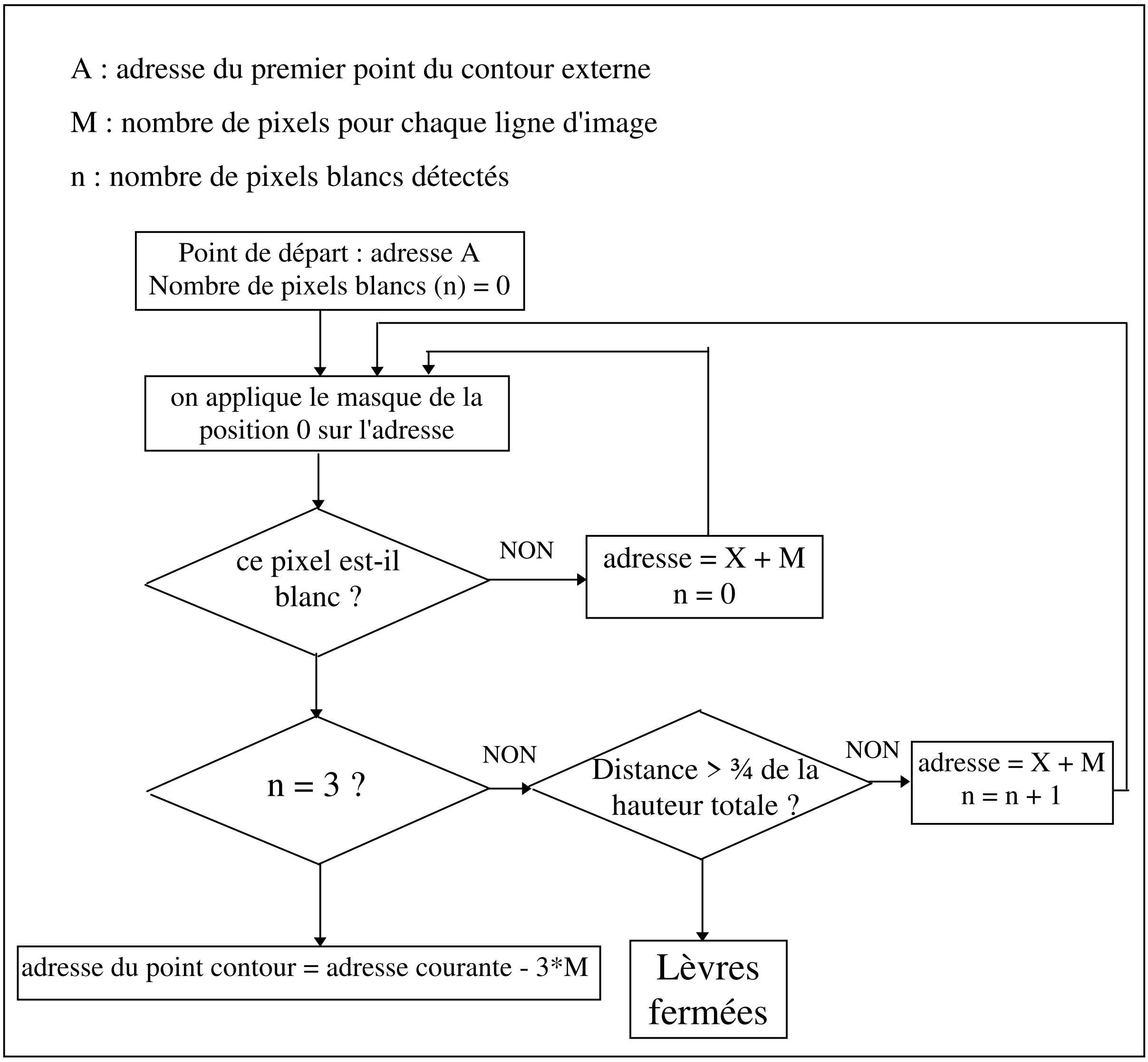


Figure 5.127 : organigramme de détection du premier point du contour interne.

2.3.4.2 Poursuite de contour et calcul des paramètres

Une fois le premier point contour détecté, il faut suivre le contour et calculer les paramètres relatifs à ce contour.

L'algorithme implémenté sur *Liptrack* est la poursuite de Freemann sur 4 connexes. Le schéma suivant montre le principe :

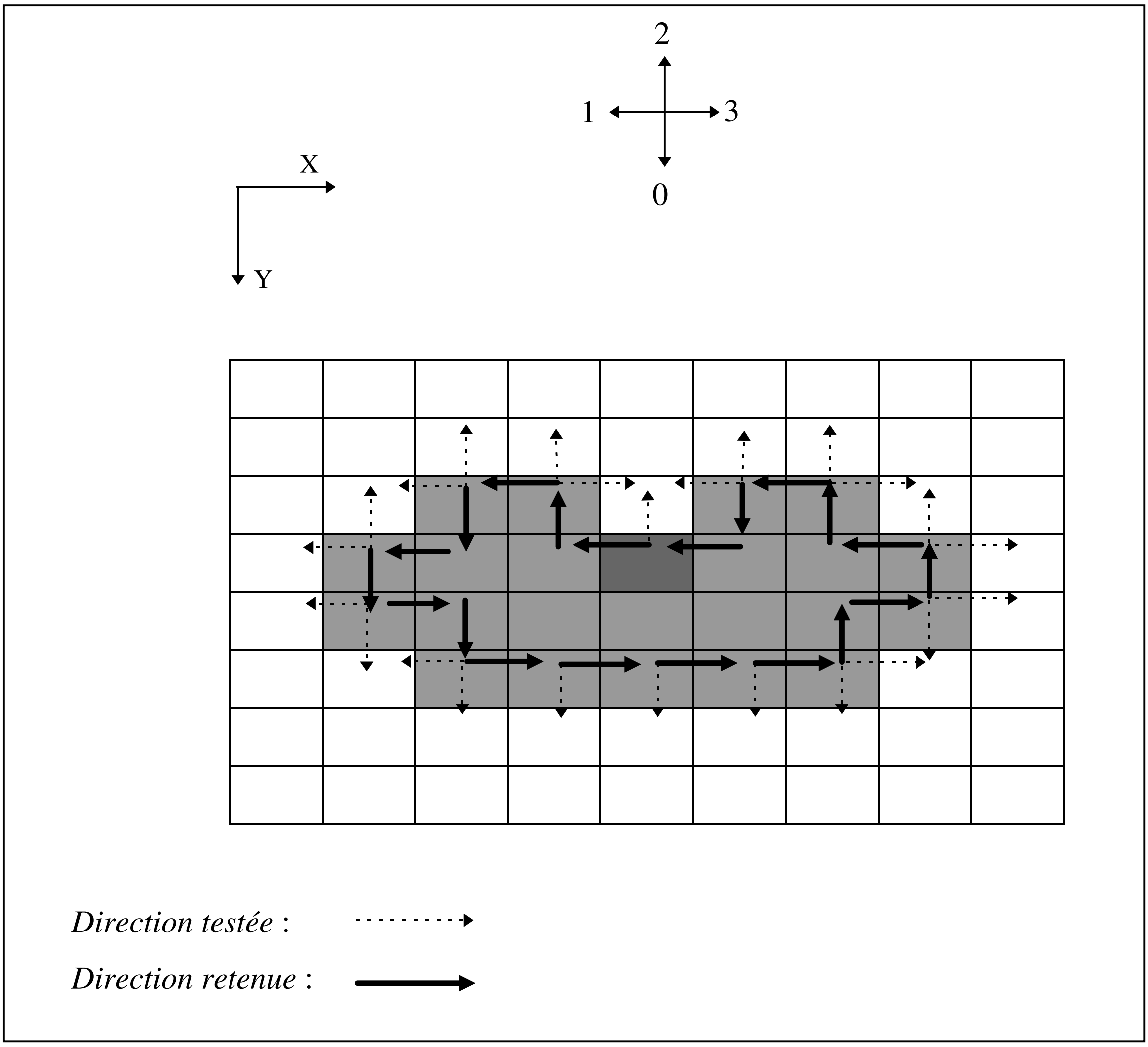


Figure 5.128 : la poursuite de contour sur 4 connexes de Freemann. Le pixel représenté en plus foncé est le pixel de départ.

Pour chaque pixel, la direction précédente et la position sont mémorisées. Pour la poursuite du contour, on incrémente la valeur de la direction courante et on teste le pixel associé. Comme dans chaque case mémoire quatre pixels ont été stockés, il est nécessaire de connaître la position du pixel à tester et d'appliquer le masque correspondant. Si le pixel est noir, on le considère comme une partie du contour, sinon on décrémente la valeur de direction et on teste le nouveau pixel et ainsi de suite. Pour pouvoir mesurer les paramètres labiaux, il est nécessaire, à chaque déplacement, de calculer les coordonnées du nouveau pixel grâce à celles du pixel précédent déjà mémorisées. L'algorithme de Freemann implémenté sur le DSP pour la poursuite du contour interne peut être représenté par l'organigramme suivant :

ADDEP : adresse de départ

ADD : adresse courante

DIR : direction courante

POS : position du pixel

PARAMA : paramètre A

PARAMB : paramètre B

XMAX et YMAX : les valeurs maximales de X et Y

XMIN et YMAX : les valeurs minimales de X et Y

X, Y : coordonnées du pixel courant en ayant le premier point du contour comme point de référence.

M : le nombre de pixel par ligne

Y = Y +1, X inchangé si DIR = 0 *X = X - 1, Y inchangé si DIR = 1*

Y = Y - 1, X inchangé si DIR = 2 *X = X + 1 , Y inchangé si DIR = 3*

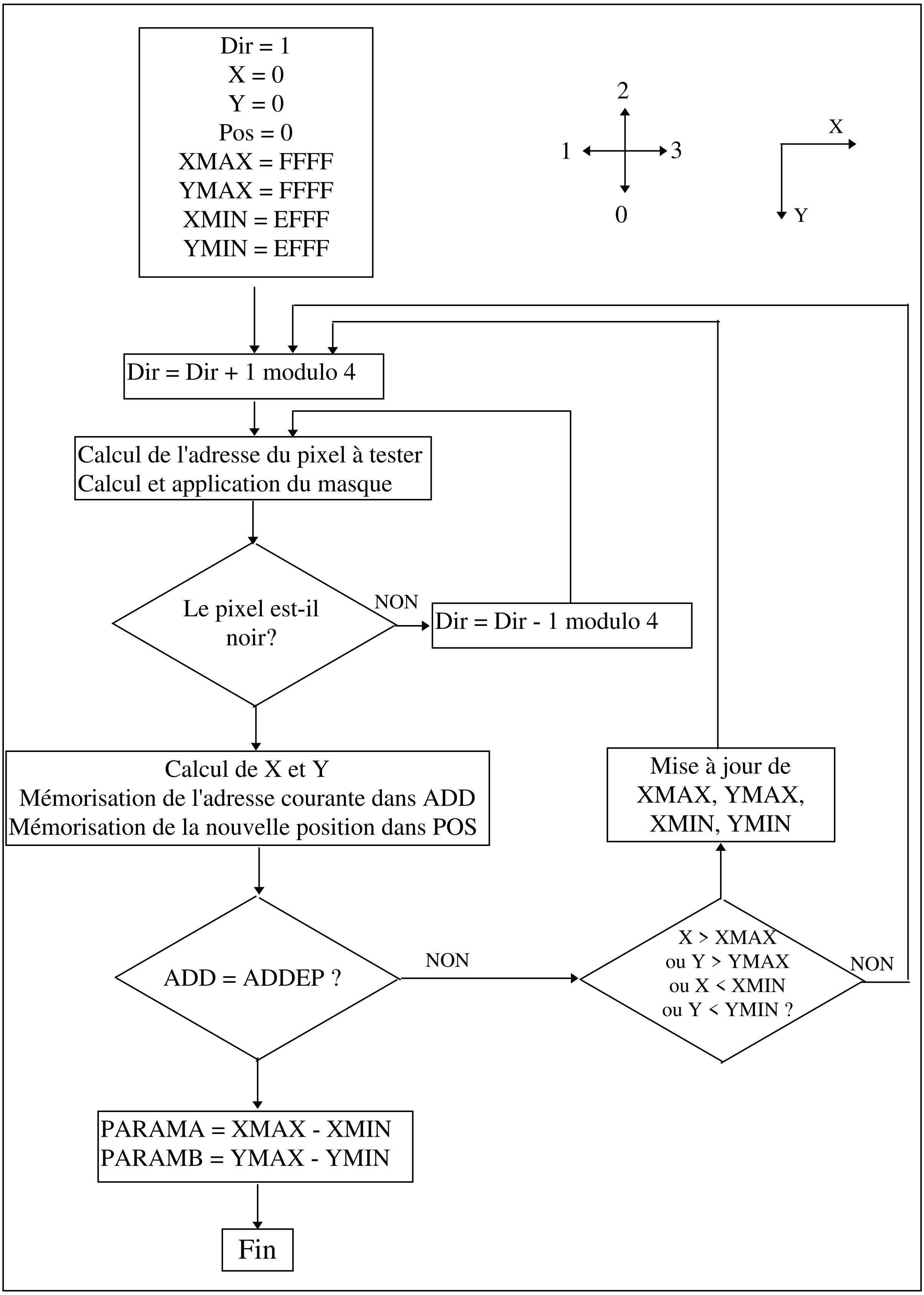


Figure 5.129 : organigramme de poursuite du contour externe des lèvres et calcul des paramètres correspondants

A chaque déplacement sur le contour, les valeurs maximales et minimales des coordonnées sont mises à jour ce qui facilite considérablement la tâche de calcul des paramètres. En effet, les paramètres A et B, A' et B' correspondent à la différence des valeurs maximales et minimales des coordonnées X et Y.

2.3.4.3 Transmission des paramètres

Comme mentionné précédemment, le Liptrack communique avec le monde extérieur grâce à son interface série RS-232. Ce type d'interface est largement utilisé dans les communications à courte distance (de l'ordre de quelques dizaines de mètres maximum) entre deux ordinateurs ou un ordinateur et un périphérique extérieur. Il est même souvent utilisé dans des protocoles de contrôle numérique entre *un maître* et *un esclave*. Pratiquement tous les ordinateurs possèdent les moyens matériels pour ce type de communication. Cependant, seules les tensions de sortie (-12 à +12 pour certains cas et -5 à +5 pour d'autres) et le codage des bits (tension positive pour 1 et négative pour 0 ou l'inverse) peuvent varier d'une machine à l'autre. Il s'agit d'une transmission asynchrone, c'est à dire que l'émetteur peut envoyer son message sans "préavis". L'horloge n'est pas transmise sur la ligne ce qui nécessite une configuration identique des deux côtés (vitesse de transmission, type de codage, configuration des trames, niveau des tensions, etc.)

Afin de garder la synchronisation, les trames de données sont précédées d'un bit de start et suivies d'un ou deux bits de stop. On peut aussi envoyer un bit de détection d'erreur qui est le bit de parité. Suivant les protocoles, celui-ci peut être à un ou zéro pour une parité paire ou impaire. Dans le cas du Liptrack, les données sont transmises avec un bit de start (niveau logique 0), huit bits de données et un bit de stop. La vitesse de transmission est facilement configurable, mais en tenant compte de la quantité d'information à transmettre sur la ligne, nous avons choisi la vitesse de 9600 bauds. La Figure 5.130 montre la composition d'une trame sur la ligne RS-232.

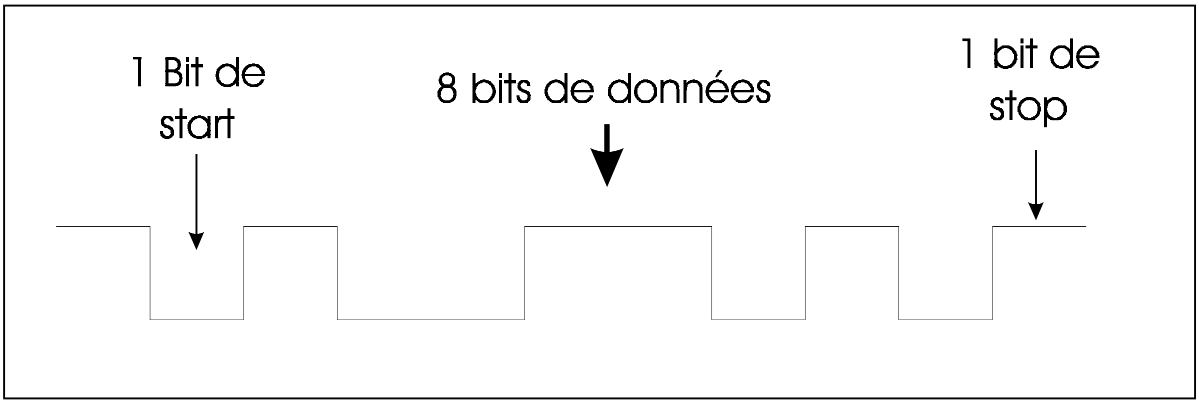


Figure 5.130 : exemple d'une trame de données sur la ligne RS-232. Elle est composée d'un bit de stop, 8 bits de donnée et un bit de start.

Un protocole de communication du type *Hand Shake* entre le Liptrack et la machine hôte a été établi. A chaque transmission de paramètre, le DSP attend l'accusé de réception de la machine hôte. La transmission est interrompue quand l'écho n'est pas renvoyé ou lorsqu'une différence existe entre la donnée transmise et celle reçue. La Figure 5.131 montre le principe de communication entre le Liptrack et une autre machine.

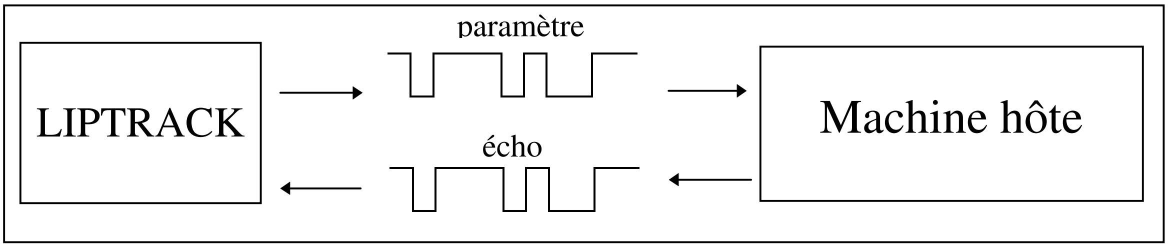


Figure 5.131 : la communication entre le Liptrack et un ordinateur sur lequel est implémenté un système de reconnaissance audiovisuelle ou un processus de synthèse de visage.

2.3.4.4 Calcul à la cadence de 50 trames par seconde

Pour avoir plus de précision dans le calcul des paramètres labiaux, chaque image doit être décomposée en deux trames. Pour cela, nous avons implémenté l'algorithme suivant sur le système Liptrack, qui permet de traiter chaque trame séparément.

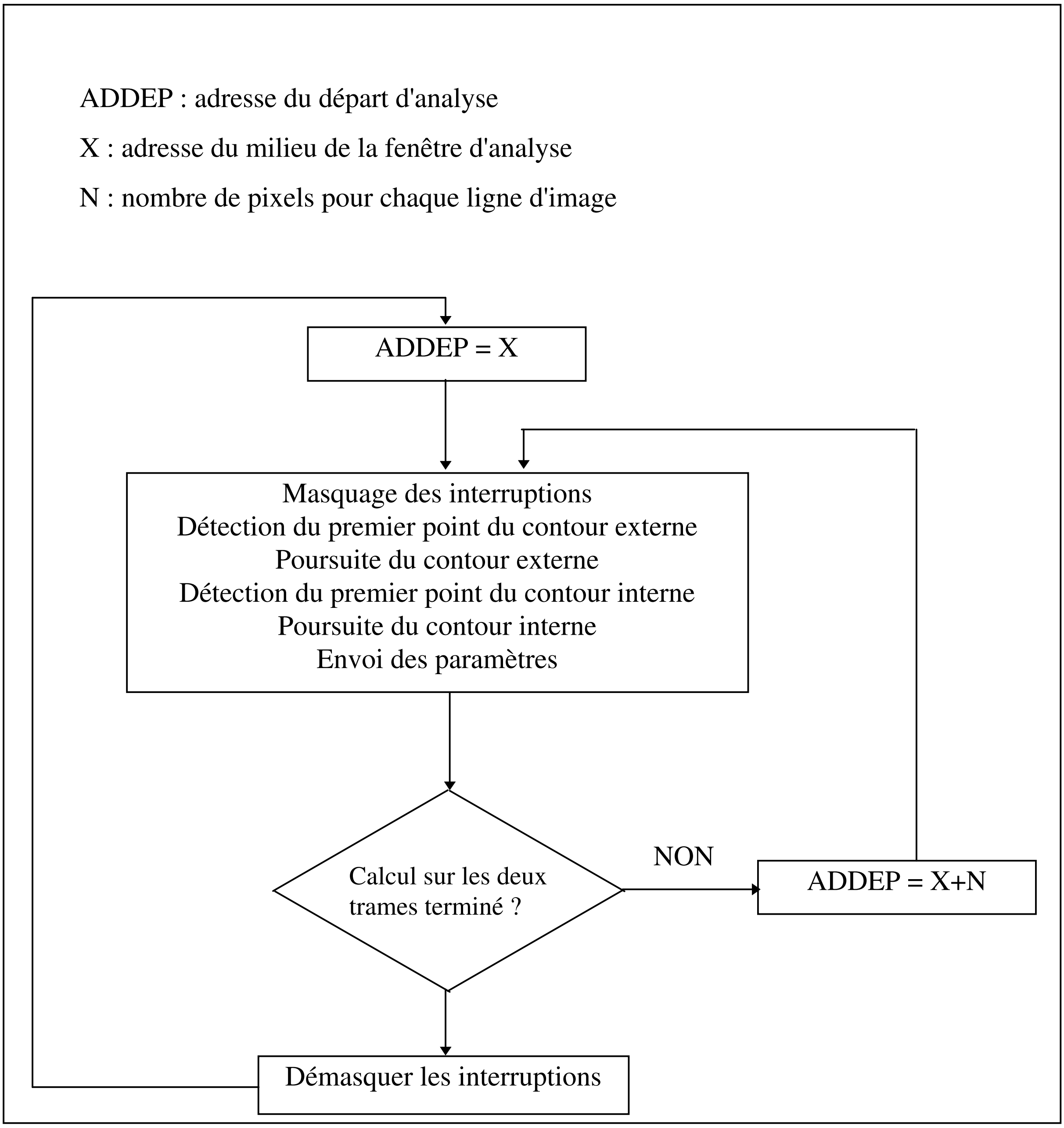


Figure 5.132 : algorithme de traitement pour deux trames séparées.

A la fin du traitement de la première trame d'une image, on recommence les calculs mais en décalant d'une ligne le point de départ de l'algorithme. En ce qui concerne les poursuites de contours, le décalage doit être de deux lignes pour traiter une seule trame à la fois.

2.3.5 Avantages

Le système de labiométrie temps réel a de nombreux avantages par rapport au poste visage-parole de Lallouache (1991) ou au système d'analyse temps réel implémenté sur Silicon Graphics (Mégard, 1995) :

1- Il peut être connecté sur une station de travail, un PC ou une autre machine ayant d'un port série du type RS232.

2- Le coût de ses composants ne dépasse pas 5 KF. Afin d'obtenir les mêmes performances, une station Silicon Graphics Indy munie d'une carte d'entrée vidéo ou une station équivalente est indispensable. Ceci nécessite un investissement matériel nettement plus important.

3- Grâce à sa taille, il peut être facilement implémenté dans une interface homme/machine. On peut envisager de connecter cet appareil à une micro-caméra et de monter l'ensemble sur un casque porté par l'utilisateur.

4- Il est adaptable aux différentes conditions d'éclairage et du maquillage des lèvres. Pour cela, il suffit de recharger la LUT préalablement calculée pour ce niveau d'éclairage.

5- Il peut utiliser plusieurs types de formats vidéo d'entrée : PAL, NTSC et Y/C.

2.3.6 Inconvénients

1. Pour le calcul de la LUT, un ordinateur muni d'une carte d'entrée vidéo est nécessaire.
2. Pour chaque prise de vues, une nouvelle LUT doit être calculée et chargée en mémoire. Il faut ajouter que le système de Lallouache (1991) ou de Mégard (1995) partagent le même inconvénient.

3. Évaluations

Pour évaluer les performances techniques du système électronique Liptrack, nous avons tout d'abord procédé à une évaluation purement théorique en fonction des capacités de calcul et de stockage d'images de ce prototype. Nous avons ensuite comparé les résultats d'analyse de ce système avec ceux obtenus par le poste visage-parole. Nous avons finalement évalué son apport d'information dans une expérience de reconnaissance visuelle de la parole.

3.1 Evaluation théorique

Dans cette partie, nous allons calculer en fonction des caractéristiques des composants du Liptrack, notamment le DSP, ses performances en nombre de calculs et ses capacités potentielles.

Prenons le cas du traitement des images entières 25 fois par seconde. Ces images ont la taille de 512x512 pixels pour un format PAL (625 lignes, 25 Hz). On peut imaginer, dans le pire des cas, que les lèvres occupent toute la largeur et toute la hauteur de chaque image. Nous allons donc estimer le temps de calcul nécessaire pour ce type d'images. Nous avons vu que la première étape dans l'algorithme de calcul des paramètres est la détection de contour. Si l'image des lèvres occupe toute la surface de l'écran, on peut facilement négliger le temps nécessaire à la détection du premier point contour. Il reste maintenant à estimer le temps de la poursuite du contour externe et interne ainsi que la détection du premier point contour du contour interne.

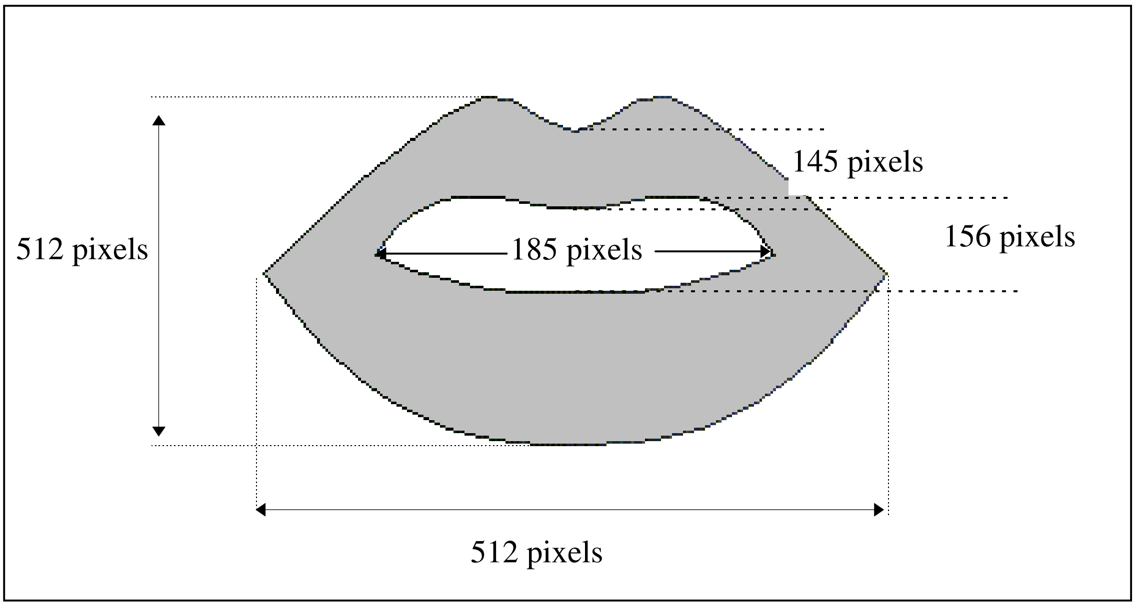


Figure 5.133 : une image des lèvres de 512x512 pixels pour l'évaluation théorique de Liptrack.

Sur l'exemple de cette image des lèvres, nous avons calculé les dimensions du contour interne. Pour calculer le nombre de pixels sur chaque contour, on peut attribuer une forme rectangulaire aux contours. Dans ce cas on a :

nombre de pixels du contour externe = 2(512 + 512) = 2048*

nombre de pixels du contour interne = 2(185 + 156) = 882*

Si on suppose que l'algorithme de poursuite de contour nécessite en moyenne 250 cycles pour chaque pixel (avec le temps d'accès mémoire) et que le test du premier point contour dure 40 cycles par pixel (ce qui est très pessimiste), on trouve :

*nombre de cycles pour la poursuite du contour externe = 250*2058 = 514500*

*nombre de cycles pour la détection du contour interne = 40*145 = 5800* *nombre de cycles pour la poursuite du contour interne = 250*882 = 220500*

nombre total de cycles = 740800

Pour 25 images par seconde, on peut trouver la fréquence d'horloge correspondant pour l'exécution de l'ensemble de ces instructions :

*fréquence d'horloge minimale = 25*740800 = 18.5 Mhz*

On voit que même en prenant le cas extrême, Liptrack doit pouvoir traiter 25 images entières par seconde.

Nous avons montré que pour avoir des mesures plus précises sur les distances géométriques des lèvres, il faut désentrelacer une image entière en deux trames. Dans ce cas, on peut reprendre les calculs pour une image de 512x256 (une trame). Ce qui donne :

nombre de pixels du contour externe = 2(512 + 256) = 1536*
nombre de pixels du contour interne = 2(185 + 78) = 526*

*nombre de cycles pour la poursuite du contour externe = 250*1536 = 384000*
*nombre de cycles pour la détection du contour interne = 40*72 = 2880*
*nombre de cycles pour la poursuite du contour interne = 250*526 = 131500*

nombre total de cycles pour une trame = 518380

Pour traiter 50 trames par seconde, la fréquence d'horloge doit être supérieure à :

*fréquence d'horloge minimale = 50*518380 ≈ 26.0 MHz*

Comme le DSP 56002 a une fréquence d'horloge de 40 Mhz, tous les paramètres géométriques des lèvres peuvent être obtenus dans le délai de 20 ms. Il faut cependant noter que dans les calculs ci-dessus, nous n'avons pas pris en compte le temps de transmission des paramètres. Nous allons donc calculer le temps restant après le calcul des paramètres.

Sachant que le DSP a une fréquence d'horloge de 40 Mhz et que les calculs doivent être effectués en 20 ms, on a :

$$\textit{temps restant après chaque calcul} = 20 - \frac{518380}{40000} \approx 7ms$$

Si chaque paramètre est codé sur deux octets, et si un paquet de transmission sur le port RS-232 a 10 bits (1 bit de start, 8 bits de données, 1 bit de stop), cela donne une vitesse minimale de transmission pour quatre paramètres qui est égale à :

$$\textit{vitesse minimale de transmission} = \frac{80}{0.07} \approx 1143 bits / seconde$$

Comme à chaque transmission de paramètre l'écho est renvoyé, la vitesse minimale doit être le double de cette valeur ; c'est à dire qu'il faut choisir la vitesse de *4800 bauds*. Cependant pour avoir une marge de temps plus grande, nous avons choisi la vitesse de *9600 bauds* pour le transfert et la réception des données. Quand il s'agit d'une application de reconnaissance ou de synthèse de visage, il est recommandé d'utiliser la vitesse maximale de transmission de la machine hôte afin de minimiser le temps de transfert.

3.2 Évaluation pratique

L'évaluation pratique de Liptrack a été effectuée en deux temps : premièrement, nous avons analysé l'ensemble du corpus de référence avec le Liptrack et comparé les résultats avec ceux obtenus par le poste visage-parole. Ensuite, nous avons évalué la robustesse des paramètres mesurés dans une application de reconnaissance visuelle seule en comparant leurs scores à ceux obtenus avec les paramètres de référence.

Pour la vérification du respect des délais de traitement pour chaque trame, nous avons utilisé le signal de commande de basculement des deux mémoires provenant du DSP. En effet, ce signal doit avoir une fréquence de 25 Hz, ce qui correspond au balayage d'une image entière. Si le DSP termine son traitement avant la fin de chargement de la nouvelle image, la fréquence de ce signal reste inchangée. Par contre, si le temps de traitement d'une image (ou deux trames) dépasse les 40 ms, le DSP ne donne pas le signal de basculement de mémoire et dans ce cas la fréquence descend au dessous de 25 Hz. Il est donc très simple de vérifier à l'oscilloscope, le fonctionnement de la carte à 50 trames par seconde. Nous avons donc utilisé cette méthode pour une première vérification et nous avons constaté que la fréquence du signal en question était toujours à 25 Hz. Ce qui signifie que dans tous les cas, le calcul des paramètres et la transmission des données pour chaque trame sont terminés en moins de 20 ms.

3.2.1 Comparaison par rapport aux mesures de référence

Pour comparer les mesures de Liptrack avec celles du poste visage-parole, nous avons analysé la totalité de notre corpus de référence avec ce système. Pour une exacte comparaison, nous avons utilisé les mêmes images en noir et blanc enregistrées sur cassette U-Matic de la sortie du chroma-key (les lèvres sont déjà en noir). La LUT chargée dans la mémoire du Liptrack était donc passante uniquement pour le noir saturé (une seule valeur égale à 1 dans toute la table).

Le problème était de retrouver le *Time Code* de la bande d'enregistrement sur les fichiers de paramètres pour pouvoir les synchroniser avec ceux du poste visage-parole. Pour cela, nous avons procédé de la manière suivante :

- Connaissant le numéro du bip de départ, nous avons pu retrouver approximativement la synchronisation.

- Pour la parfaite synchronisation, nous nous sommes basés sur les *time codes* des fichiers du poste visage-parole correspondant à la fermeture des lèvres dans "*c'est pas ...*". Les paramètres A et B sont les références de la fermeture (A=B=0).

- Théoriquement, une seule référence de synchronisation suffit pour retrouver la base de temps sur les fichiers de mesures car le Liptrack et le poste visage-parole ont la même cadence d'analyse. Afin de vérifier la synchronisation sur toute la durée de l'analyse, nous avons procédé à la visualisation des paramètres des deux sources grâce à un éditeur graphique.

Cette méthode nous a permis aussi de vérifier encore une fois la cadence d'analyse de 50 trames/seconde. Il faut ajouter que les mesures provenant du Liptrack étaient en pixel, alors que celles du poste visage-parole étaient converties en millimètre. Pour avoir les mêmes unités dans les deux cas, nous avons converti ces données en connaissant les dimensions de chaque pixel et la résolution des images dans le cas du Liptrack et du poste visage-parole. (Nous rappellons que pour cet effet, une règle est placée au niveau du visage du locuteur au début de l'enregistrement).

Nous sommes maintenant en mesure de comparer les résultats. Nous avons choisi les stimuli "ABA" et "ABU" pour la détection de la fermeture des lèvres et "UBU", "AJA" et "IZI" pour les formes extrêmes. Les figures qui suivent représentent les mesures de référence et celles du Liptrack pour ces stimuli, avec la même échelle pour chacun des paramètres.

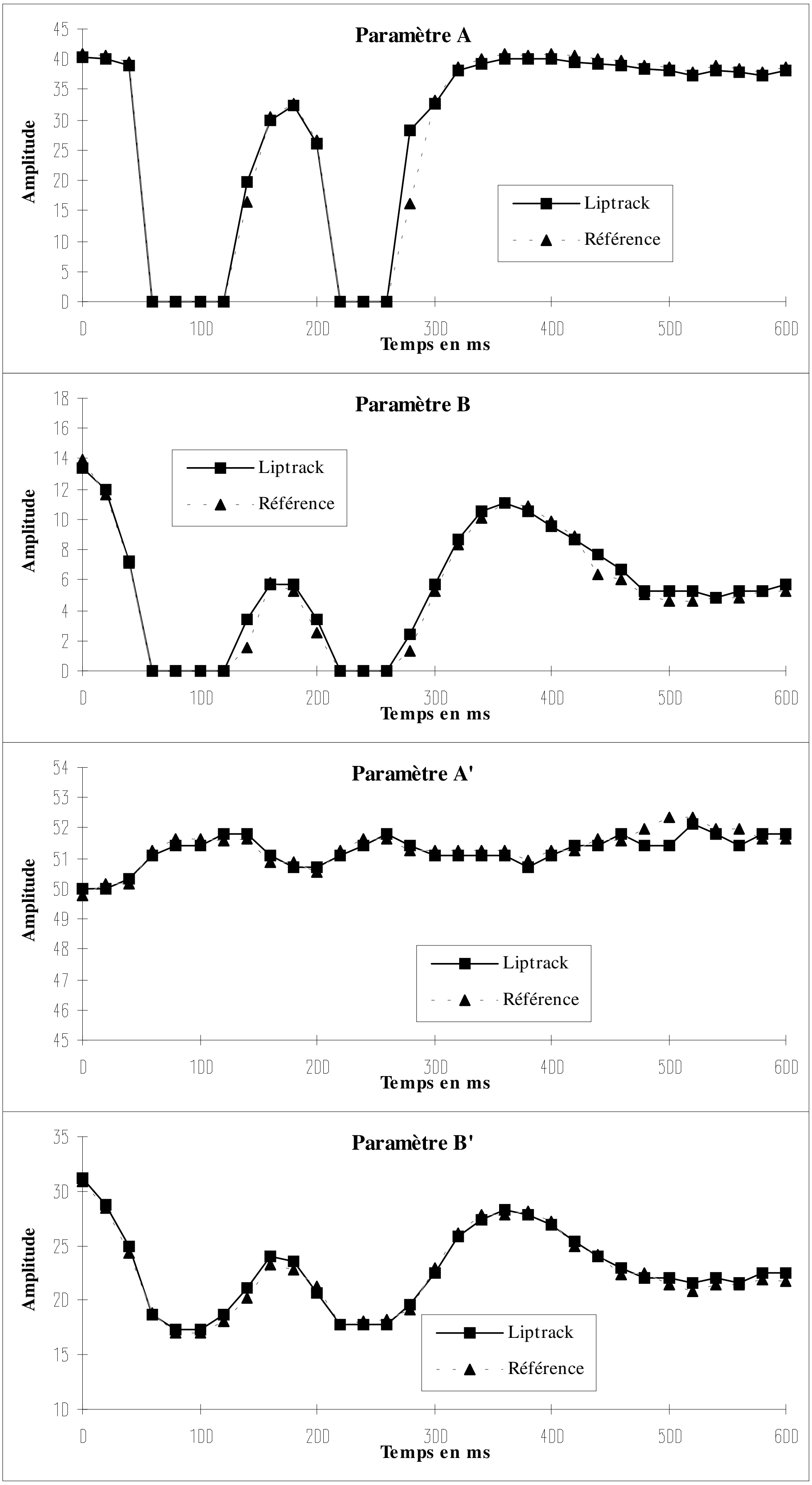


Figure 5.134 : comparaison des paramètres du Liptrack avec ceux de référence pour le stimulus ABA

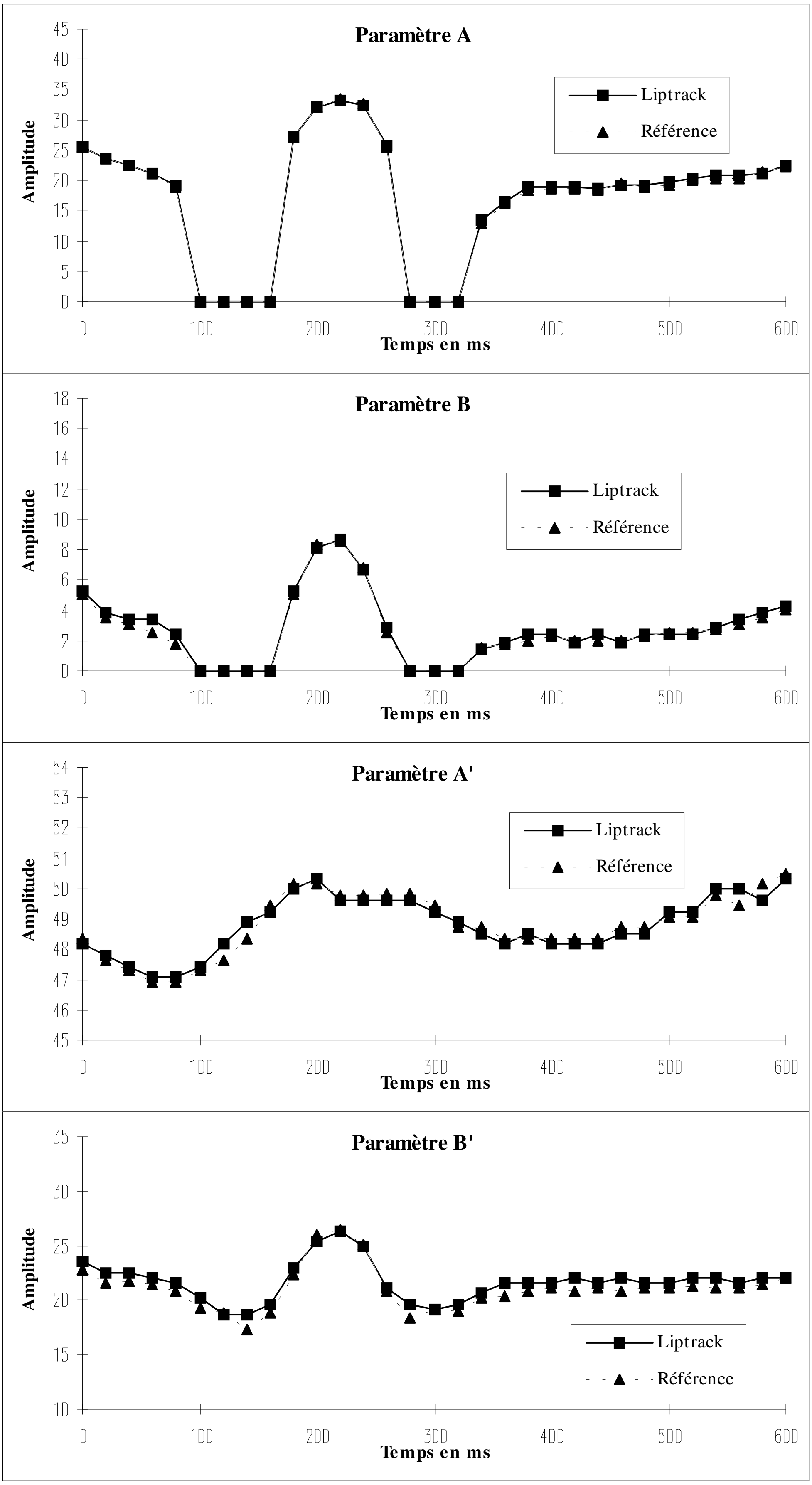


Figure 5.135 : comparaison des paramètres pour le stimulus "UBA".

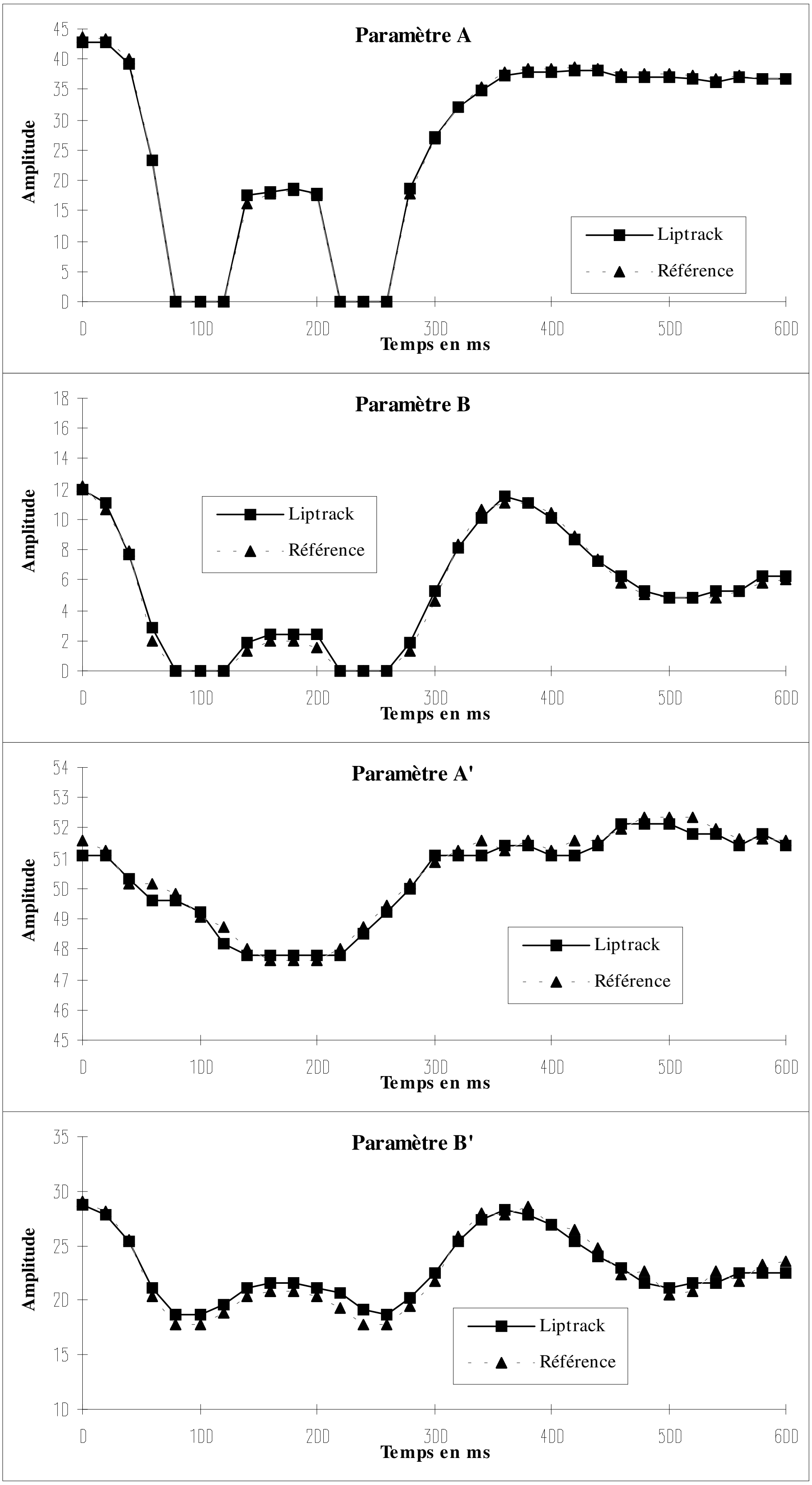


Figure 5.136 : comparaison des paramètres pour le stimulus "ABU".

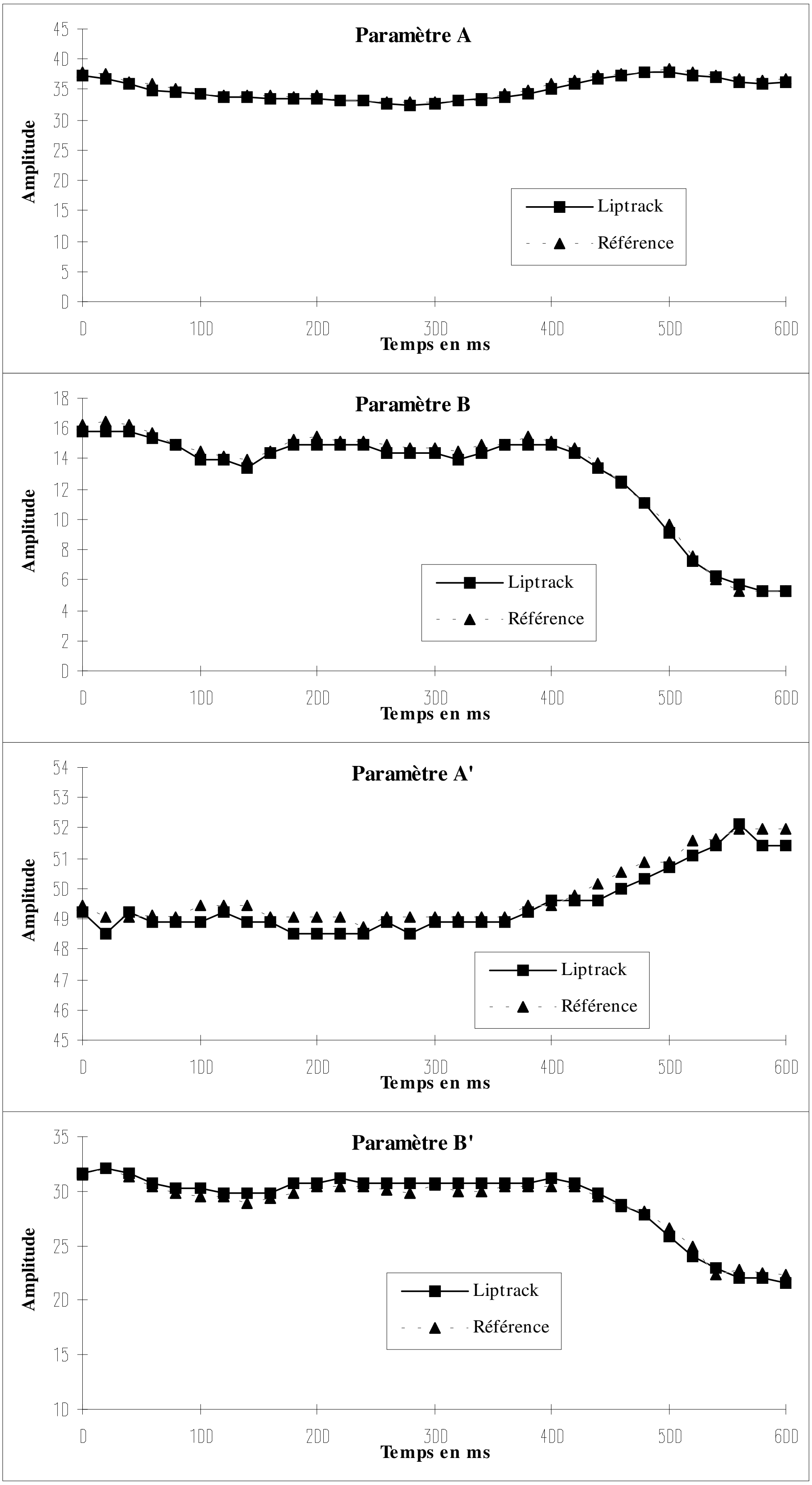


Figure 5.137 : comparaison des paramètres pour le stimulus "AJA".

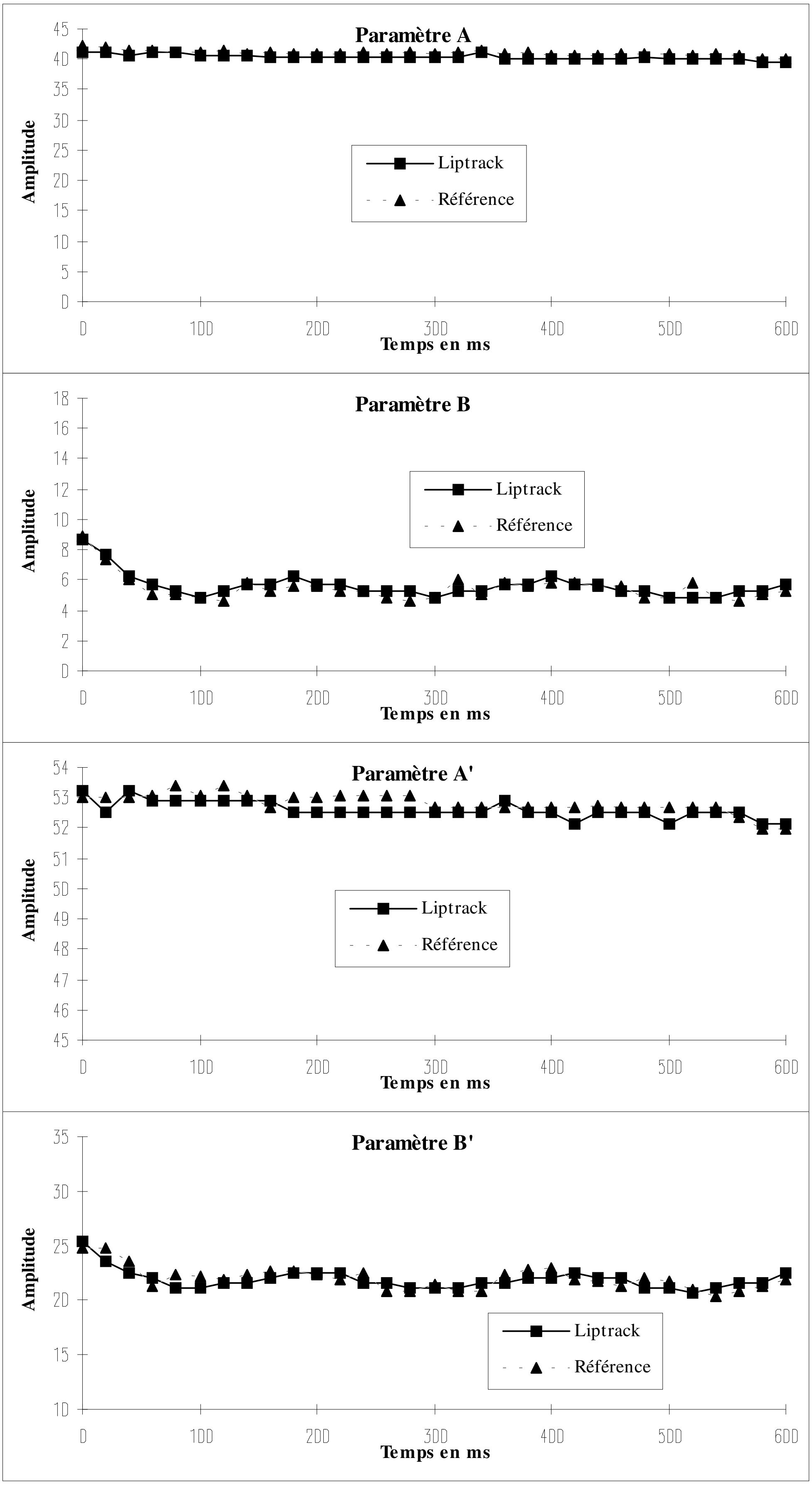


Figure 5.138 : comparaison des paramètres pour le stimulus "IZI".

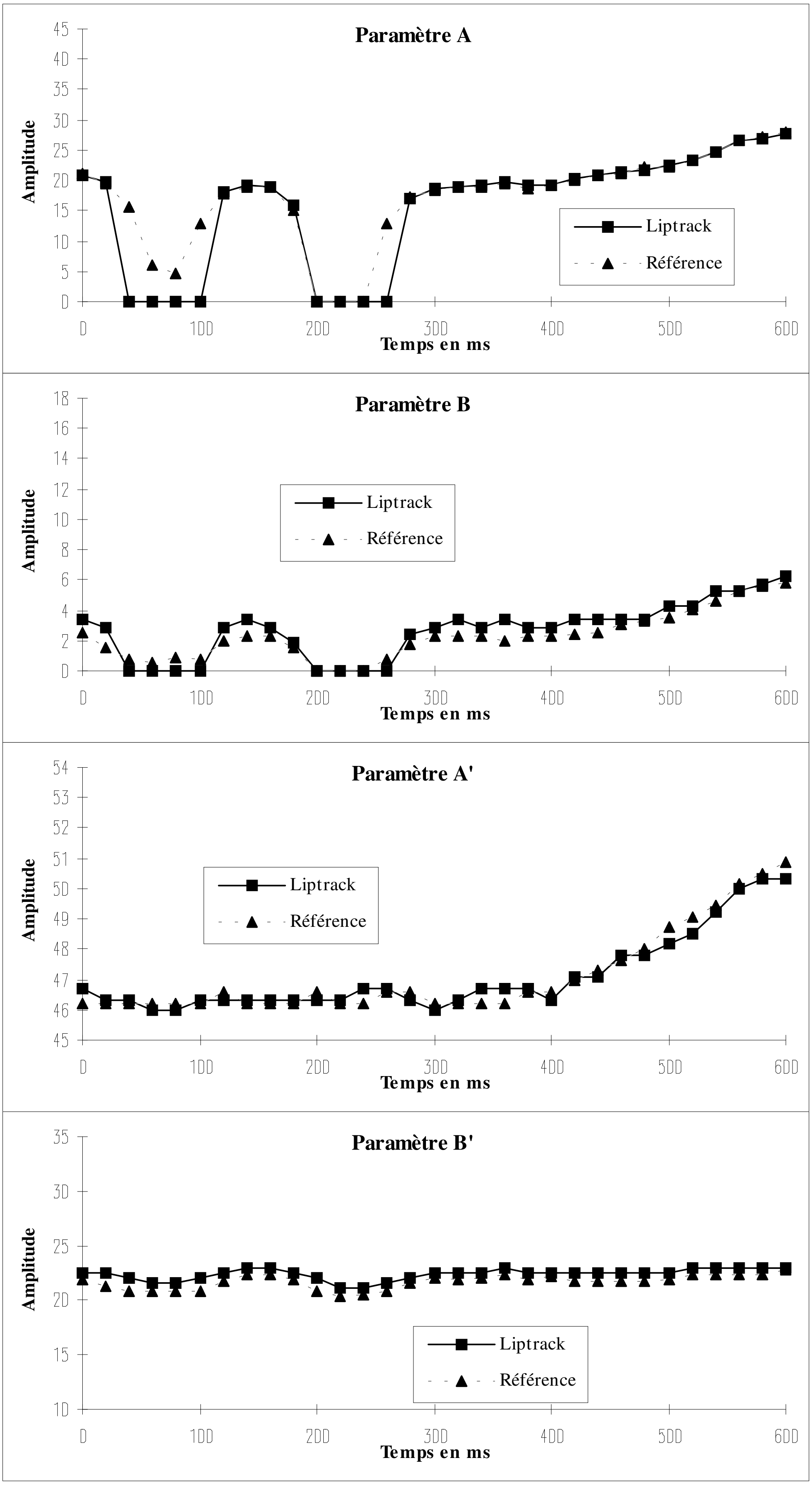


Figure 5.139 : comparaison des paramètres pour le stimulus "UBU".

3.2.1.1 Quantification de la différence de mesure

Une méthode efficace pour valider les résultats du Liptrack est de calculer, pour toutes les trames, la différence par rapport aux mesures du poste visage-parole. Pour cela, nous avons utilisé la relation suivante :

Equation 5.47
$$e = \frac{\left| p_{ref} - p_{lip} \right|}{dyn(p_{ref})} \times 100$$

dans laquelle p_{ref} est la mesure de référence, p_{lip} est la mesure donnée par le Liptrack et $dyn(p_{ref})$ représente la dynamique des paramètres de référence.

Equation 5.48
$$dyn(p) = \max_i(p) - \min_i(p)$$

où *i* est le nombre de trames analysées sur l'ensemble du corpus. On a ainsi trouvé :

dyn(A) = 45 mm

dyn(B) = 18 mm

dyn(A') = 9 mm

dyn(B') = 15 mm

Cette méthode de calcul permet de quantifier les différences relatives de mesure. Les tableaux suivants représentent les résultats de ces calculs. Les différences sont en pourcentage.

On remarque l'existence d'erreurs importantes sur certaines trames (marquées par une flèche). Les raisons de ces différences sont expliquées par la suite.

	Différence A	Différence B	Différence A'	Différence B'
	1.1	2.8	2.0	2.0
	1.1	2.0	2.0	2.4
	1.2	0.6	2.0	4.2
	0.0	0.0	2.0	1.3
	0.0	0.0	2.1	1.3
	0.0	0.0	2.2	2.1
	0.0	0.0	2.0	3.7
	7.0	10.0	1.9	5.7
	1.4	0.4	2.1	4.6
	0.6	2.4	1.9	4.8
	0.9	4.4	1.8	4.1
	0.0	0.0	2.0	0.3
	0.0	0.0	2.2	1.9
	0.0	0.0	1.8	3.4
	⇒ 27.2	6.0	2.0	3.5
	1.3	2.4	2.0	2.7
	1.2	1.6	2.2	1.4
	1.9	2.4	2.0	3.0
	1.9	0.6	2.0	3.0
	1.1	1.9	2.3	2.3
	1.9	1.5	1.9	2.2
	1.9	1.2	2.0	3.7
	1.9	7.5	2.3	1.2
	1.9	3.5	2.0	4.9
	1.2	1.2	6.1	3.3
	1.2	4.0	10.2	4.7
	1.2	4.0	2.2	5.2
	2.0	0.1	2.0	4.1
	1.2	2.6	5.9	1.4
	1.2	0.2	1.9	4.5
	1.2	2.4	1.9	5.1
min	**0.0**	**0.0**	**1.8**	**0.3**
max	**27.2**	**10.0**	**10.2**	**5.7**
moyenne	**2.1**	**2.1**	**2.6**	**3.2**

Tableau 2.14 : différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "ABA".

	Différence A	Différence B	Différence A'	Différence B'
	0.0	1.2	2.4	4.7
	0.1	1.7	1.6	6.6
	0.1	1.7	1.5	5.1
	0.1	4.6	1.4	4.4
	0.6	3.5	1.4	5.3
	0.0	0.0	1.6	5.9
	0.0	0.0	5.5	1.0
	0.0	0.0	5.7	9.6
	0.0	0.0	2.1	5.9
	0.1	1.2	2.0	5.0
	0.6	1.1	1.9	4.2
	0.6	0.1	2.1	0.9
	0.6	0.7	2.1	0.5
	0.8	1.9	2.2	1.9
	0.0	0.0	2.2	8.5
	0.0	0.0	2.3	0.5
	0.0	0.0	1.6	4.4
	1.2	0.4	2.3	3.0
	0.5	0.8	2.2	8.6
	1.4	2.0	1.7	5.5
	0.6	0.7	2.4	3.5
	0.6	0.6	2.4	8.0
	0.6	2.0	2.4	3.1
	0.2	0.6	2.3	8.0
	0.6	0.7	2.3	3.5
	0.6	0.7	1.9	3.1
	0.6	0.7	1.9	5.2
	0.7	0.5	2.0	6.5
	0.7	1.8	5.8	3.1
	0.9	1.7	6.1	4.7
	0.7	1.5	1.9	0.3
	0.0	1.3	2.0	4.7
	0.8	0.1	2.0	4.5
	0.0	1.2	2.0	4.9
min	**0.0**	**0.0**	**1.4**	**0.3**
max	**1.4**	**4.6**	**6.1**	**9.6**
moyenne	**0.4**	**1.0**	**2.4**	**4.5**

Tableau 5.15 : différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "UBA".

	Différence A	Différence B	Différence A'	Différence B'
	1.8	0.9	5.7	2.0
	1.0	2.2	1.8	2.0
	1.9	1.0	2.0	1.0
	0.1	4.7	6.1	4.9
	0.0	0.0	2.2	6.4
	0.0	0.0	1.9	6.2
	0.0	0.0	6.4	5.4
	2.9	3.6	2.3	4.9
	0.5	2.0	1.7	5.0
	0.6	2.0	1.6	5.2
	0.6	4.9	1.6	4.9
	0.0	0.0	2.4	8.5
	0.0	0.0	2.4	8.9
	0.0	0.0	2.2	6.2
	1.4	3.6	2.2	4.7
	0.9	4.0	2.1	5.0
	0.6	1.1	2.0	3.3
	1.3	3.1	5.9	4.2
	1.2	2.1	2.0	2.9
	1.2	0.6	1.9	4.8
	1.2	1.8	1.9	0.0
	1.2	1.2	6.0	7.2
	0.4	0.8	2.0	5.6
	1.2	2.3	1.8	4.9
	1.2	1.2	1.9	7.4
	1.2	0.1	2.0	4.0
	1.2	0.1	6.2	5.3
	1.2	2.6	2.1	6.8
	0.4	0.2	2.2	5.0
	0.4	2.3	1.8	4.9
	0.4	0.8	2.0	7.2
	1.2	0.5	1.9	5.7
	1.2	0.9	6.2	1.7
	1.2	2.3	1.8	4.8
min	**0.0**	**0.0**	**1.6**	**0.0**
max	**2.9**	**4.9**	**6.4**	**8.9**
moyenne	**0.9**	**1.6**	**2.8**	**4.9**

Tableau 5.16 : différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "ABU".

	Différence A	Différence B	Différence A'	Différence B'
	1.2	2.2	2.1	1.5
	2.0	3.7	6.2	0.3
	0.5	2.3	1.9	1.8
	2.1	2.1	2.3	2.2
	1.3	0.4	2.2	2.3
	0.5	3.0	6.2	5.2
	0.5	1.5	2.1	1.8
	0.5	2.8	6.2	5.9
	1.3	0.3	2.2	2.5
	0.6	1.9	6.2	5.5
	1.3	3.3	6.2	2.2
	0.6	1.8	6.2	5.2
	0.5	1.8	2.2	2.2
	0.6	3.1	2.2	3.8
	1.4	1.7	6.2	5.5
	0.6	1.7	2.2	1.3
	0.6	3.0	2.2	5.3
	0.3	3.1	2.2	5.3
	1.3	0.4	2.2	2.0
	1.3	3.3	2.1	2.0
	1.3	1.8	1.9	5.4
	1.2	1.7	2.1	2.1
	1.2	1.3	6.1	1.6
	0.4	0.4	6.0	1.0
	0.4	0.6	6.1	2.3
	1.2	2.8	2.1	4.8
	1.2	2.2	5.9	6.5
	0.4	0.9	2.1	4.9
	1.2	2.4	1.8	4.9
	1.2	0.2	6.1	2.5
	1.2	0.2	6.0	4.9
	0.4	2.4	6.1	1.0
	1.2	2.4	6.1	4.9
	0.4	1.0	2.1	4.0
min	**0.3**	**0.2**	**1.8**	**0.3**
max	**2.1**	**3.7**	**6.2**	**6.5**
moyenne	**0.9**	**1.9**	**3.9**	**3.4**

Tableau 5.17: différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "AJA".

	Différence A	Différence B	Différence A'	Différence B'
	2.7	1.2	2.5	3.8
	1.9	1.9	5.7	8.8
	1.9	0.9	2.4	6.6
	1.1	4.0	1.7	5.2
	0.3	1.2	5.6	8.1
	1.1	0.1	1.7	7.3
	1.9	4.0	5.5	2.5
	0.3	0.4	1.7	4.9
	1.9	2.6	2.1	3.6
	1.1	3.7	5.7	1.1
	1.1	1.0	5.6	0.9
	1.1	2.4	5.8	4.9
	1.9	0.2	5.8	6.5
	1.1	2.6	5.8	5.3
	1.9	4.0	5.8	2.3
	1.1	0.1	2.0	2.5
	1.9	4.5	1.9	2.0
	1.1	1.2	1.9	5.3
	2.0	0.4	2.1	4.9
	2.7	1.0	2.0	4.9
	1.2	2.3	2.0	6.4
	1.1	0.4	5.9	4.9
	1.1	1.0	2.2	2.0
	1.9	1.6	1.9	5.1
	1.1	2.6	2.0	6.5
	1.9	0.1	6.2	4.1
	1.1	5.7	2.0	2.1
	1.9	0.1	2.1	5.4
	1.1	4.0	2.1	5.2
	1.1	1.2	2.1	2.2
	1.1	2.4	2.0	4.7
	1.2	0.4	2.0	4.9
	1.1	0.5	2.0	7.2
	1.2	0.5	1.8	4.0
min	**0.3**	**0.1**	**1.7**	**0.9**
max	**2.7**	**5.7**	**6.2**	**8.8**
moyenne	**1.4**	**1.8**	**3.2**	**4.6**

Tableau 5.18: différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "IZI".

	Différence A	Différence B	Différence A'	Différence B'
	0.9	4.6	5.4	4.5
	0.6	7.5	1.3	8.3
	⇒35.0	4.2	1.3	8.4
	13.2	2.8	2.7	5.3
	10.6	4.7	2.7	5.3
	⇒28.7	4.2	1.2	8.4
	0.6	4.7	2.5	5.2
	0.5	6.0	1.2	4.7
	0.3	3.3	1.2	4.5
	2.1	2.2	1.5	4.5
	0.0	0.0	2.5	8.7
	0.0	0.0	1.2	5.4
	0.0	0.0	5.3	4.4
	⇒28.7	4.2	1.5	5.0
	0.3	3.5	2.5	2.9
	0.6	3.3	2.8	3.3
	0.2	6.0	1.2	3.9
	0.6	3.3	5.3	3.3
	0.6	7.4	5.3	4.9
	1.4	3.3	1.4	3.9
	0.2	8.9	2.7	2.9
	0.7	5.2	1.2	5.2
	0.1	4.6	2.6	5.2
	0.7	1.8	1.7	5.2
	0.9	0.4	2.4	5.2
	0.7	4.3	6.4	4.5
	0.0	1.5	6.2	5.0
	0.0	4.0	2.1	5.0
	0.0	0.2	1.9	4.5
	0.7	1.0	1.9	4.1
	0.7	2.3	5.9	1.8
	0.7	2.3	2.1	2.6
	0.7	2.3	2.0	4.4
	0.1	6.7	1.8	4.8
min	**0.0**	**0.0**	**1.2**	**1.8**
max	**35.0**	**8.9**	**6.4**	**8.7**
moyenne	**3.9**	**3.6**	**2.7**	**4.8**

Tableau 5.19: différences en % entre les paramètres mesurés par le Liptrack et ceux obtenus avec le poste visage-parole pour le stimulus "UBU".

3.2.1.2 Interprétation des résultats

On peut noter, d'après les résultats graphiques et ceux obtenus par calcul de différence, que les paramètres du contour interne, c'est à dire A et B, sont particulièrement proches de ceux de référence (Tableau 5.20). Dans tous les cas, les deux courbes sont pratiquement superposées et les différences sont très faibles. Nous avons vu que ce sont ces paramètres qui transmettent une grande partie de l'information visuelle du mouvement des lèvres.

Les courbes de comparaison des paramètres du contour externe témoignent d'une erreur plus importante. Ces paramètres ont en général peu de dynamique et même le fait de traiter les trames paires et impaires peut introduire des fluctuations.

Globalement, les principales raisons qui expliquent les différences entre le Liptrack et le poste visage-parole sont les suivantes :

• Le signal d'entrée utilisé dans cette expérience est obtenu à partir de la sortie du chroma-key analogique. Les lèvres sont donc déjà binarisées. Pour traiter ce signal, nous avons chargé une LUT qui filtre tous les pixels sauf ceux qui sont codés sur le niveau 0 (donc une seule valeur non nulle dans la LUT à l'adresse 0). Une différence de précision sur les canaux de couleur en sortie du CAN (5 bits pour R et B, 6 bits pour V) peut entraîner des erreurs sur les frontières des lèvres. De plus, les frontières externes des lèvres contiennent de très hautes fréquences (passage du niveau clair au niveau noir) qui peuvent introduire des effets de repliements de spectres à la numérisation (un filtre analogique est prévu par le constructeur du CAN pour pallier ce problème). Ce qui fait que certains pixels appartenant aux lèvres n'ont pas le niveau 0 après conversion et sont filtrés par la LUT.

• Les algorithmes de calcul implémentés sur le Liptrack sont des versions simplifiées de celles qui sont utilisées par le poste visage-parole. En effet, Lallouache utilise des filtres medians avant le traitement pour minimiser les erreurs et éliminer les pixels isolés des contours.

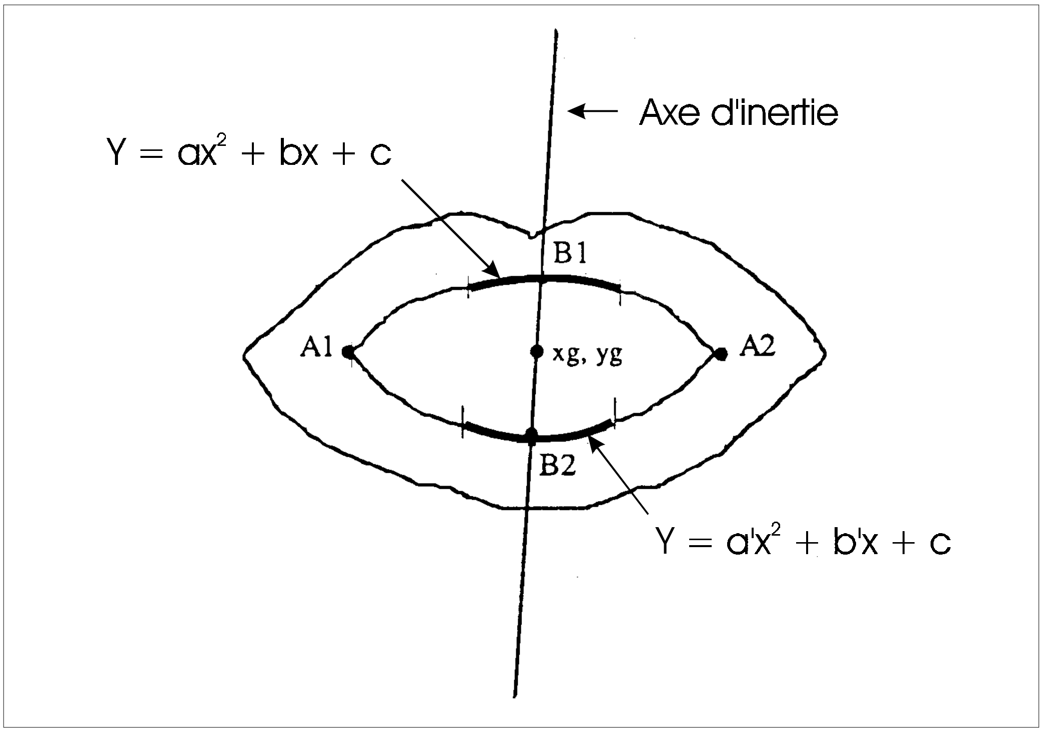


Figure 5.140: extraction des paramètres des lèvres (d'après Lallouache 1991).

Pour détecter les points B1 et B2, deux pixels du contour interne les plus proches de l'axe de l'inertie secondaire sont sélectionnés. Autour de ces deux points, une fonction polynomiale de second degré au sens des moindres carrés est appliquée. L'intersection de ces deux équations avec l'axe d'inertie donne les coordonnées des deux points B1 et B2. Cette technique permet d'obtenir des mesures lissées sur le paramètre B.

• Les paramètres du contour interne sont calculés suivant l'axe d'inertie des lèvres alors que Liptrack mesure la distance verticale entre le point le plus haut et le plus bas du contour pour le paramètre B et la distance horizontale entre les deux coins des lèvres pour le paramètre A. Compte tenu du fait que notre locuteur a des lèvres très symétriques, l'erreur due à l'utilisation des points extrêmes du contour interne sans tenir compte de l'axe d'inertie peut être pratiquement négligée. Les résultats des tableaux précédents sur les paramètres du contour interne le confirment.

• Le paramètre B' a un taux d'erreur plus important car le Liptrack et la version logicielle du poste visage-parole que nous avons utilisée ne mesurent pas exactement les mêmes distances pour ce paramètre. Pour le poste visage-parole, la mesure B' est la distance entre le point le plus bas de l'arc de Cupidon et le point le plus bas du contour externe alors que le Liptrack mesure la distance entre le point le plus haut et le plus bas du contour externe.

Comme nous connaissons cette source d'erreur, nous avons donc ajouté une valeur d'offset de deux pixels (ce qui correspond à peu près à la hauteur de l'arc de Cupidon dans une image non-entrelacée de notre corpus) au paramètre B' du Liptrack pour le rendre compatible avec la mesure du poste visage-parole. La hauteur de l'arc de Cupidon n'est pas tout à fait constante pendant le processus de la parole, donc l'erreur sur le paramètre B' ne peut pas être corrigée avec un simple décalage. Les résultats sur la moyenne des différences sur ce paramètre confirme l'existence d'une erreur plus importante par rapport aux autres paramètres.

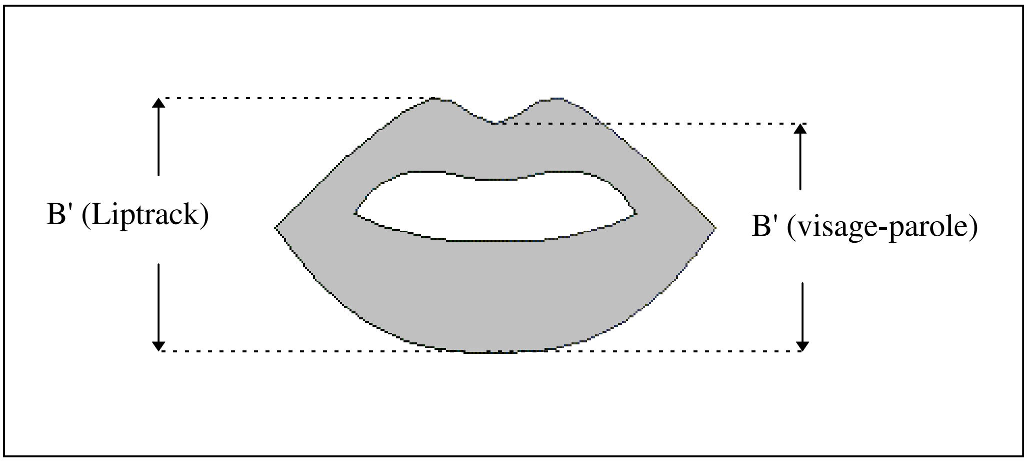


Figure 5.141 : le paramètre B' mesuré par le Liptrack et le poste visage-parole.

Par ailleurs, on peut noter l'existence d'une grande différence (35%) sur les paramètres du contour interne pour le stimulus "UBU" au moment de la première fermeture des lèvres dans la séquence "UBUBUZE". Nous avons cherché l'origine de cette grande différence car elle ne s'explique pas par des raisons évoquées plus haut. Dans cet exemple, le poste visage-parole n'a pas détecté la fermeture des lèvres alors que le Liptrack a obtenu la valeur nulle pour les deux paramètres du contour interne. Comme on l'a vu précédemment, il suffit que la hauteur du contour interne soit inférieure à trois pixels pour que le Liptrack détecte des lèvres fermées (cette condition est essentielle pour le filtrage du bruit sur les pixels des lèvres). Dans le cas du poste visage-parole, cette condition n'est pas appliquée ce qui permet de détecter les contours de moins de trois pixels de hauteur. Dans le cas de l'exemple du stimulus "UBU", la hauteur du contour interne est effectivement très faible et on mesure des valeurs nulles avec le Liptrack. La Figure 5.142 montre les lèvres au moment de la prononciation de /b/ dans "UBU". Dû à une défaillance de maquillage, au moment de la protrusion une petite zone au milieu des lèvres échappe au chroma-key. Elle est classée comme externe à la région du vermillon et c'est pour cette raison que le poste visage-parole obtient des valeurs non nulles alors qu'en réalité les lèvres sont fermées.

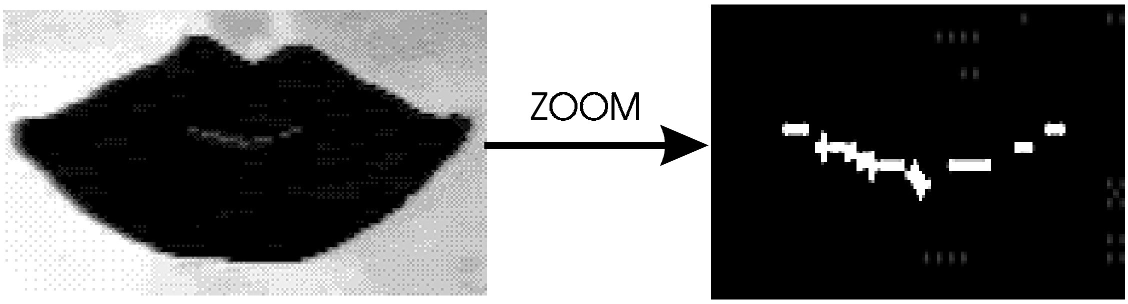


Figure 5.142 : exemple de cas où la défaillance du maquillage induit des erreurs de mesure. Les lèvres sont en réalité fermées mais à cause de la protrusion, la région non-maquillées des lèvres devient visibles et au moment du traitement, on détecte une ouverture.

Dans cet exemple, il s'agit réellement de lèvres fermées donc on peut dire que le Liptrack a "corrigé" le problème de maquillage. Mais on peut envisager le cas inverse : que se passe t-il quand le contour interne existe réellement et fait moins de trois pixels en hauteur ? Bien évidemment, le Liptrack détecte encore des lèvres fermées (on n'a pas rencontré une telle situation pour mieux illustrer le problème). Mais malgré tout cela, nous n'avons pas modifié notre algorithme pour les raisons suivantes :

• L'élimination de la condition "trois pixels consécutifs sinon rien" introduit des erreurs sur la détection du contour interne quand la région du vermillon est bruitée. Dans ce cas, les mesures A et B risquent d'être fortement bruitées.

• Le Liptrack est conçu pour être intégré sur un support lié à un casque porté par le locuteur. La région des lèvres doit en principe occuper toute la surface de l'image. Dans une telle application, une erreur de trois pixels est tout à fait négligeable.

En guise de conclusion sur cette partie, nous pouvons résumer les résultats des tableaux précédents. Nous avons calculé la moyenne et l'écart type des différences sur l'ensemble des trames des cinq stimuli de comparaison.

	Paramètre A	Paramètre B	Paramètre A'	Paramètre B'
Moyenne des différences	1.74%	2.07%	2.96%	4.12%
Ecart type des différences	4.67%	1.91%	1.77%	1.98%

Tableau 5.20 : moyennes et écart types des différences de mesures par rapport au poste de Lallouache (1991) en pourcentage.

Ces résultats montrent encore que la moyenne des différences sur les paramètres du contour interne est particulièrement faible. On note que celle relative au contour externe est globalement plus importante, mais reste tout de même raisonnable.

Pour vérifier la transmission fidèle des informations visuelles des mouvements des lèvres, nous avons testé les mesures du Liptrack dans une application de reconnaissance visuelle de la parole. La section suivante présente les résultats de cette expérience.

3.2.2 Evaluation dans une application de reconnaissance

Nous avons conçu le Liptrack avant tout pour une application de reconnaissance de la parole audiovisuelle. Il est donc très important de tester ses performances dans ce genre d'utilisation.

Afin d'obtenir des résultats de reconnaissance concluants, nous avons procédé à plusieurs expériences :

1. <u>Apprentissage de deux séries de modèles markoviens avec les données du Liptrack et ceux du poste visage-parole et test avec les données respectives</u>

La Figure 5.143 illustre le principe de cette expérience.

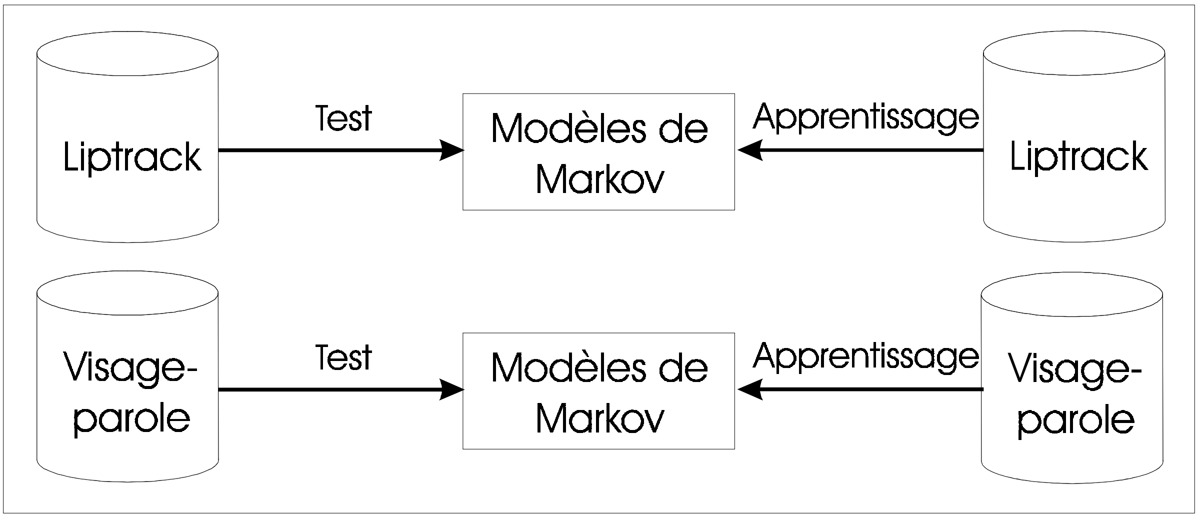


Figure 5.143 : une expérience de reconnaissance pour comparer les résultats de test. Chaque modèle est entraîné et testé avec les données d'une même origine.

Les résultats de cette expérience sont présentés sur la Figure 5.144

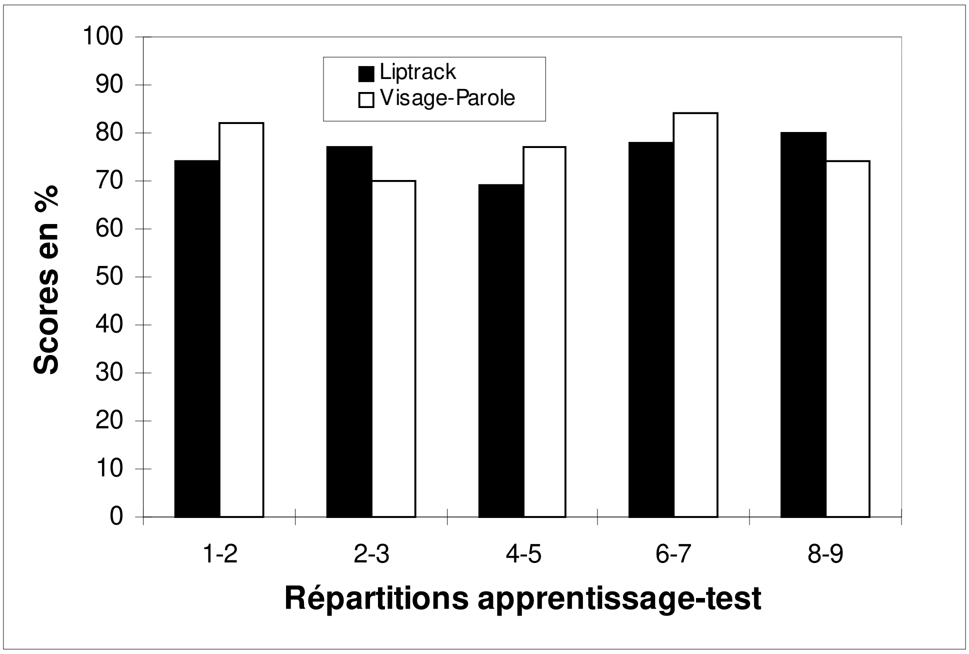


Figure 5.144 : résultats de reconnaissance visuelle. Les données de test et d'apprentissage sont de même origine.

Les moyennes et les écart-types des résultats de la Figure 5.144 sont présentés dans le Tableau 5.21 :

	Liptrack	Visage-parole
Moyenne des scores	75.6%	77.4%
Ecart-type des scores	4.28%	5.73%

Tableau 5.21 : moyennes et écart-types des scores de reconnaissance obtenus par le Liptrack et le poste visage-parole.

2. Apprentissage avec les données de visage-parole et test avec celles du Liptrack :

La Figure 5.145 montre le principe de cette expérience.

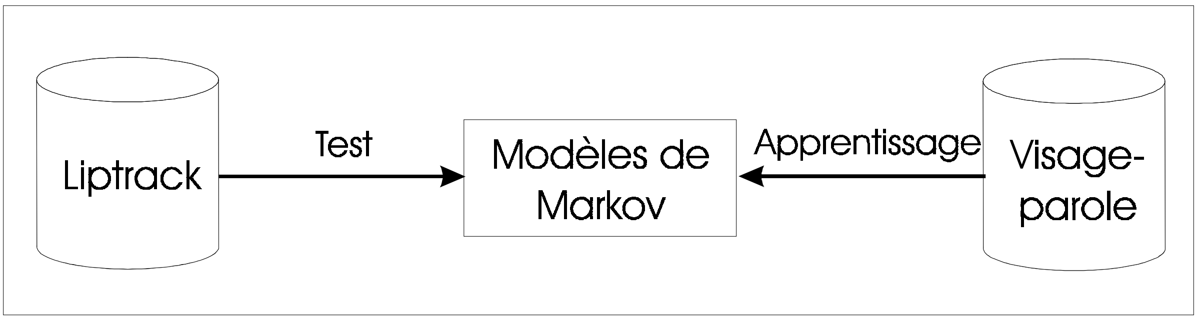


Figure 5.145 : résultats d'expériences de reconnaissance visuelle. L'ensemble d'apprentissage est constitué avec les données du poste visage-parole alors que le test s'effectue sur les données du Liptrack.

Les résultats sont présentés sur la Figure 5.146. Les scores de reconnaissance du système entraîné et testé avec les données du poste visage-parole sont reportés sur cette figure à titre de comparaison.

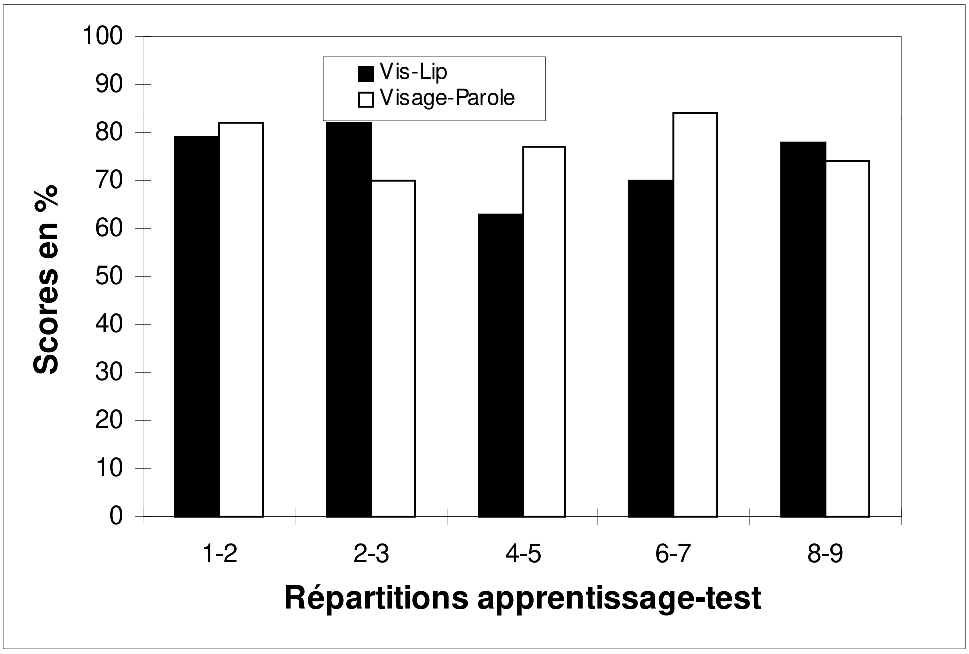


Figure 5.146 : résultats de test sur des modèles de Markov entraînés avec les données du poste visage-parole et testés avec celles du Liptrack (courbe Vis-Lip). La courbe "visage-parole" se réfère aux résultats précédents d'apprentissage et test sur les données de Lallouache.

	Visage-parole -Liptrack	Visage-parole
Moyenne des scores	74.4%	77.4%
Ecart-type des scores	7.77%	5.73%

Tableau 5.22 : moyenne et écart-type des scores de reconnaissance obtenus sur des modèles entraînés avec les données du poste visage-parole et téstés avec les données du Liptrack. La moyenne et l'écart-type des scores du visage-parole sont présentés en titre de comparaison.

3- Apprentissage avec les données du Liptrack et test avec celles du poste visage-parole.

La Figure 5.147 décrit cette expérience.

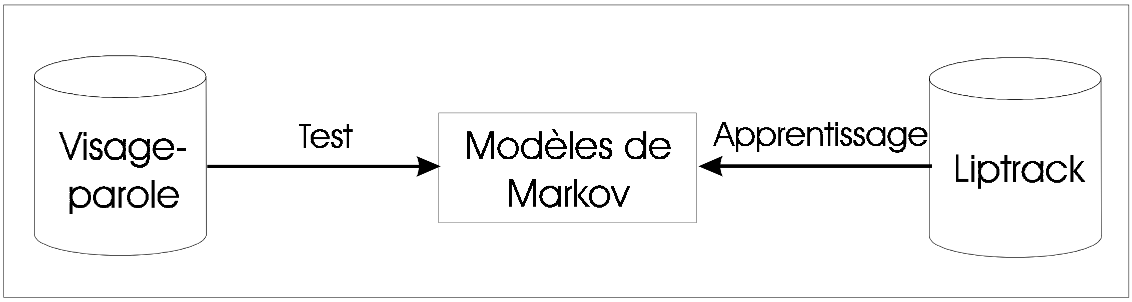


Figure 5.147 : une expérience de reconnaissance visuelle. Les données d'apprentissage proviennent du poste visage-parole et celles de test sont obtenues par le Liptrack.

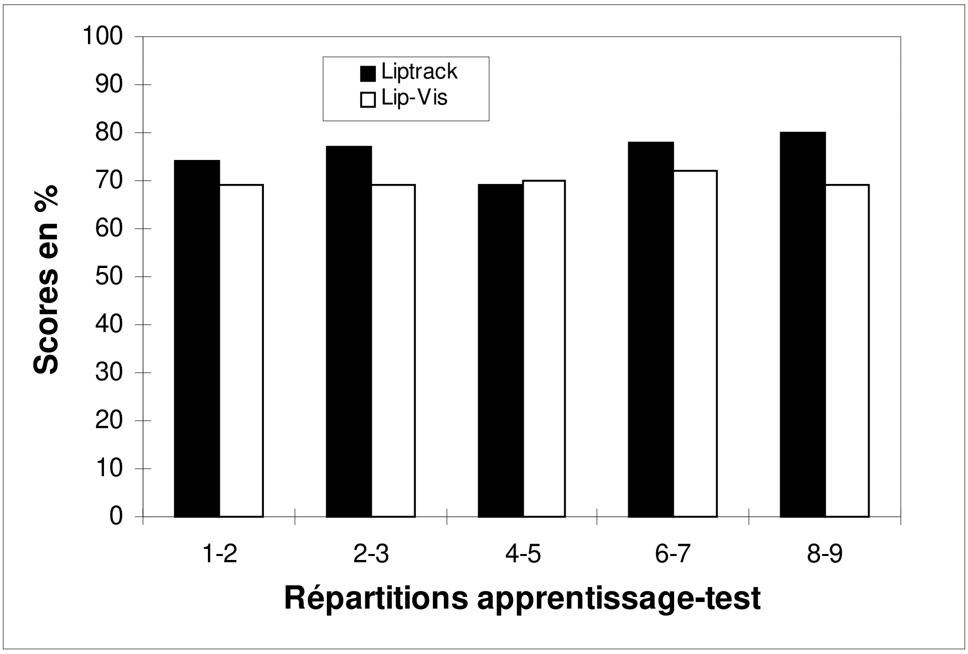


Figure 5.148 : une autre expérience de reconnaissance visuelle. L'apprentissage a été effectué sur les paramètres du Liptrack et les tests ont été élaborés sur les données du poste visage parole (courbe Lip-Vis).

Les scores de reconnaissance de cette expérience ainsi que ceux obtenus avec les données d'apprentissage et de test du Liptrack sont présentés sur la Figure 5.148.

3.2.2.1 Discussion

Nous avons effectué le calcul statistique de *t-test* ou test de Student pour comparer les résultats du Liptrack à ceux obtenus avec le poste de Lallouache (1991). La comparaison des scores de la première avec le calcul de t-test a donné le résultat suivant :

$t(4) = 0.411$ avec $p > 0.05$

ce qui signifie qu'il n'existe pas de différence significative entre les deux ensembles de données. Nous avons effectué le même test avec les données de la deuxième et de la troisième expérience, c'est à dire apprentissage et test avec des données d'origines différentes :

$t(4) = -1.193$ avec $p > 0.05$

Cette expérience prouve aussi que la différence n'est pas significative.

Les performances du système électronique Liptrack sont donc tout à fait comparables à celles du poste visage-parole de Lallouache (1991). Ce prototype est capable de restituer fidèlement les informations d'intelligibilité des mouvements des lèvres.

4. Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques techniques et les performances du système électronique Liptrack. Les différents blocs matériels ainsi que leur fonctionnement ont été détaillés. Les expériences ont montré que Liptrack est capable de traiter 50 trames par seconde (pour une entrée vidéo 625 lignes). Les résultats de l'évaluation (calcul d'erreur et tests de reconnaissance) sont très satisfaisants et témoignent d'une grande fiabilité de mesures obtenues par ce système.

Grâce à ses capacités de calcul temps réel et sa précision sur les paramètres labiaux, le Liptrack peut être intégré dans une application de reconnaissance ou de synthèse de visage parlant (applications en visiophonie). Cependant, la contrainte de maquillage des lèvres en bleu peut limiter son application. Des études sont en cours à l'ICP pour éliminer cette contrainte et concevoir un système électronique entièrement intégré capable de calculer les paramètres labiaux en temps réel.

Conclusion

Dans cette thèse, nous avons tenté de répondre à deux questions fondamentales dans le domaine de la reconnaissance audiovisuelle de la parole :

1- quelle est la meilleure architecture d'intégration audiovisuelle qui permet d'approcher au mieux les résultats de la perception chez l'homme ?

2- quels sont les paramètres les plus pertinents pour transmettre l'intelligibilité de la parole audiovisuelle ?

Pour répondre à la première question, nous avons proposé et implémenté une série d'architectures d'intégration audiovisuelle. Les résultats de nos expériences ont montré que l'intégration des deux modalités acoustique et visuelle avant classification ne peut pas donner de résultats satisfaisants. Les informations ainsi fusionnées ne permettent pas d'exploiter au mieux les propriétés audiovisuelles de la parole. Les scores de reconnaissance audiovisuelle obtenus avec ce type de systèmes ne sont pas supérieurs à ceux obtenus avec chaque modalité prise séparément, dans toutes les conditions expérimentales (niveau de bruit).

Nous avons montré que le principe d'intégration après classification peut satisfaire cette condition, à savoir AV > A et V. Nous avons donc proposé des architectures de fusion dans ce type de systèmes où le traitement des deux modalités s'effectue séparément. Les résultats ont montré que notre système à pondération de chaque canal peut s'adapter automatiquement à tous les niveaux de dégradation acoustique sans apprentissage. Il est capable de conserver les bonnes performances de reconnaissance de la modalité acoustique quand l'entrée de ce canal est claire, et de s'appuyer davantage sur une lecture labiale seule quand l'information acoustique est fortement perturbée par le bruit ambiant.

La paramétrisation des lèvres a été une étape fondamentale dans notre étude. Elle nous a permis d'élaborer le modèle de lèvres en 3D et d'extraire l'information visible des mouvements labiaux. En effectuant de nombreuses expériences de reconnaissance labiale seule, nous avons identifié les paramètres qui transmettent une grande quantité d'information sur l'intelligibilité de la parole visuelle.

Avec une réalisation technologique, nous avons montré que précision et rapidité de calcul ne sont pas incompatibles. Notre système électronique est capable de calculer avec une grande précision les paramètres labiaux et de les transmettre en temps réel vers une machine hôte, dans des applications de reconnaissance automatique ou d'animation de visages synthétiques (visiophonie à bas débit).

Par ailleurs, les perspectives de ce travail sont nombreuses. Les performances de notre système restent à être évaluées dans un contexte plus étendu, par exemple un corpus de plus grande taille prononcé par plusieurs locuteurs. Notre plate-forme logicielle (Annexe 2), organisée de façon à minimiser les interventions de l'utilisateur, constitue un outil simple et efficace pour l'élaboration et le test de nouvelles architectures de reconnaissance audiovisuelle.

Finalement, nous pensons que notre travail représente une base solide pour le développement des interfaces homme/machine impliquant la reconnaissance de la parole multimodale et les technologies multimédia.

Références bibliographiques

Abbs et Gilbert, (1973)

H. Abbs and B.N. Gilbert. A strain gauge transduction system for lip and jaw motion in two dimensions: Design criteria and calibration data. *Journal of Speech and Hearing Research,* 16:248-256.

Abry et Boë, (1986)

C. Abry and L.J. Boë. *Laws for lips*. Speech Communication, 5:97-104.

Abry et Boë, (1980)

C. Abry and L.J. Boë. *A la recherche de corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français. Labialité et Phonétique,* pages 217-237. In C. Abry, L.J. Boë, P. Corsi, R. Descout, M. Gentil and P. Graillot, editors, *Publications de l'Université des Langues et Lettres de Grenoble*, France.

Abry et Lallouache, (1991)

C. Abry and M.T. Lallouache. Audibility and stability of articulatory movements: Deciphering two experiments on anticipatory rounding in French. In proc. *of the 12th International Congress of Phonetic Sciences,* Aix en Provence, France.

Adjoudani et Benoît, (1995)

A. Adjoudani and C. Benoît. *Audio-visual speech recognition compared across two architectures*. In Proc. of the 4th EUROSPEECH Conference, volume 3, pages 1563-1566, Madrid, Spain.

Adjoudani et Benoît, (1996)

A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 461-472. Springer-Verlag, Berlin.

Adjoudani, (1993)

A. Adjoudani. *Élaboration d'un modèle de lèvres 3D pour animation en temps réel.* Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique de Grenoble, France.

Adjoudani et al., (1994)

A. Adjoudani, T. Guiard-Marigny, B. Le Goff, and C. Benoît. Un modèle 3D de lèvres parlantes. In *Actes des XXèmes Journées d'Étude sur la Parole,* pages 143-146, Société Française d'Acoustique, Trégastel, France.

Alissali et al., (1996)
M. Alissali, P. Deléglise, and A. Rogozan. Asynchronous integration of visual information in an automatic speech recognition system. *In proc. of 4th ICSLP*, Philadelphia, USA.

Angola, (1993)
O. Angola. *Analyse labiométrique en temps réel par traitement d'images vidéo.* Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique de Grenoble, France.

Benoît et al., (1991)
C. Benoît, C. Abry and L.J. Boë. The effect of context on labiality in French. In Proc. *of the 2nd EUROSPEECH Conference,* volume 1, pages 153-156, Genova, Italy.

Benoît et al., (1992)
C. Benoît, M.T. Lallouache, T. Mohamadi, and C. Abry. A set of French visemes for visual speech synthesis. *Talking Machines,* pages 485-504. Elsevier Science Publishers B.V., North-Holland, Amsterdam.

Benoît et al., (1994a)
C. Benoit, A. Adjoudani, O. Angola, T. Guiard-Marigny, and B. Le Goff. Perception, analysis and synthesis of talking lips. In *Proc. of the IMAGINA '94 Conference*, pages 142-163. Institut National de l'Audiovisuel, France.

Benoît et al., (1994b)
C. Benoît, T. Mohamadi, and S.D. Kandel. Effects of phonetic context on audio-visual intelligibility in French. *Journal of Speech and Hearing Research*, 37:1195-1203.

Bercier, (1996)
B. Bercier. *Détection des dents sur une séquence vidéo couleur*. Mémoire d'ingénieur, ENSERG, Institut National Polytechnique de Grenoble, France.

Bernstein et Benoît, (1996)
L. E. Bernstein and C. Benoît. For speech perception by humans and machines, three senses are better than one. In proc. *of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Bertelson et al., (1994)
P. Bertelson, J. Vroomen, G. Wiegeraad, and B. De Gelder. Exploring the relation between McGurk interference and ventriloquism. *International Conference on Spoken Language Processing*, II:559-562, Yokohama, Japan.

Binnie et al., (1974)
C.A. Binnie, A.A. Montgomery, and P.L. Jackson. Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research,* 17:619-630.

Bouchet et Cuilleret, (1972)
A. Bouchet and J. Cuilleret. *Anatomie topographique descriptive et fonctionnelle.* éditions Simep, Villeurbanne, France.

Bouffartigue, (1994)
C. Bouffartigue. *Acquisition du visage parlant en vidéo rapide et extraction des paramètres de face des lèvres.* Projet d'Ingénieur, ENSERG, Institut National Polytechnique, Grenoble, France.

Bregler et al., (1993a)
C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In proc. of the *International Conference on Acoustics, Speech, and Signal Processing,* volume 1, page 557-560, Minneapolis, USA.

Bregler et al., (1993b)
C. Bregler, H. Hild, S. Manke, and A. Waibel. Bimodal sensor integration on the example of speechreading. In Proceedings of the *IEEE International Conference on Neural Networks,* volume 2, pages 667-671.

Bregler et Konig, (1994)
C. Bregler and Y. Konig. Eigenlips for robust speech recognition, in proc. of *ICASSP*, pages 669-672, Adelaïde, Australia.

Brooke, (1996a)
N. M. Brooke. Talking heads and speech recognisers that can see: The computer processing of visual speech signals. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 351-372. Springer-Verlag, Berlin.

Brooke, (1996b)
N. M. Brooke. Using the visual component in automatic speech recognition. In proc. *of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Brooke et Petajan, (1986)
N. M. Brooke and E. D. Petajan. Seeing speech: Investigation into the synthesis and recognition of visible speech movement using automatic image processing and computer graphics. In proc. of the *International Conference on Speech Input and Output: Techniques and Applications*, N° 258, pages 104-109, London.

Brooke et Summerfield, (1983)
N. M. Brooke and Q. Summerfield. Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics,* 11:63-76.

Brooke et al., (1994)
N. M. Brooke, M. J. Tomlinson, and R. K. Moore. Automatic speech recognition that includes visual speech cues. *Proceedings of the Institute of Acoustics,* 16(5):15-22. Autumn Meeting, Windermere.

Cady et Kron, (1971)

J. Cady et B. Kron. *Anatomie du corps humain*. Editions Maloine, Paris, France.

Campbell, (1974)

H.W. Campbell. *Phoneme recognition by ear and by eye: a distinctive feature analysis.* PhD thesis, Katholieke Universiteit te Nijmegen, Nijmegen.

Campbell, (1988)

R. Campbell. Tracing lip movements: A theoretical approach to lipreading in speech perception. *Visible Language,* 22: 32-57.

Campbell et Dodd, (1980)

R. Campbell and B. Dodd. Hearing by eye. *Quarterly Journal of Experimental Psychology,* 32:509-515.

Cathiard, (1988)

M. A. Cathiard. *Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-rétraction des lèvres en français*. Mémoire de maîtrise, département de Psychologie, Université Grenoble II, France.

Cathiard, (1994)

M. A., Cathiard. *La perception visuelle de l'anticipation des gestes vocaliques : cohérence des événements audibles et visibles dans le flux de la parole*. Thèse de doctorat en Psychologie Cognitive, Université Pierre Mendès-France, Grenoble, France.

Cathiard et al., (1991)

M. A. Cathiard, G. Tiberghien, R. Girot-Tseva, M. T. Lallouache, and P. Escudier. Visual perception of anticipatory rounding during acoustic pauses: A cross-language study. In proc. *of the 12th International Conference of Phonetic Sciences,* volume 4, pages 50-53.

Chandramohan et Silsbee, (1996)

D. Chandramohan and P. L. Silsbee. A multiple deformable template approach for visual speech recognition. In proc. *of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Cohen et Massaro, (1993)

M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, Tokyo, Springer-Verlag.

Cohen et Massaro, (1994)

M. M. Cohen and D. W. Massaro. Development and experimentation with synthetic visible speech. *Behavioral Research Methods, Instruments and Computers*, 26(2):260-265.

Coianiz et al., (1996)

T. Coianiz, L. Torresani, and B. Caprile. 2D deformable models for visual speech analysis. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 391-398. Springer-Verlag, Berlin.

Cosi et Magno Caldognetto, (1996)

P. Cosi and E. Magno Caldognetto. Lips and jaw movements for vowels and consonants: Spatio-temporal characteristics and bimodal recognition applications. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 291-314. Springer-Verlag, Berlin.

Dalton et al., (1996)

B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 373-382. Springer-Verlag, Berlin.

Dodd, (1977)

B. Dodd. The role of vision in the perception of speech. *Perception,* 6:31-40.

Duchnowski et al., (1994)

P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lipreading. In International Conference on Spoken Language Processing, pages 547-550, Yokohama, Japan.

Erber, (1969)

N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research,* 12:423-425.

Erber, (1974)

N. P. Erber. Effects of angle, distance and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research,* 17:99-112.

Erber, (1975)

N. P. Erber. Auditory-visual perception of speech. *Journal of Speech and Hearing Research,* 40:481-492.

Escudier et al., (1990)

P. Escudier, C. Benoît, and M. T. Lallouache. Identification visuelle de stimuli associés à l'opposition /i/-/y/ : étude statique. In *Premier Congrès d'Acoustique* , volume 2, pages 541-544.

Finn, (1986)

K. E. Finn. *An investigation of visible lip information to be used in automated speech recognition* . PhD thesis, Georgetown University, Washington, DC.

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

Fromkin, (1964)
V. Fromkin. Lip positions in American-English vowels. Language and Speech, 7(3):215-225.

Fuster-Duran, (1996)
A. Fuster-Duran. Perception of Conflicting audio-visual speech: An examination across Spanish and German. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 135-144. Springer-Verlag, Berlin.

Gentil et Boë, (1979)
M. Gentil and L. J. Boë. Les lèvres et la parole : Données anatomiques et aspects physiologiques. *Rapport technique,* Institut de Phonétique de Grenoble, France.

Girin et al., (1995)
L. Girin, G. Feng, and J. L. Schwartz. Noisy speech enhancement with filters estimated from the speaker's lips. In proc. *4th EUROSPEECH Conference*, J. Pardo, editors, volume 2, pages 1559-1562, Madrid, Spain.

Goldschen, (1993)
A. J. Goldschen. *Continuous automatic speech recognition by lipreading.* PhD thesis, Department of Electrical and Computer Science, George Washington University, Washington, DC.

Goldschen et al., (1994)
A. J. Goldschen, O. Gracia, and E. D. Petajan. Continuous optical automatic speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers,* pages 572-577.

Goldschen et al., (1996)
A. J. Goldschen, O. Garcia, and E. D. Petajan. Rationale for phoneme-viseme mapping for feature selection in visual speech recognition. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 505-515. Springer-Verlag, Berlin.

Gouraud, (1971)
H. Gouraud. Computer display of curved surfaces. *Technical Report*, UTEC-CSc-71-113, U. Utah Computer Science Department.

Green, (1987)
K.P. Green. The perception of speaking rate using visual information from a talker's face. *Perception and Psychophysics*, 45: 34-42.

Green et al., (1990)
K.P. Green, E.B. Stevens, P.K. Kuhl, and A.M. Meltzoff. Exploring the basis of the McGurk effect: Can perceivers combine information from a female face and a male voice? *Journal of the Acoustical Society of America,* 87, S125.

Green et al., (1991)
K.P. Green, P.K. Kuhl, A.N. Meltzoff, and E.B. Stevens. Integrating speech information across talkers, gender, and sensory modality: Female face and male voices in McGurk effect. *Perception and Psychophysics,* 50, 524-536.

Green et Kuhl, (1989)
K.P. Green, P.K. Kuhl. The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics,* 45; 34-42.

Guiard-Marigny et Ostry, (1995)
T. Guiard-Marigny and D. J. Ostry. Three-dimensional visualization of human jaw motion in speech. In Proc. of the *129th Meeting of the Acoustical Society of America*, Washington, USA.

Guiard-Marigny, (1992)
T. Guiard-Marigny. *Animation en temps réel d'un modèle paramétrisé de lèvres.* Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Guiard-Marigny, (1996)
T. Guiard-Marigny. *Modélisation tridimensionnelle des articulateurs de la parole pour l'animation faciale : implémentation temps réel et mesure d'intelligibilité bimodale.* Thèse de doctorat, Signal Image Parole, Institut National Polytechnique de Grenoble, France.

Guiard-Marigny et al. (1994)
T. Guiard-Marigny, **A. Adjoudani**, and **C. Benoît**. 3D models of the lips for visual speech synthesis. In Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pages 49-52, New Paltz - New York, 1994.

Hennecke et al (1994)
M. E. Hennecke, K. V. Prasad, and D. G. Stork. Using deformable templates to infer visual speech dynamics. In *28th Annual Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 578-582, Pacific Grove, California, USA.

Hennecke et al. (1996)
M. E. Hennecke, D. G. Stork, and K. V. Prasad. Visionary speech: looking ahead to practical speechreading systems. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 331-350. Springer-Verlag, Berlin.

Horbelt et Dugelay, (1995)

S. Horbelt et J.L. Dugelay. Active contour for lipreading combining snakes with templates. *Quinzième Colloque GRETSI,* pages 717-720, Juan-Les-Pins, France.

HTK, (1992)
Cambridge University Engineering Department. HTK: Hidden Markov Model Toolkit V1.4, user's manual.

Jackson et al., (1976)
P. L. Jackson, A. A. Montgomery, C. A. Binnie. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19, pages 796-812.

Jacob et Sénac, (1996)
B. Jacob and C. Sénac. Un modèle maître-esclave pour la fusion des données acoustiques et articulatoires en reconnaissance automatique de la parole. *Actes des 21èmes Journées d'Etude de la Parole,* pages 363-366, Avignon, France.

Jankelson et al., (1975)
B. Jankelson, C. W. Swain, P. F. Crane, and J. C. Radke. Kinesiometric instrumentation: A new technology. *Journal of the American Dental Association,* 90, 834-840.

Jourlin, (1996)
P. Jourlin. Asynchronie dans les systèmes de reconnaissance de la parole basés sur les HMM. *Actes des 21èmes Journées d'Etude de la Parole,* pages 351-354, Avignon, France.

Klatt, (1976)
D. H. Klatt. A digital filterbank for spectral masking. *Proceedings of IEEE ICASSP*, pages 573-576.

Klatt, (1979)
D. H. Klatt. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* 7:279:312.

Lallouache, (1991)
M. T. Lallouache. *Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres.* Thèse de doctorat, Systèmes Electroniques, Institut National Polytechnique de Grenoble, France.

Le Goff et al., (1996)
B. Le Goff, T. Guiard-Marigny, and C. Benoît. Analysis-synthesis and intelligibility of a talking face. In J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, editors, *Progress in speech synthesis*, pages 235-244, Springer-Verlag, New York.

Le Goff et Benoît, (1996a)
B. Le Goff and C. Benoît. A text-to-audiovisual-speech synthesizer for French. In Proc. *of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Le Goff et Benoît, (1996b)

B. Le Goff and C. Benoît. Synthèse audiovisuelle de la parole à partir du texte. *Actes des XXIèmes Journées d'Étude sur la Parole*, pages 379-382, Avignon, France.

Le Goff et al., (1995)

B. Le Goff, T. Guiard-Marigny, and C. Benoît. *Read my lips... and my jaw! how intelligible are the components of a speaker's face?* In Proc. of the 4th EUROSPEECH conference, pages 291-294, Madrid, Spain, 1995.

Le Goff et al., (1994)

B. Le Goff, T. Guiard-Marigny, M.M. Cohen, and C. Benoît. Real-time analysis-synthesis and intelligibility of talking faces. In Proc. of the *Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA.

Le Goff, (1993)

B. Le Goff. *Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel.* Projet de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Leroy et Herlin, (1995)

B. Leroy and I. Herlin. Un modèle déformable paramétrique pour la reconnaissance de visages et le suivi du mouvement des lèvres. *Quinzième Colloque GRETSI*, pages 701-703, Juan-Les-Pins, France.

Levinson et al., (1983)

S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Technical Journal*, 62(4):1035-1074.

Liberman et Mattingly, (1985)

A.M. Liberman and I.G. Mattingly. The motor theory of speech perception revised. *Cognition,* 21:1-36.

Lindblom et Sundberg, (1971)

B. Lindblom and J. Sundberg. Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50:1166-1179.

Linker, (1982)

W.Linker. Articulatory and acoustic correlates of labial activity in vowels: a cross-linguistic study. *UCLA Working papers on phonetics*, 56.

Lisker et Rossi, (1992)

L. Lisker and M. Rossi. Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech,* 35(4), pp 391-417.

Luettin et al., (1996a)
J. Luettin, N. A. Thacker, and S. W. Beet. Active shape models for visual speech feature extraction. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 383-390. Springer-Verlag, Berlin.

Luettin et al., (1996b)
J. Luettin, N. A. Thacker, and S. W. Beet. Speechreading using shape and intensity information. *In proc. of 4th ICSLP*, Philadelphia, USA

Mac Leod et Summerfield, (1987)
A. MacLeod and Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British Journal of audiology*, 21:131-141.

Macheboeuf, (1984)
G. Macheboeuf. Evolution et progrès de l'incrustation en télévision couleur. *Radiodiffusion- Télévision,* N° 83.

Magno Caldognetto et al., (1989)
E. Magno Caldognetto, K. Vagges, N. A. Borghese, and G. Ferrigno. Automatic analysis of lips and jaw kinematics in VCV sequences. In proc. of the *2nd Eurospeech Conference*, volume 2, pages 453-456.

Mase, (1990)
K. Mase. An application of optical flow, extraction of facial expression. *IAPR Workshop on Machine Vision and Applications,* pages 195-198.

Mase et Pentland, (1991)
K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6):67-75.

Massaro, (1987a)
D. W. Massaro. Integrating Multiple sources of information in listening and reading. In *Language Perception and Production,* Academic press, New York.

Massaro, (1987b)
D. W. Massaro. Speech perception by ear and eye. In B. Dodd and R. Campbell (1987), pages 53-83.

Massaro, (1989)
D. W. Massaro. Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences,* 12:741-794.

Massaro, (1996)
D. W. Massaro. Bimodal speech perception: A progress report. In D. G. Stork and M. E. Hennecke editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 79-102. Springer-Verlag, Berlin.

Massaro et Cohen, (1993)
D. W. Massaro and M. M. Cohen. Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communications,* 12, 127-134.

Massaro et Friedman, (1990)
D. W. Massaro and D. Friedman. Models of integration given multiple sources of information. Psychological Review, 97:225:252.

McGurk et MacDonald, (1976)
H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746-748.

Mégard, (1995)
C. Mégard. *Élaboration et évaluation d'un système de traitement numérique de la couleur pour l'extraction des lèvres en temps réel.* Mémoire de D.E.A Signal Image Parole, Institut National Polytechnique de Grenoble, France.

Mégard et Butré, (1995)
C. Mégard and L. Butré. *Labiométrie en temps réel pour la télévirtualité*. Mémoire d'ingénieur, ENSERG, Institut National Polytechnique de Grenoble, France.

Mohamadi, (1993)
T. Mohamadi. *Synthèse à partir du texte de visages parlants : réalisation d''un prototype et mesures d''intelligibilité bimodale.* Thèse de doctorat, Institut National Polytechnique de Grenoble, France.

Mohammadi et Benoît, (1992)
T. Mohammadi and C. Benoît. Apport de la vision du locuteur à l'intelligibilité de la parole bruitée en français. *Bulletin de la Communication Parlée*, 2:32-41.

Montgomery et Jackson (1983)
A. A. Montgomery and P. L. Jackson. Physical characteristics of the lips underlying vowel lipreading performances. *Journal of the Acoustical Society of America*, 73:2134-2144.

Motorola, (1993)
Motorola Inc. DSP 56002, Digital Signal Processor user's manual.

Movellan, (1995)
J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems,* volume 7, pages 851-858, MIT Press, Cambridge, Massachussets.

Movellan et Chadderdon, (1996)
J. R. Movellan and G. Chadderdon. Channel separability in audio-visual integration of speech: A bayesian approach. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 473-488. Springer-Verlag, Berlin.

Neely, (1956)

K. K. Neely. Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America,* 28:1275-1277.

Nishida, (1986)

S. Nishida. Speech recognition enhancement by lip information. In proc. of *Computer Human Interface 86*, pages 198-204, Association Computing Machinery.

Panagos et Strube, (1987)

G. Panagos and H. W. Strube. Two magnetic methods for articulatory measurements. *Phonetica,* 44, 197-209.

Parke, (1974)

F. I. Parke. *A Parametric Model for Human Faces.* PhD thesis, University of Utah, Salt Lake City, UT, 1974.

Petajan, (1984a)

E. D. Petajan. *Automatic lipreading to enhance speech recognition.* PhD thesis, University of Illinois at Urbana-Champaign.

Petajan, (1984b)

E. D. Petajan. Automatic lipreading to enhance speech recognition. In *proc. of the Global Telecommunications Conference,* pages265-272. IEEE Communication Society, Atlanta, USA.

Petajan, (1985)

E. D. Petajan. Automatic lipreading to enhance speech recognition. In proc. *Computer Vision and Pattern Recognition,* pages 40-47, IEEE Computer Society Press, San Francisco, USA.

Petajan et al., (1987)

E. D. Petajan, B. J. Bischoff, D. A. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. *Technical Report* TM 11251-871012-11, Bell Labs.

Petajan et Graf, (1996a)

E. D. Petajan and H. P. Graf. Robust feature analysis for automatic speechreading and character animation. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 425-436. Springer-Verlag, Berlin.

Petajan et Graf, (1996b)

E. D. Petajan and H. P. Graf. Automatic lipreading research: Historic overview and current works. In Y. Wang, S. Panwar, S. Kim and H.L. Bertoni, editors, *Multimedia Communications and Video Coding*, pages 265-275, New York.

Plant, (1980)

G. L. Plant. Visual identification of Australian vowels and diphthongs. *Journal of Audiology*, 2:83-91.

Portel, (1996)

C. Portel. *Plate-forme multimédia pour l'EAO des langues étrangères et la lecture labiale.* Mémoire d'ingénieur, ENSERG, Institut National Polytechnique de Grenoble, France.

Prasad et al., (1993)

K. V. Prasad, D. G. Stork, and G. J. Wolf. Preprocessing video images for neural learning of lipreading. *Technical report CRC-TR-9326*, Ricoh California Center, Menlo Park, California.

Rabiner, (1989)

L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE,* 77(2):257:286.

Rao et Mersereau, (1995)

R. R. Rao and R. M. Mersereau. Lip modeling for visual speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers,* volume 1, pages 587-590, Pacific Grove, California, USA.

Reisberg et al., (1987)

D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd and R. Campbell (1987), editors, *Hearing by eye: the psychology of lipreading,* pages 97-114. Eds. Lawrence Erlbaum Associates, London.

Revéret et Benoît, (1997)

L. Revéret and C. Benoît. A viseme-based approach to labiometrics. In J. Bigün, G. Chollet and G. Borgefors, editors, *1st International Conference on Audio and Video based Biometrics for Person Authentication*, Crans-Montana, pages 335-343.

Robert-Ribes et al., (1991)

J. Robert-Ribes, P. Escudier and J. L. Schwartz. Modèles d'intégration audition-vision : une étude neuromimétique. *Rapport interne ICP.*

Robert-Ribes, (1995)

J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux liguistiques : de la perception humaine à la reconnaissance automatique des voyelles.* Thèse de doctorat, Signal Image Parole, Institut National Polytechnique de Grenoble, France.

Robert-Ribes et al., (1996)

J. Robert-Ribes, M. Piquemal, J. L. Schwartz. Exploiting sensor fusion architectures and stimuli complementary in AV speech recognition. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 193-210. Springer-Verlag, Berlin.

Rosenblum et Saldaña, (1995)

L. D. Rosenblum and H. M. Saldaña. Visual primitives for audiovisual speech integration. *Journal of Experimental Psychology: Human Perception and Performance.*

Sekiyama et Tohkura, (1993)
K. Sekiyama and Y. Tohkura. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21: 427-444.

Sekiyama et al., (1996)
K. Sekiyama, Y. Tohkura, and M. Umeda. A few factors which affect the degree of incorporating lipread information into speech perception. In proc. of *4th ICSLP*, Philadelphia, USA.

Silbernagl et Despopoulos, (1992)
S. Silbernagl and **A. Despopoulos**. Atlas de Poche de Physiologie. Edition Flammarion medecine-sciences, Paris.

Silsbee, (1993)
P. L. Silsbee. *Computer lipreading for improved accuracy in automatic speech recognition.* PhD thesis, University of Texas.

Silsbee, (1994)
P. L. Silsbee. Sensory integration in audiovisual automatic speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 561-565.

Silsbee et Bovik, (1993)
P. L. Silsbee and A. C. Bovik. Visual lipreading by computer to improve automatic speech recognition accuracy. *Technical Report TR-93-02-90,* the University of Texas Computer and Vision Research center, Austin, Texas.

Silsbee et Su, (1996)
P. L. Silsbee and Q. Su. Audiovisual sensory integration using hidden Markov models. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 489-496. Springer-Verlag, Berlin.

Smeele et al., (1992)
M. T. Smeele, A. C. Sittig, and V. J. van Heuven. Intelligibility of audio-visually desynchronised speech: Asymmetrical effect of phoneme position. *In Proc. of the International Conference on Spoken Language Processing,* pages 65-68, Bandff, Alberta, Canada.

Stork et al., (1992)
D. G. Stork, G. J. Wolf, and E. P. Levine. Neural Network lipreading system for improved speech recognition. In proc. of *the International Joint Conference on Neural Networks*, volume 2, pages 289-295, IEEE, Baltimore.

Su et Silsbee, (1996)

Q. Su and P. L. Silsbee. Robust audiovisual integration using semicontinuous Hidden Markov Models. In proc. of *the 4th International Conference on Spoken Language Processing*, Philadelphia, USA.

Sumby et Pollack, (1954)
W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212-215.

Summerfield et al., (1989)
Q. Summerfield, A. MacLeod, M. McGrath, and N. M. Brooke. Lips teeth and the benefits of lipreading. In A. W. Young and H. D. Ellis, editors, *Handbook of research on face processing,* pages 223-233. Elsevier Science Publisher B.V., North Holland.

Summerfield, (1979)
Q. Summerfield. Use of visual information for phonetic perception. *Phonetica,* 36:314-331.

Summerfield, (1983)
Q. Summerfield. Audio-visual speech perception, lipreading and artificial simulation. In M. E. Lutman and M. P. Haggard, editors, *Hearing Science and Hearing Disorders*, pages 131-182.

Summerfield, (1987)
Q. Summerfiled. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell editors, *Hearing by eye: the psychology of lipreading*, pages 3-51, Lawrence Erlbaum Associates, London, England.

Summerfield et McGrath, (1984)
Q. Summerfield and M. McGrath. Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology,* 36A(1):51-74.

Tillmann et al., (1984)
H. G. Tillmann, B. Pompino-Marshall, and U. Porzig. The effects of visually presented speech movements on the perception of acoustically encoded speech articulation as a function of acoustic desynchronisation. In M. P. R. Van Den Broecke and A. Cohen, editors, *Proceedings of the 10th International Congress of Phonetic Sciences (ICPhS),* pages 469-473.

Vogt, (1996)
M. Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In D. G. Stork and M. E. Hennecke editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 331-350. Springer-Verlag, Berlin.

Vroomen, (1992)
J. H. M. Vroomen. *Hearing voices and seeing lips: Investigations in the psychology of lipreading*. PhD Thesis, Katolieke Univ. Brabant.

Waibel et al., (1994)

A. Waibel, M. Tue Vo, P. Duchnowski, and S. Manke. Multimodal Interfaces. In *Artificial Intelligence Review Journal,* special issue.

Walker et al., (1995)

S. Walker, V. Bruce, and C. O'Malley. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception and Psychophysics*. 57, 8, pp 1124-1133.

Watanabe et Kohda, (1990)

T. Watanabe and M. Kohda. Lipreading of Japanese vowels using neural networks. In *International Conference on Spoken Language Processing*, pages 1373-1376, Kobe, Japan.

Yuhas et al., (1989)

B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural nets. *IEEE Communication Magazine*, page 65-71.

Yuhas et al., (1990)

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

Young et al., (1989)

S.J. Young, N.H. Russel, J.H.S. Thornton. Token Passing: a simple conceptual model for connected speech recognition systems. *Technical report* - CUED/F-INFENG/TR.38.

Table des matières

Chapitre 1 : les lèvres et la production de la parole

Chapitre 2 : la perception auditive et visuelle de la parole

Chapitre 3 : le traitement de la parole visuelle

Chapitre 4 : la reconnaissance audiovisuelle de la parole

Chapitre 5 : une application temps réel

Table des illustrations

Glossaire

CAN : convertisseur analogique-numérique

CNA : convertisseur numérique-analogique

FLMP : Fuzzy Logical Model of Perception

HMM : Hidden Markov Model

LUT (Look Up Table) : table de correspondance

LIPTRACK : Labiomètre Indépendant Programmable Temps Réel Avec Chroma-Key

Annexe 1

Liste des publications

Conférences avec actes et comité de lecture:

A. Adjoudani, T. Guiard-Marigny, B. Le Goff, and C. Benoît. *Un modèle 3D de lèvres parlantes.* Actes des XXèmes Journées d'Étude sur la Parole, pages 143-146, Société Française d'Acoustique, Trégastel - France, 1994.

O. Angola, B. Le Goff, T. Guiard-Marigny, **A. Adjoudani**, C. Benoît, and M.. Cohen. *Analyse-synthèse de visages parlants.* Actes des XXèmes Journées d'Étude sur la Parole, pages 147-150, Société Française d'Acoustique, Trégastel - France, 1994.

T. Guiard-Marigny, **A. Adjoudani**, and C. Benoît. *3D models of the lips for visual speech synthesis.* In Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pages 49-52, New Paltz - New York, 1994.

A. Adjoudani and C. Benoît**.** *Audio-visual speech recognition compared across two architectures*. In Proc. of the 4th EUROSPEECH Conference, volume 3, pages 1563-1566, Madrid, Spain, 1995.

Conférence invitée :

C. Benoît, **A. Adjoudani**, O. Angola, T. Guiard-Marigny, and B. Le Goff. *Perception, analyse et synthèse de lèvres parlantes.* In Actes de la Conférence IMAGINA, Institut National de l'Audiovisuel, pages 142-163, Monte Carlo, 1994.

Articles dans un livre :

A. Adjoudani and C. Benoît**.** *On the integration of auditory and visual parameters in an HMM-based ASR.* In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150, pages 461-472. Springer-Verlag, Berlin, 1996.

C. Benoît, T. Guiard-Marigny, B. Le Goff, and **A. Adjoudani**. *Which components of the face are best speechread by man and machine?* In D. Stork and M. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150, pages 315-328. Springer-Verlag, Berlin, 1996.

T. Guiard-Marigny, **A. Adjoudani**, and **C. Benoît**. *3D models of the lips and jaw for visual speech synthesis.* In Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in speech synthesis*, pages 247-256, Springer-Verlag, New York, 1996.

Annexe 2

La plate-forme logicielle de reconnaissance audiovisuelle

Cette partie est consacrée à la présentation de notre plate-forme logicielle d'apprentissage et de test à base de HMM. Comme nous l'avons précisé, les outils du logiciel HTK développés à l'Université de *Cambridge* constituent les principaux programmes de commande. Nous décrivons donc brièvement les outils que nous avons utilisés dans nos scripts (version 1.4 de HTK). Les programmes de gestion de fichiers et l'emplacement de chaque groupe de données ainsi que les commandes de traitement en série seront présentés à la suite.

1. Les outils HTK

Le logiciel de reconnaissance automatique de la parole HTK est un logiciel qui est connu mondialement pour ses nombreuses fonctionnalités. Les modèles markoviens sont créés, puis les paramètres de chaque modèle sont estimés en utilisant l'algorithme de Baum-Welch.
Les principales commandes qui sont utilisées dans notre système sont les suivantes :
HCode : il permet de traiter les fichiers audio et d'extraire les paramètres spectraux.
HInit : l'exécution de cette commande permet de fournir une première estimation de la variance et la moyenne de chaque modèle de Markov.
HRest : grâce à ce programme, les paramètres de chaque modèle sont estimés avec l'algorithme de Baum-Welch.
HVite : c'est l'outil qui calcule, grâce à l'algorithme de Viterbi (légèrement modifié), le modèle le plus probable pour un échantillon à l'entrée du système.
HResult : c'est un simple outil qui calcule le pourcentage de mots correctement identifiés. Nous avons pris en compte que le pourcentage de mots correctement identifiés et les omissions et les insertions sont négligées.

2. La base de données

Notre base de données est constituée de plusieurs types de fichiers :

- Les fichiers de données visuelles : ils sont composés de trames de paramètres qui sont respectivement A, B, S, A', B', S', C, P1, P2, M, S1, S2, d1, d2, d3, Mp. Ces fichiers ont l'extension *.res*. Ils sont stockés dans le répertoire *Reco/HTK_V1.4A/data_res.*
- Les fichiers de données audio avec l'extension *.aiff* et contiennent les échantillons acoustiques enregistrés à 16 KHz sur 16 bits. Toutes les informations concernant la fréquence d'échantillonnage et le type de codage sont stockées sur les 62 bytes d'en-tête. Ces données sont stockées dans le répertoire *Reco/HTK_V1.4A/data_aiff.*
- Les fichiers audio clairs codés sur 12 coefficients MFCC + delta + énergie avec l'extension *.mfc* sont stockés dans le répertoire *Reco/HTK_V1.4A/data_mfc*. Ils sont utilisés pour l'apprentissage des modèles markoviens acoustiques.
- Les fichiers audio bruités avec l'extension *.bruit* sont enregistrés dans le répertoire *Reco/HTK_V1.4A/data_noise*. Le fichier README dans ce répertorie indique le niveau de bruit ajouté sur l'ensemble de ces fichiers.
- Les résultats de codage des fichiers audio bruités sont stockés dans le répertoire *Reco/HTK_V1.4A/data_mfc_bruit*. Ils ont l'extension *.mfc_bruit*.
- Le logiciel HTK travaille avec des fichiers binaires pour l'entrée des modèles. Il est donc indispensable de coder les fichiers de paramètres sous format binaire pour qu'ils puissent être lus par les outils de HTK. Les fichiers de paramètres enregistrés sous le format HTK sont stockés dans le répertoire *Reco/HTK_V1.4A/data_HTK_reduit*. Ils ont l'extension *.HTK*. Ces fichiers contiennent uniquement les paramètres labiaux codés en binaire. Le répertoire *Reco/HTK_V1.4A/data_HTK_reduit_red* contient les fichiers de paramètres visuels codés en binaire avec leur dérivée première. Ces fichiers ont la même extension, c'est à dire *.HTK*.
- Pour le processus d'apprentissage et de reconnaissance, HTK a besoin des fichiers label qui permettent d'associer un mot à une classe. Ces fichiers contiennent aussi l'information sur la durée du mot. Le répertoire *Reco/HTK_V1.4A/data_lab* contient les fichiers labels correspondant aux stimuli audio. Pour les fichiers de paramètres labiaux, les fichiers label sont stockés dans le répertoire *Reco/HTK_V1.4A/data_res_lab.*
- Pour intégrer les modalités acoustiques et visuelles en utilisant les probabilités de sortie, nous avons stocké l'ensemble de ces valeurs pour tous les modèles et ceci pour tous les

niveaux de bruit de test. Ces données se trouvent dans *Reco/proba/audio* et sont stockées suivant le rapport signal sur bruit. Les probabilités des modèles de Markov visuels sont stockées dans *Reco/proba/vidéo*.

- Le répertoire *Reco/HTK_V1.4A/data_mfc_mix* contient les fichiers de coefficients cepstraux concaténés avec les paramètres visuels. Les coefficients MFCC sont le résultat du traitement sur des données acoustiques claires.
- Le répertoire *Reco/HTK_V1.4A/data_mfc_mix_bruit* contient les fichiers de coefficients acoustiques et visuels mixtes. Cependant, la différence avec le précédent répertoire est que les coefficients cepstraux sont obtenus à partir des données acoustiques bruitées.

3. Les utilitaires

3.1 Les outils de traitement de fichiers

Nous avons crée des utilitaires qui permettent de traiter des fichiers de paramètres visuels ou les fichiers binaires en format HTK

3.1.1 Les fichiers de paramètres visuels

Pour mettre les fichiers de paramètres au format binaire HTK, nous avons élaboré les programmes suivants

- Le programme *AddDelta2res* permet de transformer les fichiers de paramètres en format binaire HTK. En utilisant l'option *-p* de cet outil, l'utilisateur a la possibilité de sélectionner les paramètres. Il suffit de taper le nom du paramètre après l'option *-p* suivi de *end*. L'option *-d* permet de calculer les dérivées pour chaque trame.
- Le programme *res2HTK_mixer+del* en plus des ces fonctionnalités permet de concaténer les paramètres visuels aux paramètres MFCC du répertoire *data_mfc_bruit*.
- *HTK_labmaker* permet de créer les fichiers label.

3.1.2 Les fichiers audio

Pour manipuler les fichiers audio, nous avons développé les programmes suivants :

- *noise_adder* permet d'ajouter du bruit suivant le niveau spécifié à un fichier du format *aiff*. Le fichier de sortie est sous format *raw-data*.
- *labmaker* permet de créer les fichiers labels pour les échantillons audio.

3.2 Les outils de traitement automatisé

Afin de traiter l'ensemble des fichiers d'un répertoire nous avons élaboré des scripts en langage *csh*.

3.2.1 Les fichiers de paramètres visuels

- *auto_+del_paramselector* permet de sélectionner les paramètres spécifiés, calculer les dérivées et enregistrer le fichier final en HTK. Une fois le traitement sur tous les fichiers du répertoire terminé, les fichiers résultats sont stockés dans *data_HTk_reduit_del/sauvegarde*.
- *auto_mixer+del* permet de concaténer les paramètres visuels sélectionnés avec leur dérivée aux coefficients cepstraux obtenus à partir des signaux audio clairs. Les résultats sont stockés dans le répertoire *Reco/HTK_V1.4A/data_mfc_mix/sauvegarde.*
- *auto_mixer+del_bruit* a la même fonctionnalité à la différence qu'il permet de concaténer les paramètres visuels aux coefficients cepstraux obtenus à partir des données acoustiques bruitées.

3.2.2 Les fichiers audio

- *auto_noise_adder* permet d'ajouter du bruit blanc avec le niveau spécifié sur l'ensemble des fichiers du répertoire. Les fichiers audio bruités sont enregistrés en format *.aiff* et sont envoyés vers le répertoire *Reco/HTK_V1.4A/data_noise.*
- *auto_labmaker* permet de créer des fichiers label correspondant et de les envoyer vers le répertoire *Reco/HTK_V1.4A/data_lab*.

3.2.3 Les fichiers de coefficients MFCC

- *tr2te_mfc* est un utilitaire qui permet de créer des fichiers d'apprentissage et de test (préfixe *tr* et *te*). Le nombre de fichiers d'apprentissage et de test doit être spécifié.
- *cycle_tr2te* permet de créer des fichiers d'apprentissage et de test suivant le principe de permutation en fonction de leur numéro. Par exemple *cycle_tr2te 1 2* sélectionne les fichiers numéro 1 et 2 pour le test et le reste pour l'apprentissage.

3.2.4 Les fichiers de coefficients mixtes et HTK

Pour créer les fichiers d'apprentissage et de test, des utilitaires semblables à ceux implémentés pour les fichiers MFCC ont été créés. Ils se trouvent dans le répertoire *Reco/HTK_V1.4A/data_mfc_mix*, *Reco/HTK_V1.4A/data_mfc_mix_bruit* et

Reco/HTK_V1.4A/data_HTK_res+del pour les fichiers mixtes non bruités, les fichiers mixtes bruités et les fichiers de paramètres avec dérivées respectivement.

3.3 Apprentissage

3.3.1 Création des modèles de Markov

La commande *proto_res_red_maker* permet de créer un prototype markovien à neuf états émetteurs et de fixer le nombre de paramètres composant chaque vecteur émis. La matrice de covariance est diagonale.

3.3.2 Apprentissage audio

La commande *training1* permet d'exécuter la première phase d'apprentissage sur les fichiers MFCC obtenus avec des signaux acoustiques non-bruités. Le modèle initial s'appelle *proto* et les résultats sont stockés dans le répertoire *Reco/HTK_V1.4A/hmm.0*. Pour réestimer les paramètres de chaque modèle, la commande *training2* est utilisée pour appliquer l'algorithme de Baum-Welch sur les modèles obtenus avec la première phase d'apprentissage. Les résultats sont stockés dans le répertoire *Reco/HTK_V1.4A/hmm.1*.

3.3.3 Apprentissage sur les paramètres visuels seuls

- La commande *training1_res* permet d'effectuer l'apprentissage sur les paramètres visuels du répertoire *Reco/HTK_V1.4A/data_HTK* en utilisant le prototype de Markov appelé *proto_res*. Les résultats sont stockés dans le répertoire *Reco/HTK_V1.4A/hmmres.0*. *training2_res* effectue l'apprentissage en appliquant l'algorithme de Baum-Welch sur les modèles obtenus à partir de la première phase d'apprentissage.
- La commande *training1_red+del_red* calcule les paramètres du modèle sur les fichiers de coefficients visuels avec dérivées. Le prototype initial s'appelle *proto_res_red+del* et les résultats sont stockés dans *hmmres_red+del.0*. *training2_red+del* réestime les paramètres des modèles comme dans les cas précédents.

3.3.4 Apprentissage sur des données mixtes

- *training1_AV* permet d'effectuer la première phase d'apprentissage sur des fichiers audiovisuels mixtes non bruités. Le prototype initiale s'appelle *proto_AV*. La commande *training1_AV_bruit* est destinée à entraîner les modèles de Markov avec des fichiers mixtes bruités à partir du prototype. Comme dans les cas précédents, *training2_AV* et *training2_AV_bruit* permettent d'affiner l'estimation des paramètres.

3.4 Test

3.4.1 Test sur les modèles acoustiques

- La commande *TEST_audio* permet d'élaborer le programme de test *HVite* sur les modèles purement acoustiques. Ce test s'effectue sur les paramètres non bruités. Il faut noter que l'outil de reconnaissance *HVite* a été modifié de façon à enregistrer les probabilités de sortie de chaque modèle. L'outil *HResult* permet ensuite de calculer le pourcentage de mots correctement identifiés. Les résultats sont stockés dans le répértoire *Reco/HTK_V1.4A/test_audio*.
- *TEST_audio_bruit* effectue le test à partir des modèles entraînés sur les signaux acoustiques clairs avec des fichiers MFCC des signaux acoustiques bruités. Les résultats de reconnaissance ainsi que les fichiers de probabilités de sortie sont stockés dans le répertoire *Reco/HTK_V1.4A/test_audio_bruit-te*.

3.4.2 Test sur les modèles visuels

- Pour effectuer le processus de test sur les modèles visuels, il suffit de lancer la commande *TEST_video*. Les résultats sont transférés vers le répertoire *Reco/HTK_V1.4A/test_video*.
- La commande *TEST_video_red+del* permet d'exécuter le programme de reconnaissance sur les paramètres visuels avec dérivées. Les résultats sont stockés dans le répertoire *Reco/HTK_V1.4A/test_video_red+del/*.

3.4.3 Test sur les modèles mixtes

- *TEST_AV* est le nom de la commande qui teste les modèles mixtes stockés dans le répertoire *Reco/HTK_V1.4A/test_AV*. Ces tests sont élaborés sur des fichiers mixtes bruités avec des modèles entraînés avec des signaux acoustiques clairs.

3.5 Intégration des données acoustiques et visuelles

- Le programme *AV_bayes+sel* permet d'intégrer les deux modalités acoustique et visuelle suivant la dispersion des quatre premiers candidats de chaque source.
- Le programme *AV_selector* sélectionne le meilleur candidat de la modalité qui a un degré de confiance plus élevé.

- Le programme *AV_bayes* permet d'intégrer les deux modalités suivant la loi multiplicative sur les probabilités de sortie.
- *AV_mixte* est le programme qui intègre les informations visuelles et acoustiques suivant l'architecture d'intégration mixte.

3.5.1 Test des architectures d'intégration

- La commande *auto_selector* permet de sélectionner le candidat audio ou vidéo qui le plus grand degré de confiance sur l'ensemble des fichiers de probabilités visuelles et acoustiques. La commande *TEST_selector* permet ensuite de calculer le taux de reconnaissance de cette intégration.
- La commande *auto_bayes* calcule le produit des probabilités des deux modalités et sélectionne le candidat qui maximise ce produit. Grâce à la commande *TEST_selector*, le pourcentage de reconnaissance est calculé.
- La commande *auto_bayes+sel* est celle qui effectue la pondération des probabilités de sortie en fonction du degré de confiance de chaque canal. Pour calculer le taux de reconnaissance obtenu à partir de cette méthode, il suffit d'exécuter la commande *TEST_bayes+sel*.
- *auto_FLMP* permet de simuler l'architecture mixte. Trois types de probabilités sont utilisés dans cette configuration : les probabilités visuelles, les probabilités acoustiques mais aussi les probabilités audiovisuelles obtenues à partir de la méthode d'identification directe. Cette commande appelle le programme *AV_mixte* qui utilise le principe de pondération à trois canaux. Le pourcentage d'identification correcte est obtenu avec la commande *TEST_FLMP*.