la reconnaissance acoustique à la reconnaissance bimodale de parole

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

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

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

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

Définition de la parole

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

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

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

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

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

Le signal de la parole

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

3.2.1 Redondance du signal

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

3.2.2 Variabilité du signal

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

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

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

3.2.3 Les effets de coarticulation

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

3.3 Extraction des paramètres

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

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

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

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

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

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

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

3.3.1 Énergie du signal

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

$$E(\text{fenêtre}) = \sum_{n\ \epsilon\ \text{fen être}} |n|^2 \qquad (3.1)$$

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

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

Dans cette équation, *R* est la valeur moyenne de l'énergie du bruit. Le résultat de ce calcul tend vers 0 lorsque la portion considérée est une zone où il n'y a que le bruit de fond. Tout le problème de cette variante réside dans l'estimation du facteur de normalisation *R*.

3.3.2 Coefficients MFCC

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

Un vecteur acoustique MFCC est formé de coefficients cepstraux obtenus à partir d'une répartition fréquentielle selon l'échelle de Mel (Bogert et al. 1963) (voir figure 3.1). L'utilisation d'échelles de fréquence non-linéaires, telles les échelles de Mel (Stevens et al. 1937) ou Bark (Zwicker 1961), permettent une meilleure représentation des basses fréquences qui contiennent l'essentiel de l'information linguistique pour la majeure partie du signal de parole. La correspondance entre les valeurs de fréquence en Hertz F_{Hertz} et en Mel F_{Mel} est calculée par (O'Shaughnessy 1987) :

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

Par ailleurs, il est possible de calculer des coefficients cepstraux à partir d'une répartition fréquentielle linéaire sans utiliser une échelle de Mel mais en conservant la répartition linéaire des échelles de fréquence. Ces coefficients sont alors appelés LFCCs (*Linear Frequency Cepstral Coefficients*) (Rabiner and Juang 1993).

Afin de séparer la source spectrale de la réponse fréquentielle, l'opération de méthode cepstrale se base sur la propriété du logarithme qui permet de transformer un produit en addition. Une transformée discrète en cosinus (*Discret Cosinus Transform*, DCT) permet ainsi d'obtenir les *N* coefficients cepstraux désirés (Ahmed et al. 1974). Considérant f la fonction de transformation spectrale, le k^{me} coefficient cepstral *C(k)* est donc obtenu par :

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

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

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

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

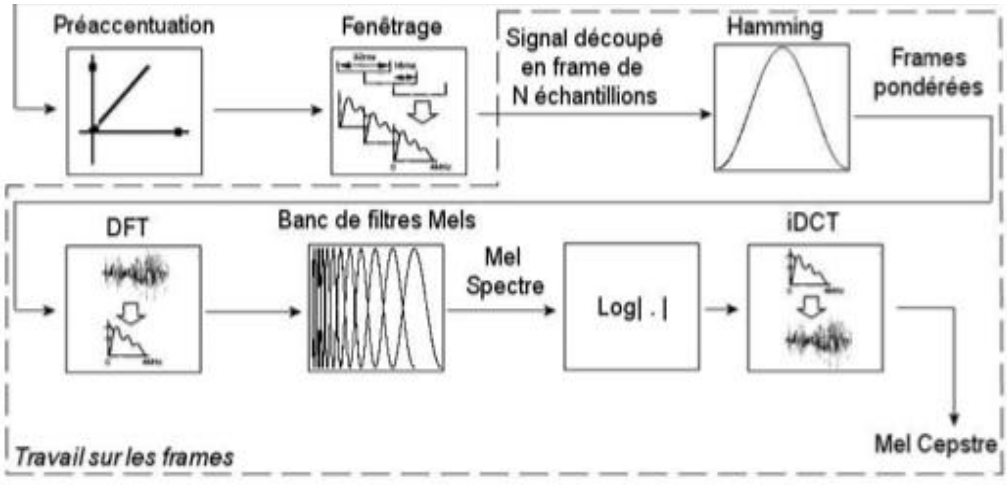

Figure 3.1 – Schéma de calcul des MFCC.

3.3.3 Taux de passage par zéro

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

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

$$bcr(\text{fenêtre}) = \sum_{n\ \epsilon\ \text{fen être}} |f(n) - f(n-1)|\ avec\ f(n) = f(n-1) \begin{matrix} 1\ si\ n > S \\ si - S \leq n\ \leq S \\ -1\ si\ n < -S \end{matrix} \quad (3.5)$$

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

3.3.4 Autres paramétrisations du signal

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

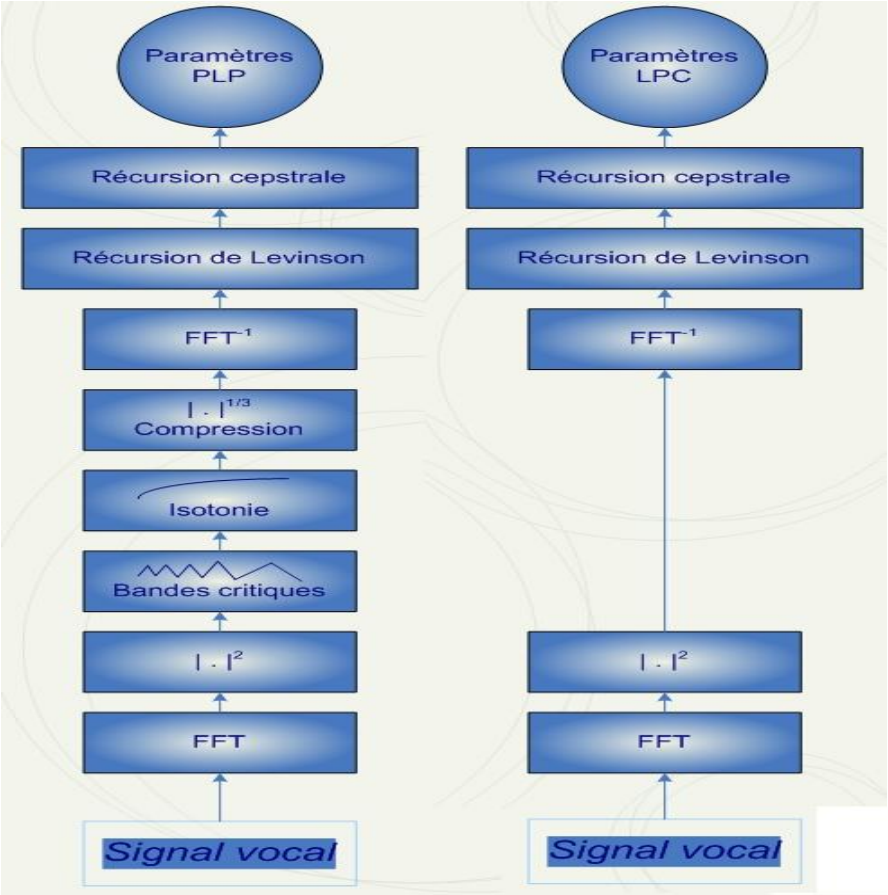

Figure 3.2 – Schémas de calcul les paramètres PLP et LPC.

3.3.5 Dérivées première et seconde

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

3.4 Réduction de l'espace de représentation

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

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

3.5 Les modes de fonctionnement d'un système de reconnaissance

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

- **Dépendant du locuteur (monolocuteur)**

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

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

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

- **Indépendant du locuteur**

Tout locuteur peut utiliser le système de reconnaissance.

- **Elocution**

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

- **Mots isolés :**

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

- **Mots connectés :**

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

- **Parole continue lue :**

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

- **Parole continue spontanée :**

  C'est le discours usuel, sans aucune contrainte.

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

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

3.6 La reconnaissance bimodale de la parole

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

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

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

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

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

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

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

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

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

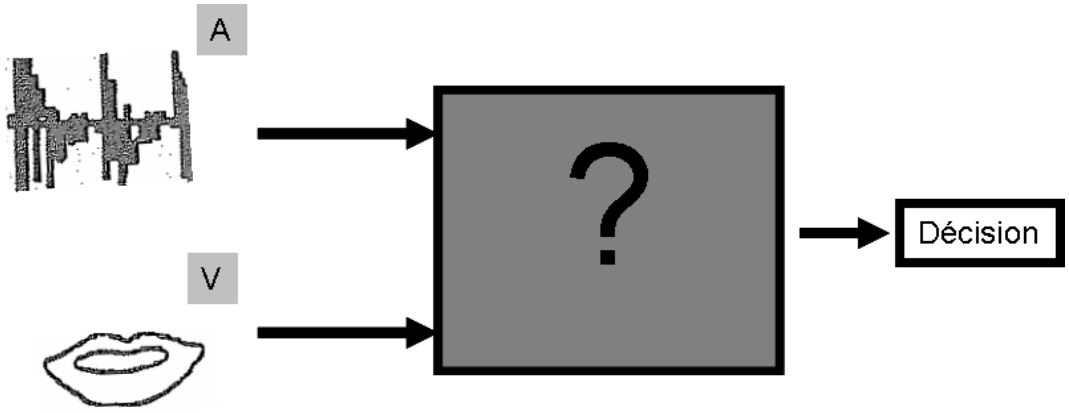

Figure 3.3 – Le noyau d'un processus d'intégration audio-visuelle dans la perception de la parole (d'après Schwartz et al. (1998)).

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

3.6.1.1 Modèle ID

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

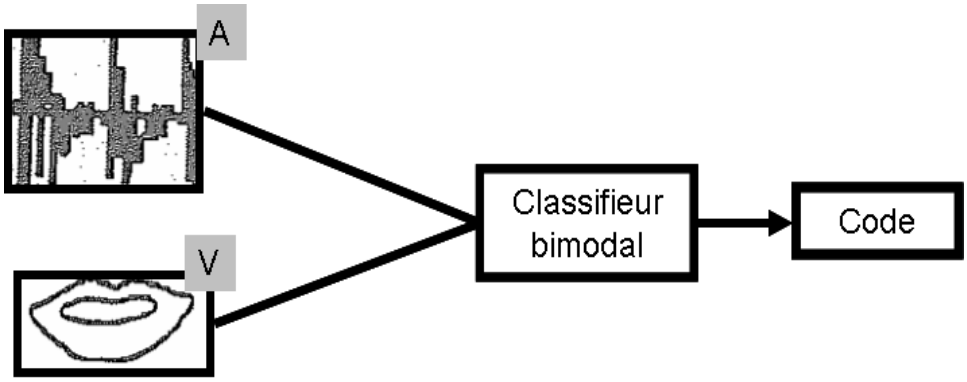

Figure 3.4 – Modèle à identification directe.

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

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

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

3.6.1.2 Modèle IS

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

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

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

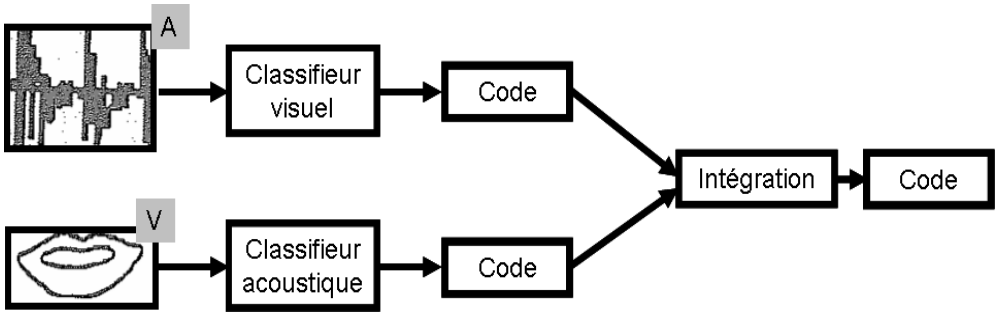

Figure 3.5 – Modèle à identification séparée.

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

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

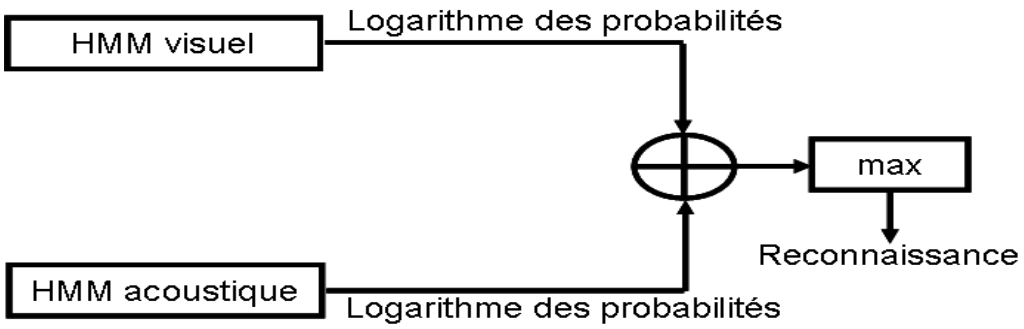

Figure 3.6 – Modèle d'intégration basé sur la maximisation des produits des probabilités conjointes (D'après Adjoudani (1998)).

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

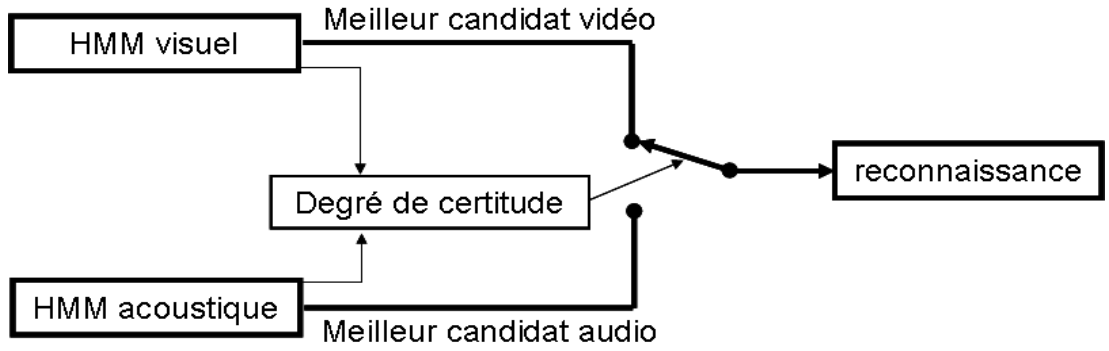

Figure 3.7 – Méthode de sélection du meilleur candidat acoustique ou visuel (D'après Adjoudani (1998)).

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

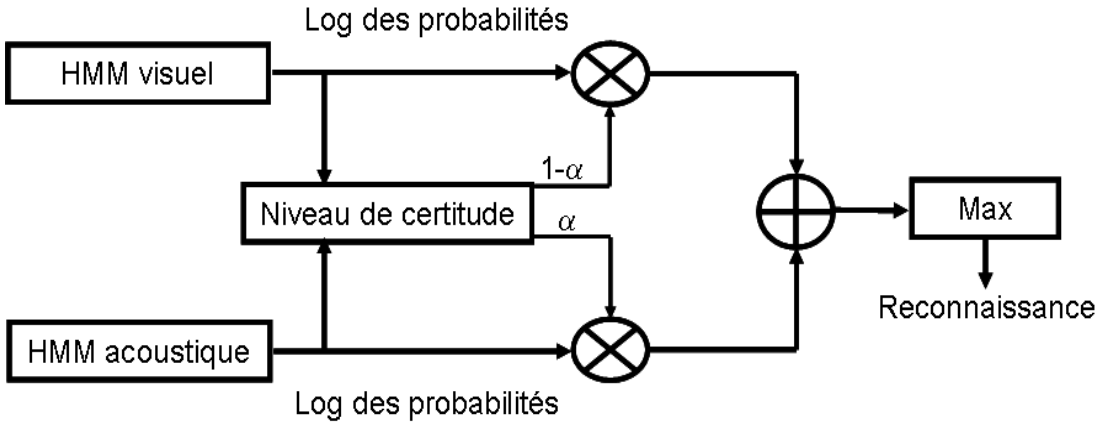

Figure 3.8 – Architecture d'intégration audiovisuelle par pondération (D'après Adjoudani (1998)).

3.6.1.3 Modèle RD

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

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

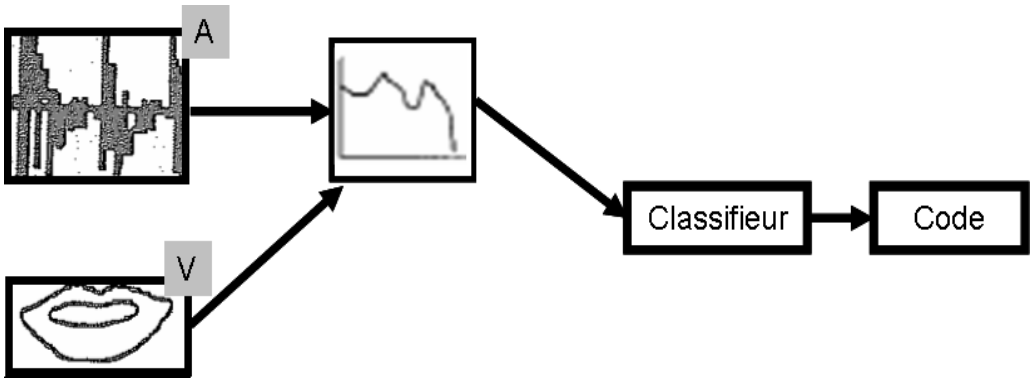

Figure 3.9 – Modèle à recodage dans la modalité dominante.

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

3.6.1.4 Modèle RM

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

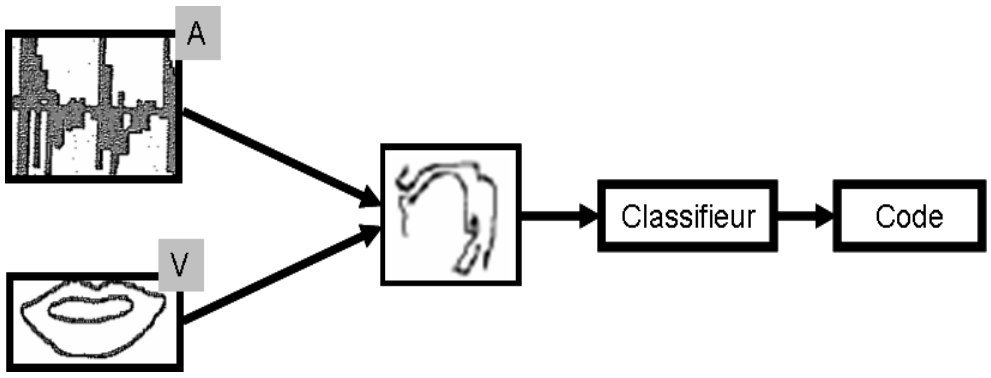

Figure 3.10 – Modèle à recodage dans la modalité motrice.

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

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

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

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

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

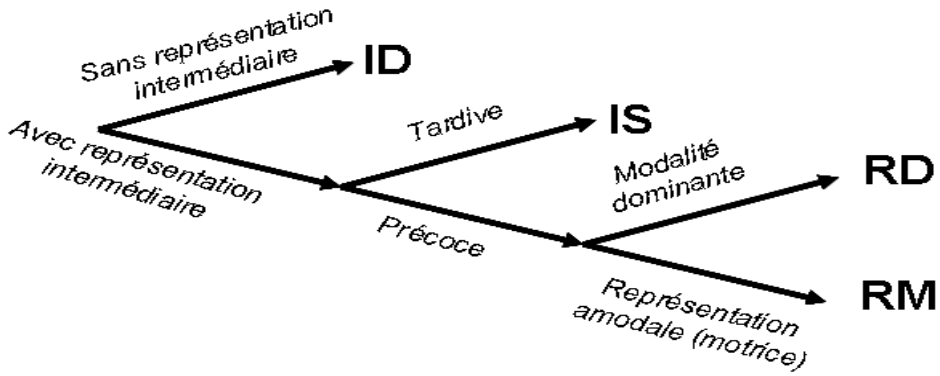

Figure 3.11 – Taxinomie des modèles d'intégration (d'après Robert-Ribès (1995)).

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

3.6.3 Etudes comparatives

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

3.6.3.1 ID vs. IS

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

Modèle ID

⊕ Modèle facile à implémenter: l'observation bimodale peut se former à partir d'une concaténation des indices des deux modalités.

⊕ Possibilité de pondérer chaque canal à condition de disposer d'un corpus d'apprentissage de taille importante (Silsbee et Su 1996).

⊖ Modèle nécessitant un corpus de taille relativement grande par rapport au modèle IS (Jacob et Sénac 1996) car la taille des modèles à apprendre est plus importante.

⊖ Nécessité d'une topologie identique des deux sources.

⊖ Conservation de la coordination temporelle entre les deux modalités durant la fusion.

⊖ Le problème de déphasage n'est pas géré.

⊖ Apprentissage adapté à chaque niveau du Rapport Signal sur Bruit (RSB) de l'entrée acoustique (Silsbee et Su 1996).

Modèle IS

⊕ Nécessité d'un corpus moins important pour l'apprentissage que pour le modèle ID grâce au traitement séparé de chaque modalité.

⊕ Les deux modalités ne demandent pas forcément d'avoir la même architecture de reconnaissance.

⊕ Le modèle s'approche plus des hypothèses faites sur la perception audiovisuelle (Robert-Ribès 1995; Massaro 1996).

⊕ Capable de traiter l'asynchronie: par exemple dans le cadre d'un mot entre son état initial et final.

⊖ Le module d'intégration peut être complexe et dépendant du corpus.

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

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

3.6.3.2 RD vs. RM

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

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

3.7 Conclusion

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

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

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

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