2.2 Visage complet ou indices visuels ?

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

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

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

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

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

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

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

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

2.3 Localisation et suivi de visages

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

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

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

2.3.1 Localisation de visages

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

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

2.3.1.1 Approches couleur

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

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

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

$$r = \frac{k.R}{R+V+B'},\ v = \frac{k.V}{R+V+B'},\ b = \frac{k.B}{R+V+B'} \tag{2.1}$$

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

$$r + v + b = k. \tag{2.2}$$

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

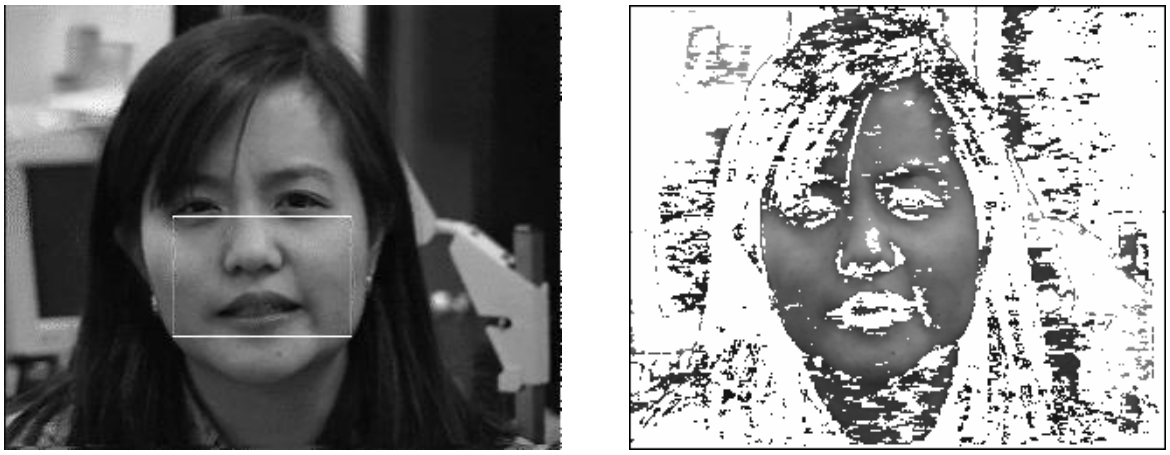
distance à une valeur centrale. L'auteur indique avoir testé un modèle générique de la couleur de la peau construit à partir de 100 images de différents sujets de la base XM2VTSDB (Messer et al. 1999), mais que les résultats étaient « assez logiquement » moins précis qu'en utilisant des modèles de la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

(a) Image à traiter (b) Ellipse de moindre coût

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

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

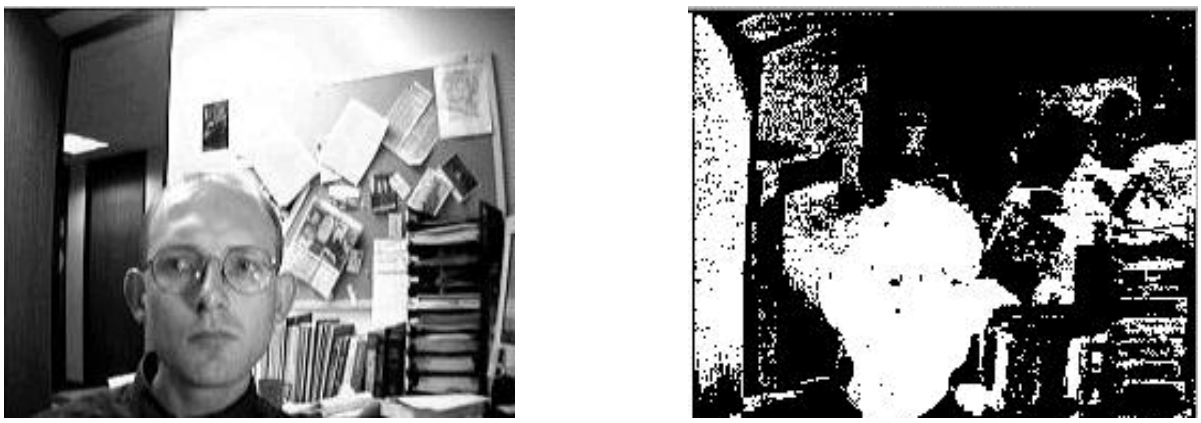
individuel (IFCC) est calculé et utilisé. Il peut être ré-estimé régulièrement pour rendre la détection du visage robuste aux changements de l'environnement.



(a) Image à traiter (couleur) (b) Régions de visage (en blanc, «fond»)

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

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



(a) Image à traiter (couleur) (b) Régions de « peau » (en blanc)

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

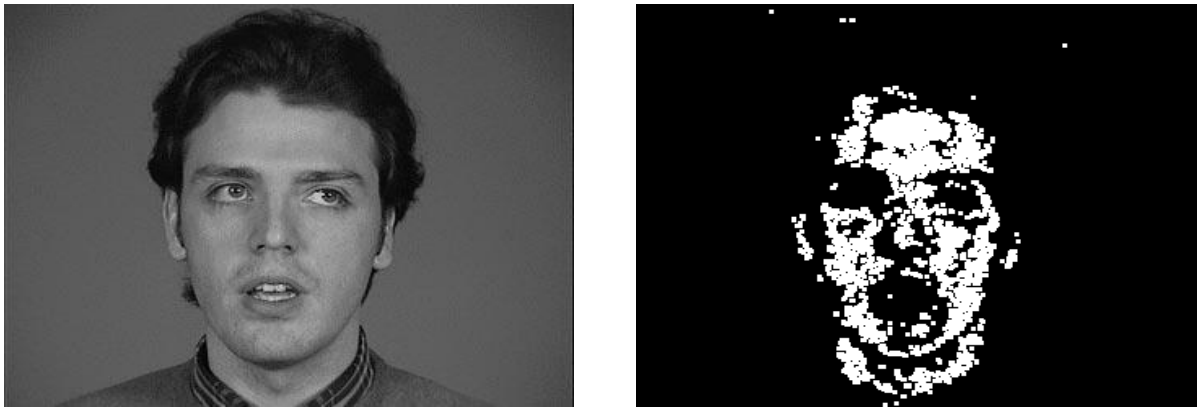
Pour repérer plusieurs visages ou des visages de tailles différentes dans une image, Senior propose une approche multi-résolution en utilisant une pyramide d'images (l'image initiale ré-échantillonnée à des résolutions inférieures) et considère chaque zone rectangulaire de m× n pixels comme un candidat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

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

$$Q_{bas} \leq Q_i \leq Q_{haut} \tag{2.3}$$

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

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



(a) Image à traiter (couleur) (b) Régions de visage (en blanc)

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

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

$$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} 0.30 & 0.59 & 0.11 \\ 0.60 & -.027 & -0.32 \\ 0.21 & -0.52 & 0.31 \end{pmatrix} \begin{pmatrix} R \\ V \\ B \end{pmatrix}. \quad (2.4)$$

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

2.3.1.2 Approches statistiques

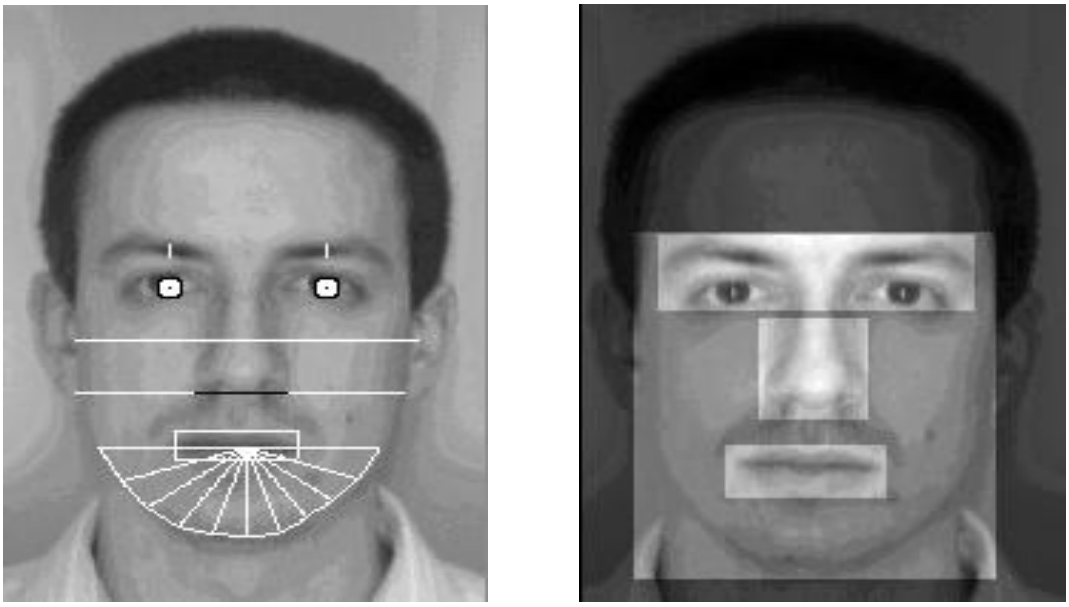
L'approche statistique pour la localisation de visages consiste à se baser sur un échantillon (des images exemples) que l'on souhaite représentatif, pour modéliser l'apparence d'un visage. L'approche peut être directe à partir d'exemples sans a priori, ou indirecte, en choisissant un espace de représentation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

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

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

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

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



(a) Éléments localisés automatiquement

(b) Éléments localisés avec l'approche « template matching »

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

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

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

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

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

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

2.3.2 Localisation de la bouche

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

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

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

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

2.3.2.1 Approches couleur

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

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

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

(a) Image à traiter (couleur) (b) Teinte filtrée (c) Détection des lèvres

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

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

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

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

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

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

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

(a) Image à traiter (couleur) (b) Seuillages de Q (c) Filtrage morphologique

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

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

$$\{A, B, U, V, hue_{ab}; hue_{uv}; chroma_{uv}\} \quad (2.6)$$

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

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