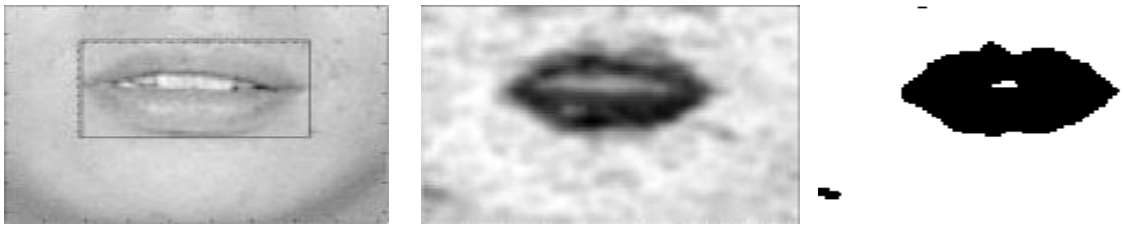


(a) Image initiale (b) Carte d'appartenance (c) Segmentation finale

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

2.3.2.2 Approches statistiques

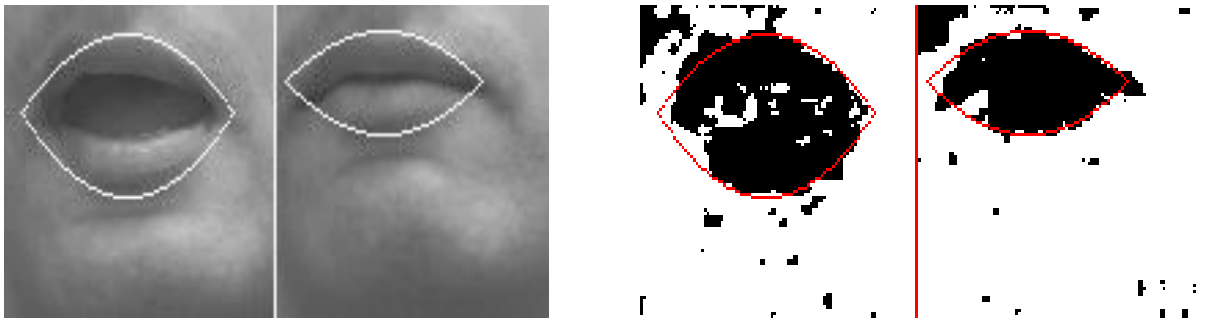
Pour les approches statistiques, comme nous l'avons déjà évoqué dans la section 2.3.2.1, l'espace de représentation (couleur) idéal pour séparer les lèvres du reste du visage sera déterminé statistiquement à partir d'exemples, au lieu d'être déterminé a priori.

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

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

$$C_i = \alpha . R_i + \beta . V_i + \gamma . B_i \, . \tag{2.7}$$

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



(a) Images initiales (b) Segmentations finales obtenues

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

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

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

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

(a) Image initiale (b) Teinte filtrée (c) Sortie du réseau de neurones

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

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

2.3.2.3 Approche par corrélation avec des patrons

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

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

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

2.3.2.4 Approches mouvement

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

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

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

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

2.3.2.5 Autres approches

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

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

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

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

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

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

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

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

2.4 Conditions « naturelles » (écologiques)

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

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

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

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

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

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

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

2.5 Comparaison image-modèle

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

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

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

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

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

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

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

2.6 Corpus existants

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

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

2.7 Conclusion

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

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