Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole

Lionel Reveret

THÈSE

présentée par

Lionel REVÉRET

pour obtenir le grade de

DOCTEUR

de l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

(Arrêté ministériel du 30 mars 1992)

Spécialité : Micro-électronique

Sujet de thèse :

CONCEPTION ET EVALUATION D'UN SYSTEME DE SUIVI AUTOMATIQUE DES GESTES LABIAUX EN PAROLE

Date de soutenance : Vendredi 28 mai 1999

Composition du Jury :

Jean-Marc CHASSERY	Président
Raja CHATILA	Rapporteur
Joseph MARIANI	Rapporteur
Eric VATIKIOTIS-BATESON	Examinateur
Patrice SENN	Examinateur
Gérard BAILLY	Examinateur
Jean-Marc DOLMAZON	Directeur de thèse
Christian BENOIT*	Directeur de thèse

Thèse préparée au sein de l'Institut de la Communication Parlée

I. Résumé

Cette thèse présente un système de suivi automatique des gestes labiaux à partir d'une séquence vidéo d'un locuteur. Le système combine une analyse ascendante et descendante de la forme des lèvres. Une première étape d'analyse chromatique, basée sur un apprentissage statistique, fournit une image en niveaux de gris où le contraste entre lèvres et peau est rehaussé. Parallèlement, un modèle linéaire 3D des gestes labiaux est appris pour un locuteur à partir de formes clés phonétiquement pertinentes. Le modèle est alors synthétisé et projeté sur l'image imposant a priori les caractéristiques de la forme des lèvres. Il est adapté sur l'image rehaussée des lèvres par optimisation de ses paramètres de contrôle. Ce système combine ainsi de manière hybride la précision de l'analyse chromatique et la robustesse imposée par le modèle. Ce système est évalué sous divers aspects : ses capacités à s'adapter à la morphologie labiale et aux stratégies articulatoire de plusieurs locuteurs, la qualité des mesures géométriques délivrées et sa rapidité d'analyse.

Le système complet a été implanté et testé en langage C sur une station de travail monoprocesseur. L'exécution est évaluée en nombre d'instructions à partir du code machine généré par le compilateur du système de la station. Ces résultats ont permis d'identifier les zones critiques de traitement pour lesquels des optimisations sont proposées. Compte tenu de ces optimisations, il apparaît que la cadence de 50 images par seconde est alors accessible sans avoir recours à une implantation matérielle spécialisée.

Mots-clés : parole audiovisuelle, suivi automatique des lèvres, analyse chromatique, modélisation articulatoire, analyse / synthèse de modèle 3D, temps réel.

I. Remerciements

* Christian nous a quitté brusquement le 26 avril 1998. Il a guidé mes premiers pas de chercheur en me communiquant sa passion pour la parole. Ces trois ans m'ont amené dans des contrées que je n'imaginais pas découvrir. Plus qu'une thèse, Christian m'a offert de nouveaux horizons, une nouvelle vie. J'ai travaillé à ses cotés avec passion. Je garde aujourd'hui l'enthousiasme à « rendre la parole visible le plus possible ».

Le travail que je présente ici est né d'une des audaces de Christian. Je remercie vivement Pierre Escudier et Jean-Marc Dolmazon de s'y être pleinement associés : Pierre, en m'acceptant à l'ICP, et Jean-Marc, en co-encadrant ma thèse.

Je remercie toute la famille ICPéenne pour la disponibilité de tous, chercheurs, techniciens, thésards. Je salue en particulier mes camarades d'équipe Thierry, Bertrand et Ali avec qui j'ai pu partager travail et bonnes rigolades, et nos trois mamans Dominique, Joëlle et Nadine, si patientes avec nos impatiences. Parmi tout le personnel scientifique de l'ICP qui m'a aidé après la disparition de Christian, je remercie en particulier Gérard Bailly, grâce à qui j'ai pu mener à bien la rédaction de ce manuscrit.

Les nouveaux horizons que m'a fait découvrir Christian m'ont amené très loin de Grenoble. Mon séjour aux laboratoires HIP-ATR au Japon fut en tout point décisif. Je remercie mille fois Eric Vatikiotis-Bateson-san de m'avoir accueilli dans son équipe... et dans sa maison. La curiosité scientifique et la générosité d'Eric lui ont fidèlement attaché un groupe de collaborateurs de haute valeur. Merci à Kuratate, Hani, Frédérique et Marc.

Cette thèse est maintenant terminée et je remercie Monsieur Chassery, président du jury, Messieurs Chatila et Mariani, mes rapporteurs, Messieurs Senn, Bailly, Dolmazon et Vatikiotis-Bateson, membres du jury, d'avoir accepté d'y prêter leur regard scientifique.

Je termine en adressant toute ma reconnaissance à mes parents. Tout le long de ce chemin, vous m'avez appris à ne jamais abandonner devant les difficultés. C'est votre exemple que j'ai suivi en terminant cette thèse.

Ces dernières lignes sont enfin pour le soleil de ma vie, Anne-Claire, toi qui a accepté d'épouser un thésard angoissé... merci pour ton soutien sans failles, merci pour ton amour.

I. Table des matières

Liste des figures

I. INTRODUCTION

L'analyse fiable des mouvements faciaux occupe aujourd'hui une place importante dans l'étude des signaux de la parole. Au sein de cette étude, les lèvres s'imposent comme un des organes visibles les plus informatifs et les plus accessibles à la mesure. Dans le cadre de la communication homme-machine, le signal visuel de lèvres parlantes peut s'appréhender à la fois comme modalités d'entrée et de sortie. La machine peut lire sur les lèvres en intégrant des paramètres labiaux dans les systèmes de reconnaissance automatique et réduire considérablement sa sensibilité au bruit ambiant. Elle peut aussi synthétiser à l'écran l'image de visages parlants. Dans les deux cas l'enjeu est d'isoler et de caractériser les gestes de parole produit par les lèvres.

Depuis 10 ans, l'Institut de la Communication Parlée peint les lèvres en bleu pour « voir la parole ». S'appuyant sur ce qui a déjà été « vu », l'objectif de cette thèse est de continuer à « voir sans maquillage » et de l'intégrer dans un système optimisé pouvant atteindre un traitement en temps réel - pour un flot vidéo standard, le temps réel signifie 50 images par seconde. Au delà de la conception micro-électronique, cet objectif relève autant de la vision par ordinateur que de la parole : de la vision par ordinateur, parce que les déformations et l'aspect de l'objet labial exigent des traitements élaborés de l'image; de la parole, parce que les lèvres, qui, du sourire à la grimace peuvent faire beaucoup plus que parler, ne sont étudiées ici *que* dans un cadre de communication langagière, ce qui limite avantageusement l'espace des gestes labiaux observables en regard de ce que la physiologie leur permet.

Beaucoup de travaux sur l'analyse labiale automatique postulent que seules des approches statistiques de la texture et de la forme peuvent résoudre convenablement les problèmes soulevés par l'immense variabilité inter- et intra- locuteurs, et ceux liés aux variations des conditions d'enregistrement (chapitre 1). Cette thèse souscrit complètement à ce point de vue. Mais dès lors qu'est choisie une approche statistique se pose le problème des données d'apprentissage : comment les modéliser, comment les sélectionner ? Beaucoup lient la représentativité d'un corpus à l'importance de sa taille. Sans réfuter ce point de vue, le travail présenté ici propose néanmoins une alternative. Ce travail défend la thèse que, en se basant sur un échantillonnage pertinent de la forme des lèvres, il est suffisant, pour retrouver les gestes labiaux sur une séquence d'images, d'utiliser un corpus constitué d'une dizaine de « formes clés », phonétiquement identifiables et pavant de manière optimale l'espace

articulatoire du locuteur. Ce corpus fournit alors la base d'apprentissage pour un codage du signal visuel de parole.

Des milliers de pixels d'une image aux quelques degrés de liberté des lèvres parlantes, c'est la possibilité d'une compression du signal visuel que cette thèse tentera d'abord de mettre en évidence, puis de mettre en œuvre dans la définition d'une architecture matérielle dédiée à la lecture labiale automatique et la mesure de paramètres géométriques. Cet enjeu théorique et technologique rejoint les objectifs fixés par la normalisation MPEG4 (1999) pour le codage audiovisuel des animations faciales.

A partir de la couleur des pixels des lèvres, une première analyse statistique permet de générer une image en niveaux de gris où le contraste entre les lèvres et le reste du visage est rehaussé (chapitre 2). Lorsque les lèvres sont maquillées en bleu, cette étape suffit à isoler convenablement la « forme » du « fond ». La proximité des couleurs des lèvres, de la peau et de la langue la rend cependant insuffisante pour une analyse géométrique précise lorsqu'aucun artifice cosmétique n'est employé. Ainsi, la définition de modèles géométriques contraignant la forme des organes et leur adaptation aux images, où cette forme n'est qu'imparfaitement estimée, fournissent une contrainte supplémentaire permettant de régulariser le problème. Cette technique, héritée de l'imagerie médicale, régularise les contours par divers critères géométriques. Quelques modèles spécifiques aux lèvres ont été proposés mais, dans un souci d'universalité, n'ont jamais suffisamment contraint la forme pour éviter qu'elle ne se perdre dans le fond. La modélisation proposée dans cette thèse se limite aux variations géométriques des lèvres propre à la parole. Le modèle s'appuie sur un échantillonnage par points de contrôle permettant de définir un modèle géométrique 3D pour toute forme de lèvres. L'analyse statique d'un corpus phonétique d'une dizaine de « formes clés » donne ensuite le codage des gestes labiaux propre à un locuteur (chapitre 3). Le suivi automatique consistera alors à inverser le modèle complet en recherchant la configuration des paramètres articulatoires qui ajustent au mieux la projection du modèle 3D à celle des lèvres estimée sur l'image 2D par l'analyse de la couleur. Une attention toute particulière est portée à l'évaluation des résultats ainsi obtenus pour la mesure labiale (chapitre 4). Après l'évaluation de la pertinence de la méthode, c'est sa structure algorithmique et sa complexité en calcul qui seront enfin étudiées en vue de son implantation sur une architecture matérielle (chapitre 5).

I. Chapitre 1. De la parole audiovisuelle à l'analyse labiale automatique

Le téléphone et la radio prouvent la capacité d'une parole purement auditive à transmettre avec efficacité une communication langagière. Néanmoins, la perception humaine tire aussi profit de l'information visuelle apportée par le visage du locuteur notamment lorsque les conditions acoustiques sont dégradées. C'est cette bimodalité intrinsèque, et le gain d'intelligibilité qu'elle apporte, qu'explore l'étude de la parole audiovisuelle. Mise en évidence pour la communication humaine, elle ouvre de nouvelles perspectives pour la communication *avec* et *par* la machine.

Bien que la communication orale engage l'ensemble du visage du locuteur, les lèvres occupent une place privilégiée : elles fournissent une source visuelle d'information pour la perception de la parole et, étant toujours identifiables, se prêtent à une analyse automatique. La capture automatique des mouvements labiaux (ou labiométrie) tend à doter l'ordinateur de paramètres intelligibles et indépendants pour contrôler des visages synthétiques parlants ou bien identifier le message énoncé par une reconnaissance audiovisuelle automatique. Les difficultés technologiques résident dans la complexité de ces mouvements et la variabilité intra- et inter- locuteurs.

I.1 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.

I.1.1 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 et 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 et 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 et 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 et Dodd, 1996). Cette fusion est très robuste aux conditions externes puisqu'elle persiste même lorsque les sujets sont prévenus de l'effet. Celui-ci 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.

I.1.2 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 ; Erber, 1969 ; Binnie et al., 1974 ; Erber, 75 ; Summerfield, 1979, 1989 ; Benoît, 1996).

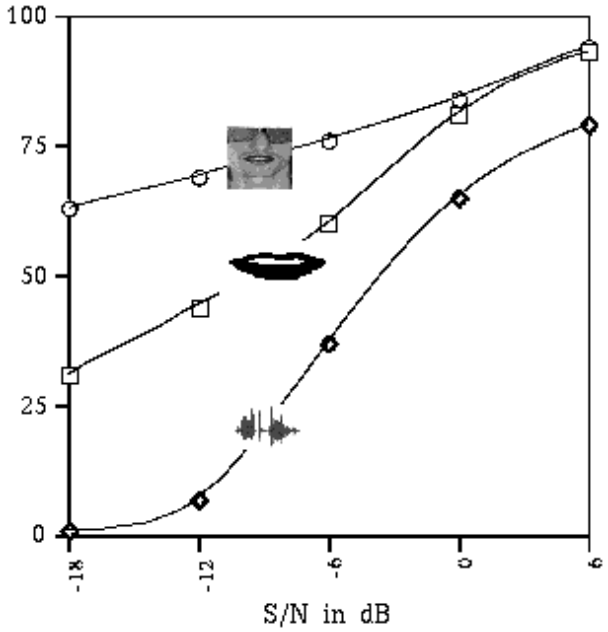


Figure 1. Comparaison de l’intelligibilité de la parole bimodale en condition bruitée en ajoutant successivement les lèvres puis tout le visage du locuteur (Benoît et al., 1996).

La Figure 1 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, 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.

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.

I.1.3 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.

I.1.3.1 La synthèse audiovisuelle de visages parlants

Le problème crucial auquel se heurte la génération de parole audiovisuelle synthétique réside dans la capacité à préserver la cohérence les signaux acoustiques et visuels, cohérence dont l'importance a été soulignée dans la section précédente si l'on veut accéder à un certain réalisme et obtenir de la machine parlante une intelligibilité audiovisuelle satisfaisante. Un moyen naturel d'animer une tête parlante synthétique consiste à rechercher directement sur un visage humain parlant les paramètres de commandes du modèle de synthèse (capture du mouvement). Ces paramètres peuvent provenir de mesures sur l'ensemble du visage (Essa et al., 1994 ; Kuratate et al., 1998) ou seulement sur les lèvres (à l'ICP : Benoît et al., 1996 ;Guiard et al., 1996 ; Le Goff, 1997). En complément, des travaux ont portés sur l'extraction de paramètres géométriques à partir du signal acoustique (Lavagetto, 1995; Yamamoto et al., 1997). En matière de visages parlants Rubin et Vatikiotis-Bateson (1998) ont récemment développé sur le serveur Web des laboratoires Haskins un site donnant un aperçu détaillé de l'état de l'art actuel (http://www.haskins.yale.edu/haskins/head.html).

Le modèle 3D de lèvres de Guiard et al. (1996) et le modèle de visage où il est intégré (Le Goff, 1997) sont contrôlés par des paramètres labiaux géométriques mesurés sur les contours externe et interne des lèvres (étirement horizontal, aperture verticale) et sur la vue de profil (protrusion du point de contact des lèvres supérieure et inférieure, protrusion des lèvres supérieure et inférieure). Les contours des lèvres d'un locuteur sont automatiquement extraits avec précision grâce à un maquillage bleu et un fort éclairage. L'animation audiovisuelle est alors réalisée en présentant la voix du locuteur et la synthèse du modèle 3D contrôlé par les paramètres labiaux extraits simultanément. Des tests sur des stimuli VCVCV ont montré un gain d'intelligibilité en présence de bruit ajouté sur la bande sonore. Les résultats obtenus n'atteignent bien sûr pas la qualité que fournit un visage humain naturel (Benoît, 1996). L'animation audiovisuelle du visage a été implantée dans un système de synthèse audiovisuelle à partir du texte fournissant à la fois la séquence audio et la trajectoire des paramètres de contrôle depuis une phrase tapée au clavier. Ici encore, un gain d'intelligibilité en présence de bruit a été constaté (Le Goff, 1997).

Ces travaux montrent l'intérêt que présente une synthèse visuelle contrôlée par les seuls mouvements labiaux. Même si toute l'information nécessaire à une animation réaliste n'est pas utilisée, ces approches ont mis en évidence un gain d'intelligibilité.

I.1.3.2 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 ; Yuhas et al., 1990 ; Stork et al., 1992 ; Waibel et Lee, 1990 ; Bregler et al., 1993 ; Rogozan et al., 1996; Luettin, 1997).

A l'ICP, 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.

I.1.3.3 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.

I.1.3.4 le rôle de la labiométrie

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 (détaillé au §I.4). 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. Ainsi, avant de dresser un bilan des différentes techniques employées aujourd'hui, les sections suivantes présentent un aperçu de l'anatomie labiale pour en mesurer la complexité et donnent quelques repères phonétiques replaçant les lèvres au centre d'une tâche de production de parole.

I.2 L'anatomie des lèvres

I.2.1 Les tissus

D'après les données anatomiques présentées dans « Labialité et Phonétique » (1980), les lèvres forment deux replis musculaires, recouverts d'une membrane, qui circonscrivent l'orifice de la cavité buccale. Ces replis supérieur et inférieur sont indépendants et se réunissent à leurs extrémités pour former les commissures labiales. La face externe des lèvres est recouverte par de la peau et la face interne par de la muqueuse composée de cellules disposées comme des pavés (l'épithélium). Les muscles se trouvent directement sous la peau.

La ligne entre la peau et la muqueuse dessine dans sa partie supérieure et, au centre, une courbe concave dénommée « arc de Cupidon ». Elle délimite une zone de transition, dite vermillon. Celle-ci se caractérise par sa haute teneur en un liquide semi-fluide qui augmente la transparence du tissu, à tel point qu'on aperçoit la teinte rouge de la couche vasculaire sous-

jacente. C’est cette caractéristique qui fait ressortir la couleur des lèvres par rapport au reste de la peau. La zone de vermillon de la lèvre supérieure montre, en son milieu, une protubérance : le tubercule.

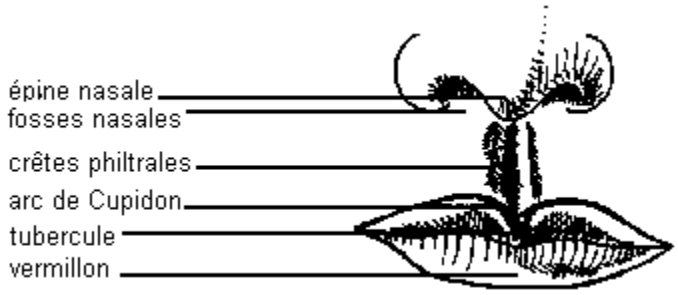


Figure 2. Aspect schématique des lèvres (d’après Zemlin, 1968).

A l’intérieur de la bouche, la muqueuse de teinte rosée rejoint les arcades alvéolo-dentaires. L’espace incurvé, ainsi délimité, forme les gouttières vestibulaires. Dans leurs parties médianes, les gouttières vestibulaires supérieure et inférieure présentent un repli muqueux : le frein de la lèvre. Celui-ci est nettement plus proéminent pour la lèvre supérieure.

I.2.2 Les muscles des lèvres

Les muscles des lèvres font partie des muscles faciaux. Ils ont tous la particularité de présenter une insertion mobile cutanée. C’est cette caractéristique qui rend possible les différentes combinaisons d’expression du visage et la souplesse des mouvements en production de la parole. Le muscle essentiel des lèvres est l’orbiculaire des lèvres qui opère comme un sphincter annulaire. Autour de celui-ci, rayonnent les autres muscles de la face dont les fibres s’imbriquent directement avec celles de l’orbiculaire.

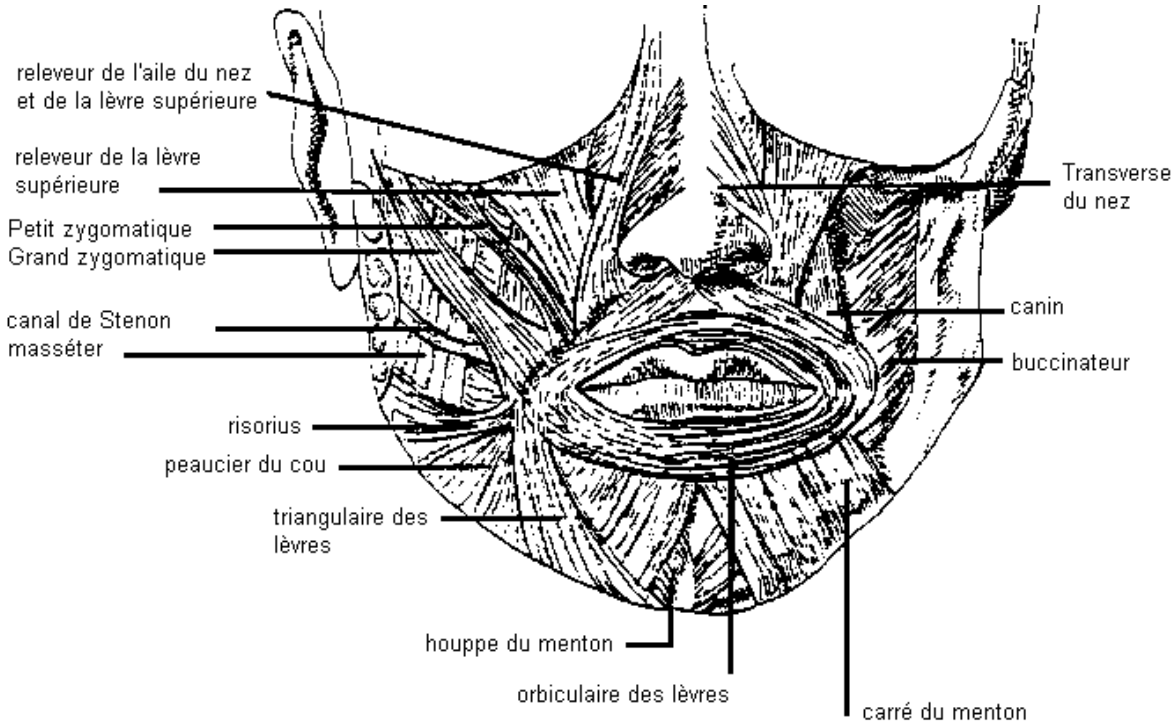


Figure 3. Les muscles de la face (d'après Bouchet et Cuilleret, 1972).

Les classifications courantes dénombrent douze muscles pour les lèvres (Zemlin, 1968 ; Rouvière et Delmas, 1972 ; Hardcastle, 1976) :

- l'orbiculaire des lèvres (orbicularis oris),
- le canin (levator anguli oris),
- le buccinateur (buccinator),
- les muscles de la houppe du menton (mentalis),
- la carré du menton (quadratus labii inferioris, ou depressor labii inferioris),
- le releveur superficiel de l'aile du nez et de la lèvre (levator labii superioris alaeque nasi),
- le releveur profond (levator labii superioris),
- le petit zygomatique (zygomaticus minor),
- le grand zygomatique (zygomaticus major),
- le risorius,
- le triangulaire des lèvres (depressor anguli oris).
- le peaucier du cou (muscle platysma).

I.2.3 Classification fonctionnelle des muscles labiaux

En complément d'études anatomiques, des mesures par electromyographie ont permis de dresser une classification des muscles labiaux suivant les mouvements qu'ils génèrent. Cette classification suit celle de Hardcastle (1976), reprise dans « Labialité et Phonétique (1980) ». Elle présente les tendances générales observées chez plusieurs sujets.

Muscles assurant l'occlusion labiale

Par contraction l'orbiculaire accole les lèvres supérieures et supérieures en abaissant la lèvre supérieure et en tirant la lèvre inférieure vers le haut. Le mouvement de la lèvre inférieure est fortement dépendent de la mâchoire. Le canin et le triangulaire peuvent aussi intervenir pour fermer les lèvres.

Muscles assurant la protrusion des lèvres

La protrusion correspond à un mouvement poussant les lèvres vers l'avant, s'accompagnant d'un rapprochement des lèvres et des commissures. C'est aussi une des fonctions principales de l'orbiculaire. La houppe du menton contribue à faire basculer la lèvre inférieure.

Muscles assurant l'arrondissement des lèvres

L'arrondissement correspond à une forme de lèvres obtenue en rapprochant les commissures. Ce geste s'oppose à l'étirement. Bien que l'arrondissement s'obtienne par une contraction de l'orbiculaire, ce geste ne s'accompagne pas forcement d'une protrusion. Des muscles comme le buccinateur ou le risorius peuvent limiter l'action de l'orbiculaire.

Muscles élévateurs de la lèvre supérieure

Comme leur nom l'indique, les releveurs supérieurs et profonds de la lèvre sont attachés à cette fonction. Du fait de leur insertion, c'est essentiellement la partie centrale de la lèvre supérieure qui est relevée.

Muscles abaisseurs de la lèvre inférieure

La lèvre inférieure est tirée vers le bas par la carré du menton. Ce muscle peut être aidé par la mâchoire. De même, le triangulaire peut aussi intervenir pour abaisser la lèvre inférieure.

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.

I.3 Repères phonétiques

I.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 ses 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 4 indique les 8 lieux d'articulation principaux identifiés en phonétique.

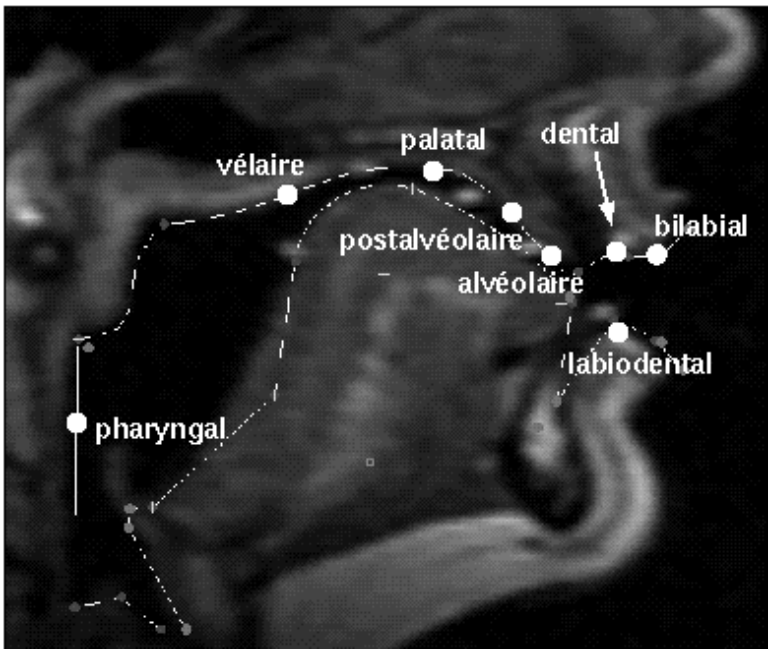


Figure 4. Le conduit vocal et les 8 lieux d'articulation principaux.

I.3.2 Des sons et des lèvres

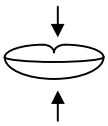
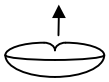
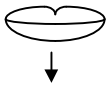
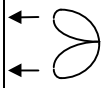
En maintenant stables et non ambiguës les différences entre les sons articulés, une représentation sensible (acoustique et visuelle) du code phonologique peut être mise en *commun* entre celui qui parle et celui qui écoute, d'où la mise en place d'une *communication*. L'ensemble fini des sons d'une langue 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 (Labialité et Phonétique, 1980) :

- voyelles arrondies (/y/, /u/, /o/, /ø/, ...), 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/, /ɛ/, /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 (/ʃ/, /ʒ /),
- 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
D.A.O. = Depressor Anguli Oris
D.L.I. = Depressor Labii Inferioris
L.A.O. = Levator Anguli Oris
LL.S. = Levator Labii Superioris
M. = Mentalis
O.O. = Orbicularis Oris
O.O.L. = Orbicularis Oris Inferior
Plat. = Platysma
Ris. = Risorius
Zyg. Maj. = Zygomaticus Major
Zyg. Min. = Zygomaticus Minor
(a) = Action antagoniste
(p) = Action protagoniste
(s) = Action synergique

Figure 5. Les réalisations articulatoires et les mouvements labiaux correspondant (d'après Labialité et Phonétique, 1980).

I.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 différents 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 (Öhman, 1966 ; Whalen, 1990).

Les études sur la géométrie labiale rassemblées dans « Labialité et Phonétique » (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 /ʃ/ ou étirées /z/. Pour illustrer l'importance de la coarticulation, il est montré par exemple que, sur 6 locuteurs prononçant une syllabe /ʃi/, la protrusion pour l'articulation du /ʃ/ 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 /ʃy/.

La modélisation de la coarticulation est un enjeu important pour la synthèse visuelle de la parole. Dans sa thèse, Le Goff (1997) a adapté pour le Français le modèle de coarticulation par fonction de dominance proposé par Cohen et Massaro (1993). Ce modèle est destiné à la synthèse de visage parlant à partir du texte. Le principe est de générer la trajectoire des paramètres articulatoires contrôlant la synthèse en superposant et en additionnant à chaque instant la contribution visuelle de chacun des phonèmes de la phrase. Cette contribution est estimée par les valeurs des paramètres articulatoires prises pour représenter visuellement le phonème, pondérées par une fonction de dominance, maximale à l'instant où le phonème apparaît dans la phrase et décroissante en aval (anticipation) et en amont (rétention).

Nous retiendrons que les formes labiales « possibles » en parole dépendent à la fois d'articulations « cibles » et de phénomènes de coarticulation.

I.4 Etat de l'art en mesure labiale ou labiométrie

Les méthodes actuelles de labiométrie sont issues de recherches sur la reconnaissance de formes en vision par ordinateur. Les techniques employées peuvent se classer selon deux grands types d'approches suivant qu'elles sont orientées « image » ou « modèle ». On désigne par orientées image des approches ascendantes, qui d'un traitement directement au niveau pixel déduisent des paramètres de haut niveau. A l'inverse, les approches descendantes dites orientées modèles utilisent une représentation géométrique des contours externe et/ou interne des lèvres que l'on cherche à adapter aux contours labiaux tels qu'ils apparaissent sur l'image.

I.4.1 Analyses orientées « image »

I.4.1.1 Séparation d'histogramme

Par une segmentation de la texture (au sens de la couleur représentée par un ou plusieurs canaux chromatiques), la région du vermillon est isolée du reste du visage afin d'obtenir les caractéristiques géométriques des contours internes et externes. Ces méthodes ont l'avantage d'être simples mais sont très sensibles aux conditions d'éclairage.

A l'ICP, le système de Lallouache (1991) utilise un traitement par chroma-key sur une séquence vidéo d'un locuteur dont les lèvres ont été peintes en bleu. Si cette méthode est précise, elle est néanmoins tributaire du maquillage et de l'éclairage. La difficulté à donner une caractérisation unique de la couleur du vermillon rend inappropriée une transposition dans des conditions sans maquillage. La zone extraite ne peut donner qu'une estimation imprécise des contours et interdit donc toute déduction de paramètres géométriques pertinents. Petajan et al. (1984, 1986) ont cependant utilisé des mesures sur un tel système de segmentation des lèvres (non maquillées). Ces paramètres ont été validés pour une tâche de reconnaissance visuelle de quelques voyelles.

Finn (1986) extrait la position de marqueurs placés sur le contour externe pour extraire avec précision des paramètres géométriques. De telles méthodes ne permettent pas d'extraire le contour interne et imposent de plus la contrainte ergonomique des marqueurs (Stork et al., 1992 ; Cosi et al., 1996). Tout comme les méthodes à base de maquillage, elles ont cependant contribué à démontrer la pertinence des paramètres visuels pour des applications de reconnaissance automatique.

Prasad et al. (1993) ont utilisé après filtrage les minima et maxima de niveaux de gris pour localiser la position des coins des lèvres et les sommets et bas du contour externe. De la même manière, Mak et Allen (1994) ont complété cette approche par une analyse cinématique en utilisant la soustraction pixel à pixel entre deux images consécutives. Ces méthodes supposent qu'il existe un fort contraste de luminance entre les lèvres et le reste de la peau, ce qui est en général peu vérifié.

Les approches par champs de Markov utilisent un lissage spatio-temporel des pixels de la région des lèvres en se basant sur un modèle statistique de proximité. Si la région segmentée ne permet pas une extraction précise de paramètres visuels pertinents comme avec un maquillage, les résultats donnent une estimation de la zone des lèvres plus robuste que les approches précédentes (Liévin et Luthon, 1998).

Ecartant le problème d'extraction de paramètres géométriques, Yuhas et al. (1990) ont présenté un système de reconnaissance où la totalité des pixels de l'image des lèvres en niveaux de gris était prise comme vecteur d'analyse dans un système de reconnaissance automatique. Pour réduire le nombre de paramètres, l'image était sous échantillonnée pour donner un vecteur de 20 pixels. Une approche similaire a été suivie par Wu et al. (1991). Inadaptée à la labiométrie, cette méthode reste limitée à l'application de reconnaissance automatique à un vocabulaire spécifique.

I.4.1.2 Flux optique

Mase et Pentland (1991) ont présenté une application du flux optique (Barron et al., 1992) à la reconnaissance audiovisuelle de la parole. Les méthodes ‡ base de flux optique sont à classer dans les approches images du fait de la représentation dense de l'information : l'analyse associe à chaque pixel un vecteur vitesse correspondant au flot d'intensité observé entre deux images consécutives.

I.4.1.3 Analyse statistique d'images en niveaux de gris

L'utilisation de l'image intégrale en niveaux de gris comme vecteur de mesure optique oblige à gérer un grand nombre de paramètres égal au nombre de pixels de l'image. Les travaux de Turk et Pentland (1991) en reconnaissance automatique de visages ont montré que la redondance d'information pouvait être réduite par décomposition d'une image sur une base de vecteurs propres obtenus par une analyse en composantes principales (ACP). L'ACP est

calculée sur quelques images en niveaux de gris d'un corpus d'apprentissage. En s'assurant que le corpus est globalement représentatif des différentes variations, la dimension des vecteurs d'analyse est alors réduite du nombre de pixels d'une image au nombre d'individus du corpus.

En utilisant des images centrées sur les lèvres et les paramètres de projection sur les vecteurs propres, cette technique a été appliquée à la reconnaissance visuelle de la parole (Brooke et al., 1994 ; Bregler et Konig, 1994 ; Murase et Sakai., 1996). Les paramètres visuels extraits par cette approche n'ont été utilisés que pour des application de reconnaissance automatique. Au chapitre 2, nous présentons une méthode originale les reliant avec une prédiction de paramètres géométriques. Il est à noter que ces méthodes restent très sensibles à des transformations géométriques simples de l'image telles que translation, rotation et homothétie qui apparaissent suite à des mouvements de tête du locuteur.

I.4.2 Analyses orientées « modèle »

Ces méthodes utilisent un modèle géométrique des contours labiaux, contrôlé par peu de paramètres. Les paramètres sont déduits par optimisation de telle sorte que le modèle s'adapte au mieux avec les contours réels. Ces derniers sont le plus souvent estimés par un gradient spatial de l'image, maximal aux sauts de texture.

I.4.2.1 Les modèles déformables

Yuille et al. (1992) présentent une technique générale de modélisation de contour par courbe polynomiale (« Deformable Templates », Yuille et al., 1992). Plus le nombre de paramètres de contrôle est réduit, plus le processus de convergence du modèle est stabilisé. Ainsi, les courbes du modèle sont souvent décrites par quelques paraboles. L'application directe aux lèvres de cette méthode a été traitée notamment par Rao et al. (1994) et Hennecke et al. (1994). L'inconvénient majeur de cette méthode tient à la simplicité des courbes qui ne permet pas toujours une modélisation précise des différentes formes.

I.4.2.2 Les contours actifs

La méthode des Snakes développée par Kass et al. (1993) utilise une description des contours sous forme de splines (« Numerical Recipes in C », Press et al., 1992). Elle s'appuie sur une régularisation des dérivées locales premières et secondes de l'abscisse curviligne d'un contour défini par des splines. La fonction à minimiser est associée à un terme d'énergie. Elle

additionne une énergie externe représentant l'adéquation du modèle avec une position estimée des contours labiaux (classiquement une zone de gradient élevé) et une énergie interne liées aux dérivées locales. Des applications aux contours labiaux ont été faites notamment par Kaucic, (1996) et Bregler et al. (1995). A l'inverse des modèles déformables, les « snakes » présentent plus de souplesse pour représenter les contours. Mais de ce fait, la forme du modèle est moins contrainte et se stabilise parfois sur de « mauvais » contours : ombre, saut d'un contour labial à un autre.

I.4.2.3 Les modèles statistiquement contraints

Cootes et Taylor (1992) ont présenté une approche générale de suivi de contours avec une énergie interne limitée statistiquement par apprentissage d'un ensemble de formes : les Active Shape Models (ASM), ou « Smart Snakes ». L'ensemble des formes d'apprentissage est obtenu par un étiquetage manuel. Il est statistiquement modélisé par une distribution gaussienne des coordonnées XY des points du modèle. Au cours du suivi automatique, les déformations du modèle sont alors contraintes à rester dans les limites de la distribution gaussienne. Ce modèle général pour le suivi automatique a été appliqué aux contours labiaux (Luettin 1996 ; Garcia et Vatikiotis-Bateson, 1997). Luettin (1997) a couplé au modèle statistique de forme, un modèle statistique des niveaux de gris. Ce système a été testé sur une base de plusieurs locuteurs dans une tâche de reconnaissance de chiffres purement visuelle.

Kaucic et al. (1996) contraignent dans le temps les déformations d'un modèle à base de splines par un filtrage de Kalman. Les déplacements des points de contrôle sont limitées a priori à un ensemble fixé de degrés de liberté au départ arbitraires puis ensuite complétés par une ACP sur les coordonnées des points de contrôle. Le filtrage de Kalman régularise l'évolution dans le temps des déformations du modèle.

I.4.3 Apprentissage et analyse-synthèse de modèle

Nous citons ici une dernière approche suivie par Basu (1997). Elle se distingue des approches classiques en ce sens qu'elle vise, avant le suivi automatique, à fournir un modèle réaliste des mouvements labiaux en parole. Le modèle de Basu est construit selon une surface tridimensionnelle. Il est défini par un réseau de 336 polygones et comprend en plus du vermillon quelques polygones pour la peau. Les déformations locales des points de la surface sont régularisées par un modèle en éléments finis basé sur des règles d'élasticité linéaire.

L’ensemble des points 3D confèrent au modèle un nombre important de degrés de liberté. Une analyse automatique a été faite pour capturer le mouvement en 3D de points marqués à l’encre sur la surface des lèvres d’un locuteur filmé selon deux vues cohérentes. Le modèle d’élasticité permet de déduire la position des autres points qui n’étaient pas filmés. Cette base d’apprentissage est alors modélisée statistiquement pour définir des contraintes sur les variations du modèle.

Ce modèle a été appliqué à un suivi automatique des mouvements labiaux en parole. L’analyse automatique, sans l’aide de maquillage ni de marqueurs, se fait en coopération avec une détection chromatique de la zone du vermillon (Oliver et al., 1998). Le suivi automatique consiste alors à adapter la projection 2D du modèle 3D pour que cette projection corresponde au mieux à la zone estimée des lèvres sur une image.

La méthode de suivi automatique présentée dans cette thèse suit une approche similaire d’analyse-synthèse. Comme Basu, nous définissons un modèle 3D qui vise à représenter de manière pertinente les mouvements d’un locuteur par apprentissage. Néanmoins dans le cas de Basu, le fait d’utiliser des marqueurs physiques pour l’apprentissage interdit toute mesure du contour interne dont le contrôle a un rôle fondamental dans l’articulation et la perception visuelle. Situés sur la surface intermédiaire du vermillon, ces marqueurs ne mesurent pas toute l’information pertinente en parole. Par ailleurs, dix modes principaux de variation ont été déduits de l’analyse statistique du corpus d’apprentissage. Un des points importants de la thèse présentée ici est de montrer que, pour le français et deux locuteurs représentatifs, l’articulation labiale peut se réduire à seulement trois degrés de liberté, ceux-là même précédemment cités dans notre introduction.

I.5 Discussion

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 video.

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 géométrique de haut niveau (contours décrit en terme 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.

Plus le modèle sera réaliste, plus il sera possible d'aller le « rechercher sur l'image ». En ce sens, l'approche de Basu nous paraît parmi toutes la plus intéressante. Nous nous en distinguerons en recherchant le réalisme à travers un modèle 3D articulatoire rassemblant les connaissances phonétiques acquises à l'ICP par l'étude systématique des mouvements labiaux en parole.

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 plupart des travaux en suivi labial automatique tendent à valider leur résultats par une application de reconnaissance visuelle automatique d'un vocabulaire de quelques mots. Une telle approche ne permet de valider au delà du vocabulaire la pertinence des paramètres extraits. Nous nous attacherons donc à fournir une validation d'ordre géométrique sur la qualité de prédiction des paramètres d'étirement et d'ouverture dont l'importance a été prouvée par ailleurs.

II. Chapitre 2. Labiométrie par analyse statistique de la couleur

L'analyse automatique de la texture doit faire face à la fois aux variations physiologiques entre individus et aux variations dues aux conditions d'éclairage. Un modèle statistique vise à apprendre pour *un locuteur* et *une situation* donnés la couleur des lèvres telle qu'elle apparaît sur l'image numérisée. La deuxième partie de ce chapitre présente une méthode où tous les pixels de l'image des lèvres d'un locuteur forment un vecteur d'observation de l'image pour une analyse statistique en composantes principales. En préparation au chapitre suivant qui aborde les degrés de liberté des lèvres, cette analyse statistique vise à présenter des premiers résultats sur la redondance d'un corpus de parole et la possibilité d'une compression du signal visuel.

II.1 Analyse statistique de la couleur des lèvres

II.1.1 Acquisitions des données

Sous forme de carte électronique ou de circuit analogique/numérique, de nombreux systèmes d'acquisition vidéo sont aujourd'hui disponibles. Ils permettent de numériser des images en couleur à partir du signal vidéo analogique. La numérisation se fait à la cadence de 25 images par secondes pour le standard vidéo PAL et 30 pour le standard NTSC. L'image est échantillonnée sur un rectangle de 768 sur 576 pixels en PAL et 640 sur 480 en NTSC. Chaque pixel est codé par trois canaux quantifiant le rouge, le vert et le bleu de la couleur. La dynamique numérique de ces canaux déterminent la finesse avec laquelle les nuances de couleur sont représentées. Les valeurs les plus courantes vont de 8 bits pour les trois canaux (256 couleurs) à 8 bits par canal (16 millions de couleurs). Le codage sur plus de 24 bits par canal s'adresse à des équipements moins standards.

Bien que la numérisation d'une image complète de 768 sur 576 pixels codés sur 24 bits est maintenant facilement réalisable à la cadence vidéo, le transfert synchronisé sur disque se heurte à des problèmes technologiques. Dans l'optique de réduire le débit de transfert pour le stockage de la séquence vidéo, souvent sont proposées des options de compression numérique des images. Tout comme la réduction du nombre de couleurs, si ces traitements continuent à laisser l'aspect visuel acceptable, ils agissent comme des filtres passe-bas et diminuent la finesse de représentation des couleurs.

La proximité des couleurs du vermillon et de la peau rend crucial ce problème de quantification pour l'identification des lèvres. Aussi, pour limiter la perte d'information et rester dans le domaine d'équipement accessible, toutes les images ont été numérisées sur 24 bits sans compression à la cadence de 25 images par seconde. Les deux systèmes de numérisation qui ont été utilisés sont les suivant :

- à l'ICP, une carte d'acquisition VINO sur station Silicon Graphics INDY. Cette configuration permet de stocker une séquence vidéo d'une dizaine d'images à partir d'un signal PAL ou NTSC, entrée composite ou Y/C . Cette carte dispose aussi d'une bibliothèque de programmation en C++ qui permet d'interfacer celle-ci avec des logiciels de traitement.
- aux laboratoires ATR-HIP, une station XTREME de la société ACCOM. Ce système permet de stocker en temps réel une séquence d'une dizaine de minutes à partir d'un signal composite PAL ou NTSC, entrée Y/C ou RGB analogique. La qualité et le coût de ce système sont sans commune mesure avec le précédent.

Ces deux systèmes permettent de générer une suites d'image au format 24 bits (3 canaux de 8 bits) propre à Silicon Graphics, sans compression, à 25 images par seconde. Chaque image numérisée se compose d'une trame paire et impaire entrelacées. Ces deux trames correspondent à deux demi images (768 sur 288 pixels) échantillonnées à 20 ms d'intervalle et sont séparables. Les paramètres d'acquisition des deux systèmes (gain, corrections, ...) sont adaptables mais la configuration de base a été utilisée.

Les valeurs RGB sont fortement influencées par les conditions de l'enregistrement initial. Trois sources d'enregistrements ont été utilisées:

- enregistrement au format PAL, entrée composite, à partir de deux caméra Tri CCD, orientées de face et de profil dont les signaux ont été juxtaposés côte à côte par une table de mixage. Ces enregistrements ont été effectués avec un éclairage artificiel de 1000W, orienté face au sujet. Les prises de vue sont cadrées sur la moitié basse du locuteur. Ce type d'enregistrement sera désigné par la suite comme *enregistrement ICP*.
- enregistrement au format PAL, entrée Y/C, à partir d'une micro-caméra identique à celle utilisée sur le casque d'acquisition audiovisuelle du projet Labiophone de la fédération ELESA. Ce casque est commercialisé par la société Ganymedia, partenaire du projet. Le casque dispose d'un éclairage autonome. Cet éclairage est de faible puissance mais la prise

de vue étant focalisée sur les lèvres sous un angle réduit, il assure l'éclairage principal. Les prises de vue sont cadrées sur les lèvres du locuteur. Ce type d'enregistrement sera désigné par la suite comme *enregistrement casque*.

- enregistrement au format NTSC, entrée RGB, à partir d'une caméra Tri CCD orientée de face et un éclairage de 2 fois 100W provenant du bas à gauche et à droite et un faible éclairage ambiant provenant du plafond. Les prises de vue sont cadrées sur la moitié basse du locuteur. Ce type d'enregistrement sera désigné par la suite comme *enregistrement ATR.*

Toutes les séquences ont été enregistrées sur cassette vidéo de type BETACAM-SP. Les enregistrements ICP et casque sont numérisés à partir des cassettes sur la station INDY à l'ICP. Les enregistrements ATR ont été numérisés sur place.

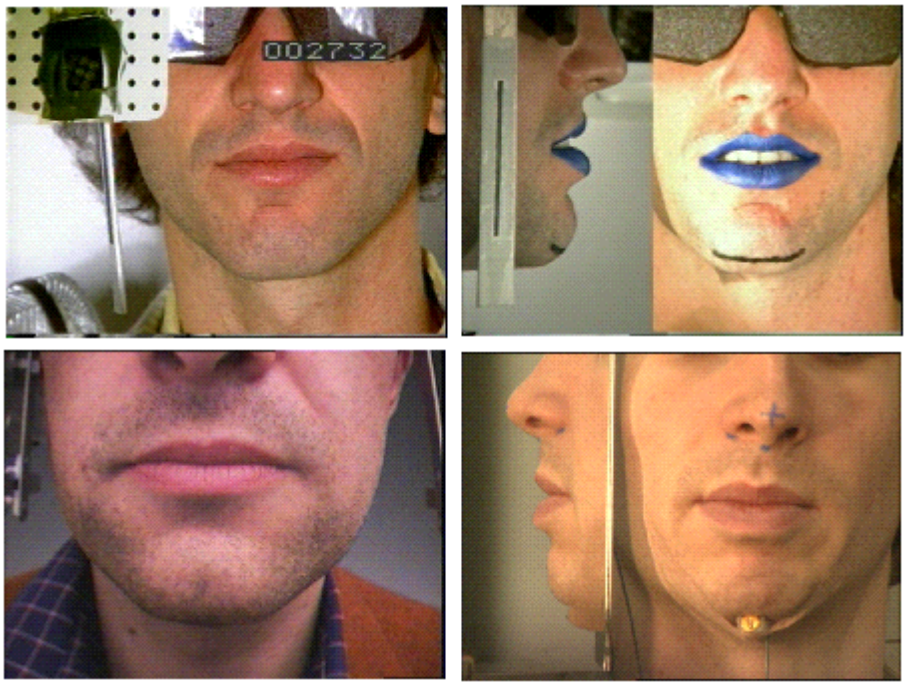


Figure 6. Les enregistrement vidéo : type ICP (en haut), type casque (en bas, à gauche), type ATR (en bas, à droite).

II.1.2 Les systèmes de représentation de la couleur

II.1.2.1 Le système RGB

La justification d'une représentation en composantes RGB s'appuie à la fois sur la physiologie de l'œil et les résultats en colorimétrie. La perception de la couleur chez l'homme est assurée au niveau le plus externe par un ensemble de cellules photosensibles situées sur la rétine. Deux grands groupes de cellules sont à distinguer : les bâtonnets et les cônes. Les bâtonnets

réagissent à des variations faible de l'intensité lumineuse et sont ainsi plutôt sollicités pour la vision nocturne. A l'inverse, les cônes interviennent pour la vision diurne et caractérisent la perception de la couleur. Il a été mis en évidence trois zones de réaction maximale des cônes en fonction de la longueur d'onde de l'onde lumineuse qu'ils reçoivent : à 415-420 nm (bleu-violet), à 530-535 nm (vert) et 560-565 nm (jaune-vert) (Encyclopédie Universalis, article Vision).

Cet aspect trichromatique de la perception de la couleur est repris en colorimétrie. Se basant sur des mesures perceptives, la colorimétrie assure que tout ensemble de trois faisceaux monochromatiques (i.e. une seule longueur d'onde) est suffisant pour représenter les couleurs par combinaison des trois fondamentales orthogonales (aucune des couleurs ne s'obtient par mélange des deux autres). Bien que l'œil humain soit plutôt spécialisé dans la distinction des verts, l'utilisation du rouge, vert et bleu est préconisée pour un meilleur échantillonnage du spectre lumineux.

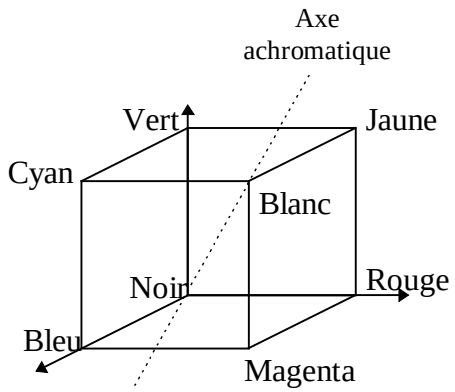


Figure 7. Représentation du cube RGB.

Dans le domaine de la vision artificielle, les systèmes d'enregistrement vidéo tri CCD (Charge Coupled Device) adoptent le plus souvent cette décomposition RGB. Leur fonctionnement consiste à filtrer la lumière simultanément en rouge, bleu et vert puis à envoyer ses trois composantes sur un plan d'éléments photosensibles se chargeant proportionnellement à l'intensité lumineuse qu'ils reçoivent. A chaque cycle de rafraîchissement des éléments, une nouvelle image est analysée. Finalement, le système traduit cet échantillonnage spatio-temporel en signal vidéo analogique. Les systèmes d'acquisition numérique des images opèrent ensuite l'échantillonnage inverse à partir du signal vidéo pour donner une représentation RGB quantifiée d'une séquence d'images. Une exploitation directe du plan des éléments CCD permettrait de s'affranchir de l'étape analogique. Ce point relève de travaux de

recherches à part entière sur les rétines artificielles. A notre connaissance aucun matériel n'est facilement disponible, à l'inverse des systèmes d'enregistrement vidéo et de numérisation.

II.1.2.2 Les systèmes séparés en luminance et chrominance

Les informations chromatique et achromatiques (nuances de gris) sont traitées séparément par l'œil. Connectées aux cônes et aux bâtonnets, les cellules bipolaires et ganglionnaires transmettent au cortex des signaux codant des oppositions de type noir/blanc, rouge/vert et jaune/bleu.

Différents systèmes colorimétriques ont été proposés pour rendre compte de cette séparation entre information chromatique et achromatique. Tous ont en commun la même caractéristique de transformer le système RGB en un autre codage sur trois valeurs où :

- une composante de luminance somme de manière équilibrée les contributions du rouge, vert et bleu,
- deux composantes chromatiques marquent des oppositions entre les couleurs de bases.

Nous exposons ci-dessous diverses représentation de la couleur suivant ce principe.

Représentation RGB normalisée

En considérant les composantes RGB d'égale amplitude, une transformation possible consiste à utiliser la somme des trois composantes RGB et deux composantes couleurs parmi les trois chacune divisée par la valeur de luminance. Il n'existe pas de nomenclature particulière pour cette transformation. Nous la désignerons dans la suite par représentation *RGB normalisée*. Les transformations à partir des composantes RGB sont les suivantes :

$$\begin{cases} L = R + G + B \\ r = R/(R+G+B) \\ g = G/(R+G+B) \end{cases}$$

Représentation YUV

Le système YUV utilisé par les standards de télévision PAL et SECAM exploite de manière plus précise la physiologie de l'œil et sa plus grande sensibilité aux radiation lumineuse situées dans le vert. Le signal de luminance Y pondère différemment les composantes RGB en

favorisant le vert. Les composantes YUV s'obtiennent par combinaison linéaire des valeurs RGB :

$$\begin{cases} Y = 0.299R + 0.587G + 0.114B \\ U = R - Y \\ V = B - Y \end{cases}$$

Représentation de la CIE

Les notions de luminance, teinte et saturation sont couramment utilisées en colorimétrie car facilement évaluable lors de tests de perception. La teinte caractérise la longueur d'onde d'un faisceau coloré et la saturation la pureté de cette couleur (du pastel au vif). S'appuyant sur des mesures perceptives, la Commission Internationale de l'Eclairage (CIE) propose plusieurs normes colorimétrique (xy, u*v* et a*b*). Ces représentations opèrent une première transformation linéaire similaire à celle du système YUV et deux transformation non linéaires. Ces deux dernières transformations sont plus difficiles à intégrer dans un système d'analyse automatique temps réel tel que nous le visons (Wyszecki et Stiles, 1982).

Représentation HSV

Couramment utilisé en infographie, le système HSV donne un codage explicite des valeurs de teinte, saturation et luminance. La représentation HSV s'interprète comme un système de coordonnées cylindrique. La valeur V de luminance désigne la projection du point couleur sur l'axe principal (axe achromatique), la saturation S donne la distance du point à cet axe et la teinte H une valeur d'angle dans le plan chromatique orthogonal à l'axe. Les valeurs HSV sont calculées à partir des valeurs RGB par des transformations non linéaires et inversibles. La non linéarité de ces transformations les rend difficiles à interpréter par rapport aux composantes RGB délivrées par les systèmes de numérisation (Hunt, 1989).

Représentation TLS

Le système TLS suit la même idée d'un codage en teinte, luminance et saturation mais il est calculé par une transformation géométrique du cube RGB. Comme pour le système HSV, la même interprétation en coordonnées cylindrique est applicable. L'axe principal joint le sommet noir (où R=G=B=0) au sommet blanc (R=G=B=1). Cet axe correspond aux points achromatiques pour lesquels R=G=B (nuances de gris). La valeur de luminance L détermine la

position du point couleur sur l'axe. Les valeurs de T et S donnent les coordonnées polaires dans le plan orthogonal à cet axe (plan chromatique). D'un point de vue géométrique, le système TLS correspond à un repère cylindrique associé au repère cartésien du cube RGB. Les formules de transformations à partir des valeurs RGB sont les suivantes :

$$T = \cos^{\langle -1 \rangle}\left(0.5 \frac{2R - G - B}{\sqrt{(R-G)^2 + (R-B)(G-B)}} \right), \text{ si } B>G, \text{ alors } T = 2\pi - T$$

$$L = \frac{R+G+B}{3}$$

La distance exacte d'un point RGB à l'axe achromatique est donnée par la formule suivante :

$$S = \sqrt{R^2 + G^2 + B^2 - RG - RB - BG}$$

Le plus souvent, une approximation (plus rapide en calcul) est utilisée :

$$S = 1 - \frac{1}{L}\min(R, G, B)$$

Cette formule s'annule sur l'axe achromatique et tend vers 0 quand la couleur tend vers le blanc. Elle présente l'inconvénient de retourner constamment la valeur maximale 1 dès qu'au moins une composante RGB s'annule. De manière générale, elle attribue une forte saturation à une coloration sombre. Lorsqu'un calcul de teinte est nécessaire, on remarquera que la première formule donnée pour la saturation est identique au terme au dénominateur de la formule de teinte. Nous n'utiliserons donc que cette première par la suite.

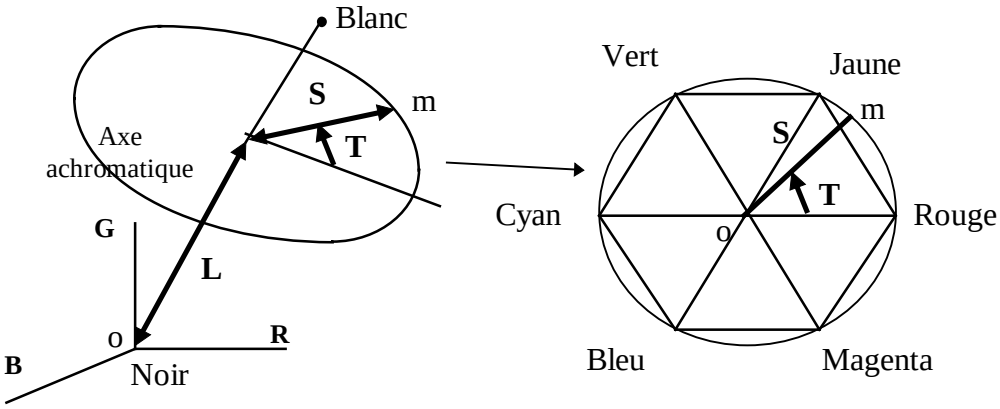


Figure 8. Représentation TLS et projection sur un plan chromatique.

La représentation paramétrique de la couleur reste une question largement débattue et d'autres systèmes de transformation existent. Leurs performances sont relativement équivalentes pour la discrimination des couleurs. Leur pertinence respective dépend essentiellement de l'application visée. Pour le type de calculs utilisés par la suite en vue de l'identification de la couleur des lèvres, nous verrons en particuliers que les performances des représentations RGB, RGB normalisé et TLS sont équivalentes.

II.1.3 Distribution statistique de la couleur du vermillon d'un locuteur

La différence de pigmentation et de vascularisation des tissus fait apparaître le vermillon des lèvres avec une couleur différente de la peau qui l'entoure. Cette couleur varie de manière importante d'un sujet à l'autre. Sa caractérisation est de plus sujette aux conditions d'éclairage de l'enregistrement (à la fois au niveau de la vidéo et de la numérisation). Une simple séparation entre luminance et chrominance ne suffit pas complètement à l'effacer. Un apprentissage de la situation est indispensable et, actuellement, seule une approche statistique peut proposer des solutions acceptables.

Des travaux ont tenté de définir des modèles statistiques universaux (indépendant du sujet et des conditions) en s'appuyant sur des bases de données importantes. Le système de détection de visage LAFTER développé par Oliver et al. (1997) comporte à la fois un modèle pour la couleur des visages et un modèle spécifique pour les lèvres. Les valeurs RGB normalisées sont modélisées par leur distribution statistique selon moyenne et covariance. La distance au modèle de chaque pixel d'une image quelconque est ensuite mesurée par la métrique de Mahalanobis associée à la distribution.

Pour une base d'apprentissage de moyenne $\vec{\mu}$ et de matrice de covariance Σ, la distance de Mahalanobis à cette base d'un point quelconque $\vec{x}$ en coordonnées RGB normalisées, se calcule par la formule suivante (vecteurs colonnes) :

$$d^2(\vec{x}) = (\vec{x} - \vec{\mu})^t \, \Sigma^{-1} (\vec{x} - \vec{\mu})$$

Soit en calculant les vecteurs propres $\vec{e}_{i=1..3}$ et les valeurs propres $\lambda_{i=1..3}$ de la matrice de covariance Σ :

$$d^2(\vec{x}) = \sum_{i=1}^{3} \frac{\left((\vec{x} - \vec{\mu}) \cdot \vec{e}_i\right)^2}{\lambda_i}$$

Ce système a prouvé son efficacité pour des tâches de détection. Il est actuellement utilisé comme support pour les travaux de suivi de gestes labiaux de Basu (1998). Notre approche par apprentissage des formes et gestes labiaux du locuteur (chapitre 3 et 4) nous a permis de disposer, pour chaque locuteur, de segmentations a priori du vermillon et de la peau sur quelques images de référence d'une séquence. Pour garantir une bonne représentativité des classes vermillon et peau, les images sont des vues prises de face. En collectant les pixels issus de ces segmentations précises, un modèle statistique de la couleur du vermillon propre

au locuteur peut être établi. Comme dans le système LAFTER, la distance de Mahalanobis mesure la vraisemblance, en tant que points du vermillon, des pixels sur les séquences d'images du locuteur. Nous comparerons des résultats obtenus à partir de plusieurs systèmes de couleur (RGB, TLS et RGB normalisé) et montrerons qu'ils donnent des résultats équivalent pour notre application. Les représentations en niveaux de gris ont été obtenues en normalisant entre 0 et 255 les valeurs extrêmes de la distance de Mahalanobis. Dans toute la suite, les valeurs proches de 255 (zone claire) correspondent à la peau et les valeurs proche de 0 (zone foncée) aux lèvres. Les zones d'incertitudes correspondent à des valeurs moyennes autour de 128 (zone grisée).

En considérant que la répartition des valeurs RGB autour de la moyenne est de type gaussien, on déduit une valeur théorique de distance couvrant 90% des données d'apprentissage. Cette valeur est alors utilisée pour fournir un seuil d'appartenance au vermillon de chaque pixel. Ce résultat est donné à titre d'exemple : c'est la valeur continue de la distance qui est utilisée comme valeur chromatique dans la suite. La Figure 9 résume les résultats calculés pour une modélisation statistique des valeurs RGB.

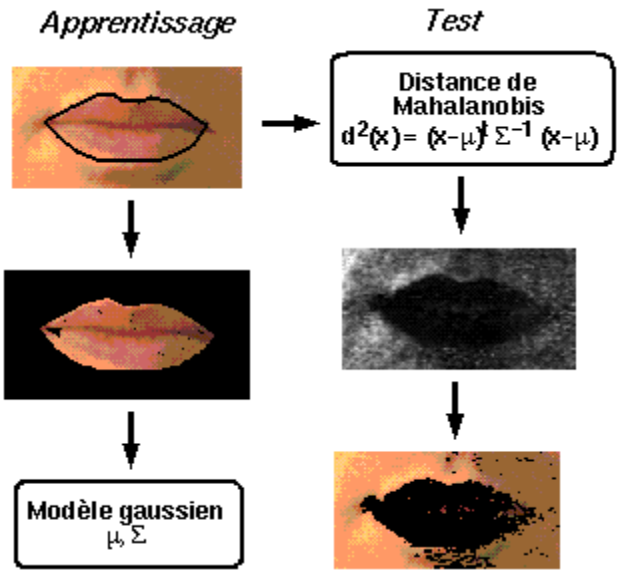


Figure 9. Apprentissage et test du modèle de couleur.

Les histogrammes suivants comparent les représentations directes selon teinte, luminance et saturation (calculée selon TLS) et les modélisations statistiques appliquées respectivement sur les composantes RGB, TLS, et RGB normalisées, la couleur étant mesurée par les distances de Mahalanobis. Comme pour les lèvres, les pixels de la peau ont été identifiés par une

segmentation explicite. Ils correspondent à une bande d'environ un centimètre autour du vermillon.

La qualité de séparation des histogrammes est calculée en affectant à chacun des 256 niveaux de quantification la classe en majorité représentée. On calcule alors pour les deux classes la proportion de points correctement affectés par cette méthode. Un indice unique de séparation est calculé en faisant le produit des deux proportions de points correctement affectés. Ainsi, plus les histogrammes des deux classes sont disjoints, plus les scores augmentent. La Figure 10 montre les histogrammes de répartition des classes lèvres et peau : d'abord suivant teinte, luminance et saturation, puis en utilisant la distance de Mahalanobis sur les systèmes RGB, TLS et RGB normalisé.

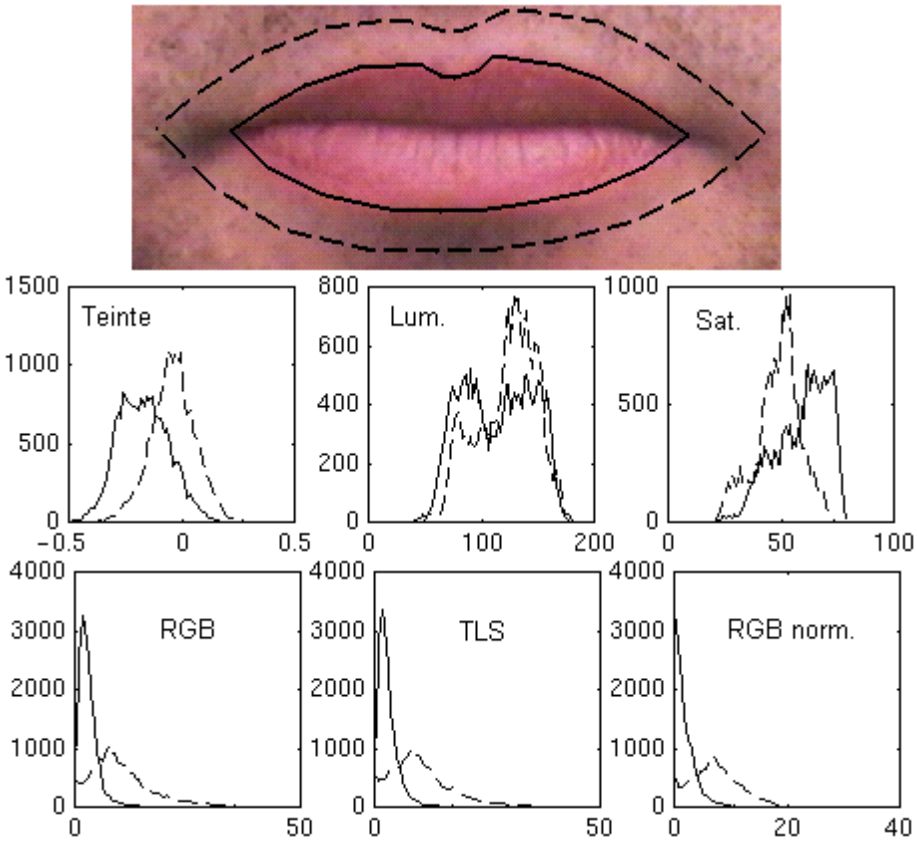


Figure 10. Histogrammes de séparation entre vermillon et peau pour un locuteur : comparaisons selon la mesure de couleur utilisée (en haut : teinte 55%, luminance 34%, saturation 51%; en bas : distance de Mahalanobis sur RGB 69%, TLS 67%, RGB normalisées 68%).

Pour juger d'une généralisation possible des résultats, trois locuteurs ont été analysés lèvres ouvertes, dans une position labiale la plus différente du cas d'apprentissage où les lèvres étaient fermées, mais issue de la même séquence (Figure 11). Ce changement de position induit des changements de couleur : l'angle d'éclairage à la surface des lèvres varie et la

muqueuse interne, visible dans cette position, donne une texture différente du reste du vermillon. Seule la mesure de la distance de Mahalanobis sur les niveaux RGB est représentée. Comme à la Figure 9, le seuillage est calculé pour une couverture de 90% de la classe des lèvres. On observe notamment que la couleur des lèvres est confondue dans tous les cas avec celle de la langue. De plus, souvent l'ombre sous la lèvre inférieure est elle aussi considérée de la même couleur que le vermillon. Enfin, des effets de reflets rejettent, à l'inverse, des points du vermillon.

Figure 11. Test de la distance de Mahalanobis sur une position différente de l'apprentissage.

II.1.4 Séparation de la peau et du vermillon par analyse discriminante

La section précédente évoquait l'amélioration de la séparation peau et vermillon obtenue grâce à une mesure statistique. Lorsque des classes sont identifiées et que l'on dispose d'échantillons a priori, l'analyse discriminante permet de trouver un ensemble de paramètres qui mesure une séparation optimale (au sens des moindres carrés) de ces classes. Dans notre cas où seulement deux classes de données à trois dimensions (composantes RGB, TLS ou RGB normalisées) sont à séparer, l'analyse discriminante donne le plan séparant au mieux les deux ensembles peau et lèvres (Liébart et al. ,1995). La mesure couleur d'un pixel RGB $\vec{x}$ se résume à centrer ce pixel sur la moyenne $\vec{\mu}$ des deux classes et effectuer le produit scalaire avec le vecteur normal $\vec{w}$ de ce plan de séparation :

$$w(\vec{x}) = (\vec{x} - \vec{\mu}) \cdot \vec{w}$$

Pour tout point, on peut ultérieurement décider de son appartenance à l'une ou l'autre de ces classes suivant le signe du produit scalaire. La Figure 12 illustre géométriquement, sur deux dimensions, les calculs mis en œuvre. L'analyse discriminante donne la moyenne μ de l'ensemble des deux classes et un vecteur directeur $\vec{w}$ de la droite de séparation Δ, orthogonale à la droite séparant au mieux les classes. Tout point *M* est affecté à l'une ou l'autre classe suivant la valeur du produit scalaire entre le vecteur $\mu\vec{M}$ et $\vec{w}$:

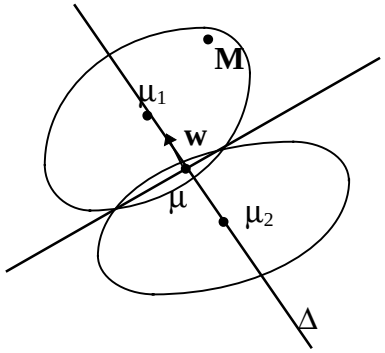


Figure 12. Séparation de deux classes par analyse discriminante.

Les calculs montrent que, pour deux classes de même effectif, plus la valeur du produit scalaire se rapproche de -1, plus la probabilité que le point appartienne à la classe 2 augmente. A l'inverse, une valeur proche de +1 dénote un point de la classe 1. Une valeur faible correspond à un point entre les deux classes. En normalisant les résultats entre 0 et 255, on obtient une image en niveau de gris où le contraste entre lèvres et peau est rehaussé. Par convention, les lèvres correspondent aux valeurs faibles (zone sombre) et la peau aux valeurs élevées (zone claire).

La Figure 13 illustre pour un locuteur sur les valeurs RGB, l'amélioration de la séparation par rapport à la méthode de la distance de Mahalanobis apportée par une analyse discriminante pondérée. Dans le cas de la distance de Mahalanobis, la segmentation est calculée avec le principe de la couverture de 90% d'une dispersion gaussienne. Dans le cas de l'analyse discriminante, le signe du produit scalaire fournit l'affectation. L'image utilisée correspond à un cas hors apprentissage.

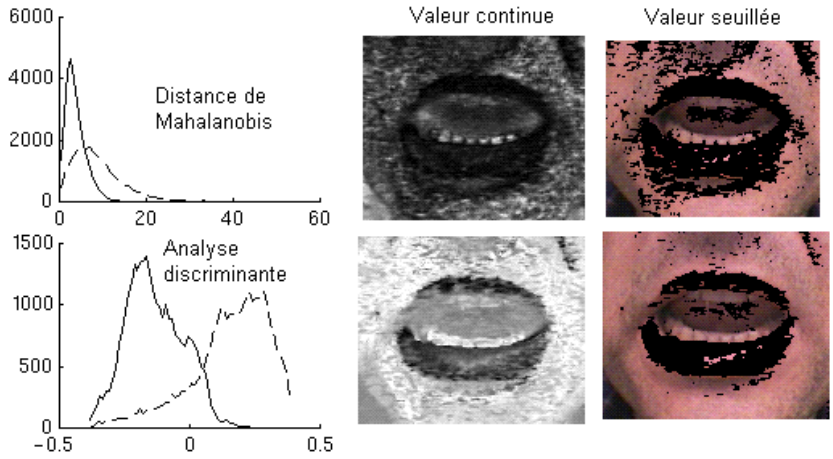


Figure 13. Comparaisons entre la séparation par la distance de Mahalanobis à la classe des lèvres et par analyse discriminante pondérée entre lèvres et peau.

II.1.5 Analyse discriminante pondérée

L'analyse discriminante fait l'hypothèse que toute donnée s'affecte nécessairement à l'une ou l'autre des classes. Numériquement cela se traduit par le fait que le produit scalaire peut rester élevé pour un point couleur situé loin des lèvres mais du même coté que celles-ci du plan de séparation. Or notre but est de donner une valeur continue de vraisemblance d'appartenance à l'une ou l'autre des classes avec la valeur du produit scalaire. Ainsi, un point très éloigné des barycentres des deux classes se verrait attribuer une forte probabilité d'appartenance à l'une ou l'autre.

Pour éviter ce problème, on pondère la valeur du produit scalaire par l'inverse de la distance de Mahalanobis sur l'ensemble lèvres et peau, augmentée de un pour éviter les problèmes de division par zéro. Ainsi, plus la distance augmente, plus le point est éloigné de l'ensemble lèvres et peau. En définissant une zone d'incertitude du produit scalaire autour de zéro, on renvoie ainsi les points éloignés dans cette zone. Cette correction peut s'appliquer pour éliminer des zones d'ombre très peu colorée qui n'apparaissent pas lors de la phase d'apprentissage :

$$p(\vec{x}) = \frac{w(\vec{x})}{1 + d^2(\vec{x})}$$

Numériquement les résultats sont ensuite normalisés entre 0 et 255 sur les minimum et maximum calculés parmi les pixels de l'image analysées.

Le tableau ci-dessous résume l'ensemble des résultats sur trois locuteurs pris respectivement dans chacune des trois conditions d'enregistrement (ICP, casque et ATR). Pour chaque

locuteur et pour chaque système de mesure, l'apprentissage est fait sur une position lèvres fermées. La position lèvres ouvertes correspond au cas de test hors apprentissage. La même technique d'affectation de classe majoritaire à un niveau de quantification est adoptée. Pour chaque classe est précisée la proportion de pixels correctement affectés. L'indice global de qualité pour la séparation des histogrammes est calculé comme le produit des proportions des deux classes correctement affectées. En gras sont repérés les couples qui maximisent ce produit.

Récapitulatif des résultats sur la mesure de la couleur

ICP : Enregistrement ICP, CAS : Enregistrement Casque, ATR : Enregistrement ATR

App. : Apprentissage

L : Lèvres, P : Peau

	Locuteur 1 : ICP				Locuteur 2 : CAS				Locuteur 3 : ATR			
	App.		Test		App.		Test		App.		Test	
	L	P	L	P	L	P	L	P	L	P	L	P
T	80	85	74	84	71	78	67	72	93	91	93	88
L	78	79	68	84	51	67	49	79	77	59	85	42
S	76	76	72	68	62	83	68	88	64	67	69	61
Mesure par distance de Mahalanobis												
RGB	89	91	83	88	90	77	83	65	90	88	91	86
TLS	90	90	83	88	89	75	80	69	90	86	92	85
RGBn	84	89	81	88	88	77	80	79	84	88	86	85
Mesure par Analyse discriminante												
RGB	**97**	**95**	93	92	93	86	88	84	94	92	93	90
TLS	**97**	**95**	93	91	91	85	86	83	**94**	**93**	**92**	**91**
RGBn	95	93	90	92	93	86	89	84	92	93	93	89
Mesure par Analyse discriminante pondérée												
RGB	**97**	**95**	**94**	**92**	94	85	95	80	95	91	**94***	**89***
TLS	**97**	**95**	95	91	94	83	92	78	**94**	**93**	94	89
RGBn	95	94	94	91	**95**	**85**	**92**	**83**	94	92	94	89

* Ce dernier résultat se distingue des deux précédents par la valeur du troisième chiffre significatif, non représenté dans le tableau.

Les résultats montrent que dans les trois cas d'enregistrement, les mesures statistiques l'emportent sur les mesures directes TLS. L'analyse discriminante pondérée l'emporte légèrement sur les autres. Nous la retiendrons par la suite pour sa garantie de robustesse. Les trois systèmes de couleurs sont équivalents. En dehors de critère évident pour en choisir un, nous retiendrons le plus simple, à savoir RGB.

Il est incorrect de prétendre que l'apprentissage de la couleur du vermillon et de la peau d'un locuteur, effectué lors d'un enregistrement, se généralise à d'autres conditions avec la *même* qualité de résultats. La caractérisation des images par chrominance seule ne suffit pas (en s'affranchissant de la luminance) à donner une caractérisation universelle de la couleur. Ainsi, un apprentissage de la couleur est à refaire pour toute nouvelle séquence. Dans la perspective du système complet (chapitre 4), cette contrainte sera adaptable au locuteur grâce à un modèle géométrique qui permet d'extraire une segmentation a priori pour toute nouvelle session : à chaque nouvelle séquence, le modèle au repos (lèvres fermées) devra être ajusté sur l'image pour donner la segmentation a priori nécessaire à l'apprentissage propre à la séquence.

II.2 Prédiction de la forme à partir d'une image en niveau de gris

De la section précédente, nous disposons d'une technique pour obtenir, à partir d'une image RGB, une image en niveaux de gris où le contraste entre lèvres et reste de la peau est rehaussé. Néanmoins, ce contraste n'est pas assez net pour permettre une segmentation précise et ainsi mesurer des paramètres géométriques. La technique présentée ici fait une association directe entre paramètres géométriques et images en niveaux de gris par régression multilinéaire. Cette régression ne se calculera pas sur les pixels directement, mais sur les composantes principales d'une ACP sur un ensemble réduit d'images. Anticipant ce calcul, l'ACP sur les images est présentée en premier.

II.2.1 De la reconnaissance de visages à la reconnaissance de formes labiales, les « eigenlips »

Cette méthode a été proposée pour la première fois par Bregler et al, 1994. Elle a été baptisée « eigenlips » par référence à la technique de reconnaissance de visages dont elle s'inspire, les « eigenfaces » de Turk et Pentland (1991). Partant des pixels en niveaux de gris comme paramètres d'analyse d'une image, ces approches montrent qu'une réduction paramétrique considérable est possible par codage sur vecteurs propres (eigenvectors) d'une ACP, calculée sur quelques images d'apprentissage en prenant chaque pixel comme un paramètre de mesure.

Dans le cas de l'expérience menée par Turk et Pentland, quelques exemples de visages à reconnaître parmi 10 personnes sont recueillis. Toutes les images ont la même taille et les prises de vue sont identiques entre visages. Toute image est alors codée par ses projections sur les images artificielles correspondant aux vecteurs propres, les « eigenfaces ». La dizaine de coefficients ainsi obtenus pour une nouvelle image sert ensuite à classer et reconnaître le visage. Les résultats ont montré que pour un ensemble d'apprentissage de 40 images (10 personnes et 4 images par personne) de 256*256 pixels, les sept premiers vecteurs propres d'une ACP sur les niveaux de gris suffisent pour reconnaître l'identité d'une personne à partir d'une nouvelle observation. Ces résultats mettent en évidence la grande redondance d'information qui existe entre des images représentant des scènes similaires, en l'occurrence ici des visages.

Les « eigenlips » transposent cette technique à l'analyse des formes de lèvres d'un locuteur. De manière similaire, la même réduction paramétrique a été observée. Le terme « eigenlips » a été introduit par Bregler (1994) pour un système de reconnaissance automatique de lettres. Par une ACP sur 4500 images de 24*16 pixels, centrées sur les lèvres d'un locuteur non maquillé, seuls les 10 premiers vecteurs propres sont conservés. Les coefficients de projection d'une image sur les 10 « eigenlips » sont alors utilisés comme paramètres visuels pour augmenter la robustesse au bruit du système de reconnaissance.

De même, Brooke (1994) a comparé les scores d'identifications par des sujets de séries de trois chiffres obtenus avec des images de 32*24 pixels centrées sur les lèvres d'un locuteur et les scores obtenus avec les mêmes images resynthétisées par un ensemble de vecteurs propres. Le corpus est constitué de 300 triplets prononcés par un locuteur anglais. L'ACP est calculé sur les images de 200 triplets, les 100 autres triplets étant utilisés pour le test. Pour un score de référence de 72% utilisant les images originales des 300 triplets, un score de 65% est obtenu avec une reconstruction à partir de 15 « eigenlips ». Ces résultats montrent la qualité du codage opéré par les projections sur vecteur propre obtenus par ACP (Revéret et Benoît, 1997 ; Revéret, 1997).

Dans la suite de ce chapitre, nous montrons que :

- les composantes extraites à partir d'un corpus d'entrainement de seulement 23 formes permettent de reconnaître en modalité visuelle pure avec de bons scores un vocabulaire de 54 mots isolés,
- mais surtout, au-delà de ces résultats, les composantes principales peuvent être directement interprétés de manière *continue* pour une prédiction de paramètres géométriques des lèvres.

II.2.2 Réduction paramétrique des images en niveaux de gris

Le corpus utilisé ici provient d'un enregistrement ICP d'un locuteur français dont les lèvres sont maquillées en bleu. En convertissant les images en niveau de gris suivant leur valeur de luminance, la qualité de contraste reste supérieure à celle obtenue précédemment par la technique d'analyse statistique de la couleur à partir de lèvres non maquillées. Cette première étude représente donc un cas idéal et vise à illustrer le lien qui existe entre des images de lèvres en niveaux de gris et des paramètres géométriques.

Le corpus total comporte 10 répétitions de 74 stimuli constitués de phrases « C'est pas $V_1CV_2CV_1z$? ». Nous nous sommes restreints à 9 répétitions de phrases où V_i est une voyelle parmi {a, i, y} et C une consonne parmi {b, v, ʒ, l, r, z}, donnant ainsi un vocabulaire de 54 mots différents. Pour chaque logatome, les images correspondant aux centres des réalisations des 3 voyelles et aux 2 consonnes ont été étiquetées sur la bande vidéo par des experts phonéticiens. Pour notre étude, seule la région des lèvres vues de face a été extraite pour donner une image sur 256 niveaux de gris de 64*48 pixels. Avec 486 phrases dont on extrait 5 images de 3072 pixels, le calcul de l'ACP pour la détermination des vecteurs propres implique la diagonalisation d'une matrice de covariance de 2430 individus et 3072 paramètres. Pour alléger ce calcul, le nombre d'individus a été réduit à un choix de 23 images. Ces 23 images correspondent aux visèmes identifiés par Benoît (1992) comme étant l'ensemble optimal de formes labiales les plus représentatives du corpus total. La représentativité est à prendre au sens de la distribution statistique de 14 paramètres géométriques mesurés sur les lèvres grâce à un système de chromakey (Lallouache, 1991). Nous reviendrons plus en détails au chapitre 3 sur ces visèmes qui sont à la base de la définition du modèle articulatoire. Notre but étant ici de mettre en évidence le lien entre les images en niveaux de gris et les paramètres géométriques, nous retiendrons seulement que ces 23 visèmes fournissent une sélection représentative pour un corpus d'apprentissage et qu'ils

permettent de réduire le calcul de l'ACP sur les images. La Figure 14 montre que 90% de la variance des 23 visèmes est atteinte avec les 9 premières composantes principales.

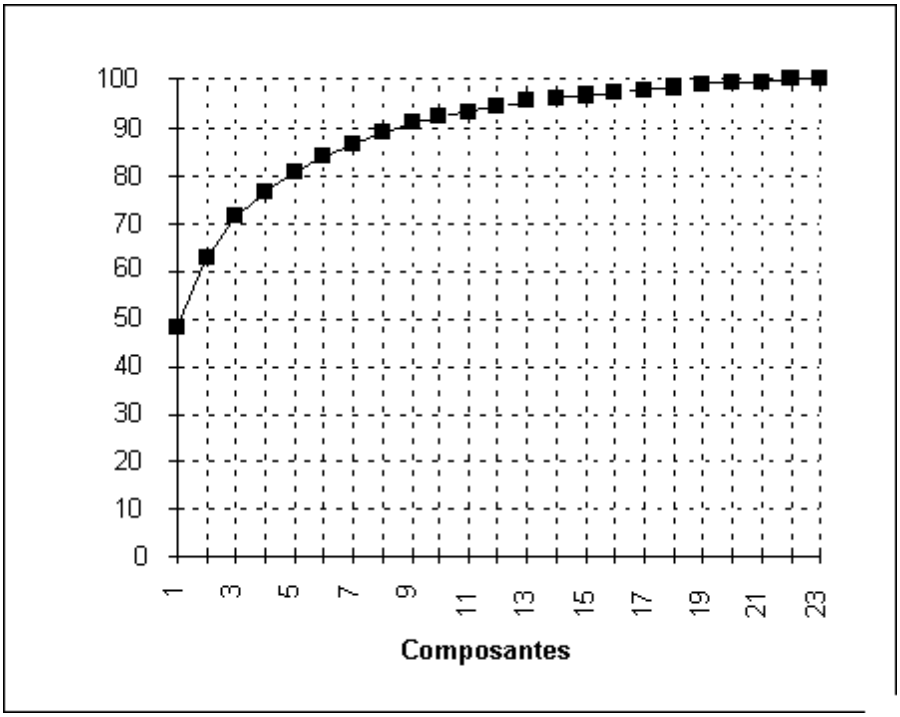


Figure 14. Evolution de la variance des 23 images en fonction du nombre de composantes principales.

La figure suivante présente la projection des 23 vecteurs images sur les deux premiers facteurs de l'ACP. La notation des visèmes suit celle spécifiée par Benoît (1992).

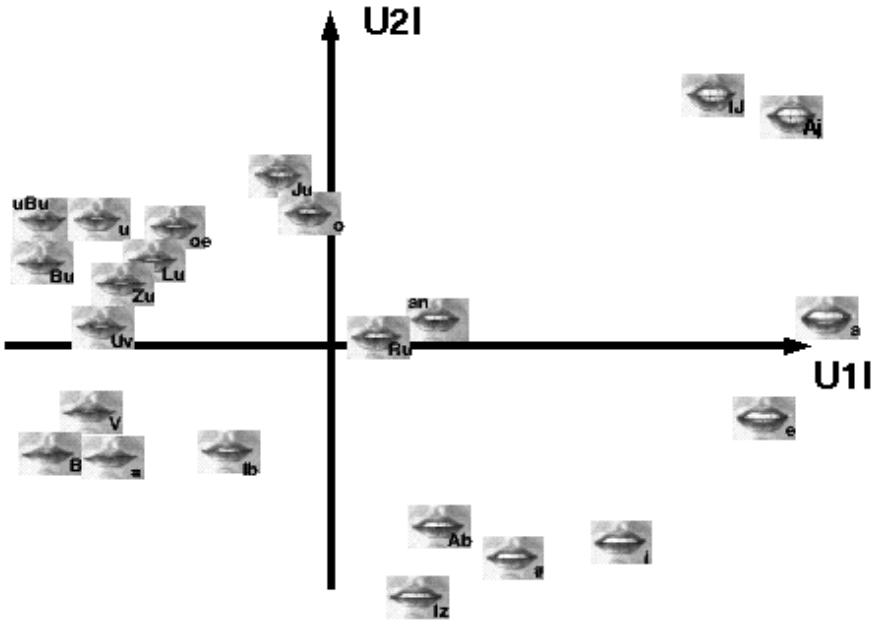


Figure 15. Projection factorielle des 23 visèmes sur les deux premiers facteurs de l'ACP calculée sur les niveaux de gris.

Une première évaluation a été faite pour juger de la stabilité de ces composantes pour coder le signal de parole. Le test consiste a reconnaître à partir des images le vocabulaire de 54 mots isolés en prenant comme paramètres les seules composantes principales. Un système à base de HMM a été utilisé pour la reconnaissance automatique (HTK) des mots $V_1CV_2CV_1z$. Sur les 9 répétitions, 7 ont servi à l'apprentissage et 2 au test. Les résultats présentés sur la Figure 16 combinent cinq permutations (test sur les séquences 1-2, 2-3, 4-5, 6-7 et 8-9). Seules les 22 premières composantes sont présentées, la 23ème n'apportant aucune variance (analyse de 23 vecteurs centrés sur la moyenne).

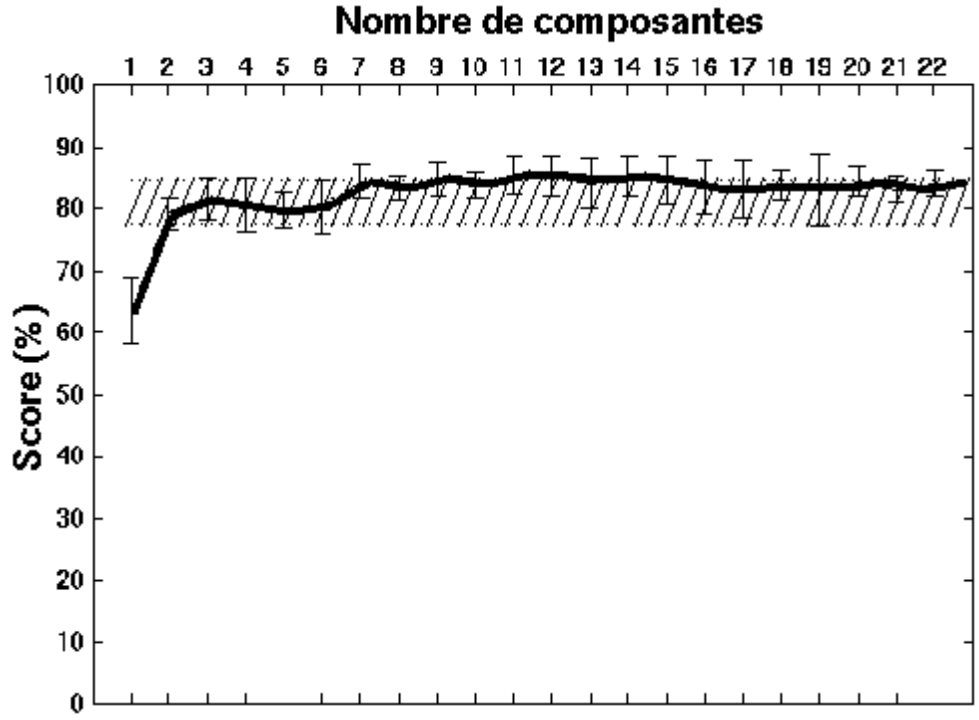


Figure 16. Scores de reconnaissance d'un vocabulaire de 54 mots à partir des seules composantes principales images.

La zone hachurée sur la Figure 16 correspond aux scores obtenus dans les mêmes conditions avec cinq paramètres géométriques mesurés de face par seuillage du maquillage bleu (voir section suivante pour la description des paramètres). Ces résultats montrent qu'un score optimal (autour de 80%) est obtenu dès les trois premières composantes. L'utilisation de composantes supplémentaires n'améliore pas significativement les résultats.

II.2.3 Prédiction des paramètres géométriques à partir des « eigenvisemes »

Dans la section précédente, les composantes principales images ont servi à discerner 54 mots. Nous les utilisons à présent de manière continue comme prédicteurs de paramètres géométriques. Ces paramètres, mesurés de face, sont:

- écartement interne A,
- ouverture interne B,
- aire interne S,
- écartement externe A',
- ouverture externe B'.

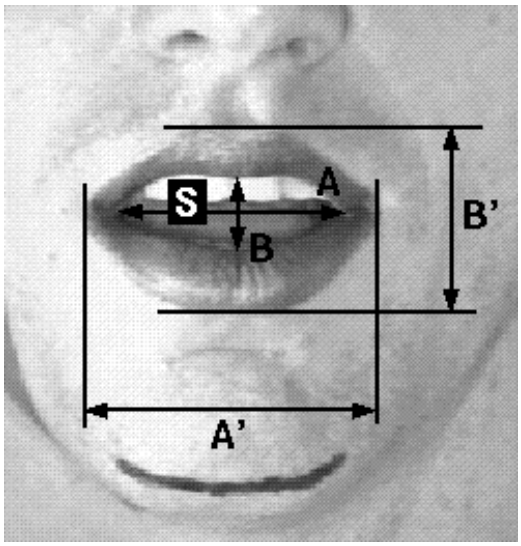


Figure 17. Les 5 paramètres géométriques mesurés de face.

Les trois premiers facteurs d'une ACP calculée sur les 5 paramètres des 23 visèmes totalisent 99.5% de la variance totale (72.1%, 22.2% et 5.2%). La Figure 18 présente la projection des 23 visèmes sur les deux premiers facteurs.

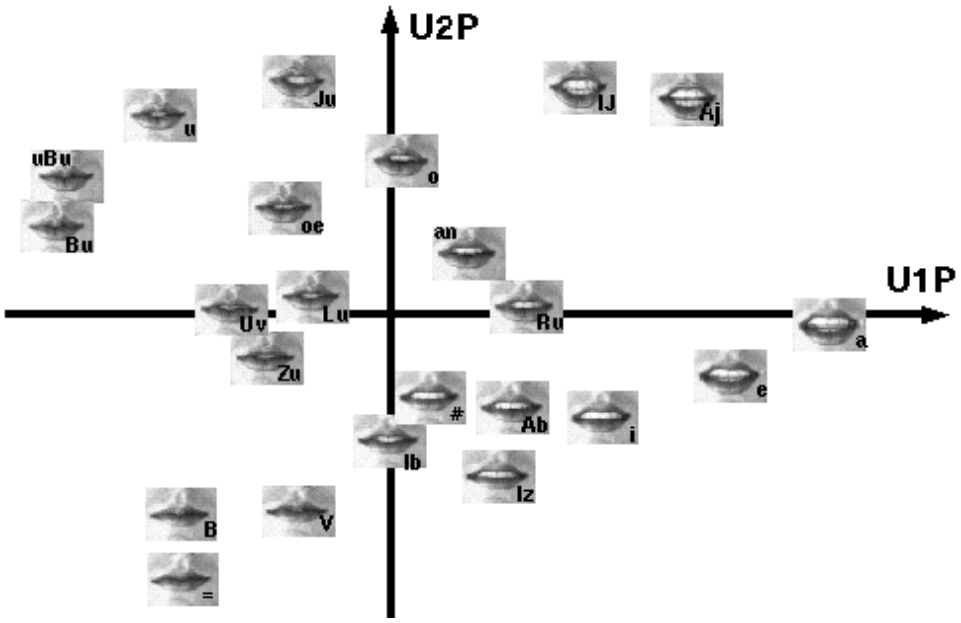


Figure 18. Projection des 23 visèmes sur les deux premiers facteurs de l'ACP calculée sur les 5 paramètres de face.

Les Figure 15 et Figure 18 mettent en évidence une similarité entre les projections factorielles des visèmes pour les deux systèmes de mesure (espace « paramètres » et espace « image »). Les corrélations entre les projections des 23 visèmes valent 0.94 sur le premier vecteur propre, 0.85 sur le deuxième vecteur propre et 0.72 sur le troisième vecteur propre. Ces résultats suggèrent déjà un fort couplage linéaire, terme à terme, entre les trois premières composantes des espaces paramètres et image.

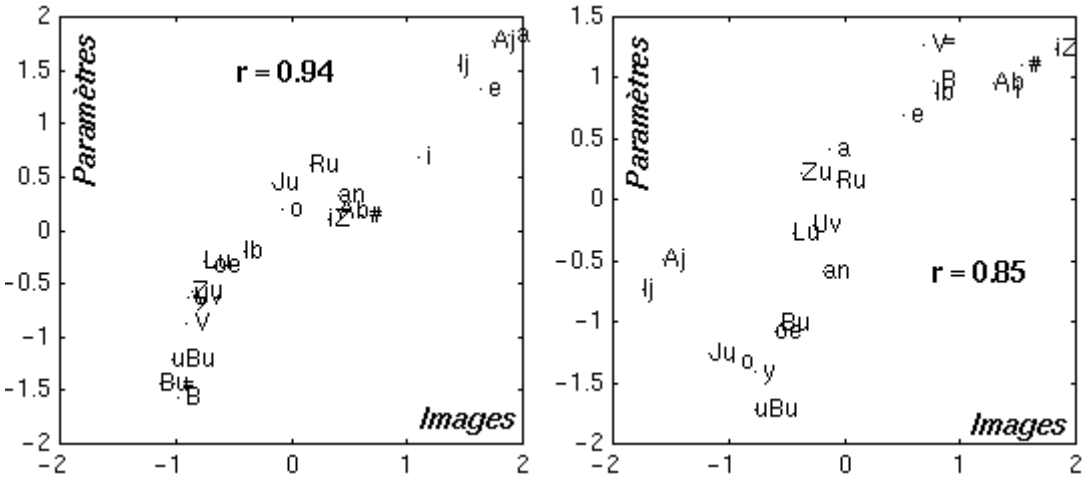
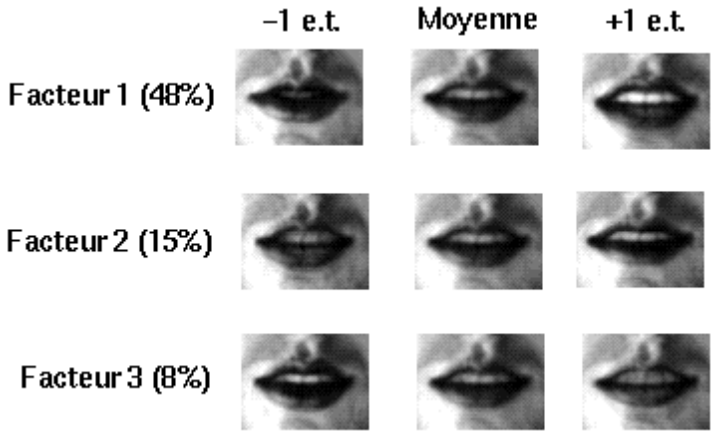


Figure 19. Corrélations entre les projections sur les deux premiers facteurs des espaces factoriels « images » et « paramètres ».

Enfin, on notera que les nomogrammes pour les deux premiers facteurs conduisent dans les deux espaces à la même interprétation articulatoire : le premier facteur s'apparente à l'ouverture des lèvres et le second à l'arrondissement. Le troisième s'apparente au relèvement

de la lèvre supérieure. Les paramètres géométriques d'ouverture B et B' ne permettent pas de faire la distinction entre le mouvement dû à la lèvre inférieure ou supérieure. Ainsi, ce troisième facteur de relèvement n'est rapporté que dans le cas de l'analyse des images en niveaux de gris.

Nomogramme images



Nomogramme paramètres

	A	B	A'	B'	S
Moyenne	27	6	50	23	163

	A	B	A'	B'	S	A	B	A'	B'	S
	Moy. - 1 e.t.					Moy. + 1 e.t.				
Facteur 1 (72%)	15	1	50	19	21	39	10	50	27	305
Facteur 2 (22%)	24	6	48	24	154	30	6	52	22	172
Facteur 3 (5%)	32	5	50	23	141	21	6	50	23	185

Figure 20. Nomogramme des espaces « images » et « paramètres ».

Tous ces résultats suggèrent qu'il existe un couplage linéaire simple entre les deux espaces qui peut être mis en œuvre par un calcul de régression multilinéaire des composantes principales vers les paramètres articulatoires. La régression multilinéaire multiple consiste à trouver un modèle linéaire qui prédit au mieux un paramètre particuliers en fonction d'un

vecteur d'observation de plusieurs paramètres. Le modèle est ajusté à partir d'un échantillon d'individus pour lesquels on connaît a priori le paramètre à estimer. L'ajustement au sens des moindres carrés conduit à des calculs linéaires simples. Si l'échantillon d'apprentissage choisi est suffisamment représentatif des variations des données, le modèle linéaire se généralise et permet alors de donner une estimation correct du paramètre recherché pour n'importe quelle observation.

Une régression multiple directement sur les pixels des visèmes est impossible puisqu'il y a moins d'individus (23 visèmes) que de paramètres (3072 pixels). En effet, le calcul de régression requiert de pouvoir inverser la matrice de covariance des individus. Dans le cas présent, cette matrice est carrée d'ordre 3072 mais de rang 22 (les 23 individus sont centrés sur la moyenne), donc singulière. La régression sur les composantes principales identifiées ci-dessus permet de contourner ce problème. Au lieu, de prendre comme paramètres les pixels de l'image, on utilise les composantes principales obtenues précédemment, correspondant aux projections sur les « eigenvisemes ».

Ainsi, à partir d'au plus 22 composantes, il est possible de calculer une régression multiple pour estimer les paramètres géométriques. Les figures suivantes donnent pour chaque paramètre la qualité de l'ajustement sur l'ensemble des 23 visèmes en fonction du nombre de composantes principales utilisées pour le calcul de régression. Les valeurs de référence sont issues de la mesure faite par le système chromakey de seuillage du bleu. La courbe moyenne et écart-type sur l'ensemble des 23 visèmes de l'erreur absolue entre la valeur de référence et la valeur prédite (trait fin, moyenne +/- écart-type) est représentée. La barre horizontale en haut correspond à la valeur de l'écart-type calculé sur les paramètres de références. La seconde indique pour les quatre premiers paramètres la valeur de référence 1 mm.

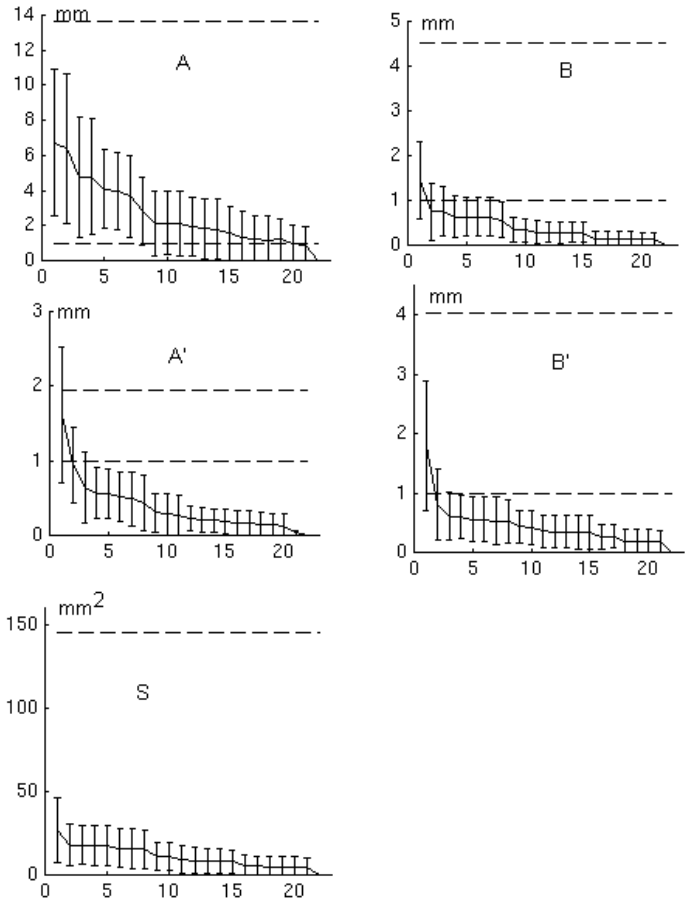


Figure 21. Evaluation de la qualité de l'ajustement pour 23 visèmes de 5 paramètres géométriques en fonction du nombre de composantes principales images utilisées pour la régression (en haut, A, B, puis, A', B' et S).

Les résultats de la régression sont ensuite généralisables, par application du modèle linéaire, pour d'autres images hors de la base d'apprentissage des 23 visèmes. Il est donc possible de donner une estimation de tout paramètre géométrique pour une image quelconque. A noter que toutes les données sont centrées sur la moyenne des 23 visèmes. De ce fait, si une image diffère trop des 23 visèmes, ses composantes principales tendront vont zéro, donnant comme estimation la valeur moyenne du paramètre géométrique.

Les figures suivantes présentent les résultats des prédictions des paramètres d'écartement interne et externe (A et A'), d'aperture interne et externe (B et B'), pour deux séquences : /ababaz/ et /azyzaz/. Ces résultats sont comparés aux mesures obtenues avec le système de seuillage du bleu. Comme la prédiction est obtenue par régression multilinéaire à partir des composantes principales, on remarquera que la prédiction de l'aperture interne est parfois négative pour la séquence /ababaz/. Les résultats obtenus en utilisant les 22 composantes principales et ceux obtenus avec seulement les trois premières sont présentées.

Le tracé du résultat du système de chromakey (trait pointillé) montre une discontinuité du paramètre d'écartement interne A. Celle-ci vient du fait que la mesure de ce paramètre est conditionnée par l'ouverture interne B qui détecte l'occlusion bilabiale. Lorsque les lèvres se ferment, la fente labiale mesurée par A passe brusquement d'une valeur élevée (40 mm) à une valeur nulle. Cette discontinuité peut perturbée les systèmes de reconnaissance automatique à base de HMM qui font l'hypothèse d'une évolution continue des paramètres d'entrées. Ce problème disparaît avec la prédiction des paramètres géométriques à partir des composantes principales images puisque le calcul reste toujours continu quelle que soit la forme des lèvres.

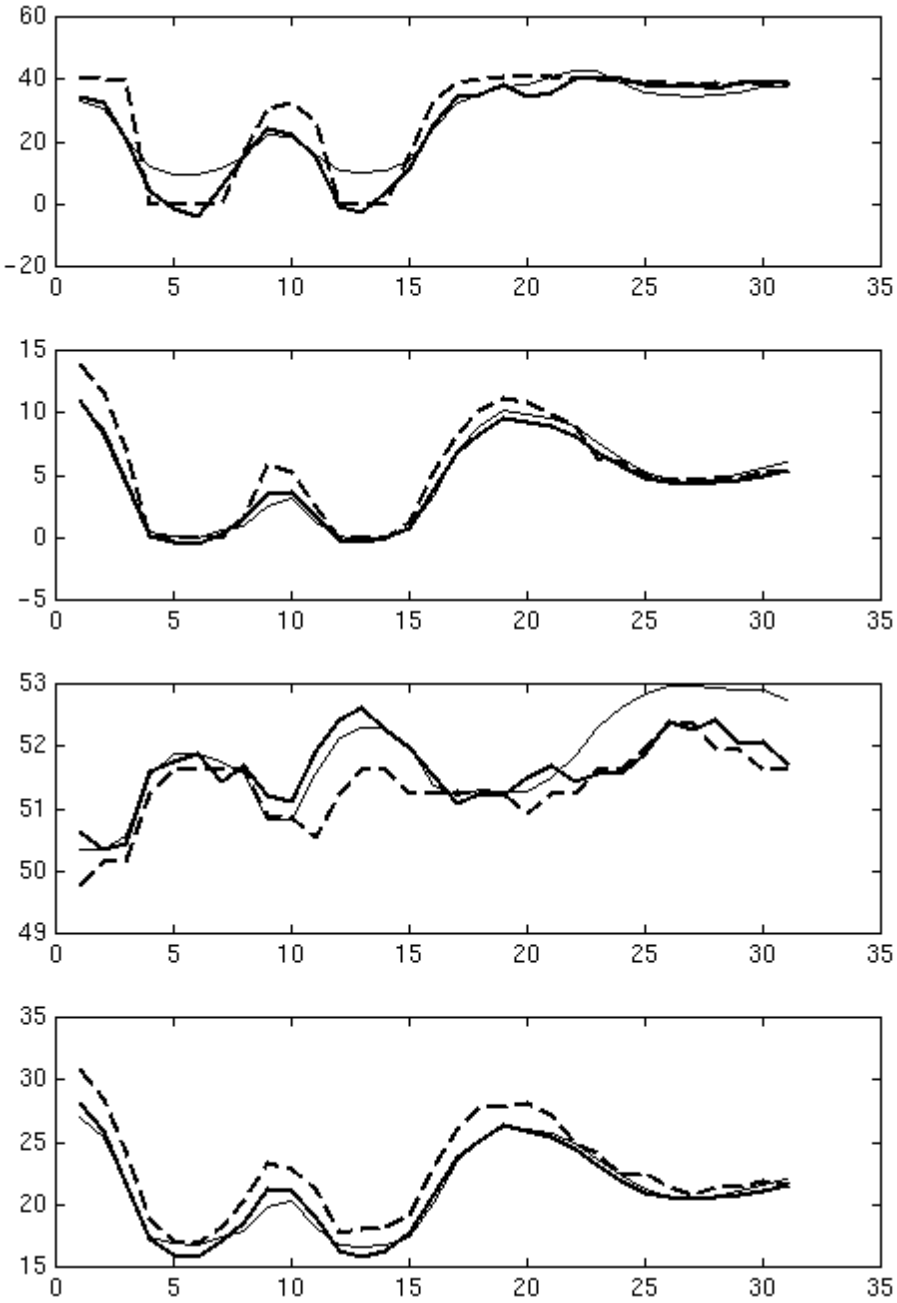


Figure 22. Résultats de prédictions pour la séquence /ababaz/. Les paramètres sont de haut en bas A, B, A' et B'.

Les mesures de référence par seuillage du bleu sont en trait pointillé. Les résultas prédits à partir de 22 composantes sont en trait continu épais. Les résultats prédits à partir de trois composantes sont en trait continu fin. En abscisse figurent les numéro de trames (50 images par seconde).

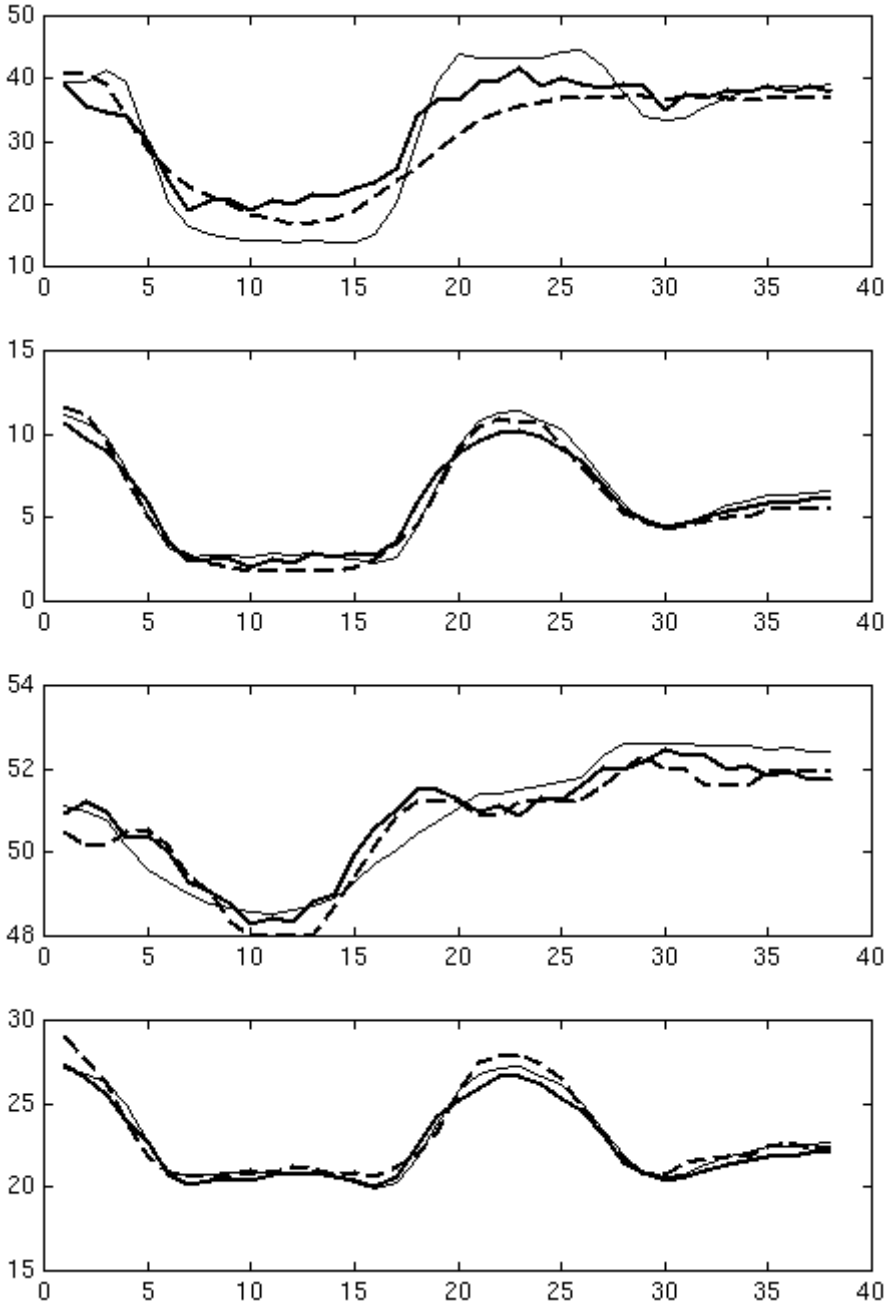


Figure 23. Résultats de prédictions pour la séquence /azyzaz/, paramètres A, B, A' et B'.

Le tableau suivant résume les résultats obtenus pour les 9 répétitions des 54 mots « $V_1CV_2CV_1z$ ». L'erreur absolue est calculée sur toutes les images des 486 mots du corpus (environ 30 images par mot) entre les valeurs de référence mesurées par seuillage du bleu et les valeurs prédites à partir de 22 composantes. On donne pour chaque paramètre la moyenne de l'erreur absolue, son écart-type, sa valeur maximale et le coefficient de corrélation. Afin de mieux évaluer l'importance de l'erreur sur un paramètre, on rappelle pour chacun son écart-type sur l'ensemble des 23 visèmes d'apprentissage.

		A	B	A’	B	S
Nb de comp. principales ↓	e.t. sur 23 visèmes →	13.7	4.5	2.0	4.0	145
3	moy. erreur	4.1	0.7	0.6	0.8	21
	e.t. erreur	2.9	0.6	0.4	0.7	22
	corrélation	.77	.96	.75	.93	.97
10	moy. erreur	3.6	0.7	0.4	0.8	21
	e.t. erreur	3.1	0.7	0.3	0.7	19
	corrélation	.75	.96	.85	.93	.97
22	moy. erreur	2.8	0.7	0.4	0.8	18
	e.t. erreur	2.7	0.6	0.3	0.7	19
	corrélation	.82	.96	.84	.93	.97

Ces résultats montrent qu’une bonne qualité de prédiction des paramètres géométriques est atteinte dès les 3 premières composantes. Ces résultats sont à rapprocher de ceux du test de reconnaissance du vocabulaire de 54 mots où, aussi, les scores n’augmentaient plus de manière sensible au delà des trois premières composantes.

En se basant sur les coefficients de corrélation, il ressort que les paramètres d’étirement A et A’ sont moins bien prédits que les paramètres d’ouverture B et B’. Ces résultats sont à relativiser par rapport à la précision que l’on peut obtenir sur ces paramètres. Le caractère discontinu de la méthode de mesure de A par chromakey a déjà été évoqué. De plus les ombres qui apparaissent aux commissures des lèvres pénalisent beaucoup la qualité de la texture pour la mesure de A’.

II.3 Discussion

L’analyse de la couleur a montré qu’une estimation probabiliste de la zone du vermillon pouvait être donnée. Basée sur une étude statistique, elle nécessite d’effectuer un apprentissage pour chaque session d’enregistrement. La contrainte d’utilisation qu’impose cet apprentissage systématique peut être contournée grâce à un modèle des lèvres du locuteur. Le

modèle, défini une seule fois, restera identique pour tous les enregistrements ultérieurs du locuteur. Ainsi, après alignement de la prise de vue de l'image des lèvres du locuteur et de son modèle personnel, l'apprentissage de la couleur pourra être relancé à partir de la segmentation indiquée par le modèle. Cette technique s'applique aisément dans le cadre des travaux présentés au chapitre suivant qui traite de la modélisation des formes labiales.

La précision obtenue par cette analyse de la couleur reste néanmoins insuffisante pour mesurer directement des paramètres géométriques par segmentation. Si le contraste entre le vermillon et la peau est augmenté, les éléments à l'intérieur de la bouche telles que la langue et les gencives peuvent avoir une couleur proche des lèvres. Ces résultats montrent à quel point les méthodes basées sur une détection de gradient peuvent rencontrer de difficultés. Il est nécessaire d'apporter de l'information a priori pour régulariser l'estimation de la région des lèvres à partir du signal image. Les chapitres suivants abordent ce problème à travers un modèle articulatoire.

Les résultats mis en évidence sur un locuteur par l'analyse des images en niveaux de gris (« eigenvisemes ») et les interprétations articulatoires qui s'y rattachent suggèrent qu'un codage réduit des gestes labiaux en parole est envisageable. Nous verrons comment le modèle articulatoire confirme ce résultat.

III. Chapitre 3. Modélisation des gestes labiaux en parole

Malgré la grande variabilité qu'offre une image de quelques milliers de pixels, nous venons de voir qu'il suffit d'un codage de cette image sur une dizaine de paramètres (projection sur les « eigenvisemes ») pour en extraire une information articulatoire. Nous appliquons ici cette idée d'une réduction paramétrique à travers un modèle de synthèse 3D des lèvres, contrôlé par un codage articulatoire linéaire des gestes labiaux (Revéret et Benoît, 1998).

Toute forme de lèvres est d'abord décrite géométriquement par une surface paramétrique 3D contrôlée par 30 points d'interpolation. Les 90 coordonnées tridimensionnelles sont déterminées, en imposant quelques contraintes de symétrie, à partir de deux vues simultanées des lèvres d'un locuteur. Par des considérations phonétiques, 10 formes particulières sont identifiées, couvrant de manière optimale l'espace articulatoire du Français. Une étude statistique sur la mesure géométrique 3D par les points d'interpolation de ces 10 formes produites par un locuteur montre qu'il est alors possible de réduire le contrôle du modèle géométrique initial à 3 paramètres articulatoires. Ces paramètres correspondent aux principaux gestes labiaux de la parole : arrondissement, abaissement de la lèvre inférieure et relèvement de la lèvre supérieure.

Notre modélisation reprend et généralise le modèle 3D de lèvres précédemment développé à l'ICP par Guiard et Adjoudani (Guiard, 1995, 1997 ; Adjoudani, 1996). En se basant sur des paramètres géométriques, identiques pour l'analyse et la synthèse, le modèle de Guiard et Adjoudani rassemblait à un même niveau de contrôle description géométrique et contrôle articulatoire. Notre démarche sépare ces aspects en deux étapes distinctes. Notre modélisation géométrique en s'appuyant sur des points dans l'espace laisse beaucoup plus de degrés de liberté pour une description morphologique adaptable à tout locuteur. Malgré des contraintes imposées sur le modèle géométrique qui réduisent ces degrés de liberté à 36, ces derniers restent en nombre trop élevé pour permettre un contrôle efficace. Les paramètres de contrôle articulatoire spécifiques à un locuteur sont alors déterminés par l'analyse statistique de 10 formes « clés ». Nous présentons en introduction quelques repères phonétiques sur l'articulation labiale qui ont guidé le choix de ces formes, puis les approches courantes en modélisation paramétrique des organes de la parole dont nous nous sommes inspirés.

III.1 Modélisation articulatoire

Bien que les lèvres fournissent l'organe de la parole a priori le plus accessible à la mesure, l'étude articulatoire de l'intérieur du conduit vocal a comparativement bénéficié de beaucoup plus de travaux. La raison en est sans doute que la recherche des lieux et degrés de constriction sont repérables sur une vue sagittale 2D du conduit interne (obtenue par IRM ou rayon X) et qu'ils suffisent à rendre compte de la plus grande partie des phénomènes acoustiques et phonétiques en production de la parole. Nous reprenons ici la nomenclature établie par Gabioud pour les différents types de modèles articulatoires (Gabioud, 1994). Elle s'applique principalement à la modélisation du conduit vocal vue sous une coupe sagittale. Les notions et les moyens mis en œuvre s'étendent cependant à toute modélisation articulatoire tridimensionnelle.

III.1.1 Modèles géométriques

Ces modèles décrivent uniquement la forme géométrique des contours des organes. Cette forme est représentée par une série de segments, arcs et courbes raccordés de manière continue (Mermelstein, 1973). Le contrôle des déformations du modèle est assuré par des paramètres tels que distance, angle, point ou rayon. Des modèles à base de segments et d'arcs de cercle ont été proposés pour la géométrie labiale.

On peut aussi citer dans cette catégorie la plupart des travaux en infographie pour la représentation du visage humain. Dans ces modèles, la surface du visage est représentée par un réseau continu de polygones (triangles le plus souvent) définis par leurs points aux sommets. Quelques points sont laissés libres pour le contrôle de la déformation. Des règles de géométrie lient le mouvement des autres points pour simuler l'élasticité de la peau.

Ces modèles présentent l'avantage de permettre une utilisation directe de la mesure du mouvement de l'organe (« motion capture ») par la position d'un marqueur physique : marqueur coloré, bobines de l'articulograph (Carstens™), marqueur infrarouge Optotrak (Northern Digital Inc.™). Cependant, cette simplicité d'utilisation se heurte à la difficulté de pouvoir donner une interprétation physique directe aux paramètres de contrôle.

III.1.2 Modèles physiologiques

Ces modèles s'appuient sur une représentation explicite de la physiologie des organes tel que l'activité musculaire ou l'élasticité des tissus. Nous citons ici deux exemples. Pour la langue, Payan (1996) a proposé un modèle 2D biomécanique de langue contrôlé par 7 muscles moteurs. La structure du modèle est décomposée en 48 éléments finis auxquels sont appliqués les lois linéaires d'élasticité pour simuler la déformation des tissus sous l'action des muscles implantés.

Pour les lèvres, Basu (1998) a proposé un modèle 3D surfacique composés d'éléments finis auxquels sont appliqués des équations d'élasticité. Contrairement au modèle de langue de Payan, celui-ci ne possède pas de paramètres de contrôle explicites simulant une action musculaire. Les hypothèses d'élasticités contribuent juste à régulariser la distribution des points de la surface des lèvres lorsque n'importe quel point est géométriquement déplacé.

Les approches par décomposition en éléments finis présentent l'avantage d'intégrer par définition des contraintes de forme liée à la biomécanique telle que la conservation de volume, propre à un hydrostat comme la langue. Bien que ce type de modèle aborde une description beaucoup plus juste de la nature réelle des articulateurs, leur utilisation en suivi automatique est complexe : il est généralement difficile de retrouver de manière stable l'état des paramètres de contrôle à partir du signal (acoustique ou visuel). Basu propose quelques exemples d'application de son modèle en suivi automatique mais ce travail n'a jusqu'à présent été confronté à aucune évaluation sur des séquences de parole continue.

III.1.3 Modèles linéaires

L'analyse factorielle donne une représentation de données multidimensionnelles sous la forme d'une distribution centrée autour d'un point et répartie selon plusieurs axes issus de ce point. En particulier, l'analyse en composantes principales fournit la distribution autour de la moyenne des observations selon les différents axes d'inertie du nuage de données (au sens quadratique). Les coordonnées selon ces axes orthogonaux fournissent un nouveau jeu de paramètres *décorrélés*, les composantes principales. Elles sont classées par ordre d'importance de leur axe d'inertie associé. Par les propriétés des axes d'inertie, l'ACP supprime ou réduit les redondances qui peuvent exister entre les données et les paramètres initiaux. Les premiers facteurs suffisent à donner une représentation globale des variations des données, les facteurs subséquents réalisant seulement quelques ajustements locaux.

Cette analyse statistique a trouvé une application dans la modélisation articulatoire pour réduire à quelques paramètres orthogonaux de variation principaux, les observations géométriques des organes en production de la parole. Le modèle articulatoire correspond alors à un modèle linéaire contrôlé par des paramètres issus des composantes principales et quantifiant des variations de formes par rapport à la configuration moyenne. Cette approche implique deux choix préliminaires : la caractérisation géométrique de l'organe étudié et le corpus d'apprentissage collectant les observations. Les deux sont d'égale importance : la caractérisation géométrique initiale doit être capable de capturer l'ensemble des variations de formes sans contraintes trop fortes. De même, pour prétendre à un caractère d'intérêt général, le corpus doit *couvrir* suffisamment de situations représentatives. Pour limiter la taille du corpus à analyser, il est donc nécessaire de procéder à un choix sélectif des formes à mesurer.

Pour le conduit vocal, le modèle de Maeda s'appuie sur les points d'intersection d'une grille polaire (Maeda, 1979) avec le contour d'une vue sagittale du conduit vocal. Une grille, orthogonale aux parois du conduit, échantillonne le « tube », en une série de sections, depuis le larynx jusqu'aux alvéoles. La grille présente une partie courbée entre le palais mou et le haut du pharynx, raccordant les parties buccale et pharyngale. La forme géométrique du conduit est alors définie par l'ensemble des positions des points situés à l'intersection des sections de la grille avec les deux parois du conduit. Les images d'analyse ont été recueillies par enregistrement vidéo aux rayons X - 50 trames par seconde - de 10.4 secondes d'un corpus constitué de 10 phrases phonétiquement équilibrées. Au total, 519 trames de la vue sagittale du conduit vocal ont été mesurées géométriquement par la grille polaire.

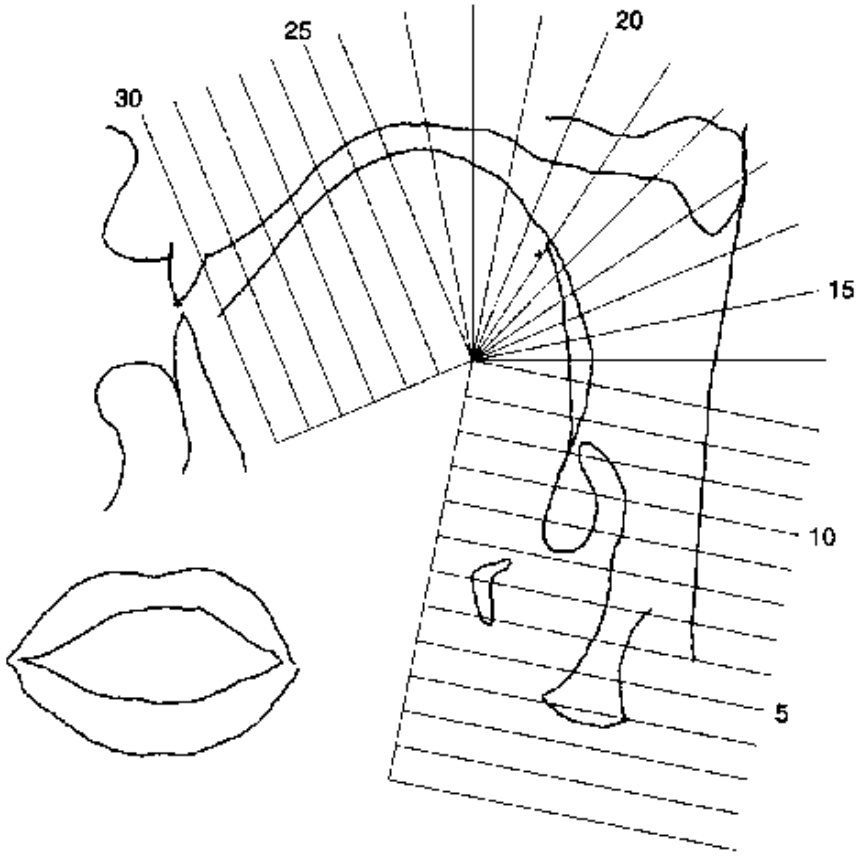


Figure 24. La grille de mesure du conduit vocal utilisée dans le modèle de Maeda (1979).

Les mouvements de la mâchoire influencent directement la position de la langue. Ainsi, pour ne conserver que le mouvement propre de la langue, l'effet de la mâchoire a été préalablement soustrait aux données. Il représente 15% de la variance totale du corpus et a été retenu comme paramètre indépendant. Ensuite, par analyse en composantes principales des positions de la langue, trois paramètres articulatoires ont été retenus. Ils correspondent aux trois premiers facteurs et sont interprétables d'un point de vue articulatoire :

1. le premier paramètre (43% de la variance) est identifié à la mesure de l'avancée / recul de la langue dans la cavité buccale,

1. le second (23% de la variance) à la hauteur du dos de la langue,

1. le troisième (7% de la variance) à la hauteur de la pointe de la langue (apex).

En comptant la contribution de la mâchoire, le modèle statistique final porte 88% de la variance totale des données recueillies et mesurées. Au paramètre de mâchoire et aux trois paramètres de la langue, trois autres ont été ajoutés : la hauteur du pharynx, l'ouverture et la protrusion des lèvres. Ainsi, le modèle est contrôlé en tout par 7 paramètres articulatoires.

Cette approche n'est bien sûr pas unique : pour la mesure géométrique, d'autres types de caractérisation du contour des parois ont été utilisés telle qu'une échelle régulière de l'abscisse curviligne (Kaburagi et Honda, 1994). De même pour le choix du corpus, une alternative courante consiste à ne prendre que quelques logatomes explorant l'espace articulatoire maximal (Badin et al., 1995).

Dans notre cas, la mesure géométrique des formes labiales est faite par un ensemble de 30 points XYZ caractéristiques d'un lieu anatomique. Cette modélisation géométrique est appliquée sur un corpus réduit de 10 formes particulières, produites par un locuteur, pour déduire ensuite par ACP un modèle articulatoire contrôlé par 3 paramètres. Le choix de ce corpus de 10 formes fait suite à une étude phonétique, menée par Benoît (1992), qui avait eu pour résultat d'identifier des formes labiales clés, appelées visèmes, prenant en compte les phénomènes de coarticulation. Nous détaillons maintenant nos modélisations géométrique et articulatoire.

III.2 Modélisation géométrique 3D des lèvres

III.2.1 La modélisation par surface paramétrique

De manière schématique, la géométrie des lèvres peut s'assimiler à la sortie d'un tuyau épais traversant la peau du visage. Cette situation se formule géométriquement comme la moitié supérieure d'un tore coiffant les deux cylindres qui figurent les parois internes et externes du tuyau. Une formulation en surface paramétrique du tore est donnée par un faisceau continu de cercles, ou d'ellipses par extension; l'axe des z indiquant le sens de la cavité, on ne garde que la moitié où z est positif :

$$z = a \times \sin(u)$$
$$x = R_x(u) \times \cos(v) \text{ , avec } R_x(u) = R_x + a \times \cos(u)$$
$$y = R_y(u) \times \sin(v) \text{ , avec } R_y(u) = R_y + a \times \cos(u)$$
$$\text{où } u \in [0, \pi] \text{ et } v \in [0, 2\pi]$$

La surface du tore est décrite par la circulation d'une ellipse de rayons $R_x(u)$ $R_y(u)$. Par analogie avec les lèvres, le contour externe correspond alors à l'ellipse $u = 0$ et le contour interne au cercle $u = \pi$. Bien sûr, les lèvres ne sont pas aussi arrondies qu'un tore. Ce principe de génération de la surface des lèvres par circulation d'un contour fermé s'appuyant sur les contours externe et interne illustre cependant bien le cadre géométrique dans lequel s'inscrit à la fois le premier modèle 3D de lèvres développé à l'ICP par Guiard et Adjoudani (Guiard,

1996 ; Adjoudani, 1996) et le nôtre qui le généralise. Dans les deux cas, la surface des lèvres est décrite par un contour circulant, défini par une courbe 3D fermée, construite de la même manière le long d'un trajet allant du contour externe au contour interne.

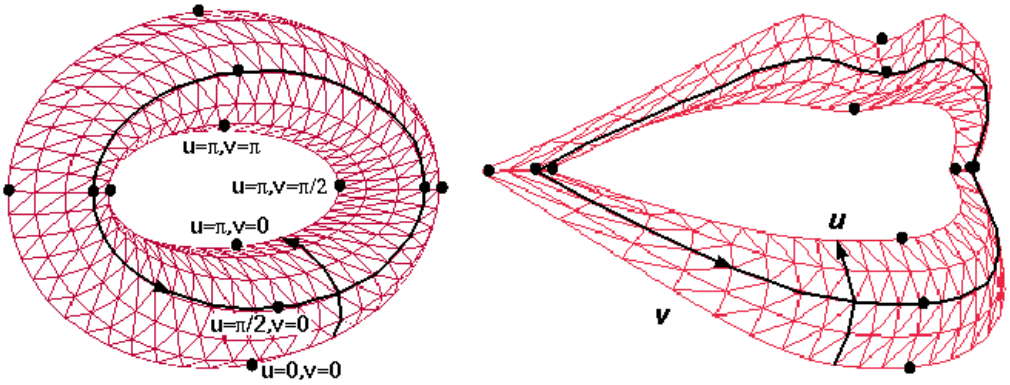


Figure 25. Les lèvres décrites comme une surface paramétrique de structure torique.

Pour des lèvres réelles, le contour externe caractérise la limite entre le vermillon et la peau l'entourant. Il est généralement identifiable à la différence de couleurs entre les tissus (voir chapitre 2). Le contour interne quant à lui n'a pas de réalité anatomique. Dans une prise de vue de face, il est néanmoins identifiable comme la limite visible de la surface des lèvres, limite définie géométriquement par les points où la normale à la surface est orthogonale à la direction de la vue. Le contrôle de la géométrie de ce contour présente un intérêt acoustique important. En effet, toujours dans le cadre d'une prise de vue de face, il coïncide avec la sortie du « tube » acoustique que constitue le conduit vocal (Motoki et al, 1994). L'aire interne aux lèvres conditionne ainsi directement les composantes fréquentielles des sons (Fant, 1968). De plus, sous la même définition géométrique de lieu où les normales sont orthogonales à la direction de visée, le contour interne caractérise le contact entre lèvres inférieure et supérieure lors d'une occlusion bilabiale. La vue de face étant le contexte de communication - et donc d'analyse - le plus courant, on voit donc que la géométrie du contour interne est bien un paramètre essentiel du contrôle en production.

Une fois contours interne et externe définis, il serait possible de générer une première surface paramétrique joignant ces deux courbes tridimensionnelles, par exemple, par interpolation linéaire entre les points des contours. Or, dans le cas du tore, le trajet du contour circulant suit non pas des droites mais des arcs des cercles de rayon *a* (figure précédente). Pour rendre compte de manière réaliste du repli des lèvres, on astreint donc le parcours du contour circulant à passer par un ou plusieurs contours intermédiaires entre l'externe et l'interne. Ces

contours intermédiaires n'interviennent ainsi que pour définir le repli des lèvres. Ils ne sont pas identifiables dans une vue de face, sinon par marquage, et nécessitent d'avoir recours à un autre angle de vue pour déterminer leur position.

Bien qu'ils suivent un cadre géométrique similaire, notre modélisation géométrique et celle de Guiard et al s'appuient sur des principes différents. Le modèle de Guiard et al utilise directement pour la synthèse les paramètres d'analyse de face et profil. Parmi tous les paramètres, 5 sont ensuite gardés pour le contrôle, les autres en étant déduits par des régressions linéaires simples. Ces paramètres imposent par définition des contraintes géométriques qui entraînent une perte de généralité des possibilités de représentation. De plus, la mesure du profil est réduite à seulement trois paramètres ce qui a contraint les auteurs à choisir de manière arbitraire un certain nombre de coefficients pour bâtir la forme du modèle.

La définition des courbes de notre modèle s'appuie sur des points de contrôle de courbes paramétriques. Par des interpolations polynomiales de troisième degré, ces points suffisent à générer des lèvres réalistes vues comme une surface paramétrique. Les points de contrôle sont définis dans l'espace XYZ et correspondent à des lieux anatomiques présents et identifiables quelque soit la forme des lèvres. Ainsi, notre modèle peut s'adapter à un locuteur quelconque et permettre une meilleure représentation de la diversité des formes labiales. Nous détaillons maintenant plus en détails les caractéristiques du modèle de Guiard et al et l'apport de notre modèle.

III.2.2 Le modèle de Guiard et Adjoudani (Guiard et al., 1996)

Un premier modèle 2D des contours internes et externes avait été défini à partir de 17 paramètres sur un locuteur maquillé en bleu : 8 pour le contour interne et 9 pour le contour externe (Guiard, 1996). Au moyen de règles de symétrie, ces paramètres permettent de déduire un ensemble d'équations polynomiales pour la construction des contours externe et interne. Par une étude statistique effectuée sur le même locuteur, seuls trois paramètres indépendants ont été gardés et servent de paramètres de contrôle. Les autres paramètres de mesure, nécessaire à la construction du modèle, sont déduits par régression linéaire. Les trois paramètres de contrôle sont :

- l'ouverture interne,
- l'écartement interne (ou aperture),

- la protrusion du contact labiale,

Les courbes de contour sont décrites par des équations de la forme $y = x^n$. Elles ont été proposées en premier par Lindblom et Sundberg (1971) pour décrire le contour interne vu de face. Elles présentent la particularité de lier le paramètre de courbure n à une corrélation forte - observée dans plusieurs études (Fromkin, 1964; Abry et Boë, 1980; Benoît, 1992) - entre l'aire interne et le produit des largeur et hauteur interne. En effet, si S est la surface interne, A la largeur et B la hauteur, en négligeant l'arc de cupidon, un rapide calcul montre que n se déduit du rapport S sur A*B selon :

$$S = AB - 4\int_0^{A/2} \frac{B}{2}\left(\frac{x}{A/2}\right)^n dx = \frac{n}{n+1} AB$$

L'ensemble des résultats sur différentes langues ont rapporté que le coefficient de régression entre S et A*B vaut entre 0.7 à 0.8, donnant une estimation de la valeur de n entre 2.33 et 4.

Le contour interne est ainsi décrit par 6 morceaux continus de courbes : quatre de type $y = x^n$, les coefficients n étant déduits par régression linéaire, et deux courbes de type $y = \cos(a.x)$ pour l'arc de cupidon. Le contour externe est construit sur les mêmes principes géométriques.

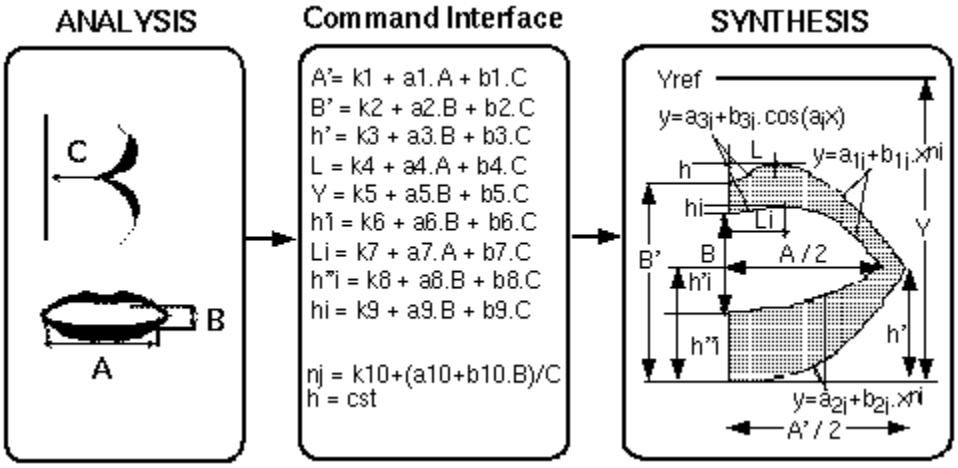


Figure 26. Le modèle 2D initial (d'après Guiard, 1996).

Aux contours interne et externe, ont été ensuite ajoutés trois contours intermédiaires pour l'extension à la troisième dimension (Adjoudani, 1993 ; Guiard et al., 1996). Les paramètres de face nécessaires à la définition des coordonnées XY de ces contours ont été interpolés linéairement à partir de ceux donnés par la mesure pour les contours interne et externe. Chaque contour est ensuite décrit dans le plan axial XZ, orthogonal au plan coronal XY du modèle 2D, par quatre courbes de type $z = x^m$. Des coefficients fixés *ad hoc* (déduits de vues de profil) et trois mesures de profil utilisées comme paramètres de contrôle règlent les

paramètres des quatre équations de chaque contour. Ainsi, aux trois paramètres de commande du modèle 2D précédent, deux paramètres de profil ont été ajoutés : le protrusion de lèvre supérieure et la protrusion de la lèvre inférieure. Par échantillonnage de la valeur de X, les courbes $y = x^n$ et $z = x^m$ donnent ainsi un ensemble de points XYZ pour les cinq contours. La surface du modèle est ensuite générée par un réseau de polygones entre les points obtenus.

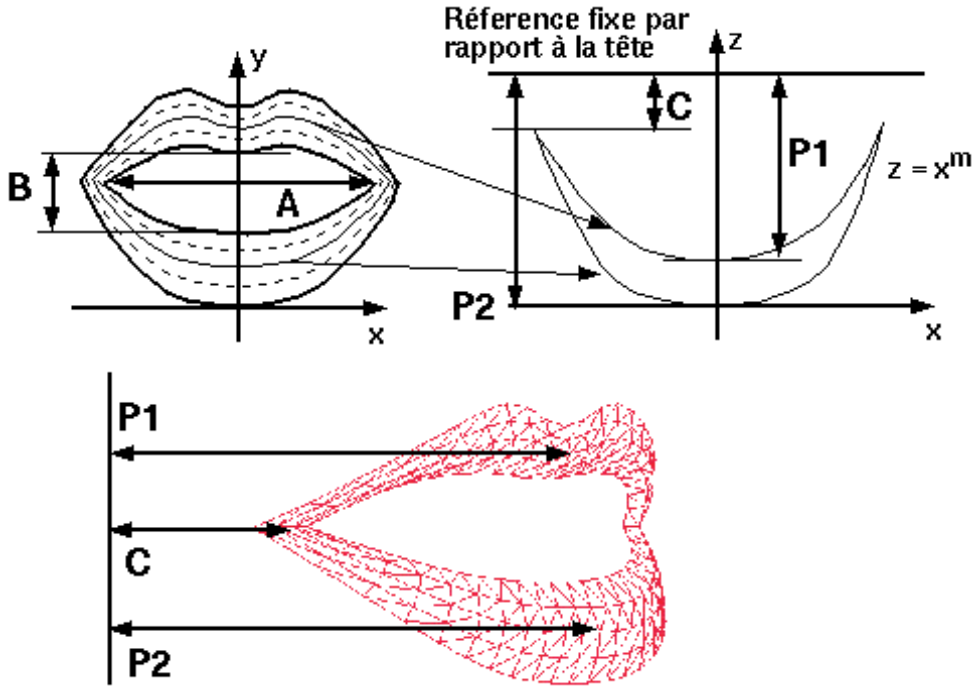


Figure 27. Le modèle 3D de Guiard et Adjoudani.

Le but était de fournir un outil de synthèse visuelle de la parole qui soit intelligible. Si les paramètres de contrôle restent généralement mesurables quel que soit le locuteur, l'étude statistique et les coefficients qu'elle fixe imposent une morphologie unique au modèle. Elle suit celle du premier locuteur étudié et ne garantit donc pas une bonne adéquation à toute forme de lèvres. Refaire pour chaque locuteur une étude similaire ne suffit pas non plus à représenter fidèlement tous les cas. Par exemple, les paramètres géométriques de base imposent une symétrie exacte entre les cotés gauche et droit des lèvres. De même un seul paramètre d'aperture contrôle le déplacement des lèvres supérieure et inférieure sans distinguer la contribution de chacune. Enfin, les coefficients pour l'extension à la troisième dimension des courbes ont été déterminés arbitrairement, les données suivant l'axe Z étant réduites aux trois paramètres de profil, ce qui est insuffisant pour décrire toute la forme.

III.2.3 Une nouvelle approche par points de contrôle géométrique

Le but de notre modélisation étant davantage tourné vers l'analyse, il était donc nécessaire d'adopter une approche géométrique offrant plus de liberté de mesure et ainsi de représentation. La forme du modèle de Guiard et al. a d'abord été échantillonnée par l'ensemble minimal de points dont les coordonnées XYZ permettent de retrouver les paramètres de construction du modèle complet, les symétries étant respectées. En interprétant ensuite les courbes du modèle comme des fonctions d'interpolation de ces points de contrôle, la contrainte de symétrie est levée puisque les points peuvent ensuite se déplacer indépendamment les uns des autres.

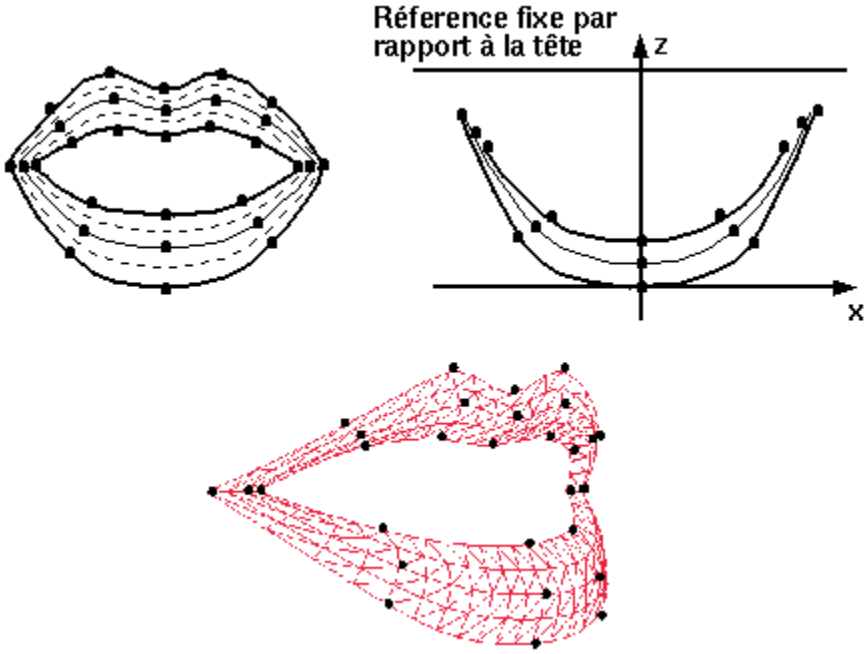


Figure 28. Interprétation du modèle de Guiard et Adjoudani par interpolation de points de contrôle (5 contours fixes).

Dans le plan coronal XY, les équations $y(x)$ d'un contour quelconque (interne, externe ou intermédiaire) se retrouvent à partir de dix points de contrôle. En plus des contours interne et externe, nous n'avons gardé qu'un seul contour intermédiaire pour décrire le repli des lèvres. Par une seconde interpolation des points de contrôle, orthogonale à celle définissant un contour, sur les trois contours interne, externe et intermédiaire, il est possible de générer de manière continue une infinité de groupes de 10 points de contrôle pour la génération de contours intermédiaires. Ceci revient à paramétrer automatiquement le trajet de circulation du contour fermé, évitant la recherche *ad hoc* de coefficients spécifiques comme c'était le cas des trois contours intermédiaires du modèle de Guiard et al. Parmi les dix points d'un contour

quelconque, identifiés pour la reconstruction dans le plan XY, huit permettent de retrouver dans le plan XZ les quatre équations z(x) servant à décrire le repli des lèvres.

Les 30 points d'interpolation nécessaires à la définition des trois contours de base (10 points par contour) constituent ainsi les points de contrôle de notre modélisation géométrique. La donnée de leur position XYZ définit tout le modèle, lui conférant ainsi, théoriquement, 90 degrés de liberté.

Pour les dix points de contrôle de chaque contour, on donne la nomenclature suivante :

- deux *points aux extrémités gauche* et *droite* (E_G et E_D) correspondant aux commissures,
- deux *points inférieur* et *supérieur* (I et S) pour les intersections du contour avec le plan sagittale,
- deux *points de l'arc de cupidon* (A_G et A_D) aux sommets gauche et droit de l'arc,
- quatre *points de courbure* pour spécifier l'allure des deux courbes à gauche et à droite entre les commissures et le point inférieur (points C_{IG} et C_{ID}), et entre les deux autres courbes entre les commissures et les points de l'arc de cupidon (points C_{SG} et C_{SD}).

Enfin, on notera respectivement X^{ext}, X^{med} et X^{int}, avec X étant l'un des dix points ci-dessus, tout point sur les trois contours de base externe, intermédiaire (ou médian) et interne. Sans cette précision de contour, le point désignera l'ensemble des 10 points d'interpolation d'un contour quelconque. Le symbole X^* désigne l'ensemble X^{ext}, X^{med} et X^{int} des trois points X pris sur les contours internes, médian et externe.

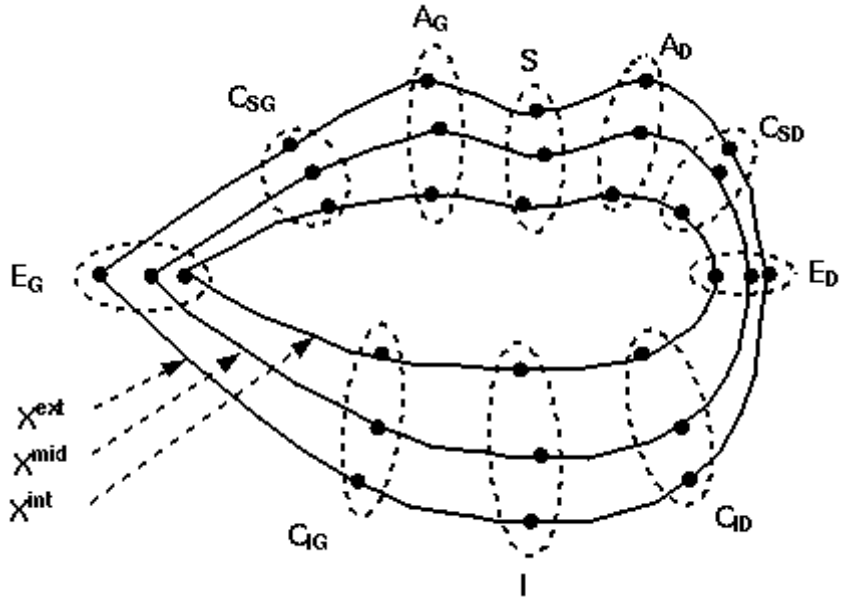


Figure 29. Les trois contours de base et les 30 points de contrôle.

III.2.4 Génération du modèle géométrique à partir des points de contrôle

La surface 3D est générée par un faisceau continu de contours 3D décrits par des polynômes de troisième degré passant par les 3 contours de base interne, externe et intermédiaire. Pour définir chaque contour 3D, les équations de type $y = x^n$ et $z = x^m$ utilisées dans le modèle de Guiard et Adjoudani (Figure 28) ont été remplacées par des interpolations polynomiales cubiques. L'obtention, à partir des points de contrôle, des coefficients de courbure *n* et *m* pour les courbes $y = x^n$ et $z = x^m$ s'est avérée lourde en calcul et impose de plus que les points de courbure C_* soient situés exactement entre les points les encadrant :

$$y = y_0 + (y_1 - y_0)\left(\frac{x - x_0}{x_1 - x_0}\right)^n \Rightarrow n = \frac{\ln\left(\frac{y_i - y_0}{y_1 - y_0}\right)}{\ln\left(\frac{x_i - x_0}{x_1 - x_0}\right)} \text{ au point } (x_i \quad y_i).$$

Dans le contexte d'une analyse automatique, cette formulation peut amener des instabilités numériques importantes qu'il aurait fallu gérer à chaque calcul du modèle. Les polynômes d'interpolation de degré trois n'imposent aucune contrainte a priori sur la disposition relative des points. Par ailleurs, l'étude statistique menée par Guiard et al ont montré que le coefficient n des courbes de type $y = x^n$ était proche de trois dans le cas du locuteur étudié.

Trois points d'interpolation laissent à un polynôme du troisième degré un dernier paramètre libre qui sera utilisé dans notre cas pour fixer une valeur de tangente aux points de raccordement des contours. La tangente nulle, implicite dans le cas des courbes de type $y = x^n$ sera donc imposée explicitement ici (Figure 30).

Nous avons conservé le principe de la séparation des équations de chaque contour 3D selon le plan XY coronal (interpolation $y(x)$) et le plan XZ axial (interpolation $z(x)$). Par mise en correspondance de ces deux plans, les mêmes points d'interpolation permettent de définir toutes les équations 3D d'un même contour.

Dans le plan XY, chaque contour est construit à partir de 10 points d'interpolation. Il est composé de 6 polynômes de degré trois $y(x)$ raccordés C^1 (tangentes continues), sauf aux commissures E_* où le raccord est seulement C^0 (tangentes non continues). Les coordonnées X sont d'abord calculées en interpolant pour chacune des 6 courbes les coordonnées X des points d'interpolation du contour. Les coordonnées Y sont ensuite interpolées par des

polynômes de degré trois sur les valeurs de X en imposant des tangentes nulles au point inférieur *I* et aux deux points de l'arc de cupidon A_* .

Dans le plan axial XZ, chaque contour est construit à partir de 8 points d'interpolation pris parmi les 10 (les valeur en Z des deux points A_* de l'arc de cupidon ne sont pas utilisées, elles sont automatiquement déduites). Les coordonnées Z sont interpolées par quatre polynômes de degré trois sur les valeurs de X en imposant des tangentes nulles aux points supérieurs *S* et inférieur *I*. Le parcours le long d'un contour, correspondant à l'échantillonnage des valeurs de X, donne le premier paramètre de la surface.

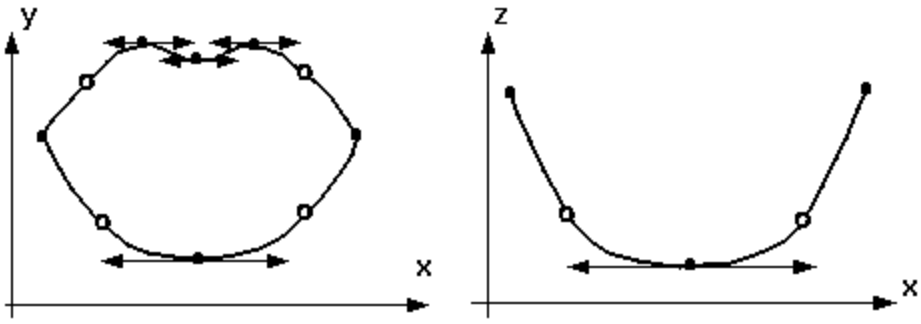


Figure 30. Courbes $y(x)$ et $z(x)$ d'un contour 3D quelconque. Les cercles pleins délimitent les portions de courbes.

Seuls les 10 points d'interpolation de chacun des trois contours de base sont donnés a priori par les 30 points de contrôle géométrique. Ainsi, afin de générer n'importe quel contour intermédiaire entre les trois, les 10 points d'interpolation de n'importe quel contour sont eux-mêmes déduits par interpolation. Chacun des 10 points nécessaires à la construction d'un contour 3D est interpolé par un polynôme de degré trois à partir des trois points correspondant au même lieu géométrique sur les trois contours de base. L'échantillonnage sur ces 10 interpolations simultanées donne le second paramètre de la surface. Le nombre de pas d'échantillonnage correspond au nombre de contours intermédiaires.

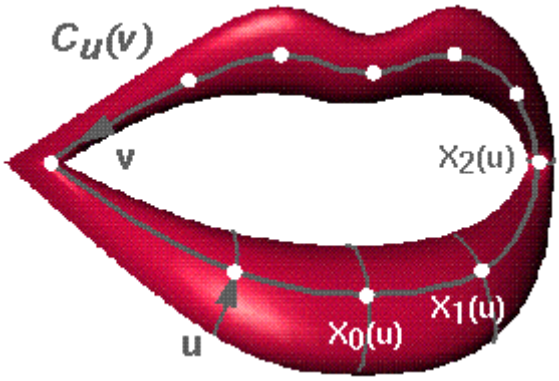


Figure 31. Modèle 3D de lèvres par surface paramétrique (9 contours).

La paramètrisation en u donne un ensemble de 10 points de contrôle $X_i(u)$ permettant de définir un contour 3D complet $C_u(v)$ dont le parcours est échantillonné par v.

Nous présentons par la suite une méthode d'édition manuelle des points de contrôle géométrique et du modèle à partir d'images vidéo de lèvres prises simultanément sous deux vues différentes. Les points sont placés de telle sorte que le modèle 3D projeté sur 2 vues d'analyse coïncide « au mieux » (selon le manipulateur) avec l'image des lèvres. Avant d'aborder l'édition des positions XYZ des points de contrôle, nous spécifions comment calibrer la projection du modèle 3D sur les vues 2D.

III.2.5 Calibrage des projections 2D sur les vues d'analyse

Différents niveaux de détails peuvent être pris en compte pour modéliser la projection d'un point 3D sur le plan image d'une caméra : effet de perspective, non linéarité et asymétrie de la lentille... Dans le cadre des enregistrements ICP et ATR (ch. 2, §1.1), les lèvres sont proches du centre de l'image et leur taille est très faible par rapport à la distance de la caméra au sujet (5 centimètres pour 150 et 250 cm). De plus, les déplacements dans l'axe de la caméra sont de faible amplitude. En suivant le critère proposé dans la thèse de Vézien (1995), nous nous placerons donc sous l'hypothèse d'un modèle orthographique de caméra, où les effets de perspectives sont négligés : toutes les projections sont alors parallèles entre elles et orthogonales au plan de l'image. La position (trois valeurs de translation et trois angles de rotation) et une précision en nombre de pixels par millimètres (un seul coefficient isotropique) définissent les sept paramètres pour notre modèle de caméra. Les relations entre les coordonnées (x, y, z) dans un référentiel 3D donné et les coordonnées (u, v) sur l'image 2D sont donc les suivantes :

$\vec{X} = R_1 \vec{X}_1 + T_1$, où
$\vec{X} = \begin{bmatrix} x & y & z \end{bmatrix}^t$ sont les coordonnées dans le système de référence,
$\vec{X}_1 = \begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^t$ les coordonnées dans le repère local de la caméra,
R_1 la matrice de rotation de la caméra et T_1 son vecteur de translation dans le système de référence.

Par le facteur d'échelle *a* en pixels par millimètres de la caméra, on a la position (*u*, *v*) du point à l'écran :

$$\begin{pmatrix} u \\ v \end{pmatrix} = a \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$

Différentes techniques existent pour calculer automatiquement les paramètres de la caméra (R_1, T_1 et *a* dans notre cas). Nous avons adopté une méthode simple de calibration manuelle sur une cible plane de dimension connue et munie de plusieurs points de repère. La cible est d'abord présentée la plus parallèle possible au plan de vue de la caméra pour donner une première estimation du paramètre *a*. Ensuite, par ajustement des angles de rotation et des valeurs de translation, un modèle projeté de la cible est aligné au plus près avec l'image réelle de la cible. Les paramètres donnent donc une interprétation de la position de la caméra dans un repère local de la cible. Dans le modèle orthographique, la projection est invariante par translation selon l'axe de la caméra et rend impossible l'ajustement du paramètre de translation suivant cet axe. Ainsi, ce dernier paramètre de translation est fixé égal à la distance de la caméra au sujet (150 cm pour ICP et 250 cm pour ATR).

En utilisant une image de la cible visible sur deux caméras au même moment, on peut déduire la position relative de l'une par rapport à l'autre. Les deux caméra sont alignées dans le même repère local de la cible de calibration :

$\vec{X} = R_1 \vec{X}_1 + T_1 = R_2 \vec{X}_2 + T_2$
où $\vec{X}_1$ et $\vec{X}_2$ s ont les coordonnées locales aux caméra 1 et 2 du même point $\vec{X}$.

En éliminant la dépendance à la cible de calibration, on passe alors de la caméra 2 à la caméra 1 par :

$\vec{X}_1 = \left[R_1^t R_2 \right] \vec{X}_2 + \left[R_1^t \left(T_2 - T_1 \right) \right] = R_c \vec{X}_2 + T_c$
ou sous une forme vectorielle en posant $R_c = \begin{bmatrix} \vec{u} & \vec{v} & \vec{w} \end{bmatrix}$ et $T_c = \begin{bmatrix} \vec{t} \end{bmatrix}$
$\vec{X}_1 = \begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^t = x_2 \vec{u} + y_2 \vec{v} + z_2 \vec{w} + \vec{t}$

Pour tester la validité des approximations du modèle de projection et de la calibration des caméras, un exemple de reconstruction 3D a été testé par mise en correspondance des deux vues. Pour un point quelconque se projetant en (u_1, v_1) sur la vue 1 et (u_2, v_2) sur la vue 2, ses coordonnées dans les repères locaux des caméra sont :

$$\vec{X}_1 = \begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^t \text{, avec } \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \frac{1}{a_1}\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} \text{ et } z_1 \text{ inconnu}$$

$$\vec{X}_2 = \begin{bmatrix} x_2 & y_2 & z_2 \end{bmatrix}^t \text{, avec } \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \frac{1}{a_2}\begin{pmatrix} u_2 \\ v_2 \end{pmatrix} \text{ et } z_2 \text{ inconnu}$$

Par la connaissance de z_1, on déduit la position 3D du point dans le repère local de la première caméra 1. Cette valeur ne peut être trouvée à partir de la vue 1. On recherche alors la position z_1 qui minimise l'erreur de projection du point sur la vue 2.

En isolant le terme en z_1, on a :

$$x_2(z_1) = \left(\vec{X}_1 - \vec{t}\right)\cdot\vec{u} = \left(\vec{c}_1 - \vec{t}\right)\cdot\vec{u} + z_1\left(\vec{u}\cdot\vec{e}_3\right)$$
$$y_2(z_1) = \left(\vec{X}_1 - \vec{t}\right)\cdot\vec{v} = \left(\vec{c}_1 - \vec{t}\right)\cdot\vec{v} + z_1\left(\vec{v}\cdot\vec{e}_3\right)$$
$$\text{avec } \vec{c}_1 = \begin{bmatrix} x_1 & y_1 & 0 \end{bmatrix}^t \text{ et } \vec{e}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^t$$

On cherche alors à minimiser l'erreur de projection sur la caméra 2 en fonction de z_1 :

$$e(z_1) = \left[x_2(z_1) - x_2\right]^2 + \left[y_2(z_1) - y_2\right]^2$$

Pour cela, il faut $e'(z_1) = 0$, soit en posant $u_z = \vec{u}\cdot\vec{e}_3$ et $v_z = \vec{v}\cdot\vec{e}_3$,

$$x_2'(z_1)\left[x_2(z_1) - x_2\right] + y_2'(z_1)\left[y_2(z_1) - y_2\right] = 0\text{, d'où}$$
$$u_z\left[\left(\vec{c}_1 - \vec{t}\right)\cdot\vec{u} + u_z z_1 - x_2\right] + v_z\left[\left(\vec{c}_1 - \vec{t}\right)\cdot\vec{v} + v_z z_1 - y_2\right] = 0$$

D'où on déduit le calcul de z_1 en fonction de x_1, y_1, x_2, y_2 :

$$z_1(x_1, y_1, x_2, y_2) = \frac{u_z x_2 + v_z y_2 - \left(\vec{c}_1 - \vec{t}\right)\cdot\left(u_z\vec{u} + v_z\vec{v}\right)}{u_z^2 + v_z^2}$$

Le terme du dénominateur ne s'annule que si les deux caméra sont orientées parallèlement au même axe. Ce n'était pas le cas pour nos enregistrements. Si les deux caméras sont orientées de telle sorte que leurs positions relatives ne diffèrent que d'une translation et une rotation d'axe y (cas de deux caméras parfaitement parallèles au sol), les équations se simplifient. Le terme v_z s'annule et il vient alors :

$$z_1\left(x_1, y_1, x_2, y_2\right) = \frac{1}{u_z}\left[x_2 - \left(\vec{c}_1 - \vec{t}\right)\cdot\vec{u}\right]$$

Or on a,

$$x_2\left(z_1\right) = \left(\vec{c}_1 - \vec{t}\right)\cdot\vec{u} + z_1 u_z$$
$$y_2\left(z_1\right) = \left(\vec{c}_1 - \vec{t}\right)\cdot\vec{v}$$

On retrouve donc une reconstruction exacte avec $x_2\left(z_1\right) = x_2$ et $y_2\left(z_1\right)$ indépendant de z_1.

La figure suivante illustre la qualité de reconstruction d'un parallélépipède de dimensions connues. La précision obtenue est de l'ordre de 2 millimètres.

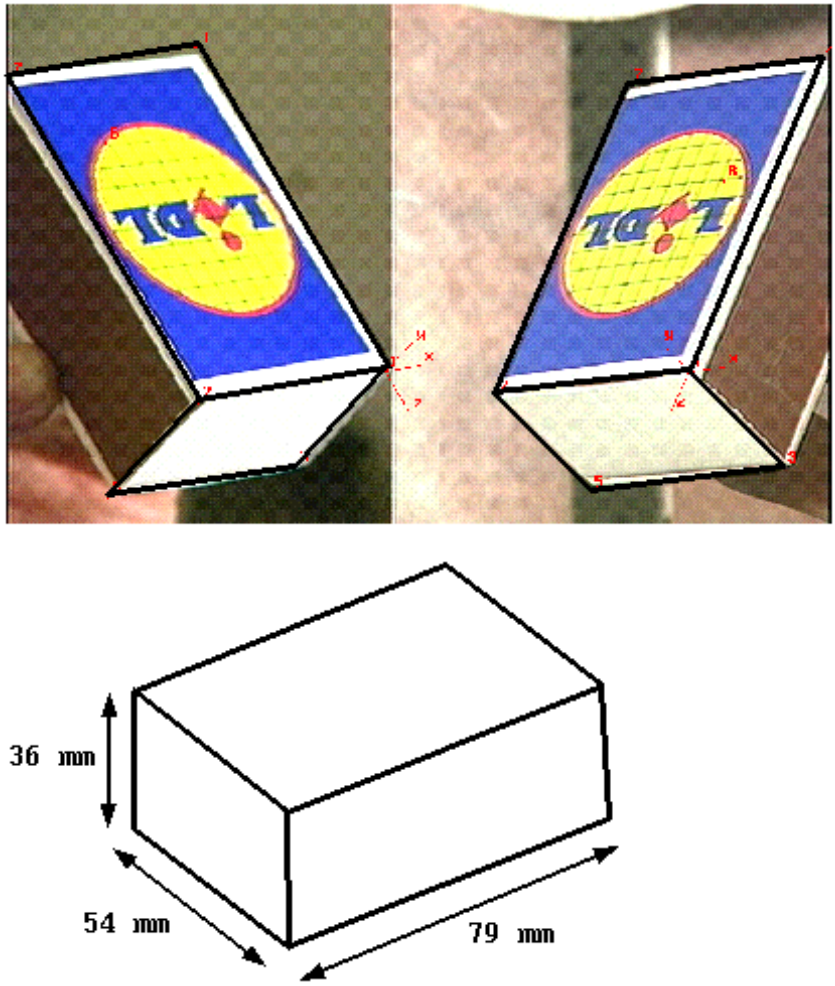


Figure 32. Reconstruction 3D d'un parallélépipède (Taille réelle = 79x55x34 mm).

Le système de coordonnées dans lequel sont exprimées les positions XYZ des points est lié aux caméras. Tout mouvement libre de la tête par rapport aux caméras introduit donc un biais pour la mesure des mouvements labiaux propre à la parole. L'utilisation du casque « Labiophone » limite ce problème puisque la tête est solidaire de la caméra. Néanmoins ce casque ne permet qu'une seule prise de vue et interdit toute utilisation pour la reconstruction

3D. Dans les autres cas d'enregistrements où l'on dispose de deux vues cohérentes du locuteur, il sera donc nécessaire de mesurer la position de la tête (sous forme d'un déplacement rigide) pour ensuite la soustraire à la position des points des lèvres.

Le déplacement rigide se mesure à partir d'une position initiale de la tête. Pour tout point $\vec{X}_{H,0}$ dans ce repère initial, la relation avec le même point $\vec{X}_H$ dans le repère lié à la tête après un mouvement rigide s'exprime par :

$$\vec{X}_{H,0} = R\vec{X}_H + T$$

En alignant un objet de dimensions connues (exprimées dans le repère lié à la tête en position initiale) sur son image par la caméra 1, on détermine la correspondance entre le repère lié à la caméra et le repère initial lié à la tête :

$$\vec{X}_1 = R_H \vec{X}_{H,0} + T_H$$

Ainsi, on obtient la relation entre un point lié à la tête (quelque soit sa position) et son image sur la caméra 1 par :

$$\vec{X}_1 = R_H\left(R\vec{X}_H + T\right) + T_H$$

Par la suite, nous supposerons que l'orientation de la tête reste fixe par rapport aux caméras et qu'elle est parallèle à celle qui correspond à la vue de face. Seuls les variations en translations XY (dans le plan de la caméra) seront corrigées par le suivi d'un point repéré sur la tête, en l'occurrence un point identifiable sur le nez.

III.2.6 Construction du modèle géométrique

III.2.6.1 Construction du modèle à partir de 2 vues calibrées

Bien que le système développé ne l'impose pas, les vues de face et profil ont été choisies pour la repérage de la position des 30 points de contrôle. La vue de face apporte de manière naturelle la plus importante information visuelle en communication orale. La vue de profil quant à elle présente l'avantage non seulement de couvrir l'information géométrique complémentaire la plus large (puisque orthogonale) pour le repérage dans l'espace mais aussi d'apporter une information articulatoire directe sur la protrusion labiale. Cette vue possède cependant aussi l'inconvénient de faire disparaître les points situés sur le profil non filmé du

locuteur. Nous aurons recours à des hypothèses de symétrie de profils pour contourner ce problème.

Pour l'entraînement de son modèle 3D à partir d'image réelles, Basu (1998) utilise un repérage stéréoscopique de points physiques marqués à l'encre sur la surface des lèvres. Malheureusement, cette méthode ne permet pas de repérer les points sur le contour interne. Il en découle une mauvaise représentation de ce contour puisque celui-ci doit ainsi être extrapolé à partir du reste du modèle. Dans notre cas, les lèvres sont analysées sans aucun maquillage pour l'identification des points de contrôle dont le placement est laissé libre au manipulateur.

Nous nous appuyons sur le modèle géométrique pour la recherche des points de contrôle géométriques : en effet, ces points ne constituent que l'interface de contrôle de la forme labiale et c'est la forme - au sens « gestalt » - que nous cherchons à mettre en correspondance avec les contours labiaux, conférant ainsi plus de robustesse au placement des points. A l'aide d'un interface graphique, l'utilisateur déplace donc les 30 points d'une configuration de départ en ajustant leurs coordonnées XYZ de telle manière que le modèle géométrique 3D, synthétisé et projeté simultanément sur les deux vues par les modèles de caméras, coïncide au plus prés avec l'image réelle des lèvres. Les deux vues sont présentées sur une même image : soit un mixeur vidéo permet d'afficher côte à côte les signaux des deux caméras (enregistrements ICP), soit la deuxième vue est obtenue par l'image d'un miroir disposé dans le plan d'une caméra unique (enregistrements ATR). A noter que dans ce dernier cas, le modèle de caméra correspond à une vue virtuelle où l'axe des X est inversé. L'orientation du miroir, inconnue a priori, est déterminée par la procédure de calibration des caméras.

Les coordonnées XY sont ajustées sur la vue de face. Les variations en XY sont reportées sur la vue de profil, où seule la coordonnée Z des points est modifiable. Le surplus d'information apporté par le tracé du modèle aide à la décision du positionnement des points. Cette aide est indispensable pour la vue de profil où certains points n'apparaissent pas alors que leur positionnement influent directement sur le contour apparent du modèle (« ombre » projetée du modèle). Un exemple typique est le repli de la peau des joues qui masque la commissure des lèvres sur le contour externe. En extrapolant le contour apparent, le point de contrôle est placé en cherchant l'adéquation non plus d'un point physique mais de la courbe du contour apparent.

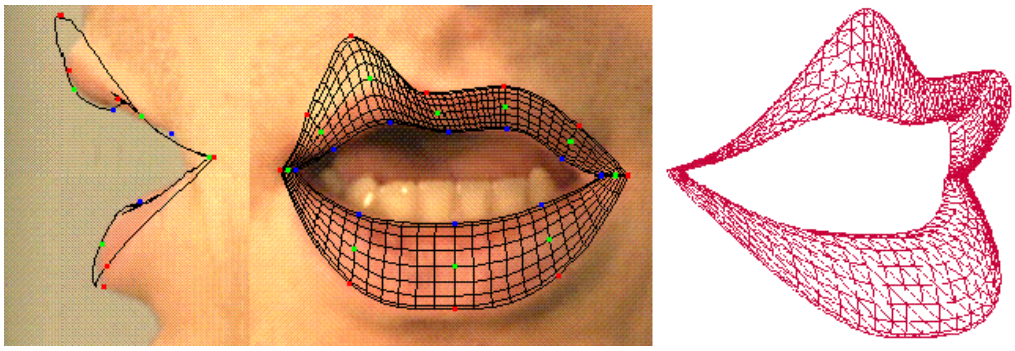


Figure 33. Edition des points de contrôle par adéquation des contours apparents. Le point à gauche de l'arc de cupidon a volontairement été placé sur une position éloignée de son emplacement réel.

Dans la vue de face, les contours apparents correspondent aux contours externes et internes définis par les équations $y(x)$. Le contour apparent de la vue de profil n'est pas accessible analytiquement et dépend de l'orientation de la caméra. Aussi, il a été déterminé par « ombrage », c'est à dire par projection sur le plan de vue de tous les points de la surface 3D paramétrique calculée par les 30 points de contrôle. Tout point XYZ de la surface paramétrique est visible dans une vue donnée si le produit scalaire de la normale à la surface en ce point et de la direction de visée de la caméra est positif. Un point de la surface est décrété appartenir au contour apparent s'il est visible et qu'au moins un de ses voisins ne l'est pas. Enfin, un segment est tracé entre deux points voisins étiquetés « contour apparent » de la surface paramétrique. Ce simple algorithme est suffisant en raison de la géométrie convexe des lèvres qui épousent de manière cylindrique la forme de la mâchoire.

III.2.6.2 Réduction des degrés de liberté par contraintes géométriques

Les contours apparents ne permettent pas toujours de déterminer le positionnement de certains points de contrôle. De face, c'est le cas de tous les points du contour intermédiaire qui n'ont aucune influence sur la définition des contours internes et externes. De profil, seul le contour externe et quelques points des contour interne et intermédiaires peuvent être placés. Toute la surface située sur l'autre profil est totalement masquée et la profondeur des points de contrôle ne peut être déduite. Pour contourner ce problème dans le cadre de l'analyse manuelle, des règles géométriques ont été introduites pour inférer la position des points inaccessibles en fonction de celles des autres points.

D'autres règles ont été introduites pour stabiliser la mesure des valeurs XY des points accessibles $I^{int,ext}$ et $C_*^{int,ext}$. Dans la perspective d'appliquer une analyse statistique sur les coordonnées XYZ des points de contrôle (modèle linéaire présenté par la suite), il était

nécessaire d'imposer dans certains cas une contrainte sur une des deux coordonnées pour mesurer les variations de forme et non pas celles liées aux ambiguïtés de positionnement auquel le manipulateur peut se heurter. Empiriquement, nous avons abouti aux 10 règles suivantes :

1. toutes les positions XY des points X^{med} du contour intermédiaire étant indécidables, celles-ci sont placés au milieu de celles des points X^{int} et X^{ext} correspondants sur les contours internes et externes ;

2. les positions Z des points du profil masqué $C_{\{I,S\}D}{}^{*}$, $A_D{}^{*}$ et $E_D{}^{*}$ sont déduites par symétrie avec leur homologue $C_{\{I,S\}G}{}^{*}$, $A_G{}^{*}$ et $E_G{}^{*}$ sur le profil visible ;

3. la position Z des points aux extrémités $E_G{}^{int}$ du contour interne est au milieu de celles des points aux extrémités $E_G{}^{med}$ et $E_G{}^{ext}$ des contours intermédiaires et externes ;

4. la position Z des points de courbure $C_{\{I,S\}G}{}^{med}$ du contour intermédiaire dépend des positions Z des points I^{*}, S^{*}, $E_G{}^{*}$ et $C_{\{I,S\}G}{}^{int,ext}$. Les quatre courbes Z en fonction de X des contours interne et externe des lèvres supérieure et inférieure sont interprétées en inversant une équation du type $z = x^m$. Les points aux extrémités ($E_G{}^{int}$-I^{int} , $E_G{}^{int}$-S^{int} , $E_g{}^{ext}$-I^{ext} , $E_g{}^{ext}$-S^{ext}) de ces courbes sont connus et les coefficients de courbure *m* internes et externes sont donnés par les points de courbure $C_{\{I,S\}G}{}^{int,ext}$. Des coefficients de courbure similaires sont ensuite déduits des coefficients *m* internes et externes pour les courbes $z = x^m$ supérieure et inférieure du contour intermédiaire. Connaissant les points aux extrémités $E_G{}^{med}$, I^{med} et S^{med} du contour intermédiaire, les deux équations pour les lèvres supérieure et inférieure sont entièrement spécifiées. En prenant ensuite la valeur de l'équation $z = x^m$ du contour intermédiaire prise en X des points $C_{\{I,S\}G}{}^{med}$ la position Z des points de courbure $C_{\{I,S\}G}{}^{med}$ du contour intermédiaire est déterminée ;

5. La tangente $y'(x)$ au point I étant nulle, le positionnement en X de ce point peut s'avérer ambiguë. Aussi, la valeur X des points $I^{int,ext}$ est fixée au milieu de celle des points $E_G{}^{int,ext}$ et $E_D{}^{int,ext}$;

6. Les points *C* sont interpolés pour déterminer l'allure générale de la courbure. Tous les couples (X, Y) d'une même courbe donnent donc un résultat identique. Aussi, la valeur en X des points $C_{I\{G,D\}}{}^{int,ext}$ est fixée égale à la moitié des valeurs X des points $I^{int,ext}$ et $E_{\{G,D\}}{}^{int,ext}$;

7. Pour les mêmes raisons, la valeur X des points $C_{S\{G,D\}}^{int,ext}$ a été fixée égale à la moitié des valeurs X des points $E_{\{G,D\}}^{int,ext}$ et $A_{\{G,D\}}^{int,ext}$;

8. La symétrie entre la gauche et la droite de l'arc de cupidon étant généralement respectée, la valeur X des points $S^{int,ext}$ a été fixée égale à la moitié de celles des points $A_G^{int,ext}$ et $A_D^{int,ext}$. Sans restreindre de manière importante la généralité de la forme, cette règle supprime ainsi un degré de liberté ;

9. La position en Z des points de l'arc de cupidon $A_{\{GD\}}^{*}$ n'intervenant pas dans le calcul des courbes $z(x)$, leur valeur est recalculée à partir des deux courbes z(x) joignant les points S^{*} et $E_{\{G,D\}}^{*}$.

10. Enfin, dans le but de traiter le cas où les commissures ne sont pas totalement accessibles, la largeur dans le plan XY des commissures a été fixée. Elle est mesurée une fois pour toutes sur une forme ouverte où aucun contact entre les lèvres supérieure et inférieure n'apparaît. De chaque côté, la distance en X entre les points des contours interne $E_{\{GD\}}^{int}$ et externe $E_{\{GD\}}^{ext}$ est ensuite maintenue fixe. La hauteur Y des points $E_{\{GD\}}^{int}$ est maintenue égale à celle des points $E_{\{GD\}}^{ext}$. Le contour interne tel qu'il apparaît ensuite est représenté par un « chevauchement » des courbes Y de X du contour interne (Figure 34). Ce procédé fournit un moyen de détecter les contacts entre les deux lèvres. Par un traitement supplémentaire, il est cependant envisageable de modéliser plus précisément ces contacts et de modifier la place des contours en conséquence. A la synthèse, la gestion des faces cachées de l'objet 3D génère intrinsèquement une limite visible de contact.

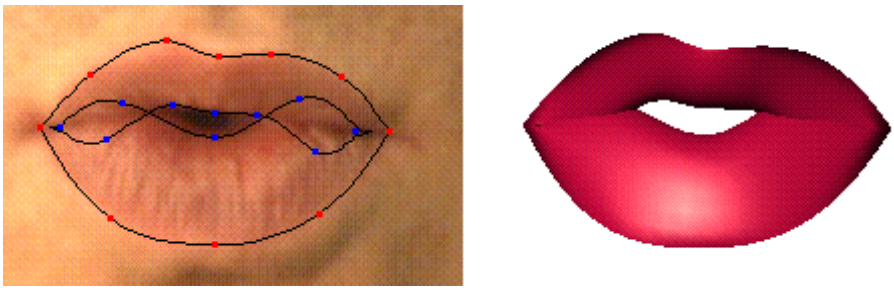


Figure 34. Exemple de modélisation des contacts pour la réalisation de petites ouvertures.

Quatre contraintes supplémentaires contrôlent la vraisemblance de positions relatives de certains points. Contrairement aux 10 règles précédentes, elles ne réduisent pas les degrés de liberté et fixent seulement des limites relatives de variations :

1. les valeurs en X des points de la partie droite doivent rester supérieures à celles des points situés à gauche,
2. la lèvre supérieure reste au-dessus de la lèvre inférieure en maintenant la valeur Y du point I^{int} inférieure à celle du point S^{int} à un millimètre près pour simuler géométriquement une occlusion complète des deux lèvres,
3. les valeurs Y des points I^{ext}, I^{med} et I^{int} sont strictement croissantes,
4. les valeurs Y des points S^{ext}, S^{med}, S^{int} sont strictement décroissantes.

Au terme de cette procédure, la positions XYZ des 30 points de contrôle est déterminée. Les 10 règles géométriques réduisent à 36 les 90 degrés de liberté initiaux. Outre la projection de la surface paramétrique, les résultats sont visualisables par un objet Open Inventor (OpenInventor) généré à partir des points de la surface paramétrique. La figure suivante donne quelques illustrations de l'adaptation du modèle à différents locuteurs.

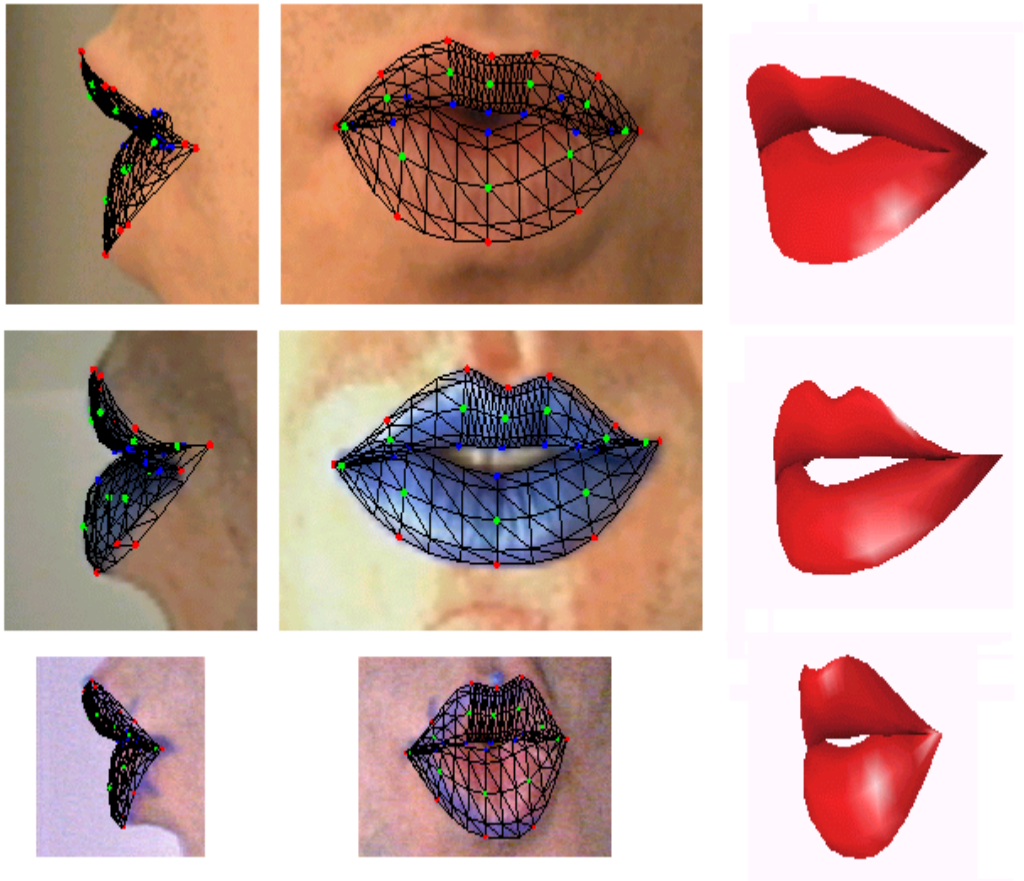


Figure 35. Adaptation du modèle à trois locuteurs différents.

III.2.7 Analyses géométriques à partir du modèle

Il est possible de déduire des mesures géométriques du type des paramètres ICP à partir de toute position du modèle géométrique . Les coordonnées XYZ des points de la surface 3D permettent de déduire directement un certain nombre de paramètres géométriques :

- ouverture interne, notée B, égale la différence des coordonnées Y des points A^{int} et I^{int},
- ouverture externe, notée B', égale la différence des coordonnées Y des points A^{ext} et I^{ext},
- écartement externe, notée A', égale la différence des coordonnées X des points G^{ext} et D^{ext},
- valeurs de protrusion avec les coordonnées Z des points A^{med} (lèvre supérieure), I^{med} (lèvre inférieure) et $E_{GD}{}^{med}$ (commissure, gauche ou droite), notées respectivement P1, P2.

La détermination des paramètres d'écartement (noté A) et de protrusion (noté C) du contact labiale nécessite une analyse indirecte de la géométrie du modèle. La méthode la plus simple consiste à synthétiser l'ombre projetée du modèle sur les plans XY et XZ et d'en déduire les paramètres. Cette méthode permet par ailleurs de calculer les paramètres de surface interne S

et externe S'. A noter que dans ce cas, la précision de l'échantillonnage de la surface 3D influence la valeur des paramètres.

III.3 Un modèle linéaire : de la géométrie à l'articulatoire

Les 36 degrés de liberté de la modélisation géométrique permettent d'appréhender pratiquement toute forme de lèvres et de locuteur. Hormis les 4 contraintes sur les positions relatives des points de contrôle, il n'est fait aucune hypothèse sur la physiologie intrinsèque des lèvres. Ainsi, notre modélisation géométrique tolère en plus des formes normales, des formes potentiellement « aberrantes » ou inaccessibles. Nous avons évoqué en III.2.3 les modèles statistiques qui ont été proposés pour représenter et contrôler de manière vraisemblable les mouvements articulatoires suivant très peu de paramètres. S'inspirant de cette approche, nous avons réduit, pour un locuteur, la variabilité des 36 degrés de liberté à un modèle statistique contrôlé par 3 paramètres. Le modèle s'appuie sur un ensemble d'apprentissage de 10 formes caractéristiques. Le choix de ces formes provient des 23 visèmes identifiées par Benoît et al. (1992) à partir d'un large corpus d'un seul locuteur maquillé. Nous traitons d'abord ce premier locuteur à partir de 10 images issues du même corpus maquillé. Nous appliquons ensuite la même méthodologie à un autre locuteur non maquillé prononçant les 10 visèmes d'apprentissage pour la construction de son modèle articulatoire.

III.3.1 Les 23 visèmes du Français

Les six classes articulatoires que nous avions citées en III.1.2 regroupent les phonèmes selon un point de vue purement phonétique en indiquant pour chaque classe son trait articulatoire distinctif. Nous avons vu que les phénomènes de coarticulation peuvent modifier de manière importante la forme que prennent les lèvres pour la prononciation d'un phonème en situation non isolée. La classification idéale en six groupes des formes labiales ne couvre donc pas toutes les possibilités de formes labiales lorsqu'elles apparaissent dans un contexte réaliste de parole continue. Des classements ont été proposés à l'issue d'expériences en perception visuelle de la parole (Fisher, 1968 ; Summerfield, 1989). Si les résultats ont montré des classifications moins restrictives, ils restent néanmoins dépendants des phénomènes de coarticulation liés aux mots choisis pour le corpus.

En 1990 à l'ICP, a été enregistré un corpus audiovisuel d'un locuteur français, filmé de face et de profil, et dont les lèvres étaient soigneusement maquillées en bleu pour détecter les contours externe et interne avec une précision de l'ordre du mm (Lallouache, 1991). Le

locuteur avait été sélectionné pour sa forte symétrie gauche et droite et le dynamisme de ses mouvements labiaux. Ce corpus a été spécialement conçu pour explorer un maximum de situation de coarticulations pouvant apparaître en Français. Il est constitué de répétitions de 74 mots sans signification réalisant des transitions entre phonèmes. Chaque mot est construit sur une syntaxe $V_iC_kV_jC_kV_iz$ avec $C \in \{b, v, z, ʒ, l, r\} \cup \{\varnothing\}$ et $V \in \{a, i, y\}$ donnant 63 combinaisons, ou sur une syntaxe $V_iV_iV_iz$ avec $V \in \{e, ɛ, ø, œ, u, o, ɔ, ɛ̃, œ̃, õ, ã\}$ donnant 11 combinaisons. Chaque mot est prononcé dans une phrase porteuse « C'est pas [mot] ? ». La forme interrogative a pour fonction d'accentuer la voyelle finale, allongée de plus par la dernière consonne [z]. Chaque mot est répété 10 fois dans un ordre pseudo-aléatoire évitant les répétitions consécutives du même mot. Les 740 stimuli ont été prononcés lors d'une même session. Pour chaque mot, des experts phonéticiens ont ensuite étiqueté les 5 images des mots $V_iC_kV_jC_kV_iz$ et les 3 images des mots $V_iV_iV_iz$ correspondants aux centres de réalisation acoustique des voyelles et des consonnes. A l'aide d'un fort éclairage (1000W) et d'un maquillage vif, un système vidéo de chromakey permet d'isoler facilement du reste du visage les contours apparants des lèvres sur une image mixant les vues simultanées de face et de profil (Lallouache, 1991). A partir des contours, 4 aires et 7 distances sont mesurées sur la face et 3 distances sur le profil.

En vue d'identifier des classes stables de formes labiales prenant en compte les effets de coarticulation, Benoît a sélectionné pour une étude statistique un corpus de 774 images prises parmi 9 répétitions des mots suivants :

- la voyelle centrale des mots $V_i\mathbf{V_i}V_iz$ avec $V \in \{e, ɛ, ø, œ, u, o, ɔ, ɛ̃, œ̃, õ, ã\} \cup \{a, i, y\}$,
- la voyelle centrale des mots $V_iC_k\mathbf{V_j}C_kV_iz$ avec $C \in \{b, v, z, ʒ, l, r\}$ et $V \in \{a, i, y\}$,
- la deuxième consonne des mots $V_iC_kV_j\mathbf{C_k}V_iz$ avec $C \in \{b, v, z, ʒ, l, r\}$ et $V \in \{a, i, y\}$.

Par ce choix, sont analysées les 14 voyelles prises dans un contexte propre, les effets de 6 consonnes sur 3 voyelles et réciproquement de 3 voyelles sur 6 consonnes. En combinant analyse factorielle et classement dynamique, 21 groupes de formes ont été identifiés. A ces 21 classes, deux classes ont été ajoutées pour les formes préphonatoires et les formes fermées au repos. En prenant pour chaque classe le représentant le plus proche du barycentre, les 23 « visèmes » ont ainsi été identifiés parmi le corpus.

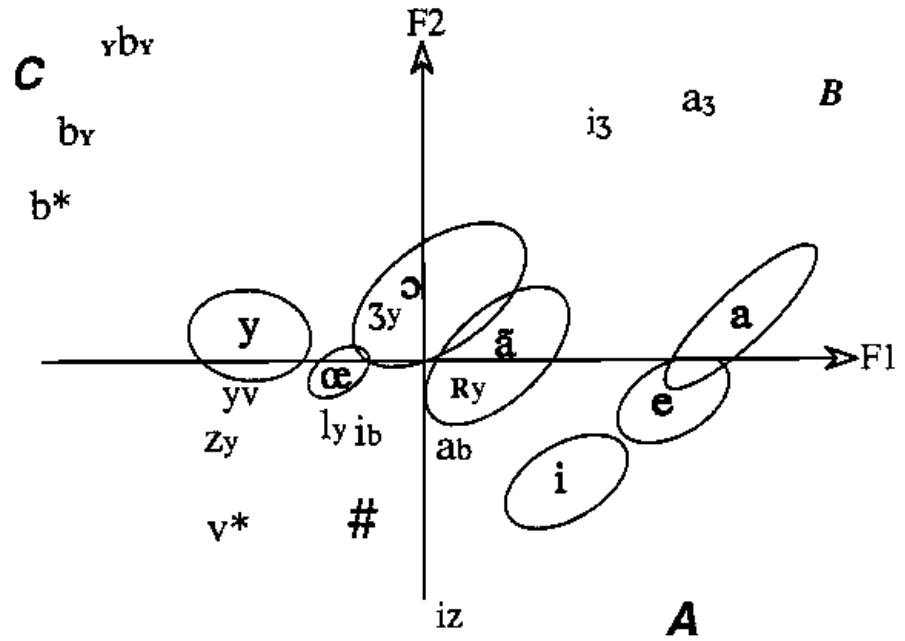


Figure 36. Projection factorielle des 23 visèmes sur les deux premiers facteurs, obtenue par analyse des correspondances (Benoît et al., 1992).

Remarque

Le terme de visèmes reste encore controversé. Il a été introduit pour la première fois par Fisher (1968) comme l'équivalent visuel du phonème (*visual phoneme*). Nous avons montré que les phénomènes de coarticulations rendent très délicate la décision d'arrêter un choix définitif de formes « cibles » prototypiques. Dans son article, Benoît mentionne cette difficulté et désigne comme visème, un représentant réel d'une classe d'allophones (ensembles des réalisations contextuelles d'un même phonème), obtenue *expérimentalement* par son analyse statistique des paramètres géométriques.

L'influence de la coarticulation se manifeste clairement pour la constrictive /ʒ/ qui impose, par exemple, la séparation du [a] en deux visèmes /a/ et /$a_{ʒ}$/ selon que la voyelle est prononcée seule ou en contexte. En complément de la classification des visèmes, une analyse discriminante fait ressortir 6 paramètres principaux comme étant les plus sélectifs : A, B, A', B', C et S.

Pour évaluer la représentativité des 23 visèmes, nous avons comparé une ACP sur les 23 visèmes et une ACP sur les 774 formes dont ils sont issus. Chaque analyse a été calculée en utilisant la même mesure sur les 6 paramètres géométriques les plus discriminants. Dans les deux cas, les trois premiers facteurs comptent pour plus de 90% de la variance totale. De plus, ils sont corrélés terme à terme respectivement à 0.998, 0.987 et 0.927. Ainsi, ces résultats

montrent que la connaissance des 23 visèmes est suffisante pour retrouver la majeure partie de l'information du corpus total de 774 formes.

III.3.2 Réduction du corpus des 23 visèmes à 10 visèmes

Une interprétation articulatoire peut être donnée aux deux premières composantes principales obtenues par ACP des 23 visèmes mesurés par les 6 paramètres géométriques. Ce résultat avait déjà était évoqué en II.2 où les mêmes paramètres géométriques (le paramètre de profil C excepté) étaient déduits des images en niveaux de gris grâce à la base de visèmes. Nous rappelons que la première composante oppose l'ouverture à la fermeture et la seconde l'arrondissement à l'étirement. Nous retrouvons là les traits articulatoires principaux des lèvres.

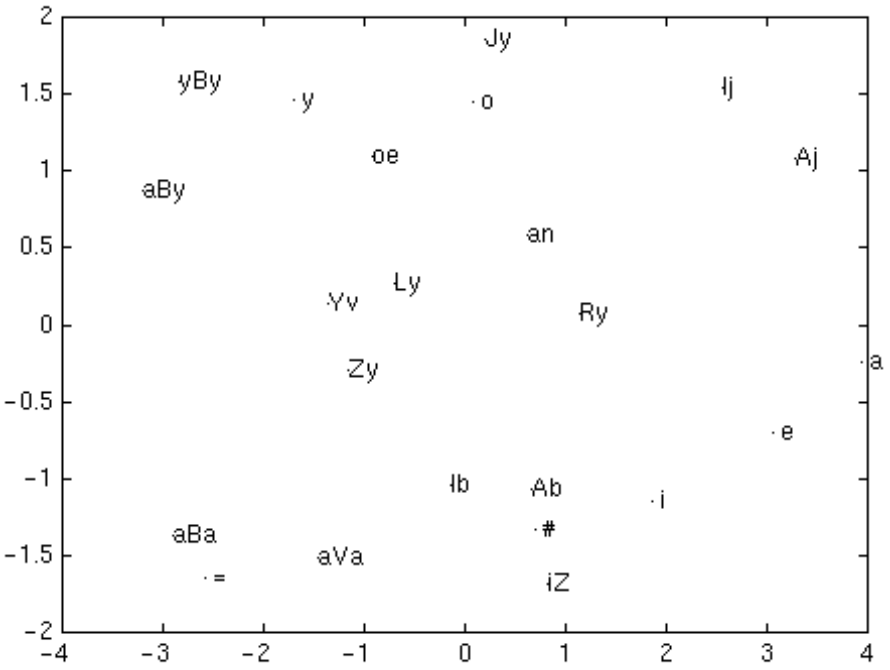


Figure 37. Projection des 23 visèmes sur les deux facteurs (70% et 24%) obtenus par ACP sur 6 paramètres (A, B, A', B', S et C).

La projection sur ces deux facteurs montre que des individus « extrêmes » se dégagent : /$a_{\text{ʒ}}$/, /i_z/, /$_y b_y$/, /o/, /a/, /$_a b_a$/. Ces individus correspondent à des cas où les variations par rapport à la forme moyenne sont les plus importantes. En reprenant l'interprétation articulatoire des composantes, ces formes sont donc produites par une articulation accentuée : /a/ est produit par une ouverture et un étirement, /$a_{\text{ʒ}}$/ par ouverture et arrondissement, /ɔ/ par arrondissement et ouverture moyenne, ... etc. Les autres formes, pour lesquelles l'articulation est moins

prononcée, se situent à l'intérieur de l'espace délimité par les formes extrêmes. Ainsi, nous ne garderons pour le corpus d'apprentissage de notre modèle statistique que ces formes extrêmes, partant du principe qu'elles explorent au plus loin l'espace articulatoire du locuteur. Les formes centrales se retrouvent par interpolation à l'intérieur de l'espace convexe décrit par les formes extrêmes. Il est intéressant de noter que, mise à part les labio-dentales, toutes les classes articulatoires de base sont représentées : voyelle arrondie /o/, voyelle étirée /a/, occlusives bilabiales /$_{a}$b$_{a}$/ et /$_{y}$b$_{y}$/ (se distinguent par la coarticulation imposée par la voyelle /y/), constrictive alvéolaire /i$_{z}$/ et post-alvéaolaire /a$_{ʒ}$/.

Les 6 formes extrêmes n'incluent pas toutes les voyelles du Français. Les cibles vocaliques fournissent des points d'ancrage fort dans l'articulation. Aussi, pour intégrer leur forme *exacte* dans le corpus nous avons ajouté quatre visèmes vocaliques importants par leur couverture : /y/, /œ/, /i/ et /ã/. Pour finir, nous avons retenu 10 formes de base pour notre corpus d'apprentissage : /a/, /i/, /y/, /o/, /œ/, /a$_{ʒ}$/, /ã/, /i$_{z}$/, /$_{a}$b$_{a}$/, /$_{y}$b$_{y}$/.

Le tableau suivant examine la perte d'information induite par la réduction du corpus des 23 visèmes à nos 10 formes. On donne la part de variance expliquée par les trois premiers facteurs dans les quatre cas selon que le corpus est constitué des 23 ou 10 visèmes et selon que la mesure de la forme labiale est faîte par les 6 paramètres géométriques ou les 90 coordonnées XYZ des 30 points de contrôle de notre modélisation géométrique.

Part de la variance et corrélation entre les vecteurs propres obtenus par ACP sur les 23 ou 10 visèmes, mesurés par les 6 paramètres géométriques ou les 90 coordonnées des 30 points de contrôle :

	6 paramètres de mesure			30 points de contrôle		
	23 visèmes	10 visèmes	Corrélation	23 visèmes	10 visèmes	Corrélation
Facteur 1	69.7 %	73.4 %	.99	66.2 %	70.1 %	.99
Facteur 2	24.1 %	20.6 %	.99	22.4 %	22.1 %	.98
Facteur 3	4.6 %	5.6 %	1.0	3.7 %	3.2 %	.81
Somme cumulée	98.4 %	99.6 %	-	92.3 %	95.4 %	-

Pour les deux systèmes de mesure, lorsque le corpus d'apprentissage est réduit de 23 à 10, les trois premiers vecteurs propres des ACP (comptant pour plus de 90% de la variance) restent fortement corrélés. Du point de vue de la modélisation articulatoire, cette réduction de corpus laisse donc le modèle linéaire invariant. Les tableaux suivant donnent la corrélation entre chaque visème et sa reconstruction par 3 facteurs seulement, obtenus par ACP sur un corpus de 23 ou 10 visèmes. Les cas d'une mesure par les 6 paramètres géométriques et par les 30 points de contrôle ont été séparés.

Reconstruction des 23 visèmes à partir des 3 premiers facteurs d'une ACP sur des corpus d'apprentissage de respectivement 23 et 10 visèmes, mesurés par les 6 paramètres :

	a	i	y	e	o	ʒy	œ	**ã**	ab	aʒ	ib	iz	iʒ	yv	aba	aby	yby	ava	zy	ry	ly	#	=
23 visèmes	**.99**	**.99**	**1.0**	.99	.99	.96	**1.0**	**.98**	.99	**1.0**	.98	**1.0**	.99	1.0	**1.0**	.99	**.99**	.98	.98	.95	.98	1.0	1.0
10 visèmes	**.99**	**.99**	**1.0**	.99	.99	.96	**1.0**	**.97**	.98	**1.0**	.97	**1.0**	.99	.99	**1.0**	1.0	**.99**	.98	.98	.96	.98	1.0	1.0

Reconstruction des 23 visèmes à partir des 3 premiers facteurs d'une ACP sur des corpus d'apprentissage de respectivement 23 et 10 visèmes, mesurés par les 90 coordonnées XYZ des 30 points de contrôle géométriques.

	a	i	y	e	o	ʒy	œ	**ã**	ab	aʒ	ib	iz	iʒ	yv	aba	aby	yby	ava	zy	ry	ly	#	=
23 visèmes	**.99**	**.98**	**.97**	.97	**.96**	.97	**.95**	**.76**	.95	**.97**	.91	**.98**	.98	.82	**.93**	.98	**.96**	.88	.95	.87	.91	.97	.97
10 visèmes	**.99**	**1.0**	**.98**	.96	**.97**	.95	**.95**	**.69**	.95	**.97**	.88	**.99**	.94	.84	**.97**	.95	**.98**	.89	.86	.82	.88	.96	.93

La restriction au corpus de 10 visèmes entraînent peu de perte d'information pour la mesure sur 6 paramètres. La mesure par les 30 points de contrôle laisse plus de degrés de liberté. Aussi dans ce cas, la qualité de reconstruction sur 3 facteurs est légèrement moins bonne.

III.3.3 Le modèle articulatoire : des formes aux gestes

Nous abordons la modélisation articulatoire de deux locuteurs différents pour lesquels les 10 formes labiales évoquées ci-dessous ont été mesurées par la position XYZ des 30 points de contrôle du modèle géométrique. Le locuteur 1 correspond à celui de l'enregistrement ICP maquillé. Le locuteur 2 correspond à un cas non maquillé d'enregistrement de type ATR.

III.3.3.1 Locuteur 1 - corpus ICP

La figure suivante présente le nomogramme du locuteur 1 pour les trois premières composantes principales. Elles totalisent respectivement 70%, 22% et 3% de la variance totale sur les 10 visèmes.

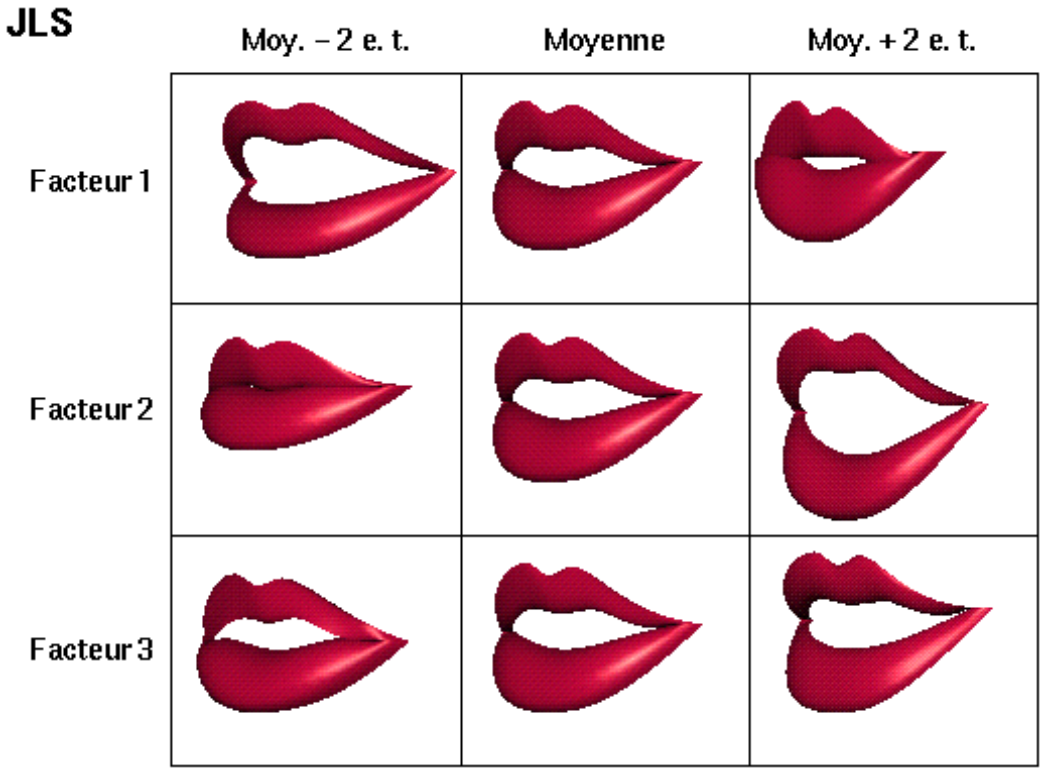
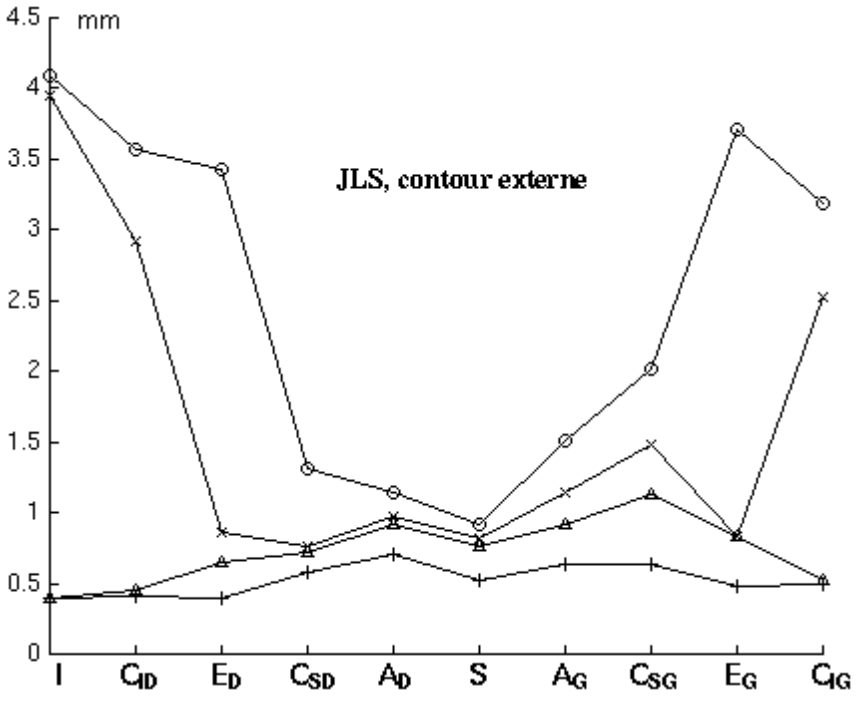


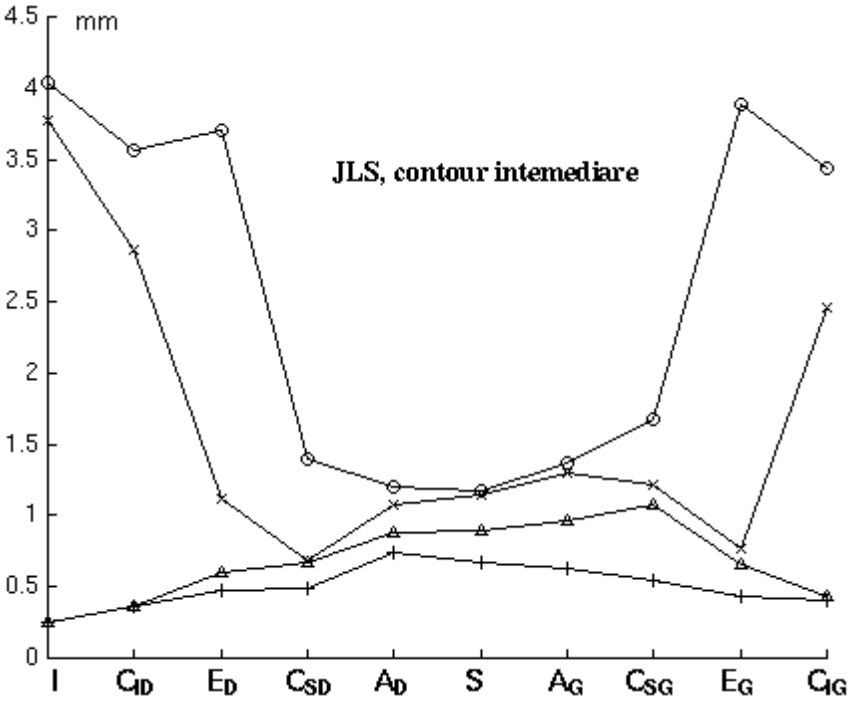
Figure 38. Les trois composantes principales du locuteur 1 (modèle synthètisé à la position moyenne ± 2 fois l'écart-type de chaque paramètre).

Le tableau suivant donne la corrélation des paramètres articulatoires avec 6 paramètres géométriques (écartement externe A', ouverture externe B', écartement interne A, ouverture interne B, protrusion de la lèvre supérieure P1 et inférieure P2) :

	A'	B'	A	B	P1	P2
paramètre 1	-0,92	-0,13	-0,51	-0,28	0,82	0,96
paramètre 2	0,34	-0,94	-0,57	-0,89	-0,20	-0,06
paramètre 3	-0,19	0,31	0,32	0,31	0,14	0,08

Les figures suivantes détaillent pour chaque point de contrôle la part de variance sur les 23 visèmes du mouvement expliquée par chacune des trois composantes principales. Pour chaque point, l'écart-type est calculé sur la norme euclidienne résumant ainsi tout le déplacement dans l'espace XYZ. Les trois contours externe, intermédiaire et interne ont été séparés. La ligne du haut indique l'écart-type initial (cercles o), la dernière l'écart-type restant après soustraction de l'effet des trois composantes (croix +).





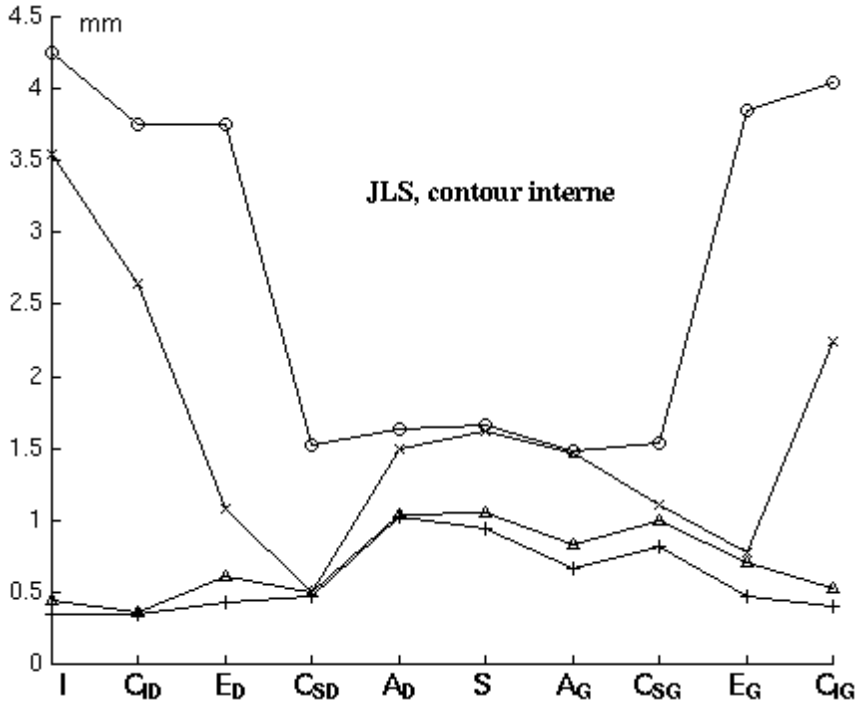


Figure 39. Ecart-type des résidus du mouvement des points de contrôle en supprimant l'effet cummulatif de chaque paramètre articulatoire.

III.3.3.2 Locuteur 2 - Corpus ATR

Sur le corpus ayant servi à identifier les 23 visèmes, les lèvres du locuteur étaient maquillées en bleu. Cette condition idéale de contraste entre vermillon et peau permet de comparer les résultats de notre système de suivi automatique à l'aide du modèle articulatoire (chapitre suivant) à ceux obtenus par une mesure directe sur les masques de seuillage. Pour évaluer le système en situation non maquillée, nous avons utilisé un corpus d'un autre locuteur pour lequel nous avons construit un modèle articulatoire de la même manière sur les 10 formes présentées. Les trois premières composantes principales totalisent respectivement 75%, 12%, et 7% de la variance totale.

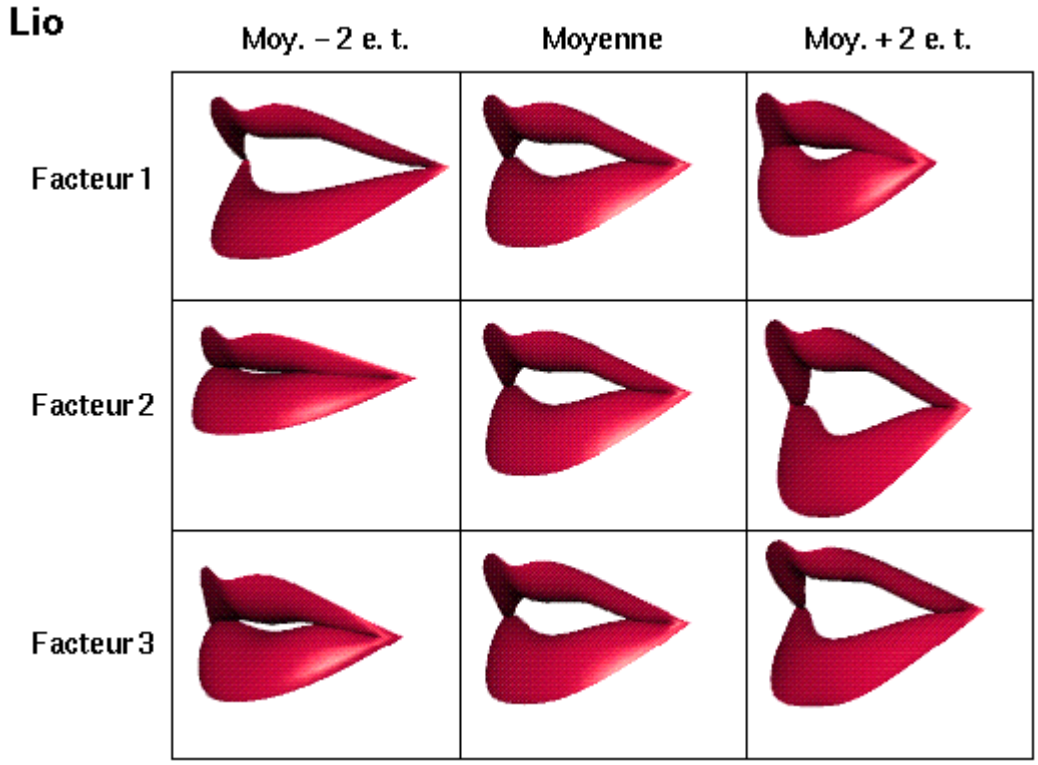
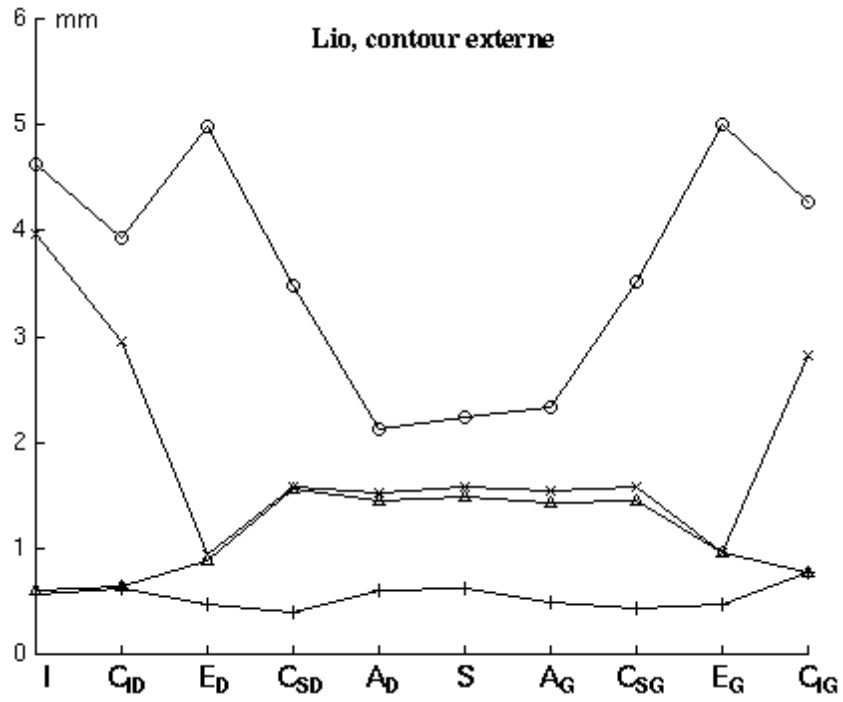


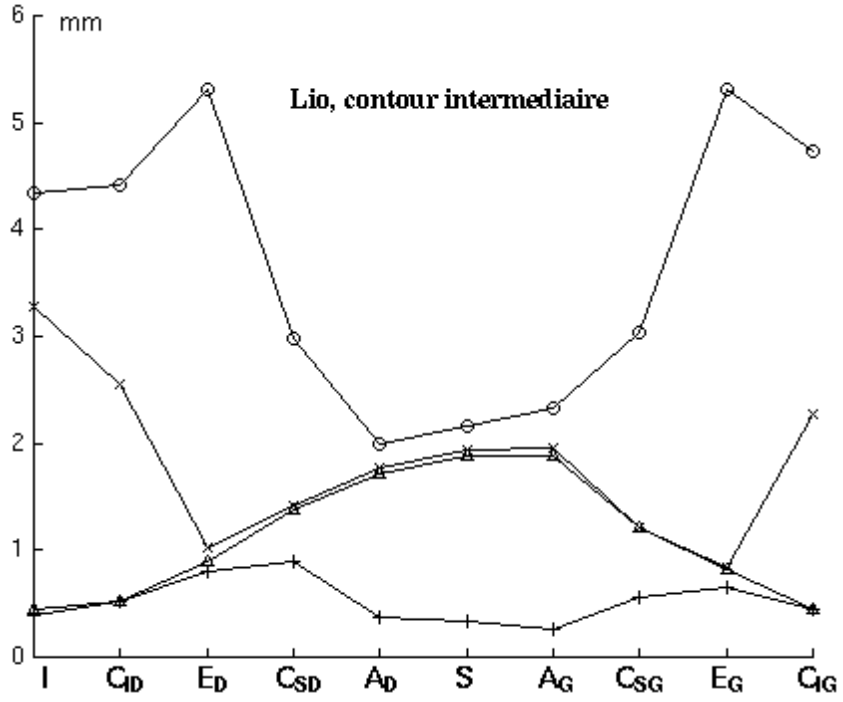
Figure 40. Les trois paramètres articulatoires du locuteur 2.

Le tableau suivant donne la corrélation des paramètres articulatoires avec 6 paramètres géométriques (écartement externe A', ouverture externe B', écartement interne A, ouverture interne B, protrusion de la lèvre supérieure P1 et inférieure P2) :

	A'	B'	A	B	P1	P2
paramètre 1	-0.95	-0.26	-0.80	-0.64	0.94	0.95
paramètre 2	0.21	-0.89	-0.18	-0.56	-0.23	0.22
paramètre 3	-0.09	-0.35	-0.47	-0.51	-0.14	-0.03

De la même manière, nous avons évalué individuellement la part de la variance du mouvement des points de contrôle expliquée par chaque paramètre articulatoire. Les graphiques suivant séparent contours externes, intermédiaires et internes pour le second locuteur.


mm
Lio, contour externe
6
5
4
3
2
1
0
I
CID
ED
CSD
AD
S
AG
CSG
EG
CIG


mm
Lio, contour intermediaire
6
5
4
3
2
1
0
I
CID
ED
CSD
AD
S
AG
CSG
EG
CIG

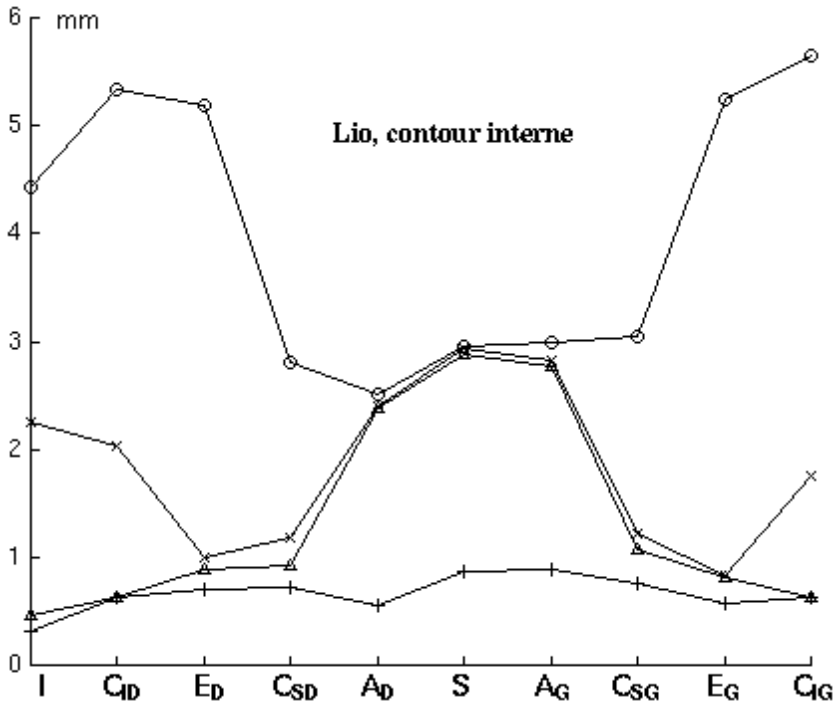


Figure 41. Ecart-type des résidus du mouvement des points de contrôle en supprimant l'effet cummulatif de chaque paramètre articulatoire pour le locuteur 2.

Encore plus clairement pour ce locuteur, il ressort que la première composante principale détermine le mouvement des points aux extrémités E_*, la seconde composante celui des points I, C_{I*} de la lèvre inférieure et la troisième celui des points C_{S*}, A_* et S de la lèvre supérieure.

Pour les deux locuteurs, les variations de formes contrôlées par les trois premières composantes principales ont une interprétation articulatoire (Figure 38 et Figure 40) :

1. la première composante (Locuteur 1 : 70%, Locuteur 2 : 75%) réalise un geste qui oppose arrondissement (lié à une protrusion) et étirement,
2. la seconde (Locuteur 1 : 22%, Locuteur 2 : 12%) correspond à un abaissement de la lèvre inférieure,
3. la troisième (Locuteur 1 : 3%, Locuteur 2 : 7%) correspond à un relèvement de la lèvre supérieure.

Nous avons directement conservé les composantes principales comme commandes articulatoires du modèle pour chaque locuteur. Sans prétendre avoir ainsi retrouver les commandes motrices des lèvres, ces paramètres de contrôle se rapprochent néanmoins des degrés de liberté imposés par les muscles labiaux (voir chapitre 1) :

- l'occlusion labiale, opérée par l'orbiculaire, est réalisée par la coopération des deuxième et troisième paramètres,
- la protrusion et l'arrondissement, opéré par l'orbiculaire et la houppe du menton, est réalisée par le premier paramètre,
- le relèvement de la lèvre supérieure, opéré par les releveurs de l'aile du nez, est réalisée par le troisième paramètre,
- l'abaissement de la lèvre inférieure, opéré à la fois par le carré du menton et le déplacement de la mâchoire, est réalisée par le second paramètre,
- l'écartement des commissures, opéré par le buccinateur, est réalisée par le premier paramètre.

Chaque modèle articulatoire est spécifique au locuteur étudié. Néanmoins, nous venons de voir que, pour deux locuteurs différents, les gestes représentés par les trois composantes principales sont semblables et font référence à des commandes musculaires communes. Numériquement, nous avons donc comparé les vecteurs propres associés aux modèles respectifs des deux locuteurs. Les trois premiers vecteurs propres s'identifient aux paramètres articulatoires. Le tableau suivant donne la matrice de corrélation des 5 premiers vecteurs propres.

Locuteur 1 → Locuteur 2 ↓	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5
Facteur 1	**.98**	-.02	-.02	-.03	-.04
Facteur 2	0	**.90**	-.21	-.01	-.02
Facteur 3	.04	.25	**.80**	-.20	.17
Facteur 4	.04	.26	-.17	**.24**	-.08
Facteur 5	.02	0	-.16	-.26	**.29**

Comparaison entre les deux locuteurs sur les vecteurs propres obtenus par ACP calculée sur les 10 formes de base, mesurées par les 30 points de contrôle.

Il est intéressant de noter, d'abord, l'importance des corrélations entre les trois premiers facteurs, identifiés aux paramètres articulatoires. Ensuite, ces corrélations chutent brusquement à partir du quatrième facteur qui ne correspond plus à un paramètre articulatoire clairement interprétable. Ainsi, même si la morphologie des deux locuteurs est différente, les déformations contrôlées par les paramètres articulatoires semblent proches.

III.4 Discussion

La modélisation géométrique est suffisante pour décrire la plupart des formes de lèvres mais ne contraint pas la physiologie propre des lèvres. Nous avons réduit son contrôle à 3 paramètres articulatoires qui rendent comptent d'une gestuelle labiale cohérente en production de la parole. La modélisation articulatoire impose pour chaque locuteur que 10 formes particulières soient extraites pour construire le modèle.

Contrairement au modèle linéaire classique de Maeda, notre analyse ne prend pas en compte l'information explicite du mouvement de la mâchoire. Elle est apparaît implicitement dans le

second paramètre articulatoire. Cette simplification peut entraîner cependant des problèmes important pour certaines formes telles que les labio-dentales, absentes de notre corpus de 10 visèmes, où le recul de la lèvre inférieur est expliqué par la mâchoire. Ce point sera à prendre en compte pour des modélisations futures plus précises.

Nous aborderons à présent l'application de notre modèle articulatoire au suivi automatique de mouvements labiaux. L'objectif consiste à retrouver les paramètres articulatoires qui donnent au modèle 3D complet après synthèse une apparence la plus proche possible de l'image réelle des lèvres. Se plaçant dans une optique sans maquillage, seul le modèle du second locuteur sera utilisé.

IV. Chapitre 4. Labiométrie par analyse-synthèse d'un modèle articulatoire

Nous avons vu qu'en raison de la proximité des couleurs du vermillon et de la peau, les contours labiaux sont difficiles à distinguer. L'analyse faite au chapitre 2 a montré que même si le contraste entre vermillon et peau est rehaussé, il reste trop flou pour permettre une mesure directe. Au prix d'un grand nombre de degrés de liberté, notre modélisation géométrique 3D donne une définition précise des contours labiaux. Pour deux locuteurs particuliers, nous avons vu que le contrôle d'un modèle 3D reproduisant la gestuelle labiale en parole spécifique à chacun peut cependant se réduire à trois paramètres articulatoires. Comme annoncé au chapitre 1, notre approche de la labiométrie consiste alors à mettre en correspondance l'estimation de la zone du vermillon obtenue par analyse statistique de la couleur (chapitre 2), avec la projection du modèle 3D du locuteur filmé (chapitre 3) en adaptant par optimisation les trois paramètres articulatoires.

IV.1 Suivi de contours et inversion articulatoire

Etant défini comme une surface 3D, le modèle fournit plus d'information que les seuls contours (Revéret et al., 1997). Ainsi, nous nous intéressons dans ce chapitre à l'adéquation entre « l'ombre » projetée du modèle et *toute* la région du vermillon, estimée par l'analyse statistique de la couleur. Notre approche de la labiométrie consiste à retrouver à partir de l'image des lèvres, les commandes articulatoires du modèle qui projette au mieux - suivant l'angle de vue - le modèle sur l'image. Transposé dans la modalité visuelle, ce problème se rattache à celui de l'inversion articulatoire, classiquement réalisée à partir d'un signal acoustique. Le but dans ce cas est de retrouver les paramètres d'un système de synthèse articulatoire, basé sur un modèle du conduit vocal (voir III.3.2), afin qu'il produise les caractéristiques acoustiques d'un donné.

Les algorithmes pour le suivi de contour par modèle et l'inversion articulatoire fonctionnent sur le même principe : les paramètres de contrôle du modèle de synthèse (acoustique ou visuelle) sont ajustés par itérations successives jusqu'à ce que le résultat de la synthèse coïncide au mieux - par rapport à une distance et un critère d'arrêt - avec l'observation d'une cible sensible (son ou image). Ceci implique la mise en place de trois éléments : le modèle de

synthèse, la mesure de l'adéquation du modèle avec l'observation du signal et la mise à jour des paramètres de contrôle du modèle pour améliorer cette adéquation (Figure 42).

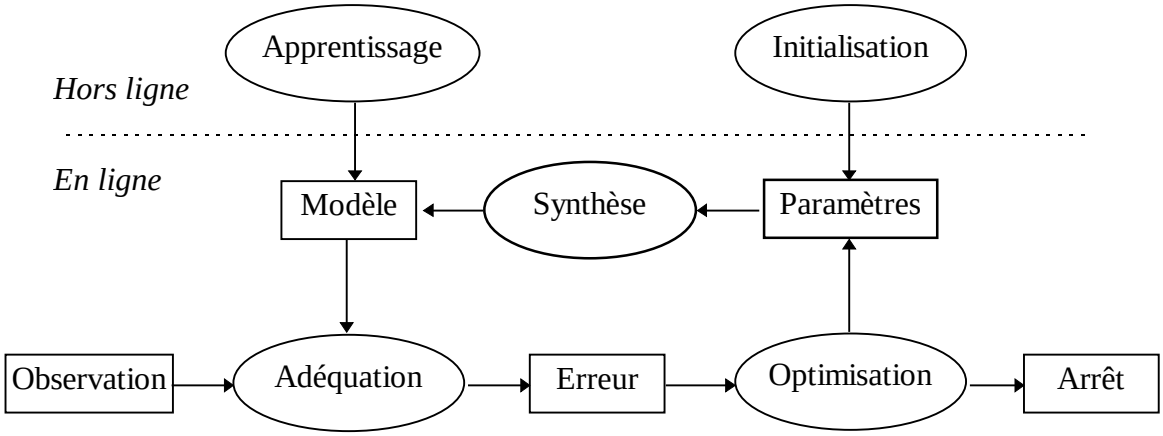


Figure 42. Les étapes pour l'inversion des modèles de synthèse par boucle d'optimisation.

Dans le cas du suivi automatique d'une séquence vidéo, à chaque nouvelle image analysé, les paramètres de modèle sont habituellement positionné dans l'état atteint à l'image précédente pour débuter la boucle d'optimisation.

IV.2 Apprentissage des modèles

Cette phase décrit les conditions pour utiliser le modèle articulatoire d'un locuteur dans une autre prise de vue que celle de son apprentissage initial. Nous décrivons comment générer un nouveau modèle d'analyse de la couleur des lèvres.

IV.2.1 Le modèle 3D du locuteur

La même procédure d'apprentissage a été appliquée à deux locuteurs différents et a abouti pour chacun à un modèle articulatoire contrôlé par trois paramètres. Par la phase de calibration, les mesures géométriques du modèle sont données en millimètres et sont alignées sur un repère fixe par rapport à la tête. Ainsi, pour chaque locuteur, le modèle est réutilisable dans une autre prise de vue, hors de la phase d'apprentissage. Il faut pour cela calibrer la nouvelle prise de vue en suivant une procédure similaire à celle décrite au III.3.5. Le modèle étant en 3D, il présente l'avantage de s'adapter à un angle de vue différent.

La position initiale du modèle peut être déterminée en alignant la projection du modèle dans un état de repos (lèvres fermées) sur deux ou une seule vue. Cette phase d'alignement permet de déterminer la position du point d'origine lié à la tête. Ensuite, soit la tête doit être

maintenue solidaire avec la caméra (enregistrement casque type Labiophone), soit sa position dans l'espace doit être mesurée ou estimée au même titre que les paramètres articulatoires.

IV.2.2 Le modèle de couleur

Nous avons vu au chapitre 2 que la séparation entre vermillon et peau par analyse discriminante nécessite de connaître a priori au moins une segmentation pour collecter les données d'apprentissage (voir II.1.4). Les variations de conditions d'éclairage et éventuellement de systèmes d'acquisition imposent que cet apprentissage soit refait à chaque nouvelle session. La phase initiale d'alignement du modèle décrite précédemment fournit la segmentation a priori. Une zone de peau entourant le vermillon est automatiquement générée en calculant, à partir du modèle, une bande plane d'un centimètre de longueur, tangente à la surface et s'appuyant sur le contour externe. Pour les deux locuteurs étudiés, cette largeur d'un centimètre confère à la classe « peau » un nombre de pixels a peu près équivalents à celui des pixels de la classe « lèvres ».

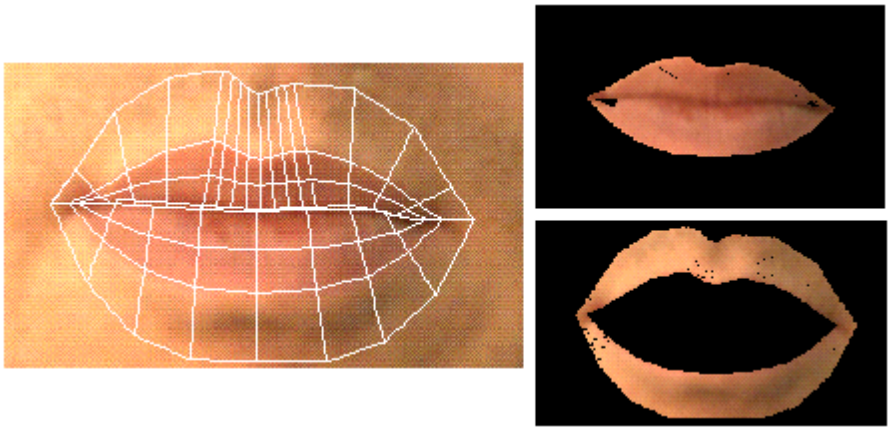


Figure 43. Génération d'une bande de peau autour des lèvres et constitution des classes pour l'apprentissage de la séparation.

IV.3 Inversion optico-articulatoire

IV.3.1 Mesure de l'adéquation entre le modèle et l'image des lèvres

Le traitement de la couleur par analyse discriminante permet de fournir pour chaque pixel RGB une estimation statistique, selon une seule valeur scalaire, de son appartenance aux lèvres ou à la peau. Sous les conditions que nous avons choisies, cette valeur est faible pour des couleurs proches des lèvres et élevée pour des couleurs proches de la peau. Des valeurs moyennes dénotent une incertitude pour les cas où l'affectation à l'une ou l'autre de ces deux

classes est ambiguë (couleur très différente des deux classes ou bien d'égale similitude aux deux classes).

La projection calibrée des points XYZ de la surface paramétrique 3D sur l'image analysée donne, au niveau pixel, « l'ombre » du modèle synthétisé. On mesure alors l'adéquation de cette ombre avec la région des lèvres en sommant les estimations de couleur pour chaque pixel couverts. La surface projetée étant variable, on normalise la somme des estimations par le nombre de pixels considérés. Ce résultat sert alors de mesure d'adéquation entre modèle et image. Plus la valeur est faible, meilleure est l'adaptation du modèle. C'est cette grandeur que l'on cherchera à minimiser.

La contribution des pixels de la bande entourant le contour externe est éventuellement comptée avec un signe opposé. L'utilisation de cette bande de peau vise à lever certaines ambiguïtés pour la détermination de la position du modèle, dues au fait que la couleur de la langue et des gencives se rapproche davantage de celle des lèvres que celle de la peau. Comme ces éléments peuvent apparaître adjacents au vermillon, la mesure d'adéquation ne permet pas de déduire un positionnement correct si on ne prend en compte que la surface des lèvres. En revanche, si la contribution de la peau est ajoutée, elle forcera la bande a quitter la zone du vermillon et stabilisera la limite du modèle entre lèvres et peau au niveau du contour externe.

Dans les tests présentés plus bas, l'ajout de la bande de peau a permis en effet une amélioration de la prédiction depuis la vue de face des paramètres liés à l'étirement lorsque les lèvres sont fermées. En effet, lors d'une occlusion, une forme exagérément arrondie reste couverte par une forme plus étirée. En rajoutant la bande de peau, les parties situées sur les commissures sont alors contraintes à quitter la zone de vermillon. Les figures suivantes montrent sur un exemple de suivi automatique l'amélioration apportée par la bande de peau. La première figure montre l'image originale où le contour externe a été manuellement repéré en trait pointillé. La seconde figure montre le résultat de suivi sans utilisation du contour peau (il a été néanmoins dessiné). Enfin la dernière figure montre le résultat d'un suivi utilisant le contour de peau.

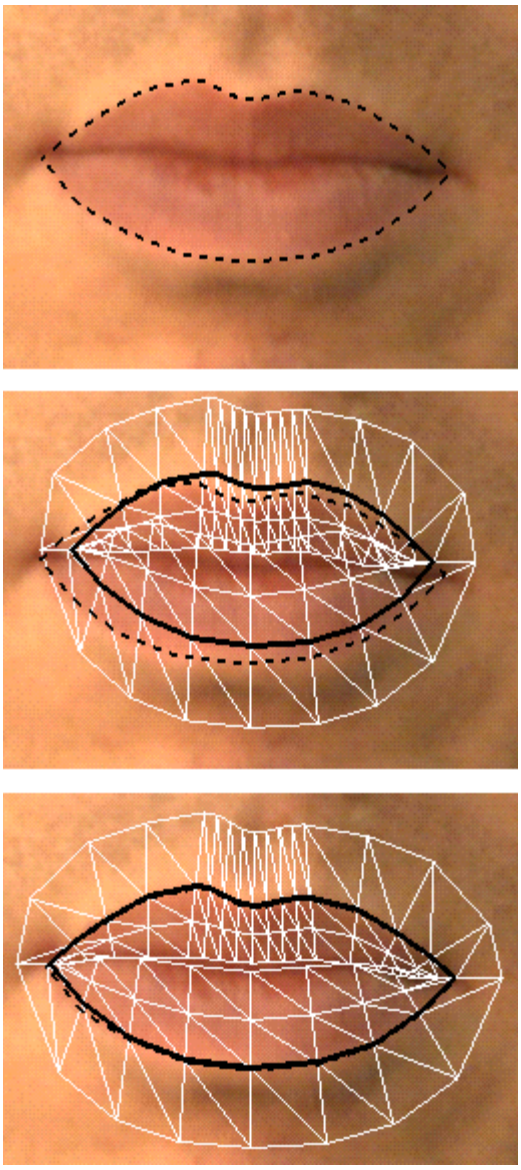


Figure 44. Contribution du contour de peau au suivi automatique à partir d'une vue de face.

Cet exemple montre clairement le rétablissement opéré par la bande de peau pour déterminer l'étirement des lèvres. On observe cependant en même temps que cette technique entraîne une ouverture erronée des lèvres en attirant trop bas la limite entre la peau et la lèvre inférieure. Ce problème, abordé dans la suite des résultats, sera résolu en ne prenant que la moitié supérieure de la bande de peau.

IV.3.2 Optimisation des paramètres articulatoires du modèle

Les problèmes d'optimisation non linéaire visent à trouver le minimum d'une fonction d'une ou plusieurs variables de manière itérative. Un grand nombre de techniques a été proposé et la pertinence de chacune s'accorde en fonction du cas traité et des exigences sur le compromis

entre précision et temps de calcul (« Numerical Recipes in C », Press et al., 1992). Il ressort que le principal problème réside dans le risque de convergence vers des solutions sous-optimale lorsque la fonction à minimiser présente des minima locaux. Ce risque augmente avec le nombre de paramètres de la fonction et nécessite des algorithmes d'optimisation lourds en coût de calcul. Dans notre cas, la mesure d'adéquation à minimiser est fonction des trois paramètres articulatoires seulement. Ce faible nombre d'arguments nous a conduit à nous limiter à une approche simple dichotomique, le risque de minima locaux étant faible. Les paramètres sont déduits par des évaluations successives de l'adéquation du modèle et de l'image. Les détails de l'implantation de l'algorithme sont donnés au dernier chapitre.

IV.4 Résultats et évaluation

Le paradigme d'évaluation du système de labiométrie consiste à comparer les résultats obtenus par la méthode automatique d'inversion optico-articulatoire avec un étiquetage manuel de référence des 30 points de contrôle (face et profil). Pour tout positionnement de ces points, il est possible de déduire, pour les deux locuteurs modélisés au chapitre 3, les trois paramètres articulatoires de la forme par projection du vecteur des 90 coordonnées XYZ sur les trois vecteurs propres associés. Les comparaisons ont d'abord porté sur ces paramètres articulatoires. Par ailleurs, la qualité de prédiction de différents paramètres géométriques importants est évaluée. Ces paramètres sont du type de ceux mesurés à l'ICP par le système de chromakey (ouverture et écartement interne et externe, protrusion).

Trois séries différentes ont été testées : les 23 visèmes du premier locuteur maquillé en bleu (locuteur 1 du chapitre 3, enregistrement type ICP), les 10 visèmes du second locuteur non maquillé et une phrase complète prononcée par ce même locuteur (locuteur 2 du chapitre 3, enregistrement type ATR). Dans le cas des visèmes isolés, le suivi automatique consiste à converger vers le visème en partant de la forme moyenne du modèle articulatoire (paramètres de commande à 0). Le suivi de la séquence met à jour la position des paramètres à chaque image en utilisant les valeurs atteintes à l'image précédente comme valeurs d'initialisation. Le suivi utilise en tout premier lieu les prises de vues de face. Cependant, des tests ont été faits en intégrant à la fois un suivi à partir des vues de face et de profil.

Pour le suivi de face, il n'est pas utile d'interpoler des contours supplémentaire entre les 3 contours de base (interne, externe et intermédiaire). Le suivi de profil nécessite par contre une

plus grande précision, aussi trois contours supplémentaires ont été introduits par interpolation. Dans les deux cas, chaque contour est échantillonné sur 24 points.

Dans tous les tests présentés, la position de la tête du locuteur dans l'espace a été déterminée par une procédure externe (suivi manuel d'un point de référence lié au nez). Tous les enregistrements comportent une vue de face et de profil. Seule la translation 3D de la tête a été prise en compte, les mouvements de rotation par rapport aux caméras ont été négligés. La tête des locuteurs est alors considérée dans l'axe de la caméra de la vue de face.

IV.4.1 Série 1 : Locuteur maquillé - 23 visèmes

Ce locuteur étant maquillé en bleu, il ne présente pas d'ambiguïtés entre couleur du vermillon et intérieur de la bouche. Aussi, les tests sur ce locuteur n'intègrent pas l'estimation de la bande de peau entourant les lèvres dans la mesure de l'adéquation du modèle avec l'image. Les figures suivantes comparent les résultats selon que le suivi se fait avec la vue de face ou avec une combinaison face et profil. Parmi les 23 visèmes, 10 ont servi à construire le modèle du locuteur. Les 13 restant ne figurent pas dans l'apprentissage. Les résultats présentés rassemblent les 23 visèmes comme un seul jeu de test.

IV.4.1.1 Paramètres articulatoires

Le tableau suivant donne les coefficients de corrélation sur la séquence entre les paramètres articulatoires prédis et les paramètres de référence pour les deux techniques de suivi (face, face + profil).

	Arrondissement	Abaissement lèvre inférieure	Relèvement lèvre supérieure
suivi de face	0.76	0.95	0.87
suivi de face + profil	0.99	0.99	0.88

Chaque figure compare, pour un paramètre articulatoire, le résultats de prédiction de ce paramètre par les deux techniques de suivi (face, face + profil). Les 23 visèmes sont présentés dans l'ordre suivant, en gras figurant les visèmes d'apprentissage :

1	**2**	**3**	4	**5**	6	**7**	**8**	9	**10**	11	**12**	13	14	**15**	16	**17**	18	19	20	21	22	23
a	i	y	e	o	$ʒ_y$	œ	**ã**	a_b	$a_ʒ$	i_b	i_z	$i_ʒ$	y_v	$_ab_a$	$_ab_y$	$_yb_y$	$_av_a$	z_y	r_y	l_y	#	=

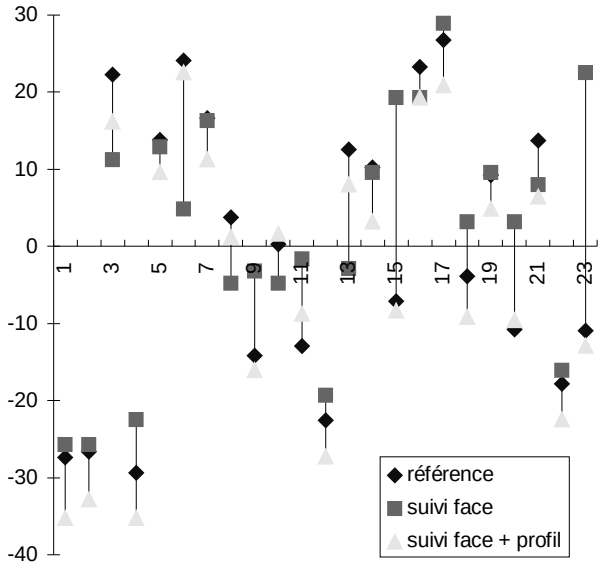


Figure 45. Prédiction du premier paramètre articulatoire pour le locuteur maquillé, $r_{Face} = 0.76$, $r_{Face/Profil} = 0.99$ (coefficient de corrélation sur les 23 visèmes).

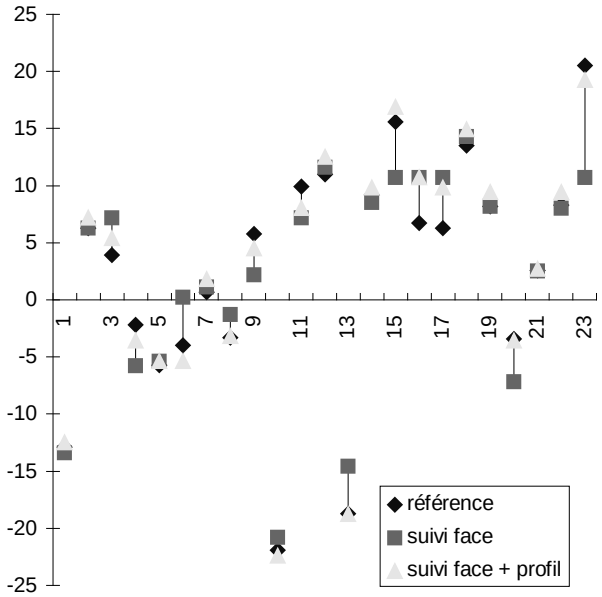


Figure 46. Prédiction du second paramètre articulatoire pour le locuteur maquillé, $r_{Face} = 0.95$, $r_{Face/Profil} = 0.99$.

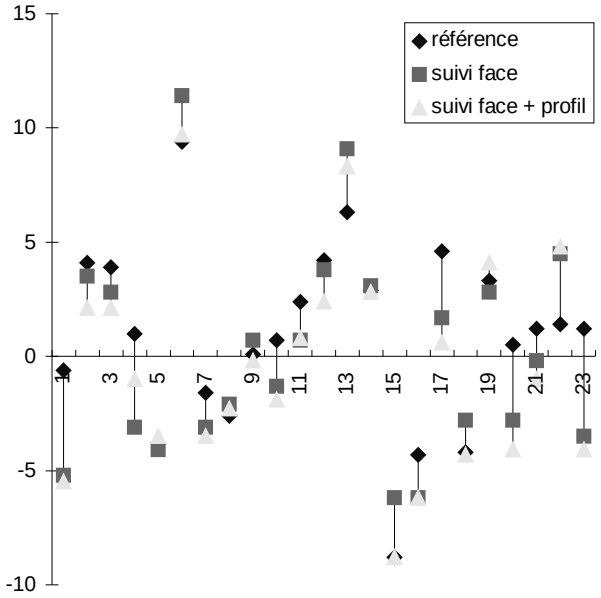


Figure 47. Prédiction du troisième paramètre articulatoire pour le locuteur maquillé, $r_{Face} = 0.87$, $r_{Face/Profil} = 0.88$.

On constate que la vue de profil stabilise la prédiction du premier paramètre et évite des erreurs importantes. Ce paramètre étant lié à l'arrondissement, il est normal d'observer cette amélioration sur certains visèmes protrus ($ʒ_y$[6], l_y[21] et r_y[20]). Lors du suivi de face, des erreur pour la prédiction de ce paramètre apparaissent aussi lors de formes fermées (a_b[9], i_b[11], $_ab_a$[15] et =[23]). Ceci vient du fait que l'algorithme d'optimisation tend à favoriser le premier paramètre pour réaliser une fermeture par un arrondissement excessif. L'aire des lèvres pour une forme très protruse étant couverte par la zone réelle du vermillon, l'optimisation se stabilise dans cette position.

IV.4.1.2 Paramètres géométriques

Les paramètres articulatoires étant définis en 3D, il est possible de donner une estimation des paramètres de profil P1 et P2 même à partir d'un suivi de face uniquement. Les erreurs sont données en millimètres.

Par étiquetage manuel des points de contrôle et projection, on obtient pour toute forme la valeur de ses paramètres articulatoires. Par synthèse à partir de ces paramètres, on génère une forme qui ne correspond pas exactement à la forme initialement étiquetée puisque les trois paramètres réduisent nécessairement l'information géométrique. Les tableaux présentent donc,

en plus des résultats de suivi automatique, l'erreur faite sur les paramètres géométriques lors de la re-synthèse à partir des valeurs de référence des paramètres articulatoires.

Erreur moyenne (mm) - Locuteur maquillé - 23 visèmes

	A'	B'	A	B	P1	P2
re-synthèse	0.7	0.5	2.6	0.8	0.8	0.5
suivi de face	1.1	0.7	2.5	0.7	1.7	1.3
suivi de face / profil	0.7	0.6	3.0	0.5	1.2	0.6

Ecart-type de l'erreur (mm) - Locuteur maquillé - 23 visèmes

	A'	B'	A	B	P1	P2
re-synthèse	0.5	0.4	4.4	0.5	1.9	0.5
suivi de face	1.2	0.6	4.6	0.5	2.2	1.2
suivi de face / profil	0.5	0.5	4.4	0.5	2.1	0.5

Erreur maximum (mm) - Locuteur maquillé - 23 visèmes

	A'	B'	A	B	P1	P2
re-synthèse	1.9	1.8	18.2	1.9	9.2	1.6
suivi de face	5.4	2.0	21.8	2.1	10.0	5.4
suivi de face / profil	1.7	2.4	21.3	1.6	10.3	1.9

La moyenne des corrélations entre la forme étiquetée et la forme atteinte par le suivi automatique pour le vecteur des 90 coordonnées XYZ des 30 points de contrôle sont respectivement :

- re-synthèse, r = 0.93,
- suivi de face, r = 0.72,
- suivi de face et profil, r = 0.90 .

IV.4.2 Série 2 : Locuteur non maquillé - 10 visèmes

De la même manière, le locuteur non maquillé modélisé au chapitre 3, a été testé sur les 10 visèmes ayant servi à construire son modèle. Les mêmes tests ont été effectués : suivi de face et suivi combinant face et profil. De plus, ce locuteur n'étant pas maquillé, est aussi présentée l'évaluation d'un suivi de face incluant l'estimation de la bande de peau autour du vermillon lors d'un suivi de face.

IV.4.2.1 Paramètres articulatoires

Le tableau suivant donne les coefficients de corrélation sur la séquence entre les paramètres articulatoires prédis et les paramètres de référence pour les trois techniques de suivi (face, face + profil, face + peau).

	Arrondissement	Abaissement lèvre inférieure	Relèvement lèvre supérieure
suivi de face	0.90	0.99	0.81
suivi de face + profil	0.88	0.87	0.98
suivi de face + peau	0.97	0.99	0.98

Chaque figure compare, pour un paramètre articulatoire, le résultats de prédiction de ce paramètre par les trois techniques de suivi (face, face + profil, face + peau). Les 10 visèmes sont présentés dans l'ordre suivant :

1	2	3	4	5	6	7	8	9	10
$_{a}b_{a}$	a	$a_{ʒ}$	œ	i	i_{z}	ã	o	$_{y}b_{y}$	y

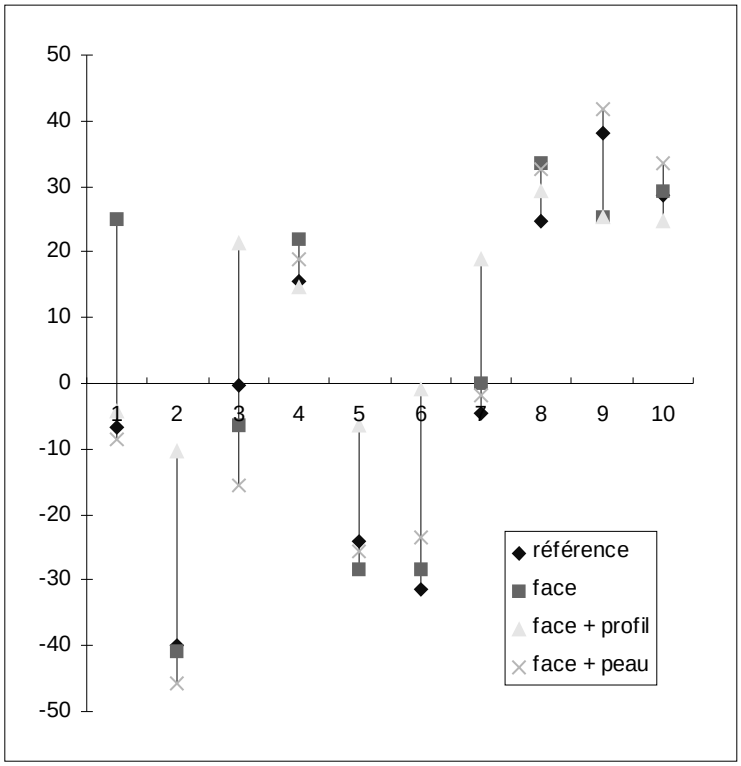


Figure 48. Prédiction du premier paramètre articulatoire pour le locuteur non maquillé, $r_{Face} = 0.90$, $r_{Face/Profil} = 0.88$, $r_{Face/Peau} = 0.97$ (coefficient de corrélation sur les 10 visèmes).

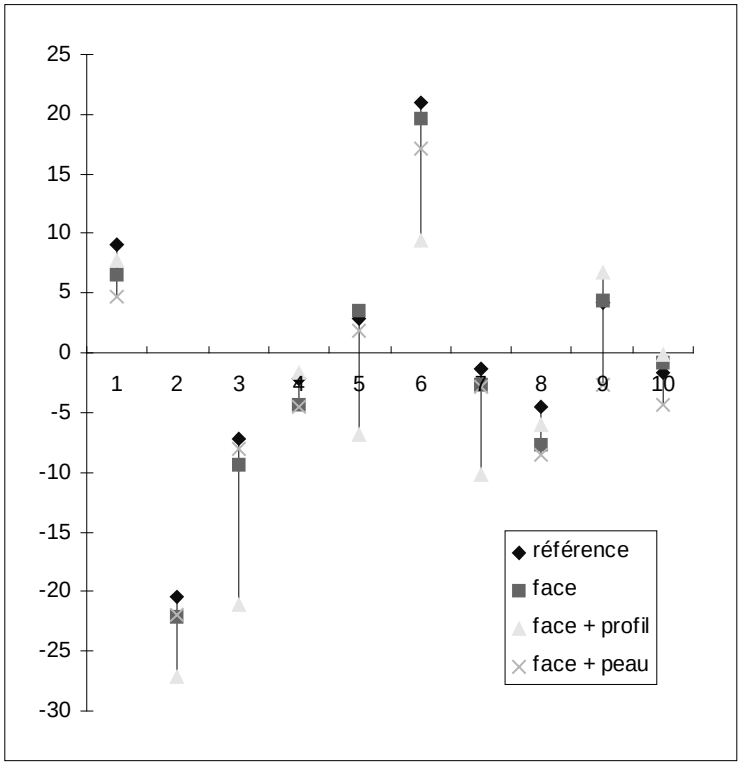


Figure 49. Prédiction du second paramètre articulatoire pour le locuteur non maquillé, $r_{Face} = 0.99$, $r_{Face/Profil} = 0.87$, $r_{Face/Peau} = 0.99$.

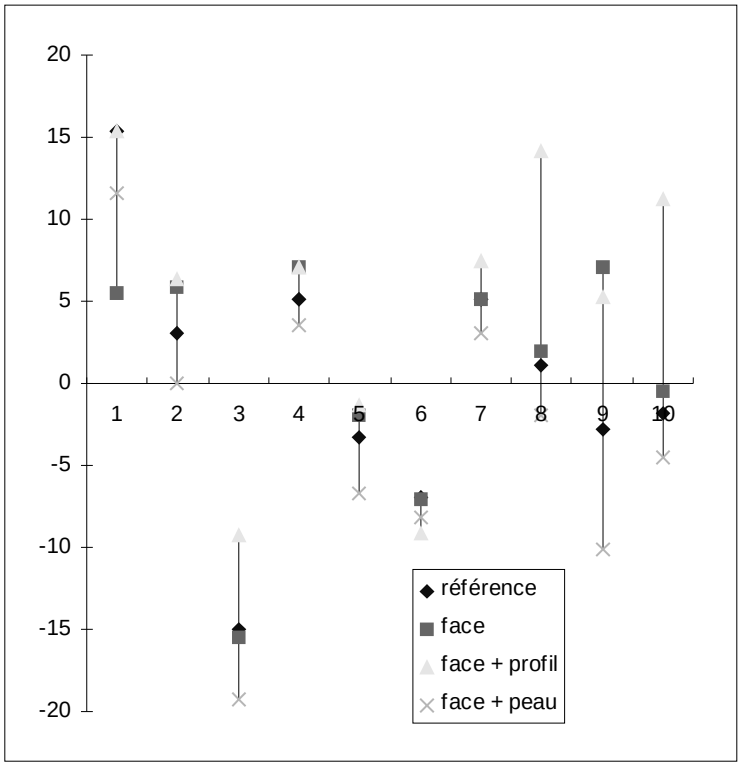


Figure 50. Prédiction du troisième paramètre articulatoire pour le locuteur non maquillé, $r_{Face} = 0.81$, $r_{Face/Profil} = 0.81$, $r_{Face/Peau} = 0.98$.

L'apport du profil, bénéfique dans le cas du locuteur maquillé, s'avère quelque fois très pénalisant pour ce locuteur non maquillé. Le problème vient, d'une part, que dans un visème tel que a_3[3] par exemple, l'épaisseur de vermillon devient extrêmement faible. Le traitement couleur ne fait alors apparaître qu'un très faible nombre de pixels comme appartenant aux lèvres. Par ailleurs, dans certains cas, la peau des joues du locuteur vient recouvrir et obstruer partiellement la vue de profil du vermillon à partir des commissures (visème étiré i_z en particuliers). En l'absence de modèle explicite de cette situation morphologique, l'algorithme d'optimisation cherche à écarter le vermillon de cette zone ayant une forte probabilité d'appartenance à la peau. Dans l'optique de tirer bénéfice d'un suivi combinant deux vues, il apparaît que pour le cas non maquillé, une vue légèrement de trois quart améliorerait sans doute les performances en laissant plus de pixels de lèvres visibles.

En remplacement du profil, la bande de peau stabilise correctement le suivi (les coefficients de corrélation sont rehaussés pour les trois paramètres articulatoires). On observe une légère tendance à sous-estimer le second paramètre articulatoire. Ce problème vient du fait que l'estimation de la couleur déborde légèrement le vermillon sur la lèvre inférieure par rapport à l'étiquetage manuel. Ainsi, la limite lèvre inférieure / peau est légèrement décalée. Bien que numériquement peut significatif, ce décalage peut néanmoins entraîner des conséquences

importantes puisqu'un visème fermé tel que $_{y}b_{y}$ peut être analysé avec les lèvres légèrement entrouvertes. Ce problème est préjudiciable à l'intelligibilité puisque l'occlusion labiale est un trait articulatoire visible très important.

Pour y remédier, seul la limite lèvre supérieure / peau a été conservée. En conservant une même qualité globale de résultats (coefficients de corrélation à 0.95, 0.99 et 0.96), la fermeture sur $_{y}b_{y}$ est en plus préservée. On observe cependant un problème déjà rencontré sur le visème $_{a}b_{a}$: l'occlusion est essentiellement assurée par une surestimation du paramètre d'arrondissement.

IV.4.2.2 Paramètres géométriques

Erreur moyenne (mm) - Locuteur non maquillé - 10 visèmes

	A'	B'	A	B	P1	P2
re-synthèse	0.9	0.6	1.9	0.4	0.5	0.5
suivi de face	1.4	0.9	1.3	0.4	1.1	1.1
suivi de face / profil	3.5	2.0	7.8	1.4	2.0	1.3
suivi de face + peau	1.0	1.7	6.8	1.6	0.8	0.8

Ecart-type de l'erreur (mm) - Locuteur non maquillé - 10 visèmes

	A'	B'	A	B	P1	P2
re-synthèse	0.4	0.4	1.5	0.4	0.4	0.4
suivi de face	1.9	0.7	1.1	0.4	1.4	1.0
suivi de face / profil	2.4	1.4	5.6	1.2	1.4	1.0
suivi de face + peau	0.6	0.5	8.7	1.2	0.5	0.6

Erreur maximum (mm) - Locuteur non maquillé - 10 visèmes

	A’	B’	A	B	P1	P2
re-synthèse	1.3	1.3	4.4	1.2	1.2	1.4
suivi de face	6.8	2.6	3.4	1.1	4.3	3.5
suivi de face / profil	6.9	3.7	15.1	3.1	3.9	3.0
suivi de face + peau	1.8	2.6	28.0	4.2	1.7	2.0

La moyenne des corrélations entre la forme étiquetée et la forme atteinte par le suivi automatique pour le vecteur des 90 coordonnées XYZ des 30 points de contrôle vaut respectivement pour les quatre méthodes :

- re-synthèse, r = 0.95,
- suivi de face, r = 0.82,
- suivi de face et profil, r = 0.68,
- suivi de face et peau, r = 0.91

On constate la faiblesse du score de profil et le score optimal apporté par le suivi face et peau.

IV.4.3 Série 3 : Locuteur non maquillé - phrase

La méthode de suivi automatique a été testée sur une séquence de type ATR où le locuteur non maquillé prononce une phrase issue d’un corpus phonétiquement équilibré :

« Il se garantira du froid avec ce bon capuchon. »

La séquence comporte 90 trames de même parité. L’enregistrement initial provient d’un standard NTSC : les trames sont donc séparées de 33.3 ms. Trois méthode de suivi ont été testées : suivi de face, suivi de face et peau (supérieure et inférieure), suivi de face et peau supérieure uniquement. Les problèmes déjà évoqués dans la section précédente se retrouvent dans la séquence : seule la partie supérieure de la bande de peau est en mesure de garantir la détection des occlusions au cours du suivi automatique de face. C’est la raison pour laquelle nous l’avons adoptée.

IV.4.3.1 Paramètres articulatoires

Le tableau suivant donne les coefficients de corrélation sur la séquence entre les paramètres articulatoires prédis et les paramètres de référence.

	Arrondissement	Abaissement lèvre inférieure	Relèvement lèvre supérieure
suivi de face	0.92	0.96	0.83
suivi de face + peau	0.91	0.97	0.91
suivi de face + peau sup.	0.89	0.98	0.93

Les figures suivantes présentent l'évolution au cours de la séquence de la valeur de référence des paramètres articulatoires (trait pointillé) et la valeur prédite par le suivi de face et peau supérieure (trait plein).

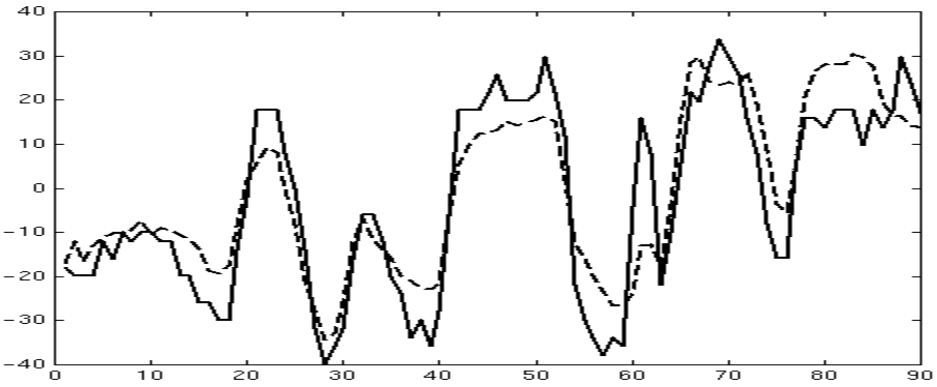


Figure 51. Prédiction du premier paramètre articulatoire (arrondissement), r=0.89.

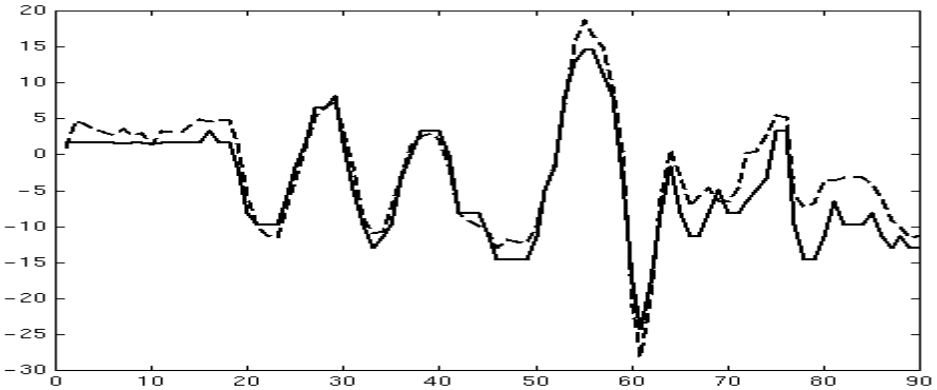


Figure 52. Prédiction du second paramètre articulatoire (relèvement de la lèvre inférieure), r=0.98.

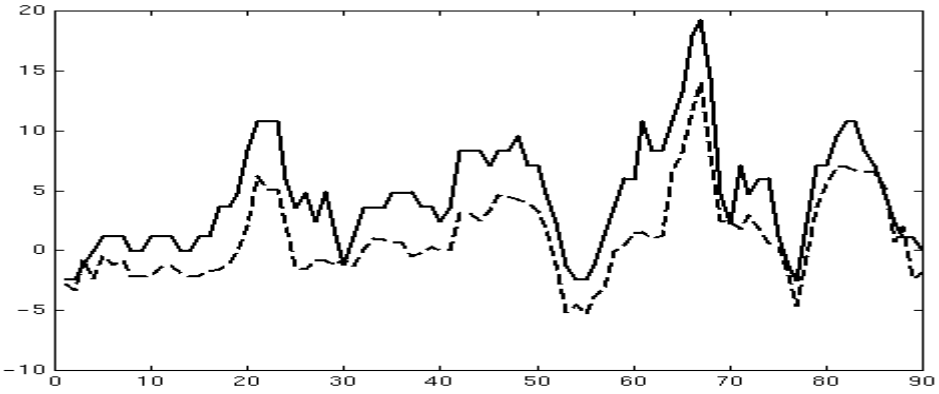


Figure 53. Prédiction du troisième paramètre articulatoire (relèvement de la lèvre supérieure), r=0.93.

IV.4.3.2 Paramètres géométriques

Coefficients de corrélation sur la séquence entre les paramètres géométriques prédis et les paramètres de référence.

	A’	B’	A	B	P1	P2
re-synthèse	0,71	0,95	0,85	0,98	0,93	0,88
suivi de face	0,59	0,96	0,93	0,97	0,77	0,74
suivi de face + peau	0,61	0,98	0,91	0,96	0,79	0,74
suivi de face + peau sup.	0,56	0,97	0,95	0,98	0,74	0,71

Erreur moyenne (mm) de prédiction sur la séquence

	A’	B’	A	B	P1	P2
re-synthèse	1,8	1,2	4,7	0,6	0,5	1,0
suivi de face	2,5	1,2	5,1	0,6	1,1	1,5
suivi de face + peau	2,7	2,4	6,3	1,8	1,0	1,4
suivi de face + peau sup.	2,3	2,0	3,7	1,2	1,3	1,8

Ecart-type de l'erreur (mm) de prédiction sur la séquence

	A'	B'	A	B	P1	P2
re-synthèse	1,2	0,9	5,3	0,5	0,5	0,9
suivi de face	1,5	0,8	5,2	0,6	0,8	1,2
suivi de face + peau	1,4	0,6	4,7	1,0	0,7	0,9
suivi de face + peau sup.	1,8	0,7	3,4	0,7	1,0	1,5

Erreur maximale (mm) de prédiction sur la séquence

	A'	B'	A	B	P1	P2
re-synthèse	5	3,7	33,9	2,2	2	4,2
suivi de face	8,7	3,1	25,6	3	3,2	6
suivi de face + peau	6,2	4	33,9	4,3	2,8	5
suivi de face + peau sup.	7,5	3,5	18,3	2,7	4,3	6,3

Les figures suivantes présentent l'évolution au cours de la séquence de la valeur de référence des paramètres géométriques (trait pointillé) et la valeur prédite par le suivi de face et peau supérieure (trait plein).

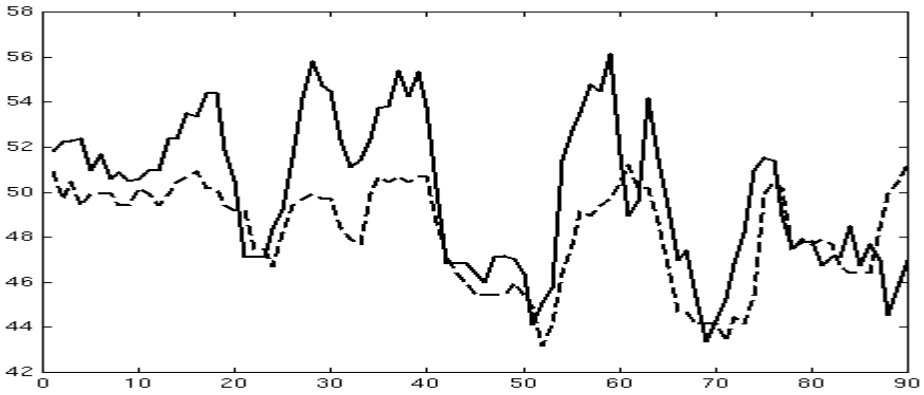


Figure 54. Prédiction du paramètre d'écartement externe A', r=0.56.

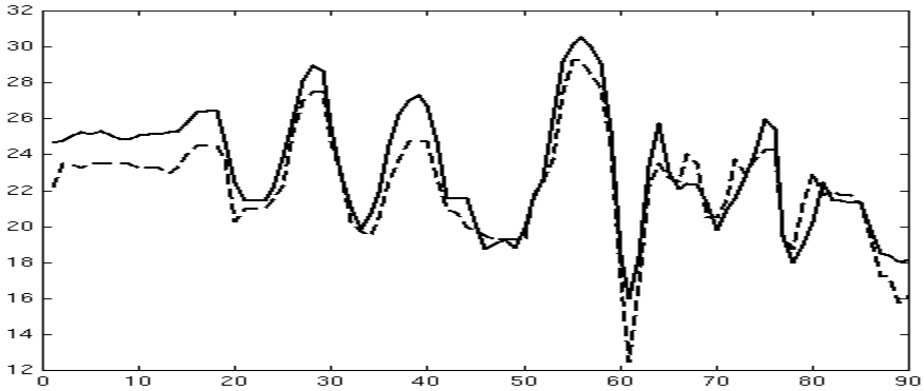


Figure 55. Prédiction du paramètre d'ouverture interne B', r=0.97.

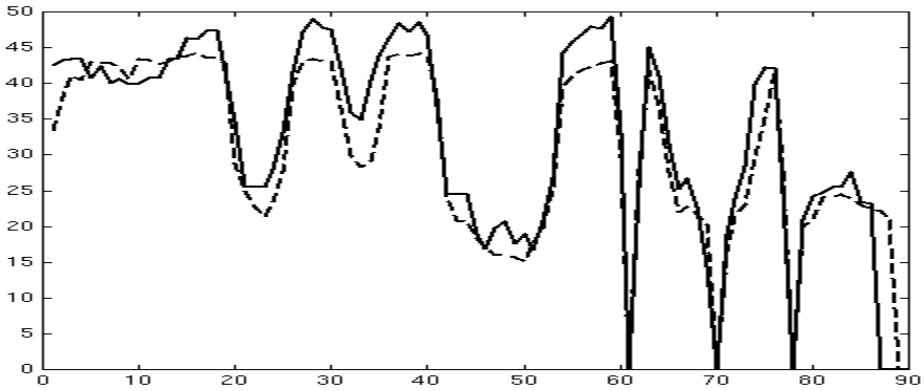


Figure 56. Prédiction du paramètre d'écartement interne A, r=0.95.

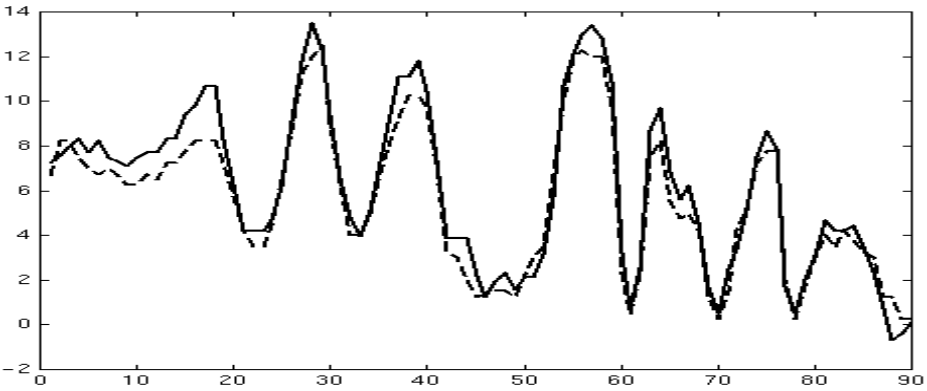


Figure 57. Prédiction du paramètre d'aperture interne B, r=0.98.

IV.5 Discussion

IV.5.1 Bilan

Les résultats montrent une fiabilité des résultats de l'ordre de 2 mm pour les paramètres d'ouverture et de protrusion, et de l'ordre de 4 mm pour les paramètres d'écartement. Au niveau articulatoire, ceci se retrouve par la meilleure prédiction du second paramètre par rapport au premier. On peut attribuer cette différence au fait que :

- d’une part, la couleur des commissures est faiblement marquée, rendant la première phase d’estimation de leur couleur moins fiable que pour la détermination de la séparation entre lèvres et peau aux niveau du plan sagittal où sont mesurés les paramètres d’ouverture,
- d’autre part, le manipulateur se heurte aussi à ce problème lors de l’étiquetage de référence. Ainsi, la valeur de référence pâtit nécessairement d’une stabilité moins importante.

IV.5.2 Insertion d’une étape de prédiction des paramètres

Faisant le lien avec la méthode de prédiction de paramètres à partir d’images en niveaux de gris, cette section présente une amélioration possible du système de suivi automatique. Elle n’a pas été implantée dans le système actuel sur lequel a porté l’évaluation précédente.

Au sein d’une séquence, à chaque nouvelle image, la position du modèle à l’image précédente est prise comme valeur d’initialisation. En cas de brusque changement de forme, ce choix peut pénaliser l’algorithme de recherche : plus la position initiale est loin du but à atteindre, plus cela laisse de place au risque de convergence vers une solution sous-optimale. En II.2, nous avons présenté une méthode de prédiction de paramètres de forme directement à partir d’images en niveaux de gris. Cette approche est moins précise que le suivi de contours mais ne dépend pas d’une estimation initiale. Nous l’appliquons ici à une prédiction initiale des paramètres articulatoires à partir d’une nouvelle image.

IV.5.2.1 Construction du modèle statistique d’images en niveaux de gris

Au II.2, le corpus d’image en niveaux de gris avait été constitué à partir des 23 images transformée en noir et blanc (valeur de luminance) d’un locuteur maquillé en bleu prononçant les 23 visèmes de Benoît. Nous utilisons ici pour le second locuteur non maquillé, le corpus réduit de 10 visèmes qui a servi à l’apprentissage du modèle articulatoire (voir III.4.2). Pour le mouvement de la tête, seules les translations dans le plan de la caméra peuvent être prises en compte. Les rotations sont négligées. Les images des lèvres doivent être alignées sur un point fixe par rapport à la tête. Par une ACP sur les 10 images en niveaux de gris, on obtient alors une base de 9 vecteurs propres images, les « eigenvisemes », propres à la session considérée.

La variation des conditions d’éclairage impose que les 10 images en niveaux de gris soient extraites à nouveau pour chaque nouvelle session. Dans la pratique, on peut envisager d’afficher successivement à l’écran les 10 formes synthétisées. En s’alignant au mieux sur la

forme affichée, le locuteur pourra ainsi extraire dix nouvelles images caractéristiques de la session.

IV.5.2.2 Prédiction des paramètres articulatoires

De la même manière que les paramètres géométriques ont été prédit au II.2.3, les trois paramètres articulatoires pour une image du locuteur sont prédits par régression linéaire à partir des projections de cette image sur les « eigenvisemes ». Pour calculer le modèle de régression linéaire sur les 10 visèmes, il est nécessaire de connaître pour chacun la valeur des trois paramètres articulatoires associés. On obtient ces valeurs pour un visème en projetant ses 90 coordonnées XYZ sur les trois premiers vecteurs propres articulatoires. Le modèle de régression linéaire est ensuite généralisable à toute nouvelle image de la session comme il a été décrit au II.2.3. La valeur prédite des paramètres articulatoires peut servir alors d'initialisation à la boucle d'optimisation décrite ci-dessus.

IV.5.3 Vers une adaptation morphologique automatique

Un tel système de suivi automatique nécessite la détermination hors ligne du modèle articulatoire propre au locuteur. Bien que les paramètres articulatoires dépendent du locuteur, ils ne peuvent que suivre la même tendance que ceux que nous avons extrait pour deux locuteurs : arrondissement, abaissement de la lèvre inférieure et relèvement de la lèvre supérieure. Ces mouvements découlent de la physiologie intrinsèque des muscles labiaux. Des différences de stratégie entre locuteurs et entre langages peuvent faire varier les caractéristiques et l'importance relative de ces paramètres mais sans cependant s'en éloigner complètement. Ainsi, nous pensons que cette approche par apprentissage articulatoire du locuteur est généralisable.

Bien que 10 formes particulières suffisent pour deux locuteurs, le placement manuel des 30 points de contrôle sur chacune représente un travail fastidieux et demande de surcroît une bonne précision de la part du manipulateur. Dans une optique d'utilisation grand public, la détermination du modèle doit donc s'automatiser. Pour cela, nous envisageons trois perspectives de développement pour une extraction automatique.

IV.5.3.1 Apprentissage à partir d'une séquence maquillée

En maquillant les lèvres en bleu sur un corpus d'apprentissage, on s'affranchit du problème de détection. En analysant la forme des lèvres de face et de profil, on peut déduire la position des

points de contrôle. Cette méthode reste tributaire de la qualité du maquillage et court le risque de fournir un modèle du maquillage plutôt que des lèvres.

IV.5.3.2 Séparation des paramètres de parole et de morphologie

De la même manière que les variations articulatoires ont été apprises, les variations morphologiques entre locuteurs peuvent peut-être suivre une paramétrisation similaire. En suivant des contraintes géométriques, nous avons vu que notre modélisation géométrique pouvaient se réduire à 36 degrés de liberté. La nature des contraintes géométriques laisse la possibilité d'une adaptation à n'importe quel locuteur. Ainsi, partant du principe que les paramètres de parole restent au nombre de 3, cela laisse au plus 33 paramètres morphologiques à déterminer. Le travail mené par Le Chevalier (1998) sur plusieurs locuteurs a dégagé 5 paramètres morphologiques en plus des paramètres articulatoires incluant notamment un paramètre de dissymétrie gauche / droite. Certains paramètres extraits se sont révélés cependant très corrélés avec les paramètres articulatoires.

IV.5.3.3 Du modèle de référence au modèle spécifique par convergence locale

En l'absence de maquillage, on peut partir du principe que le modèle actuel représente une référence en matière de locuteur. Appliqué à un autre locuteur, l'inversion du premier modèle articulatoire fournit une première estimation robuste de la position des contours. Cette estimation peut ensuite être affiné par une méthode locale de type « Snakes » pour s'adapter plus fidèlement à la morphologie du second locuteur. Par une série d'apprentissage / test sur les images des 10 formes de base produite par le nouveau locuteur, le processus converge vers l'obtention du modèle articulatoire spécifique de ce locuteur.

V. Chapitre 5. Architecture logicielle et évaluation du timing

Ce dernier chapitre définit les spécifications d'implantation du système, depuis une image RGB issue d'un flot vidéo jusqu'à l'extraction des paramètres articulatoires. La première section identifie les différentes données mise en jeu et les traitements algorithmiques. Le système complet a été implanté et testé en langage C sur une station de travail monoprocesseur. L'exécution est évaluée en nombre d'instructions grâce à un logiciel capable d'analyser le code machine généré par le compilateur du système de la station. Ces résultats permettent d'identifier les zones critiques de traitement pour lesquels des optimisations matérielles sont proposées.

V.1 Flot de données et découpage algorithmique

L'application globale consiste à déterminer les trois paramètres articulatoires d'un locuteur à partir d'une image 2D de ses lèvres. Cette opération est réalisée par l'algorithme d'optimisation présenté au chapitre 4. Il effectue un bouclage de retour sur la chaîne complète de synthèse. La figure suivante présente le synoptique complet des flux de données mis en œuvre par le système de suivi automatique.

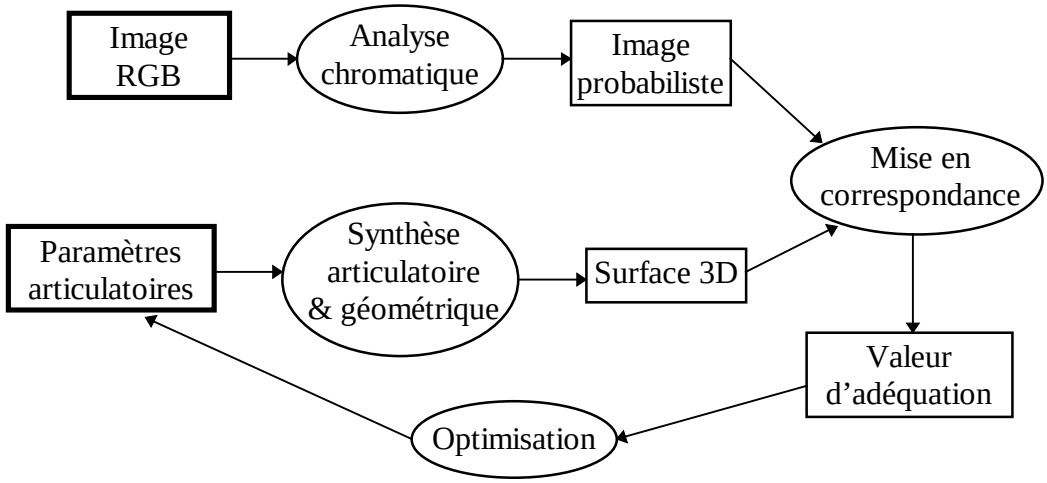


Figure 58. Synoptique des flots de données.

V.1.1 Analyse chromatique

Une première analyse chromatique de l'image RGB fournit une image où à chaque pixel est associé une valeur d'estimation probabiliste d'appartenance aux lèvres ou à la peau (chapitre 2). L'image probabiliste est centrée sur la région des lèvres et correspond à une fenêtre rectangulaire repérée par son origine et sa taille par rapport à l'image RGB.

L'image RGB correspond à un tableau à deux dimensions (I_x, I_y) de pixels RGB contenant chacun trois valeurs chromatiques. Les images RGB sont dimensionnées au format PAL (768x576 pixels). Chaque image provient d'une trame (768x288 pixels) agrandie au format PAL en interpolant linéairement deux lignes consécutives.

L'image probabiliste correspond à un tableau de deux dimensions de (F_x, F_y) pixels contenant une seule valeur. On note (O_x, O_y) l'origine de l'image probabiliste dans l'image RGB. La fenêtre d'analyse mesure 384x288 pixels. Pour une précision de l'ordre de 4 pixels par mm mesurée sur les différentes conditions de test, cette taille correspond à un rectangle de 96x72 mm.

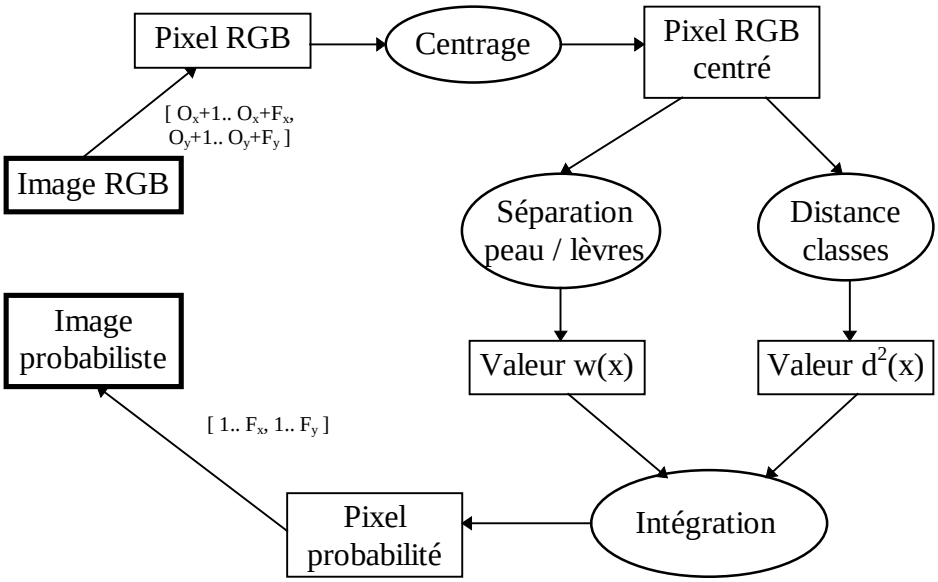


Figure 59. Synoptique de l'analyse chromatique.

Le traitement se fait pixel par pixel sur l'étendue (F_x, F_y) de l'image probabiliste. Chaque pixel RGB $\vec{x}$ est d'abord centré sur la moyenne RGB $\vec{\mu}$ des deux classes d'apprentissage lèvres et peau. Le calcul de la distance de Mahalanobis utilise les trois vecteurs propres RGB $\vec{e}_{i=1..3}$ de la distribution des deux classes. Chaque vecteur propre est divisé hors ligne par la racine carrée de sa valeur propre associée $\lambda_{i=1..3}$. Ainsi, la distance totale $d^2(x)$ est égale à la somme des produits scalaires, élevés au carré, du pixel centré avec chaque vecteur propre pondéré :

$$d^2(\vec{x}) = \sum_{i=1}^{3} \frac{\left((\vec{x}-\vec{\mu})\cdot\vec{e}_i\right)^2}{\lambda_i} = \sum_{i=1}^{3} \left((\vec{x}-\vec{\mu}) \cdot \left(\frac{\vec{e}_i}{\sqrt{\lambda_i}} \right) \right)^2$$

Le calcul de l'analyse discriminante pour la séparation peau / lèvres correspond au produit scalaire $w(x)$ du pixel avec le vecteur de séparation $\vec{w}$ des deux classes lèvres et peau. Les valeurs $w(x)$ et $d^2(x)$ sont intégrées en divisant le résultat du produit scalaire par la distance de Mahalanobis, augmentée de un pour éviter tout problème de division par zéro. Pour chaque pixel RGB $\vec{x}$ on réalise le calcul suivant :

$$w(\vec{x}) = (\vec{x} - \vec{\mu}) \cdot \vec{w}$$
$$p(\vec{x}) = \frac{w(\vec{x})}{1 + d^2(\vec{x})}$$

Le calcul de l'image probabiliste nécessite la connaissance a priori des paramètres suivants qui sont déterminés hors ligne :

- la moyenne RGB des deux classes d'apprentissage lèvres et peau,
- les trois vecteurs propres RGB pondérés de la distribution de la classe lèvre,
- le vecteur RGB de séparation des deux classes.

V.1.2 Synthèse articulatoire et géométrique

Le modèle correspond à une surface paramétrique 3D échantillonnée suivant deux abscisses notées *u* (parcours orthogonal aux contours désignant un contour fermé particuliers, ou *abscisse section*) et *v* (parcours le long d'un contour fermé, ou *abscisse contour*). Les points du modèle sont donc représentés par un tableau à deux dimensions de points 3D définis par leurs coordonnées XYZ. Les paramètres articulatoires sont représentés par trois scalaires.

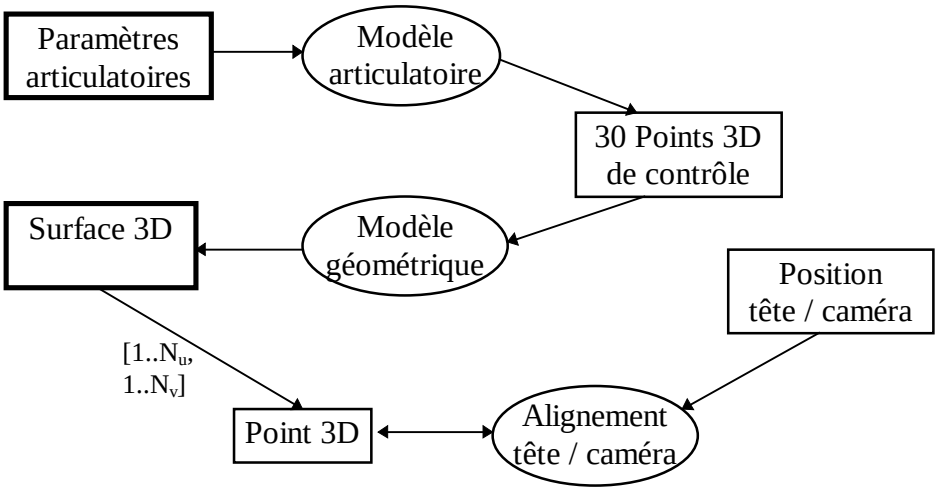


Figure 60. Synoptique de la synthèse du modèle 3D.

Une première phase de synthèse articulatoire détermine à partir des trois paramètres articulatoires $\boldsymbol{a_{i=1..3}}$ les coordonnées XYZ des 30 points 3D de contrôle $\boldsymbol{X_{i=1..30}}$ exprimées en millimètres. La synthèse suit un modèle linéaire défini pour chaque point $\boldsymbol{X_i}$ de contrôle par un point 3D moyen $\boldsymbol{M_i}$ et trois vecteurs propres 3D unitaires $\boldsymbol{V_{i,j=1..3}}$ associés à chaque paramètre articulatoire :

$$X_i = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = M_i + \sum_{j=1}^{3} a_j * V_{i,j} \quad \text{, avec } i = 1..30$$

La moyenne $\boldsymbol{M}$ et les trois vecteurs propres $\boldsymbol{V_{\bullet,j=1..3}}$ sont calculés hors ligne par une ACP sur les 10 formes de bases, chacune représentée par un vecteur à 90 paramètres correspondant aux coordonnées XYZ des 30 points de contrôle (chapitre 3).

Par interpolation polynomiale le modèle géométrique construit la surface 3D passant par les 30 points de contrôle. Le détail des interpolations polynomiales donnant la surface 3D complète est donné au chapitre 3. Elles ont été implantées à partir de l'algorithme d'Aitken (« Numerical Recipes in C», 1992), modifié pour prendre en compte des contraintes de tangentes. Cet algorithme fournit une solution rapide pour évaluer la valeur d'un polynôme d'interpolation de n'importe quel ordre à partir des points interpolés, sans avoir à calculer les coefficients du polynôme.

Nous présentons ici le calcul d'une interpolation cubique en x de trois points $(x_i , y_i)_{i=1..3}$ avec une tangente égale à t en (x_1, y_1) :

$$a(x) = \frac{x_1 - x}{x_1 - x_2}, \quad b(x) = \frac{x_2 - x}{x_2 - x_3}, \quad c(x) = \frac{x_1 - x}{x_1 - x_3}$$

$$P_{12}(x) = y_1 + a(x) * (y_2 - y_1)$$
$$P_{23}(x) = y_2 + b(x) * (y_3 - y_2)$$
$$Q_1(x) = y_1 + t * (x - x_1)$$

$$P_{123}(x) = P_{12}(x) + c(x) * [P_{23}(x) - P_{12}(x)]$$
$$Q_{12}(x) = Q_1(x) + a(x) * [P_{12}(x) - Q_1(x)]$$

$$y(x) = Q_{12}(x) + c(x) * [P_{123}(x) - Q_{12}(x)]$$

En plus des points du modèle, on génère un contour de points de peau d'une largeur 10 mm prolongeant le contour externe tangentiellement à la surface. Ce contour de peau est représenté par un tableau de points 3D. On désigne par $\phi(u, v)$ les points 3D de la surface et $\phi_p(v)$ les points 3D du contour de peau :

$$\Phi_p(v) = \Phi(u_{ext}, v) + \alpha \times \left(\frac{\partial \Phi}{\partial u}\right)(u_{ext}, v)$$

La phase articulatoire est calculée pour une position fixe de la tête (visage dans le plan XY, tourné vers l'axe des Z positif, point origine fixé sur la pointe du nez). Aussi, le mouvement rigide de la tête par rapport à la caméra est reporté sur les points du modèle pour avoir un système de coordonnées aligné avec la caméra. Dans notre cadre, cette donnée est représentée par trois paramètres de translation et elle est considérée fournie par un autre système (suivi du mouvement d'un marqueur sur la pointe du nez). Les coordonnées des points du modèle sont ensuite converties en pixels par un facteur d'échelle ***s*** (en pixels/mm), déterminé hors ligne à partir de la prise de vue d'un objet de dimensions connues. Le plan de la caméra est supposé parallèle avec le plan du visage pendant toute la séquence. Pour tolérer des mouvements de rotation de la tête par rapport à la caméra, ce calcul est à modifier en introduisant une matrice de rotation:

$$X'(t) = s*\left(R_0[R(t)X(t) + T(t)] + T_0\right)$$

en l'absence de paramètres de rotation, on a $R(t) = R_0 = id$, d'où la simplification :

$$X'(t) = s*\left(X(t) + T(t) + T_0\right)$$

Les coordonnées *X* représente un point 3D dans le repère lié à la tête exprimé en mm (origine sur la pointe du nez). Les coordonnées *X'* représente un point 3D dans le repère lié à la caméra exprimé en pixels (origine au centre de l'image). Le paramètre T_0 représente une position initiale du point origine de la tête par rapport à la caméra. Le paramètre $T(t)$ représente l'évolution au cours du temps de la translation de la tête par rapport à cette position initiale où a été mesuré le paramètre T_0.

V.1.3 Mise en correspondance entre l'image et le modèle

Cette opération consiste à accéder à l'ensemble des pixels couverts par l'ombre visible de la surface 3D projetée sur le plan XY de la caméra. L'échantillonnage pour la synthèse de la surface paramétrique 3D peut être ajusté de telle sorte que l'écart entre chaque projection d'un point XYZ coïncide avec la précision d'un pixel de l'image. Cette solution implique, pour chaque pixel du modèle, un grand nombre d'évaluations d'interpolation polynomiale. Les enregistrements vidéo utilisés présentent une précision de l'ordre de 4 pixels par mm. En assimilant les lèvres vues de face à une ellipse de largeur 5 cm et de hauteur 2 cm, le modèle devra alors être échantillonné sur environ 12500 points. Chaque évaluation de polynôme impliquant plusieurs multiplications et divisions par point et sachant que plusieurs évaluations différentes du modèle sont nécessaires pour l'optimisation, avant même toute évaluation, il est évident que cette approche entraînerait un coût de calcul prohibitif.

La surface n'est donc échantillonnée que sur 144 points : 3 contours intermédiaires de 48 points chacun. Cet échantillonnage s'est avéré suffisant pour le suivi automatique à partir d'une vue de face. Le suivi de l'ombre projetée sur une vue de profil nécessiterait davantage de précision. La mesure d'adéquation consiste alors à décomposer l'ombre visible du modèle 3D complet dans le plan XY en un réseau de facettes triangulaires constituées de trois points adjacents de la surface. La visibilité d'un polygone dépend de son orientation par rapport à l'angle de vue de la caméra. On extrait alors de chaque polygone visible la liste des pixels couverts et on somme leur valeur issue de l'image probabiliste, avec un signe négatif pour les pixels de la bande de peau. La somme est pondérée par le nombre de points couverts.

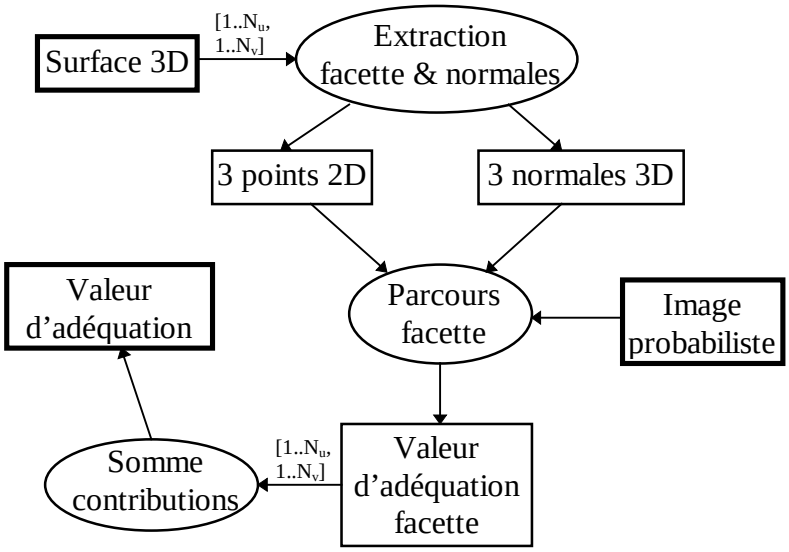


Figure 61. Synoptique de la mise en correspondance du modèle et de l'image.

Pour être comptabilisée, une facette doit être visible. Une facette est décrétée visible si ses trois points sont visibles. Un point de la surface est décrété visible si le produit scalaire de sa normale et de la direction de visée de la caméra est positif. La normale en un point se calcule par le produit vectoriel des deux dérivées partielles de la surface en ce point. Comme les coordonnées de la surface sont alignées avec le plan XY de la caméra, seule le signe de la composante en Z de la normale est à calculer.

Tous les pixels d'une facette sont atteints grâce à une interpolation linéaire du triangle joignant les trois points. La figure suivante illustre les calculs mis en œuvre pour le parcours d'une facette.

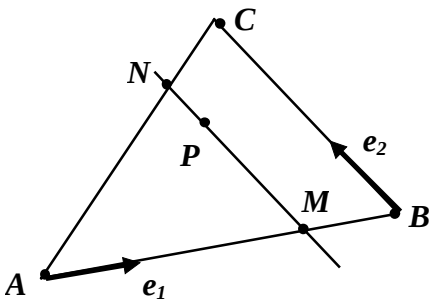


Figure 62. Parcours d'une facette triangulaire 2D.

Soient $\vec{e}_1 = \frac{A\vec{B}}{\|A\vec{B}\|}$ et $\vec{e}_2 = \frac{B\vec{C}}{\|B\vec{C}\|}$ les vecteurs directeurs unitaires des droites (AB) et (BC). Pour tout point M de [AB] et tout point N de [AC] tels que les droites (MN) et (BC) soient parallèles, on a la relation $MN = AM * \frac{BC}{AB}$. On parcourt donc tout le triangle suivant la relation :

$$\vec{P}(i, j) = \vec{A} + i * \vec{e}_1 + j * \vec{e}_2$$

avec i variant de 1 à AB, et j variant de 1 à $(i * \frac{BC}{AB})$, en valeurs entières puisque les coordonnées finales correspondent à des pixels. Les coordonnées 2D du point P permettent ainsi d'accéder au pixel de l'image probabiliste calculée préalablement. Lors du parcours, toutes les valeurs des pixels sont additionnées. Le nombre de points parcourus est aussi comptabilisé. Cette méthode est lourde en calcul mais elle présente l'avantage d'être applicable quelle que soit l'orientation du modèle par rapport au plan de vue de la caméra.

V.1.4 Algorithme d'optimisation pour l'extraction des paramètres articulatoires

Après le calcul d'une nouvelle image probabiliste, une première valeur d'adéquation est déterminée à la position courante des paramètres articulatoires (position atteinte à l'image précédente ou position arbitraire). Pour chaque paramètre, les valeurs d'adéquation sont ensuite comparées aux deux points extrêmes d'une fenêtre d'analyse : à la position courante plus une déviation positive du paramètre, à la position courante plus une déviation négative du paramètre. Si l'une ou l'autre des déviations améliore la valeur d'adéquation, elle est prise comme nouvelle position courante. Si la position courante reste la meilleure des trois, la largeur de la fenêtre de recherche du paramètre autour de la position courante est divisée par deux. On passe ensuite au paramètre suivant. Après mis à jour du dernier paramètre, la recherche est relancée à partir du premier.

La boucle s'arrête lorsque les fenêtres d'analyse des trois paramètres tombent toutes en dessous de 1. Cette valeur correspond à un déplacement vectoriel égal au vecteur propre associé au paramètre articulatoire. Ces vecteurs étant unitaires, le mouvement correspond à un déplacement d'au plus 1 millimètre dans une direction X, Y ou Z. Les valeurs initiales des fenêtres de recherche de chaque paramètre sont calculées en fonction de leur valeur propre respective. Pour chaque paramètre, la valeur propre correspond à la variance (carré de l'écart-type) de ce paramètre pour les 10 formes d'apprentissage. Ainsi, la fenêtre d'analyse d'un paramètre est initialement positionnée égale au tiers de son écart-type sur l'ensemble des 10 formes.

La Figure 63 décrit l'algorithme implanté. La variable *Image* correspond à l'image probabiliste. La fonction *Modèle* combine les calculs du modèle articulatoire et géométrique et l'alignement dans le plan de vue de la caméra. La fonction *Adéquation* détermine la valeur d'adéquation du modèle 3D à l'image courante issue de l'analyse statistique de la couleur.

initialisation de la valeur courante d'adéquation entre Modèle et Image
P_c désigne la position courante des paramètres articulatoires
P_+ et P_- désigne les positions de recherche

initialisation des fenêtres de recherche
$d = \frac{1}{3}\left[\sigma_1 \quad \sigma_2 \quad \sigma_3\right]$;

valeur initiale d'adéquation
e_c = *Adéquation*(*Modèle*(P_c), *Image*) ;

Boucle d'optimisation
répéter,
    stop = Vrai ;
    pour *i* = 1 **à** 3,
        $P_+[i] = P_c[i] + d[i]$;
        e_+ = *Adéquation*(*Modèle*(P_+), *Image*) ;
        $P_-[i] = P_c[i] - d[i]$;
        e_- = *Adéquation*(*Modèle*(P_-), *Image*) ;

        si (($e_c \leq e_+$) **et** ($e_c \leq e_-$)) **alors**,
            $d[i] = 0.5 \cdot d[i]$;
            si ($d[i] > 1$) **alors**,
                stop = Faux ;
            fin_si ;
        sinon_si ($e_- < e_+$) **alors**,
            $P_c[i] = P_+[i] = P_-[i]$;
            $e_c = e_-$;
            stop = Faux ;
        sinon,
            $P_c[i] = P_-[i] = P_+[i]$;
            $e_c = e_+$;
            stop = Faux ;
        fin_si ;
    fin_pour ;
tant_que (*stop* ≠ Vrai) ;

Figure 63. Algorithme d'extraction des paramètres articulatoires.

V.2 Complexité et estimations des temps de calcul

Tous les modules ont été écrits en langage C à la norme ANSI (Drix, 1993), portable vers différentes plates-formes. Les tests ont été effectués sur une station de travail Silicon Graphics Indy, dotée d'un processeur MIPS R4000 cadencé à 150 MHz et d'un coprocesseur pour l'arithmétique en calcul flottant. Cette machine est construite selon une architecture RISC en pipeline. La station fonctionne avec un système de type UNIX fourni par Silicon Graphics

(IRIX, version 5.3). Ce système dispose d'utilitaires de mesures temporelles permettant d'évaluer le coût d'une application sur ce type d'architecture à la fois en nombre de cycles et en nombre d'instructions (programmes *prof* et *pixie*).

Pour des raisons d'ordonnancement des opérations, nombre d'instructions et nombre de cycles ne correspondent pas puisqu'il faut parfois plusieurs cycles pour compléter une instruction dans certaines situations : instructions complexes ou attente de disponibilités d'opérandes par exemple. Certaines instructions liées spécifiquement à l'implantation logicielle sont comptabilisées dans les résultats présentés : appel et retour de procédure, accès à des coefficients stockés en tableaux de variables, branchements de programme. Une implantation plus orientée matériel sera en mesure de limiter ce ralentissement. Par ailleurs, toutes les opérations sont effectuées en virgule flottante. Un passage en virgule fixe nécessite une étude précise de la dynamique des grandeurs utilisées.

Les temps de calcul sont d'abord donnés en nombre d'instructions. Cette première estimation permet de détecter les modules de traitement les plus coûteux à savoir l'analyse chromatique et la mesure d'adéquation entre modèle et image. Ces deux traitements sont influencés de manière importante par la taille de l'image et dans une moindre mesure par l'échantillonnage de la surface 3D. Une évaluation finale estime le gain de temps et les éventuelles pertes de qualité engendrés par une diminution de ces paramètres. Les résultats finaux sont exprimés en nombre d'images par seconde qu'il est alors possible de traiter sur la station.

V.2.1 Analyse chromatique

Pour chaque image, l'analyse chromatique se fait pixel par pixel sur l'étendue $[F_x, F_y]$ de la fenêtre d'analyse. Le temps de calcul est donc directement proportionnel à la taille F_x*F_y de la fenêtre. L'évaluation fait ressortir un temps de :

$$T_1 = 93.F_x.F_y$$

Pour la dimension initiale de 384x288 pixels, le temps de calcul est donc de :

$$T_1 = 10285056 \text{ instr.}$$

V.2.2 Calcul du modèle articulatoire

Le calcul du modèle articulatoire répète le même calcul linéaire sur les 30 points de contrôle. Sauf à changer le nombre de points de contrôle, ce temps est fixe par appel. Pour les 30 points de contrôle, le temps de calcul est donc de :

$$T_2 = 1581 \text{ instr.}$$

V.2.3 Calcul du modèle géométrique

Quatre types d'interpolations polynomiales sont utilisées pour le calcul des points de la surface 3D : linéaire (deux points), parabolique (trois points), deux types de cubique (trois points et une tangente, deux points et deux tangentes). Le nombre d'évaluation de ces interpolations dépend des valeurs d'échantillonnage N_u et N_v choisies pour les deux abscisses de la surface paramétrique (contour et section). Chaque contour réunit, dans les plans XY et XZ, six courbes polynomiales chacune échantillonnée sur N_v' valeurs (N_v=6*N_v'). Le tableau suivant présente le nombre d'instructions pour chaque interpolation et son nombre d'évaluations en fonction de N_u et N_v'. Le calcul du contour de peau dépend de la valeur de N_u. La gestion des boucles pour accéder aux points de la surface, ainsi que les instructions d'appel de procédure, sont comptabilisées dans la rubrique *Boucle*. Elle est traitée comme un appel unique mais elle a été évaluée en fonction de N_u et N_v'.

	Nombre d'instructions par appel	Nombre d'appels
interp. linéaire	13	$2.N_u.N_v'$
interp. parabolique	38	$N_u.(20+4.N_v')$
interp. cubique 1	52	$N_u.(10+10.N_v')$
interp. cubique 2	33	$2.N_u.N_v'$
contour peau	48	$6*N_v'$
Boucle	$650.N_u + 388.N_u.N_v'$	1

Pour le modèle géométrique complet, le temps de calcul est donc de :

$$T_3 = 1930.N_u+288.N_v'+1512.N_u.N_v'$$

Dans le cadre des tests présentées, la surface s'appuie sur 3 contours (N_u=3) de 6*8 points (N_v'=8 et N_v=48). Le temps de calcul est donc dans ce cas de :

$$T_3 = 44382 \text{ instr.}$$

V.2.4 Alignement par rapport de la tête par rapport à une vue d'une caméra

Chacun des $(N_u+1)*N_v$ points de la surface (dont le contour de peau) subit deux rotations et deux translations pour être aligné par rapport à la vue de la caméra. Le tableau suivant donne pour un point le nombre d'instructions nécessaires par point. Par rapport à la rotation / translation liée au mouvement de la tête, l'alignement par rapport à la caméra comptabilise une multiplication pour passer des millimètres aux pixels :

Rot. / Trans. tête	39
Rot. / Trans. caméra	62
Somme pour un point	101

En conservant N'_v comme paramètre ($N_v=N'_v*6$), le temps de calcul pour le modèle géométrique complet, est donc de :

$$T_4 = 606.(N_u+1)*N_v$$

Pour $N_u=3$ et $N_v'=8$, le temps de calcul est donc de :

$$T_4 = 19392 \text{ instr.}$$

V.2.5 Evaluation de l'adéquation entre la projection du modèle et l'image

Deux calculs majeurs sont impliqués à cette étape : le calcul des normales aux points de la surface et le parcours des facettes. Le premier calcul est directement lié au nombre de points N_u*N_v de la surface. Ce n'est pas le cas pour le nombre d'instructions pour le parcours des facettes triangulaires. Quel que soit le nombre de points, et donc de facettes, la surface à parcourir reste sensiblement la même. Le tableau suivant donne un premier résultat pour le calcul des normales par point. La gestion des boucles pour accéder aux points de la surface, la somme des valeurs d'adéquation de chaque facette, ainsi que les instructions d'appel de procédure, sont comptabilisées dans la rubrique *Boucle*.

Normale	60
Boucle	140
Somme pour un point	200

Pour $N_u=3$ et $N_v'=8$, la somme est donc de $200*N_u*(6*N'v) = 28800$ instr.

Le tableau suivant donne pour différentes valeur de N_u et N'_v le nombre *total* d'instructions exécutées pour le parcours de *toutes* les facettes du modèle :

	$N'_v = 2$	$N'_v = 4$	$N'_v = 8$
$N_u = 3$	1867262*	1559868	1832010
$N_u = 6$	2642246	1954880	1998092

*Cas particuliers où le modèle géométrique n'est pas utilisé. Les facettes relient directement les 30 points de contrôle.

Ce résultat illustre bien que l'ordre de grandeur est conservé quels que soient les paramètres N_u et N_v'. De manière évidente, cette étape dépend avant tout de la taille de l'image. Nous évaluons plus bas l'influence de ce paramètre. Pour le cas N_u=3 et N_v'=8, on donne donc le temps de calcul :

$$T_5 = 1860810 \text{ instr.}$$

V.2.6 Algorithme d'optimisation

Le temps de calcul de l'algorithme en soi dépend du nombre d'itérations n_{it} de la boucle d'optimisation. Il est égal à :

$$T_6 = 167 * n_{it}$$

Le nombre d'itérations qu'il engendre détermine le nombre d'évaluations du modèle et de son adéquation avec l'image, opération caractérisée par un temps de calcul égal à $T_2+T_3+T_4+T_5$. Le nombre moyen d'itération sur la séquence est de 8 avec un maximum à 12. Dans le cas des tests sur les visèmes isolés, le nombre d'itérations est plus important puisque, partant de la forme moyenne, le suivi accède aux formes les plus extrêmes. Dans ce cas, le nombre moyen d'itérations est de 10 avec un maximum à 18. On prend le nombre moyen de 10 itérations comme cas moyen dans la suite des calculs. La valeur estimée du temps de calcul est donc :

$$T_6 = 1670 \text{ instr.}$$

V.2.7 Bilan

L'évaluation de l'analyse chromatique est unique par image. Par contre, tous les autres temps de calcul liés au modèle sont à multiplier par le nombre d'itérations. A chaque itération, 6 appels sont effectués pour le calcul du modèle articulatoire, géométrique, son alignement et l'adéquation avec l'image (2 positions de recherche pour 3 paramètres articulatoires). En comptant 10 itérations par image, il faut donc multiplier les temps T_2, T_3, T_4 et T_5 par 60.

Dans la configuration initiale présentée (F_x=384, F_y=288, N_u=3 et N_v'=8), les temps de calculs sont présentées par décroissant de nombre d'instruction par image.

	Nombres d'instructions par appel	Nombres d'appels par image	Nombre d'instruction par image
T_5	1860810	10*6	111648600
T_1	10285056	1	10285056
T_3	44382	10*6	2662920
T_4	19392	10*6	1163520
T_2	1581	10*6	94860
T_6	167	10	1670

Il ressort immédiatement que le facteur le plus limitant est l'évaluation de l'adéquation entre le modèle et l'image T_5. Viennent ensuite l'analyse chromatique T_1 et enfin le calcul du modèle géométrique T_3 dans une moindre mesure.

Les paramètres pouvant modifier les valeurs de temps de calcul sont :

- la taille de l'image probabiliste, F_x et F_y,
- l'échantillonnage de la surface 3D, N_u et N_v'.

Pour simplifier l'étude, nous limitons les modifications à un facteur de réduction d'image ***p*** et un facteur de réduction d'échantillonnage ***n*** d'un contour de la surface - on garde le principe de 3 contours intermédiaires, i.e. N_u=3, davantage de points n'a d'influence que sur le suivi de profil et on se limite ici à un suivi de face. Les réductions se font dans un facteur 2 à partir des dimensions initiales (384x288 pixels et N_v'=8 points d'échantillonnage par portion de contour). Dans ces conditions, les paramètres F_x , F_y et N_v' valent :

$F_x(p) = 384*2^{-p}$ $\quad$ $F_y(p) = 288*2^{-p}$

$N_u = 3$ $\quad$ $N_v'(n) = 8*2^{-n}$

L'influence sur T_1 (analyse chromatique) est directe puisque l'on a une réduction dans un facteur 4^p : $T_1(p) = \frac{T_1(0)}{4^p}$

De même, le calcul du modèle géométrique T3 devient, en fonction de n, égale à :

$$T_3(n) = 5790 + 4824.N'v(n) = 5790 + 38592*2^{-n}$$

Soit en comptant le facteur 60 dû au nombre d'itérations,

$$60*T_3(n) = 347400 + 2315520*2^{-n}$$

Le temps de calcul T_5 d'adéquation image et modèle dépend surtout de la taille de l'image. En effet, si cette dernière est sous échantillonnée par un facteur 2^p, il suffit de parcourir chaque facette tous les 2^p pixels dans les deux directions. Les lèvres n'étant pas rectangulaires, cela ne se traduit cependant pas directement par une réduction de temps de calcul dans un facteur 4^p. L'influence des paramètres a donc été testée par rapport aux performances atteintes sur toute la phrase de test composée de 100 trames consécutives (chapitre 4).

Pour contrôler si les réductions de paramètres entraînent une perte de résultats, les figures suivantes montrent pour la phrase de test l'évolution en fonction des paramètres ***p*** et ***n*** des coefficients de corrélations des paramètres articulatoires et les erreurs moyennes et maximales sur les paramètres géométriques par rapport à l'étiquetage manuel.

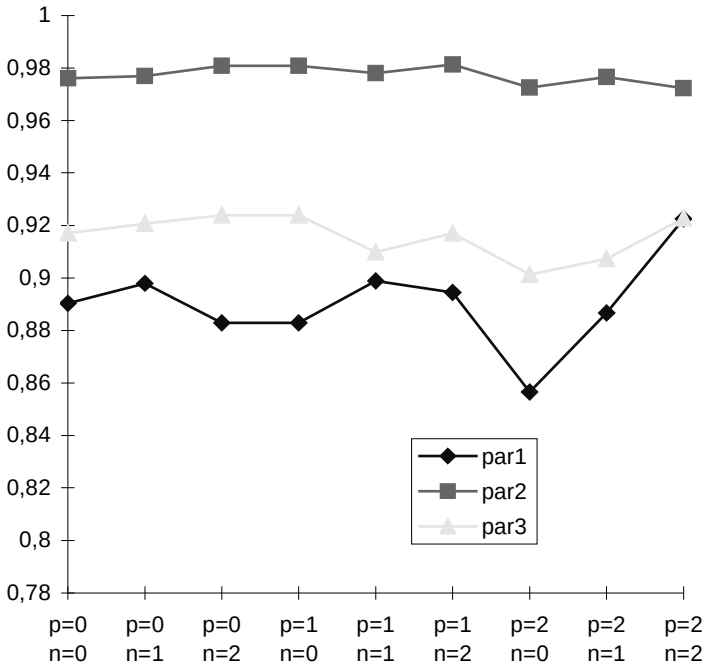


Figure 64. Evolution des coefficients de corrélation des paramètres articulatoires en fonction des paramètres ***p*** et ***n***.

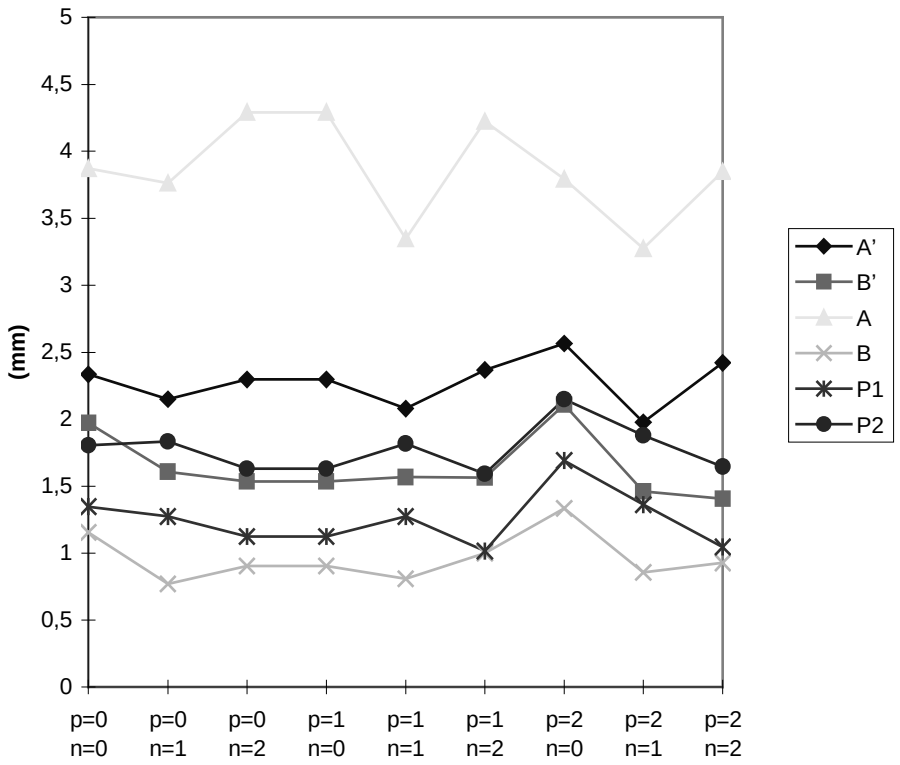


Figure 65. Evolution de l'erreur moyenne en fonction des paramètres ***p*** et ***n***.

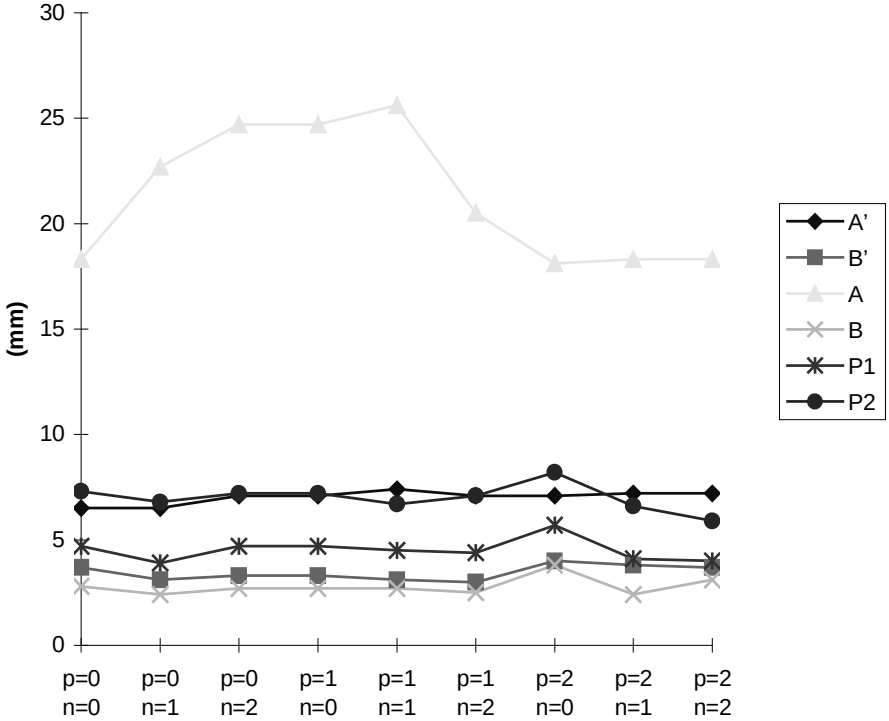


Figure 66. Evolution de l'erreur maximale en fonction des paramètres ***p*** et ***n***.

Ces résultats montrent que, même si des variations sont observées, il n'y a pas de pertes de résultats corrélées avec les deux paramètres ***n*** et ***p***. A noter que les cas où n=2 correspondent à une absence de modèle géométrique : les facettes de la surface 3D relient directement les 30

points de contrôle. Dans la séquence testée, la résolution spatiale est de l'ordre de 4 pixels par millimètres. Le cas n=2 correspond donc à une mesure d'adéquation entre l'image et le modèle sur un échantillonnage de l'ordre du millimètre.

En se basant sur les estimations données en nombres de cycles et en prenant une fréquence de 150MHz, on donne ici les temps de calcul en millisecondes. En plus des temps $T_{couleur}$, $T_{modèle}$ et $T_{adéquation}$ (respectivement T_1, T_3 et T_5), le temps global T_{total} est donné. Ces temps correspondent à la moyenne pour une trame, mesurée sur les 100 trames de la séquence. Sur la courbe de temps total, figure la cadence de traitement atteinte en nombre de trames par seconde.

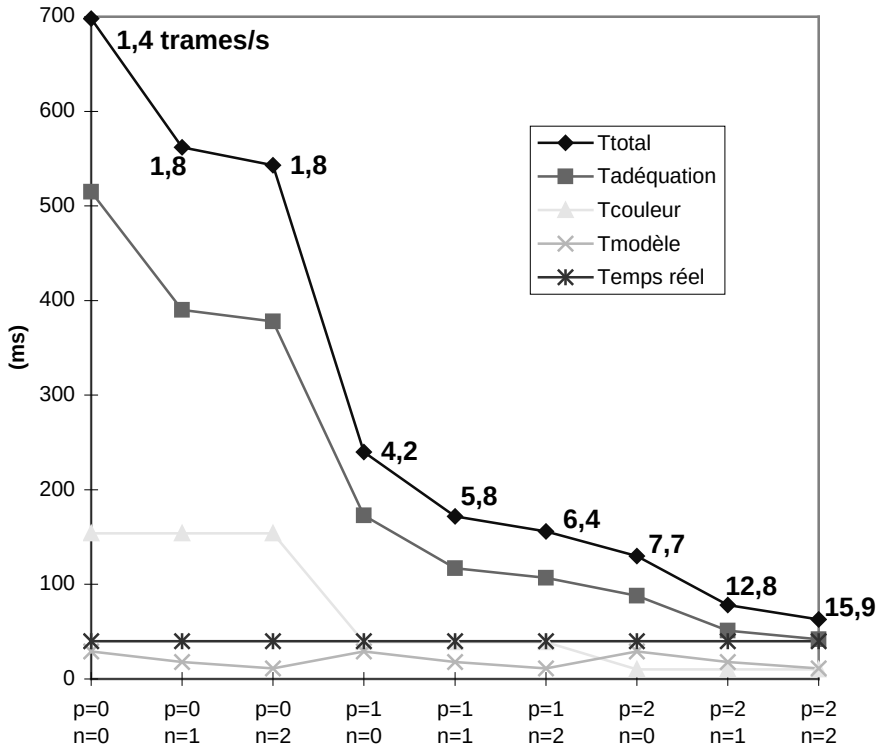


Figure 67. Résultats de l'optimisation temporelle.

V.3 Discussion et optimisations matérielles envisageables

Les deux étapes les plus limitantes que sont l'analyse chromatique et la mesure d'adéquation peuvent être directement mises en parallèle avec un décalage d'une image : après avoir effectué l'analyse chromatique d'une image, l'analyse de l'image suivante peut débuter immédiatement pendant que l'algorithme de suivi automatique est lancé. Cette optimisation est réalisable sur une architecture dotée d'au moins deux processeurs.

En considérant cette mise en parallèle effectuée, la cadence de traitement à p=2 et n=2 atteint 18,9 trames par seconde. A noter que pour les cas n=2, le modèle géométrique n'est plus utilisé. Néanmoins, les appels de procédures au sein de l'implantation restent comptabilisées. Ainsi, pour le cas p=2 et n=2, en supprimant le temps de calcul du modèle géométrique, le système atteint la cadence de 23,8 trames par seconde, soit un peu moins que deux fois le temps réel. La station étant cadencée à 150MHz, il suffit d'utiliser la même implantation sur une machine similaire à 300MHz (performances actuelles des PC) pour atteindre la cadence du temps de réel de 50 trames par seconde.

I. CONCLUSION

La difficulté majeure des systèmes de suivi automatique des gestes labiaux réside dans la capacité à traiter l'immense variabilité des conditions, due à la fois à l'enregistrement vidéo, les différences entre locuteurs et la diversité de postures que permettent les lèvres d'un même locuteur. Les travaux présentés dans cette thèse ont montré qu'il est possible de mettre en œuvre à travers un système de suivi automatique, une modélisation des gestes labiaux propres à un locuteur sur *trois* paramètres seulement. Cette considérable réduction fournit une grande robustesse au système en gardant des résultats satisfaisants. Ceci a pu être obtenu en isolant deux sources de variabilité : la morphologie du locuteur et les mouvements de la tête au cours des enregistrements vidéo. Ainsi, c'est bien la seule variation due à la parole qui a été caractérisée et exploitée à travers le modèle articulatoire contrôlé par trois paramètres. Au delà de l'objectif de suivi automatique, c'est un nouveau modèle 3D de lèvres qui est ainsi disponible. Se basant sur une modélisation géométrique plus large, il a permis d'explorer la gestuelle labiale en parole plus loin que le modèle précédent de l'ICP, en séparant notamment les mouvements des lèvres supérieure et inférieure pour l'ouverture. En gardant la même méthodologie, le modèle d'un locuteur pourra être complété et affiné en augmentant le corpus des visèmes d'apprentissage - entraînant éventuellement l'apparition de paramètres supplémentaires.

Souvent laissée pour compte à cause de son aspect fastidieux, nous avons apporté dans ce travail un intérêt tout particulier à l'évaluation objective de notre système de suivi automatique. Si une « seule » phrase a été testée, elle représente cependant un étiquetage manuel de 100 images, fournissant une référence quantifiable pour évaluer la qualité de prédiction géométrique des formes labiales. La prochaine étape d'évaluation sera donnée par l'intelligibilité audiovisuelle que pourra fournir notre modèle animé par les paramètres extraits du suivi automatique.

Si les restrictions « monolocuteur » et « tête fixe » nous ont permis de dégager un codage efficace du signal visuel de parole, elles limitent cependant la convivialité du système, convivialité que vise la suppression de la contrainte de maquillage. Des pistes ont été évoquées au chapitre 4 pour envisager une adaptation automatique à la morphologie d'un nouveau locuteur en partant d'un des modèles présentés ici - modèle *bootstrap*. La confiance dans la généralisation possible des résultats obtenus vient de l'adéquation entre les

caractéristiques des paramètres extraits des données et les observations bien connues de la phonétique sur l'articulation labiale. La limitation sur la position de la tête du locuteur est dans un premier temps résolue par le casque micro-caméra de la société Ganymedia. Par ailleurs, beaucoup de travaux de recherche proposent des solutions intéressantes sur le suivi automatique des mouvements de la tête à partir de la vidéo. Ces recherches nous ont confortés dans le choix de séparer clairement mouvement de la tête et mouvement de parole.

Le bilan dressé au dernier chapitre sur les mesures de timing ont conduit à des spécifications atteignant la cadence du temps réel (50 trames par seconde) sur une architecture matérielle conventionnelle. L'amélioration des performances des PC en traitement vidéo tend à les rendre équivalents voire supérieurs à la station de travail sur laquelle notre système a été développé. A la conclusion de cette thèse, il apparaît donc que, contrairement à l'hypothèse initiale qui a suggéré ce projet, dans le cadre des méthodes que nous avons choisies, une implantation micro-électronique sur circuit spécialisé ne s'est pas imposée pour traiter le suivi automatique sans maquillage des gestes labiaux en parole.

A l'image de tout mon cursus au sein de l'INPG, cette thèse présente un bilan largement pluridisciplinaire qui dépasse le cadre de sa spécialité. Par sa connexion à la fois à la modélisation articulatoire et à l'analyse/synthèse d'images, elle a conduit à dégager un codage des gestes labiaux en parole, applicable aux technologies multimodales du langage qui se développent actuellement.

I. Références bibliographiques

[Abry et Boë, 1986]
Abry, C., Boë, L.J., « Laws for Lips », Speech Communication, 5:97-104, 1986.

[Adjoudani, 1993]
Adjoudani, A., « Elaboration d'un modèle 3D pour animation en temps réel », Mémoire de DEA Signa-Image-Parole, INPG, 1993.

[Adjoudani et Benoît, 1995]
Adjoudani, A., Benoît, C., « Audio-visual speech recognition compared across two architectures », Proc. of the 4th EUROSPEECH Conference, Madrid, Espagne, pp. 1563-1566, septembre, 1995.

[Badin et al., 1995]
Badin P., Gabioud B., Beautemps D., Lallouache T.M., Bailly G., Maeda S., Zerling J.P., Brock G., « Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model », in Proc. of the 15th International Conference on Acoustics, volume IV, pp. 349-352, Trondheim, Norvège, Juin 1995.

[Baron et al., 1992]
Barron, J.L., Fleet, D.J., Beauchemin, S.S., « Performance of Optical Flow Techniques », Rapport technique RPL-TR-9107, Robotics and Perception Laboratory, Queen's University Kingston, Ontario, Canada, 1992.

[Basu et Pentland, 1997]
Basu S. et Pentland A., « A Three-Dimensional Model of Human Lip Motions Trained from Video » in Proc.of the IEEE Non-Rigid and Articulated Motion Workshop at CVPR'97, San Juan, Puerto Rico, Juin 16, 1997 (MediaLab TR#441).

[Basu et al., 1998]
Basu S., Oliver N., et Pentland A., « 3D Modeling and Tracking of Human Lip Motions », Proc. of ICCV'98, Bombay, India, Jan. 4-7, 1998 (MediaLab TR#442).

[Benoît et al., 1992]
Benoît C., Lallouache M.T., Abry C., « A set of French visemes for visual speech synthesis », pp. 485-504, Talking Machines : Theories, Models and Designs, G. Bailly and C. Benoît (eds.), Elsevier Science Publishers, 1992.

[Benoît et al., 1996]
Benoît C., Guiard-Marigny T., Le Goff B., Adjoudani A., « Which Components of the Face Do Humans and Machines Best Speechread ? », in Speechreading by Humans and Machines, D. Stork and M. Hennecke (eds.), Springer-Verlag, Berlin, pp. 351-372, 1996.

[Burnham et Dodd, 1996]
Burnham, D., Dodd, B., « Auditory-visual speech perception as a direct process : the McGurk effect in infants and across languages », Speechreading by Humans and Machines, Stork et Hennecke (eds.), Springer-Verlag, Berlin, pp. 103-114, 1996.

[Binnie et al., 1974]
Binnie C.A., Montgomery A.A., Jackson P.L., « Auditory and visual contributions to the perception of consonants », Journal of Speech & Hearing Research, 17, pp. 619-630, 1974.

[Bouchet et Cuilleret, 1972]
Bouchet, A., Cuilleret, J., « Anatomie topographique descriptive et fonctionelle », Villeurbanne, Simep éditions, 1972.

[Bregler et al., 1993]
Bregler, C., Hild, H., Manke, S., Waibel, A., « Improving connected letter recognition by lipreading », Proc of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, IEEE, 1 :557-560, 1993.

[Bregler et Konig, 1994]
Bregler C., Konig Y., « Eigenlips for Robust Speech Recognition », Proceedings IEEE - ICCASSP, pp. 669-672, 1994.

[Bregler et Omohundro, 1995]
Bregler C., Omohundro S.M., « Nonlinear manifold learning for visual speech recognition », in IEEE International Conference on Computer Vision, pp. 494-499, Piscatawayt, NJ, USA, 1995.

[Brooke et Scott, 1994]
Brooke N.M., Scott S.D., « PCA Image Coding Schemes and Visual Speech Intelligibility », Proc. of the Institute of Acoustics, Autumn Meeting, Windermere, UK, 16(5), pp. 123-129, 1994.

[Cathiard, 1989]
Cathiard, M.A., « La perception visuelle de la parole : aperçu des connaissances », Bulletin de l'Institut de Phonétique de Grenoble, 18 :109-193, 1989.

[Cohen et Massaro, 1993]
Cohen, M.M., Massaro, D.W., « Modeling Coarticulation in Synthetic Visual Speech », in Models and Techniques in Computer Animation, N.M. Thalmann & D. Thalmann (eds.), Springer-Verlag, pp. 139-156, 1993.

[Cootes et Taylor, 1992]
Cootes T.F., Taylor C.J., « Active Shape Models - 'Smart Snakes' », British Machine Vision Conf., pp. 266-275, Leeds, UK, 1992.

[Cosi et al., 1996]
Cosi, P., Dugatto, M., Ferrero, F., Magno Caldognetto, E., Vagges, K., « Phonetic recognition by recurrent neural networks working on audio and visual information », Speech Communication, 19, pp. 245-252, 1996.

[Dodd et Campbell, 1987]
Dodd, B., Campbell, R. (eds.), Hearing by Eye: The Psychology of Lipreading, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[Drix, 1993]
Drix, P., « Langage C, norme ANSI », 2e édition, Masson, Paris, 1993.

[Erber, 1969]
Erber N.P., « Interaction of audition and vision in the recognition of oral speech stimuli », Journal of Speech & Hearing Research, 12, pp. 423-425, 1969.

[Erber, 1975]
Erber N.P., « Auditory-visual perception of speech », Journal of Speech and Hearing Disorders, 40, pp. 481-492, 1975.

[Escudier et al., 1990]
Escudier, P., Benoît, C., Lallouache, M.T., « Identification visuelle de stimuli associés à l'opposition /i/ - /y/ : étude statistique », Proceedings of the First French Conference on Acoustics, Lyon, France, 541-544, 1990.

[Essa et al., 1994]
Essa, I., Darell, T., Pentland, A., « Tracking Facial Motion », Proc. of the Workshop on Motion and Non-rigid Articulated Objects, pp. 36-42, 1994.

[Fant, 1973]
Fant G., « Speech Sounds and Features », M.I.T. Press, Cambridge, USA, 1973.

[Finn, 1986]
Finn K., « An investigation of Visible Lip Information to be used in Automatic Speech Recognition », PhD dissertation, Georgetown University, Washington, D.C., 1986.

[Fisher, 1968]
Fisher C.G., « Confusion among visually perceived consonants », Journal of of Speech and Hearing Research, 15, pp. 474-482.

[Fromkin, 1964]
Fromkin, V., « Lip positions in American-English vowels », Language and Speech, 7(3):215-225, 1964.

[Gabioud, 1994]
Gabioud B., « Aritucaltory Models in Speech Synthesis », in Fundamentals of Speech Synthesis and Speech Recognition, Eric Keller (eds.), Willey, pp. 215-230, 1994.

[Garcia et Vatikiotis-Bateson, 1997]
Garcia, F., Vatikiotis-Bateson, E., « Active Shape Model for Lip Tracking », ATR Symposium on Face and Object Recognition, 20-23 Jan., Kyoto, Japan, pp. 58-59, 1997.

[Guiard et al., 1996]
Guiard-Marigny T., Adjoudani A., Benoît C., « 3D models of the lips and jaw for visual speech synthesis », in J.P.H. van Stanten, R. W. Sproat, J.-P. Olive and J. Hirschberg, eds., Progress in speech synthesis, Springer-Verlag, New York, 1996.

[Guiard, 1996]
Guiard-Marigny T., « Modélisation Tridimensionnelle des Articulateurs de la Parole », Thèse de Doctorat de l'INPG, nov. 1996.

[Hardcastle, 1976]
Hardcastle, W.J., « Physiology of Speech Production », Academic Press, Londres, 1976.

[Hennecke et al., 1994]
Hennecke M.E., Prasad K.V., Stork D.G., « Using deformable templates to infer visual speech dynamics », in 28th Annual Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov., 1994.

[Hunt, 1989]
Hunt R.W.G., « Measuring Color », Ellis Horwood series in applied science and industrial technology, Halsted Press, New York, 1989.

[Kant, 1787]
Kant, E., « Critique de la Raison Pure », Presses Universitaires de France, 11ème edition, 1944, édition originale, 1787.

[Kaburagi et Honda, 1994]
Kaburagi T., Honda M., « Determination of sagittal tongue shape from the positions of points on the tongue surface », Journal of Acoustic Society of America, Vol. 96, No 3, pp. 1356-1366, 1994.

[Kass et al., 1993]
Kass M., Witkin A.,. Terzopoulos D., « Snakes : Active Contour Models », International Journal of Computer Vision, 15:11, pp. 321-331, 1993.

[Kaucic et al., 1996]
Kaucic R., Dalton B., Blake A., « Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications », Proc. European Conf. on Computer Vision, pp. 376-387, Cambridge, UK, 1996.

[Kuhl et Meltzoff, 1982]
Kuhl, P.K., Meltzoff, A.N., « The bimodal perception of speech in infancy », Science, 218, pp. 1138-1141, 1982.

[Kuratate et al., 1998]
Kuratate, T., Yehia, H., Vatikiotis-Bateson, E., « Kinematics-based synthesis of realistic talking faces », Proc. of the International Conference on Auditory-Visual Speech Processing, AVSP'98, pp. 185-190, Terrigal, Australie, 4-7 decembre, 1998.

[Labialité et Phonétique, 1980]
Abry C., Boë L.-J., Corsi P., Descout R., Gentil M., Graillot P., « Labialité et Phonétique », publications de l'Université des langues et lettre de Grenoble, 1980.

[Ladefoged, 1979]
Ladefoged P., « Articulatory parameters », W.P.P. 45, U.C.L.A., pp. 25-31.

[Lallouache, 1991]
Lallouache M.T., « Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres », PhD. dissertation, INPG, Grenoble, France, 1991.

[Lavagetto, 1995]
Lavagetto F., « Converting Speech into Lip Movements : A Multimedia Telephone for Hard of Hearing People », IEEE Transactions on Rehabilitation Engineering, Vol. 3:1, 1995.

[Le Chevalier, 1998]
Le Chevallier, L., « Un modèle 3D paramètrique de lèvres », Mémoire de DEA Robotique-Image-Vision, INPG, juillet 1998.

[Le Goff, 1997]
Le Goff B., « Synthèse à partir du texte de visage 3D parlant français », Thèse de doctorat, INPG, oct. 1997.

[Liévin et Luthon, 1998]
Liévin M., Luthon F., « Lip features automatic extraction », in Proc. of the IEEE Conference on Image Processing, Chicago, USA, oct., 1998.

[Lindblom et Sundberg, 1971]
Lindblom B., Sundberg J., « Acoustical consequences of lip, tongue, jaw, and larynx movement », Journal of the Acoustical Society of America, 59:1166-1179, 1971.

[Luetin et al., 1996]
Luettin J., Thacker N.A., Beet S.W., « Speechreading using shape ans intensity information », in Proc. of the 4th International Conference on Spoken Language Processing, vol. 1, pp. 58-61, 1996.

[Luetin et Thacker, 1997]
Luettin J., Thacker N.A., « Speechreading using probabilistic models », Computer Vision and Image Understanding, 65(2):163-178, Fev. 1997.

[Luetin, 1997]
Luettin J., « Visual Speech And Speaker Recognition », PhD dissertation, Université de Sheffield, mai 1997.

[M2VTS, 1997]
Pigeon S., Vandendorpe L., « The M2VTS multimodal face database », in Proc. of the 1st International Conference on Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science, Springer Verlag, 1997.

[Maeda, 1979]
Maeda S., 1979, « An articulatory model based on statistical analysis », Journal of the Acoustical Society of America, 65, 1979.

[Mak et Allen, 1994]
Mak M.W., Allen W.G., « Lip-motion analysis for speech segmentation in noise », Speech Communication, 14(3):279-296, 1994.

[Mase et Pentland, 1991]
Mase K., Pentland A., « Automatic Lipreading by Optical-Flow Analysis », Systems and Computers in Japan, vol. 22, No. 6, 1991.

[Massaro, 1987]
Massaro, D.W., 1989, « Multiple book review of Speech perception by ear and eye », Behavioral and Brain Sciences, 12, 741-794.

[McGurk et McDonald, 1976]
McGurk H., McDonald J., « Hearing Lips and Seeing Voices », Nature, 264:746-748, 1976.

[Mermelstein, 1973]
Mermelstein P., « An articulatory model for the study of speech production », Journal of the Acoustic Society of America, Vol. 53, pp. 1070-1082, 1973.

[Motoki et al, 1994]
Motoki K., Badin P., Miki N., « Measurement of acoustic impedance density distribution in the near field of the labial horn », Proceedings of the 3rd International Conference on Spoken Language Processing, vol. 2, Yokohama, Japon, Septembre 1994.

[MPEG4, 1999]
MPEG-4, version 2, description de la norme sur le serveur Web du groupe MPEG (Moving Picture Experts Group) : http://drogo.cselt.stet.it/mpeg

[Murase et Sakai, 1996]
Murase, H., Sakai, R., « Moving object recognition in eigenspaces representation : Gait analysis and lip reading », Pattern Recognition Letters, 17(2), pp. 155-162, 1996.

[Neely, 1956]
Neely K.K., « Effect of visual factors on the intelligibility of speech », Journal of the Acoustical Society of America, 28, pp. 1275-1277, 1956.

[« Numerical Recipes in C », Press et al., 1992]
Press, W.H., Teukolsky, S.A., Vetterling W.T., Flannery B.P., « Numerical Recipes in C », Cambridge University Press, 1992.

[Öhman, 1966]
Öhman S.E., « Coarticulation in VCV utterances : Spectrographic measurements », Journal of the Acosutical Society of America, 41, pp. 310-320, 1966.

[Oliver et al., 1997]
Oliver N., Pentland A.S., Bérard F., Coutaz J., « LAFTER: Lips and Face Tracker », Computer Vision and Pattern Recognition Conference '97, 1997. (MediaLab TR#396).

[OpenInventor, 1994]
Wernecke J., « The Inventor Mentor », Open Inventor Architecture Group, Addison-Wesley Publishing Company, nov. 1994.

[Payan, 1994]
Payan Y., « Modèle biomécanique et contrôle de la langue en parole », Thèse de Doctorat, INPG, déc. 1996.

[Petajan, 1984]
Petajan E., « Automatic lipreading to enhance speech recognition », PhD. dissertation, Univ. Illinois at Urbana-Champagne, 1984.

[Petajan et Graf, 1996]
Petajan E., Graf H.P., « Robust Face Feature Analysis for Automatic Speachreading and Character Animation », in Speechreading by Man and Machine, D. Stork and M. Hennecke (eds.), Springer-Verlag, Berlin, pp. 425-436, 1996.

[Prasad et al., 1993]
Prasad K.V., Stork D.G., Wolff G., « Preprocessing video images for neural learning of lipreading », Ricoh California Research Center, Technical Report, CRC-TR-93-26, 1993.

[Rao et Mersereau, 1994]
Rao R.R., Mersereau R.M., « Lip modeling for visual speech recognition », in 28th Annual Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov., 1994.

[Reisberg et al., 1987]
Reisberg, D., McLean, J., Goldfield, A., « Easy to hear but hard to understand : A lip-reading advantage with intact auditory stimuli », in Hearing by Eye : the psychology of lip-reading, B. Dodd et R. Campbell (eds.), Lawrence Erlbaum Associates, Hillsdale, New Jersey, 97-114.

[Revéret et Benoît, 1997]
Revéret L., Benoît C., « Lip parameter extraction based on projection of raw images onto référence shapes », Proc. of the 1st IEEE Workshop on Multimedia Signal Processing, Princeton, NJ, USA, 23-25 June, 1997.

[Revéret, 1997]
Revéret L., « From raw images of the lips to articulatory parameters : A Viseme-based Prediction », Proc. of the Fifth EUROSPEECH Conference, Rhodes, Greece, Sept. 22-25, 1997, vol. 4, pp. 2011-2014.

[Revéret et al., 1997]
Revéret L., Garcia F., Benoît C., Vatikiotis-Bateson, E., « An hybrid image processing approach to lip tracking independent of head orientation», Proc. of the Fifth EUROSPEECH Conference, Rhodes, Greece, Sept. 22-25, 1997, vol. 3, pp. 1663-1666.

[Revéret et Benoît, 1998]
Revéret L., Benoît C., « A new 3D lip model for analysis and synthesis of lip motion in speech production », International Conference on Auditory-Visual Speech Processing, Terrigal-Sydney, Australie, pp. 207-212, Déc. 4-7, 1998.

[Rogozan et al., 1996]
Rogozan, A., Deléglise, P., Alissali, M., « Intégration asynchrone des informations auditives et visuelles dans un système de reconnaissance de la parole », Actes des 21èmes Journées d'Etudes sur la Parole, Avignon, pp. 359-362, 1996.

[Rouvière et Delmas, 1973]
Rouvière, H., Delmas, A., « Anatomie humaine descriptive, topographique et fonctionnelle », Tome 1, Tête et cou, 11ème édition, 1973.

[Rubin et Vatikiotis-Bateson, 1998]
Rubin P., Vatikiotis-Bateson, E., « Talking Heads », International Conference on Auditory-Visual Speech Processing, Terrigal-Sydney, Australie, pp. 231-236, Déc. 4-7, 1998. (http://www.haskins.yale.edu/haskins/head.html).

[Stork et al., 1992]
Stork, D.G., Wolf, G.J., Levine, E.P., « Neural network lipreading system for improved speech recognition », Proc. of the International Joint Conference on Neural Networks, Baltimore, USA, IEEE, 2 :289-295, 1992.

[Sumby et Pollack, 1954]
Sumby W.H., Pollack I., « Visual contribution to speech intelligibility in noise », Journal of the Acoustical Society of America, 26, pp. 212-215, 1954.

[Summerfield, 1979]
Summerfield Q., « Use of visual information for phonetic perception », Phonetica, 36:314-331, 1979.

[Summerfield, 1987]
Summerfield Q., « Some preliminaries to a comprehensive account of audio-visuel speech perception », in Hearing by Eye : The psychology of lipreading, B. Dodd and R. Campbell, eds., 1987.

[Summerfield et al., 1989]
Summerfield Q., MacLeod A., McGrath M., Brooke M., « Lips, teeth, and the benefits of lipreading », in Handbook of Research on Face Processing, A.W. Young and H.D. Ellis (eds.), Elsevier Science Publishers, pp. 223-233, 1989.

[Tulips, 1995]
Movellan J.R., « Visual speech recognition with stochastic networks », in G. Tesauro, D. Touretzky, T. Leen, eds., Advances in Neural Information Processing Systems, vol.7, MIT press, Cambridge, 1995.

[Turk et Pentland, 91]
Turk M., Pentland A., « Eigenfaces for Recognition », Journal of Cognitive Neuroscience, 3, 1, 1991.

[Vézien, 1995]
Vézien J.M., « Techniques de Reconstruction Globale par Analyse de Paires d'Images Stéréoscospiques », Thèse de Doctorat, Université Paris VII, 1995.

[Waibel et Lee, 1990]
Waibel A., Lee K.-F. (eds), « Readings in Speech Recognition », San Mateo, CA: Morgan Kaufmann, 1990.

[Walden et al, 1977]
Walden, B. E., Prosek, A., Montgomery, « Effect of training on the visual recognition of consonants », Journal of Speech and Hearing Research, 20 :130-145, 1977.

[Whalen, 1990]
Whalen D.H., « Coarticulation is largely planned », Journal of Phonetics, Vol. 18, No 1, pp. 3-35, 1990.

[Wu et al., 1991]
Wu, J., Tamura, S., Mitsumoto, H., Kawai, H., Kurosu, K., Okazaki, K., « Neural network vowel recognition jointly using voice features and mouth shape image », Pattern Recognition, 18(6), pp. 636-642, 1991.

[Wyszecki et Stiles, 1982]
Wyszecki G., Stiles W. S., « Color Science : concepts and methods, quantitative data and formulae », The Wiley series in pure and applied optics. John Wiley, Inc., New York, 1982.

[Yamamoto et al., 1997]
Yamamoto E., Nakamura S., Shikano K., « Speech to lip movement synthesis by HMM », Proc. of the First ESCA Workshop on Audio-Visual Speech Processing, Rhodes, Greece, Sept. 26-27, 1997, pp. 137-140.

[Yuhas et al., 1990]
Yuhas B. P., Goldstein M. H., Sejnowski T.J., Jenkins R.E., « Neural network models of sensory integration for improved vowel recognition », Proc. IEEE, 78(10):1658-1668, Oct. 1990.

[Yuille et al., 1992]
Yuille A. L., Hallinan P. W., Cohen D.S., « Feature extraction from faces using Deformable Templates », International Journal of Computer Vision, 8:2, 99-111, 1992.

[Zemlin, 1968]
Zemlin, W.R., « Speech and Hearing Science : Anatomy and Physiology », New Jersey, Prentice-Hall, 1968.