**Muscles étirant les commissures**

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

**Muscles abaisseurs des commissures**

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

**Muscles élévateurs des commissures**

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

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

## 1.3 Repères phonétiques

### 1.3.1 Acoustique et articulation

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

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

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

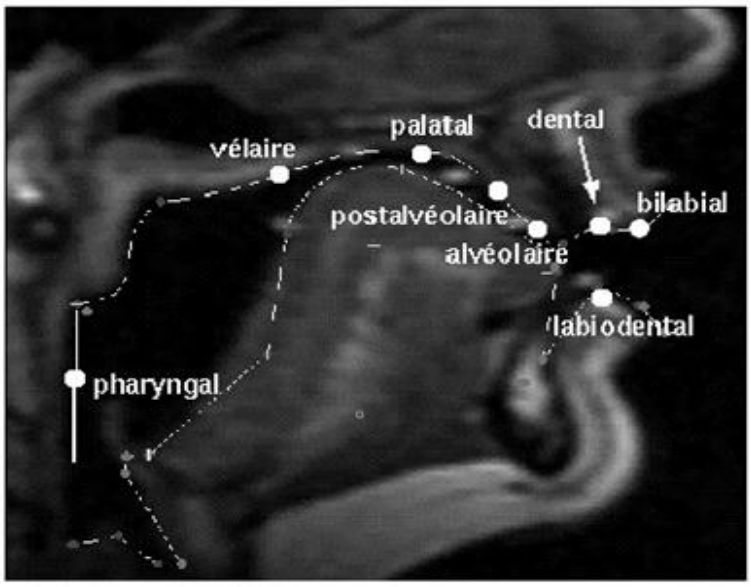



*Figure 1.6 – Le conduit vocal et les 8 lieux d'articulation principaux.*

### 1.3.2 Des sons et des lèvres

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

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

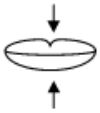
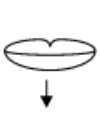
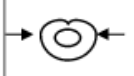
bilabiale, les sons se distinguant par leur mode d'articulation (respectivement non voisé, voisé et nasal).

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

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

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

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

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

**Liste des Abréviations**

| | | |
|---|---|---|
| Buc. = Buccinator<br>D.A.O. = Depressor Anguli Oris<br>D.L.I. = Depressor Labii Inferioris<br>L.A.O. = Levator Anguli Oris<br>LL.S. = Levator Labii Superioris | M. = Mentalis<br>O.O. = Orbicularis Oris<br>O.O.L. = Orbicularis Oris Inferior<br>Plat. = Platysma<br>Ris. = Risorius | Zyg. Maj. = Zygomaticus Major<br>Zyg. Min. = Zygomaticus Minor<br>(a) = Action antagoniste<br>(p) = Action protagoniste<br>(s) = Action synergique |

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

### 1.3.3 La coarticulation : cibles en contexte

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

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

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

## 1.4 La parole audiovisuelle et ses applications en communication

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

### 1.4.1 La bimodalité intrinsèque de la parole

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

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

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

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

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

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

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

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

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

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

### 1.4.2 L'intelligibilité de la parole audiovisuelle

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

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

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

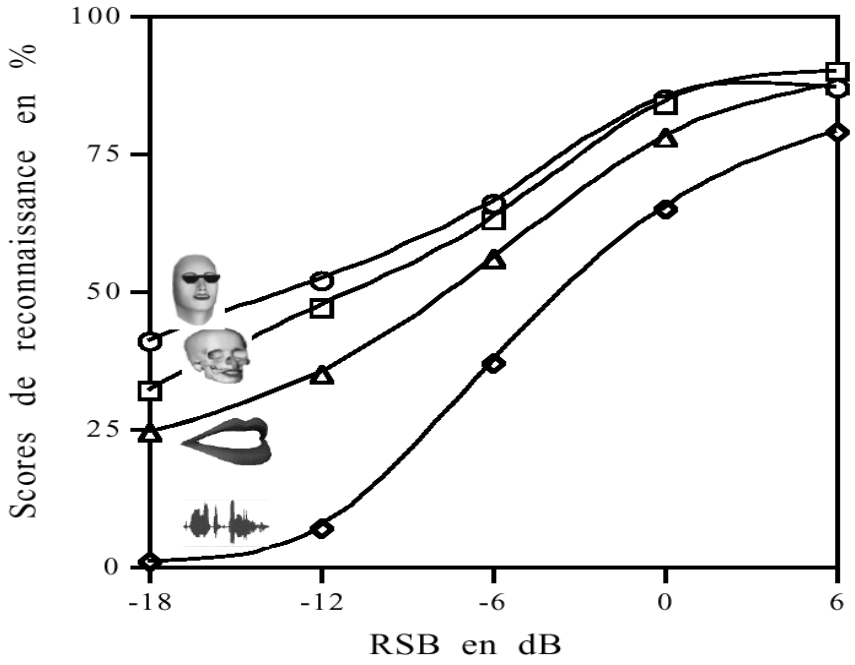



*Figure 01.8 – Comparaison de l'intelligibilité de la parole bimodale en condition bruitée en ajoutant successivement les lèvres, le mouvement de la mâchoire puis tout le visage du locuteur (Benoît et al., 1996).*

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

### 1.4.3 Perspectives pour la communication homme-machine

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

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

#### 1.4.3.1 Reconnaissance automatique de la parole audiovisuelle

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

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

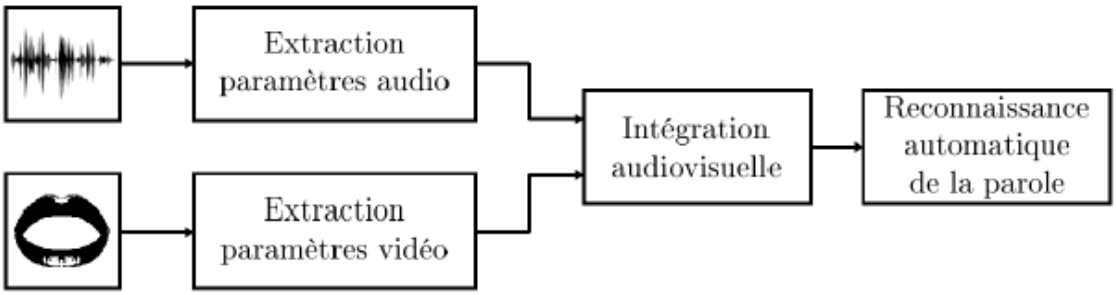



*Figure 01.9 – Schéma de principe de la reconnaissance automatique de la parole.*

### 1.4.3.2 Codage spécifique de la parole : la norme MPEG4

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

### 1.4.3.3 Le rôle de la biométrie

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

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

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

## 1.5 Conclusion

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

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

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

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

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

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

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

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

# La reconnaissance visuelle de la parole 2

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

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

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

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

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

## 2.1 Influence de l'angle de vue

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

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

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

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