INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

# THÈSE

*pour obtenir le grade de*

Docteur de l'INP Grenoble

Spécialité SIGNAL, IMAGE, PAROLE, TÉLÉCOMS

N° : ␣␣␣␣␣␣␣␣␣␣␣

# ESTIMATION DES MOUVEMENTS DU VISAGE D'UN LOCUTEUR DANS UNE SÉQUENCE AUDIOVISUELLE

*par*

Matthias ODISIO

Thèse soutenue le 12 décembre 2005 devant la commission d'examen :

| | |
|---|---|
| Président | Pierre-Yves COULON |
| Rapporteurs | Catherine PELACHAUD |
| | Franck DAVOINE |
| Directeur de thèse | Gérard BAILLY |
| Examinateurs | Jean-Luc DUGELAY |
| | Igor S. PANDZIC |

# ESTIMATION DES MOUVEMENTS DU VISAGE D'UN LOCUTEUR DANS UNE SÉQUENCE AUDIOVISUELLE

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

# THÈSE

*pour obtenir le grade de*

Docteur de l'INP Grenoble

Spécialité SIGNAL, IMAGE, PAROLE, TÉLÉCOMS

# ESTIMATION DES MOUVEMENTS DU VISAGE D'UN LOCUTEUR DANS UNE SÉQUENCE AUDIOVISUELLE

*par*

Matthias ODISIO

Thèse soutenue le 12 décembre 2005 devant la commission d'examen :

| | |
|---|---|
| Président | Pierre-Yves COULON |
| Rapporteurs | Catherine PELACHAUD |
| | Franck DAVOINE |
| Directeur de thèse | Gérard BAILLY |
| Examinateurs | Jean-Luc DUGELAY |
| | Igor S. PANDZIC |

À ma mère

# Remerciements

Tout d'abord j'exprime ma gratitude à Pierre Escudier, directeur de l'ICP lors de mon premier séjour dans ce laboratoire : il m'a accueilli puis m'a permis de faire cette thèse. Cette gratitude s'étend à son successeur, Jean-Luc Schwartz, qui a constamment fait preuve d'une grande disponibilité.

Gérard Bailly m'a amené à ce sujet de recherche et a dirigé ma thèse. Je le remercie vivement pour tout ce que j'ai appris durant ces années, pour la suite.

Je suis très reconnaissant à Pierre-Yves Coulon, Catherine Pelachaud, Franck Davoine, Jean-Luc Dugelay et Igor S. Pandzic qui m'ont marqué leur estime en constituant le jury de cette thèse et en critiquant scientifiquement mon travail. Ce mémoire a bénéficié de leurs remarques — notamment celles de F. Davoine.

Pour leur concours technique ou administratif, leurs conseils, leur confiance, j'adresse un *tutti frutti* de mercis à Christian Abry, Alain Arnal, Pierre Badin, Gérard Bailly, Denis Beautemps, Sophie Bechon, Frédéric Berthommier, Nadine Bioud, Christian Bulfone, Marie-Agnès Cathiard, Frédéric Elisei, Gang Feng, Yvette Gaude, Guillaume Gibert, Laurent Girin, Bernard Guérin, Bleicke Holm, Christian Lavergne, Hélène Lœvenbruck, Nino Medves, Joëlle Miguet, Marc Sato, Christophe Savariaux, Jean-Luc Schwartz, Annemie Van Hirtum et Pauline Welby.

Plus généralement, l'ambiance de travail et la bonne humeur qui règnent à l'ICP et à l'ENSERG sont le fait de toutes et tous.

Amis, amies : je vous remercie de vous.

*Il vient un moment où nous ne pouvons plus éluder les conséquences de nos théories, où tout ce que nous avons pensé exige d'être vécu, où toutes nos idées comme toutes nos fantaisies se convertissent en expériences, — et c'est alors que le jeu finit et que commence l'épreuve.*

CIORAN, *Cahiers.*

# Introduction

L'objectif de cette thèse est de construire et évaluer les composants d'un système qui, à partir d'une séquence vidéo d'une personne, capture les mouvements tridimensionnels de son visage. Dans le cas général, la grande variabilité des images source complique beaucoup cette tâche. De nombreux facteurs causent cette variabilité : l'orientation et la position du capteur vidéo, ainsi que ses propriétés intrinsèques (p. ex. : propriétés photométriques, déformation radiale) ; les facteurs environnementaux et notamment les conditions d'éclairage ; enfin, les facteurs, potentiellement interagissants, liés à la personne, à savoir la morphologie de son visage, ses stratégies articulatoires, l'*état* dans lequel elle est, et ce qu'elle dit.

Comme beaucoup de travaux du domaine, nous avons adopté une approche basée sur des modèles de la forme — ici, tridimensionnelle — et de l'apparence ; cela permet de contraindre et simplifier la tâche, sous l'hypothèse que chaque image de la séquence vidéo contient une projection du modèle. L'estimation des mouvements faciaux et des mouvements de tête revient à ajuster le modèle 3D de telle sorte que sa projection soit conforme aux informations de l'image.

Il est possible de retranscrire cette problématique selon le schéma de la figure 1, inspiré du cadre classique pour la synthèse audiovisuelle, où le modèle de contrôle, chargé de la génération des paramètres, pilote un modèle de la forme du visage et un modèle de son apparence. L'environnement, qui comprend ici également les autres partenaires de la communication, ne peut accéder qu'à l'apparence dans sa perception de la scène ; il peut influer sur l'apparence, p. ex. si

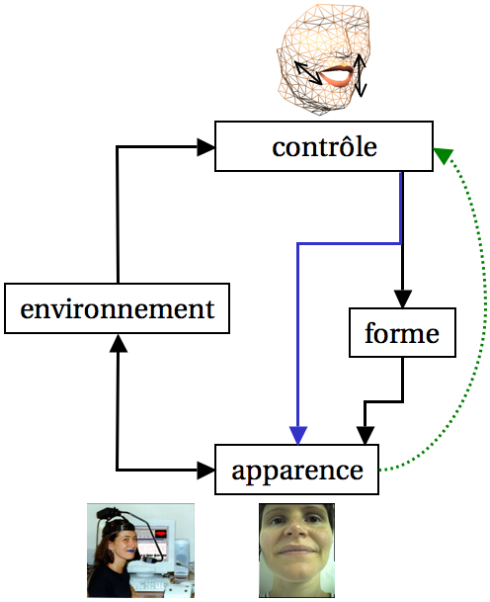



FIGURE 1 – Problématique de la thèse. Un modèle de contrôle, chargé de la génération des paramètres, pilote un modèle de la forme du visage et un modèle de son apparence. Pour une communication parlée face à face dont les conditions sont connues, les changements dans l'image sont des effets directs des mouvements articulatoires et nous proposons un contrôle articulatoire du modèle d'apparence. La tâche du système développé revient à retrouver les paramètres de contrôle depuis l'image (flèche en pointillés).

l'éclairage varie ou si la chaleur ambiante provoque une sudation ou des rougeurs chez la personne filmée.

Nous avons fait le choix de nous restreindre à un cas moins général, celui des scènes où un locuteur ou une locutrice connu est engagé avec un utilisateur ou une utilisatrice dans une communication parlée face à face, dans un contexte expressif neutre. Suivant ce paradigme, les changements dans l'image sont des effets directs des mouvements articulatoires, c'est pourquoi nous proposons un contrôle articulatoire du modèle d'apparence. La tâche du système revient à retrouver les paramètres de contrôle depuis l'image ; cela est stylisé par la flèche en pointillés sur la figure 1.

## SITUATIONS

Cette étude s'intéresse aux gestes faciaux de la parole et nous rappelons ci-dessous : quelques propriétés importantes des signaux de parole ; puis, quelques exemples d'applications pour des systèmes tels que celui développé dans cette thèse, en s'attachant enfin plus particulièrement à un projet RNRT auquel nous avons contribué.

# De la parole audiovisuelle

Le caractère multimodal, et principalement audiovisuel, de la production et de la perception de la parole a été établi durant ces cinquante dernières années par de nombreuses études portant, entre autres, sur l'ontogenèse, la phylogenèse ou l'analyse de scène.

On sait que les deux modalités — audio et vidéo — transmettent des informations complémentaires. La figure 2 reproduit les résultats de deux expériences *princeps* sous forme d'arbres de confusion perceptive. La tendance générale [1] est que le signal audio est plus efficace pour déterminer le mode d'articulation (sur l'arbre de gauche, le trait de voisement est particulièrement distingué) tandis que le signal visuel permet de distinguer les lieux d'articulation. Cette complémentarité audio-visuelle procure un gain d'intelligibilité ; un exemple d'illustration pour le français est représenté sur la figure 3, où l'on voit que la vision aide à la compréhension, même en parole claire. Si la vidéo des seules lèvres semble apporter dans cette condition autant d'information que la partie basse du visage, le gain perceptif diminue quand le bruit augmente ; cela pourrait s'expliquer par l'exploitation des redondances visuelles sur l'ensemble du visage pour l'extraction des informations labiales, processus qui nécessiterait alors une analyse visuelle suffisamment détaillée pour tirer parti des relations entre les mouvements subtils des lèvres et du reste du visage.

Par ailleurs, les deux modalités sont intégrées, fusionnées, lors de la percep-

1. Il est cependant connu maintenant que ce constat ne reflète pas la réalité des processus cognitifs : Schwartz (J.-L.), F. Berthommier et C. Savariaux. Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, 93:B69–B78, 2004 ; Ojanen (V.). *Neurocognitive Mechanisms of Audiovisual Speech Perception.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2005.

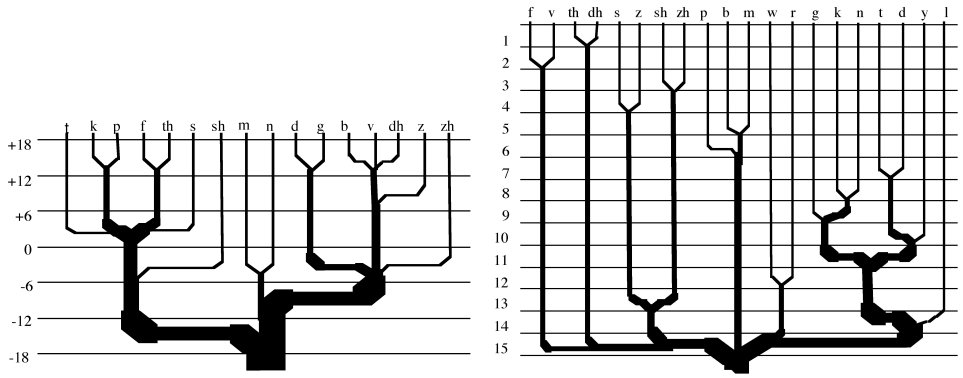



FIGURE 2 – Arbres de confusions consonantiques perceptives construits d'après les résultats d'identification de logatomes [Ca] pour l'anglais. À gauche, les confusions auditives (tiré de MILLER (G. A.) et P. E. NICELY. An analysis of perceptual confusions among some English consonants. *J. of the Acoustical Society of America*, 27(2):338–352, mars 1955) ; à droite, les confusions visuelles (tiré de WALDEN (B. E.), R. A. PROSEK, A. A. MONGOMERY, C. K. SCHERR et C. J. JONES. Effects of training on the visual recognition of consonants. *J. of Speech and Hearing Research*, 20:130–145, 1977).

tion ; cela est montré par une illusion robuste, l'effet McGurk [2] où, par exemple, des présentations audio-visuelles incongruentes [ba–ga] sont généralement perçues comme [da].

Pour ce qui nous concerne, retenons que les mouvements visuels de parole sont des mouvements d'une grande finesse dont la redondance sur tout le visage est exploitée sur le plan perceptif. C'est pourquoi nous avons considéré dans cette étude l'ensemble du visage et, pour capturer au plus fin ses mouvements, nous proposons une modélisation spécifique au locuteur ou à la locutrice. Cette spécificité perceptive nous a en outre conduit à mettre en œuvre des expériences d'intelligibilité pour l'évaluation de notre système.

2. MCGURK (H.) et J. MACDONALD. Hearing lips and seeing voices. *Nature*, 264:746–748, décembre 1976.

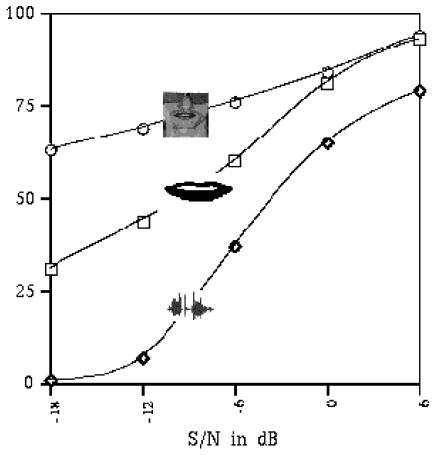



FIGURE 3 – Taux d'identification audiovisuelle de logatomes [VCVCV] pour le français en fonction du niveau de bruit audio pour trois conditions visuelles : visage entier, lèvres seules, pas de signal visuel (tiré de LE GOFF (B.), T. GUIARD-MARIGNY et C. BENOÎT. Read my lips... and my jaw ! How intelligible are the components of a speaker's face ? *In Proc. of the European Conf. on Speech Communication and Technology*, pages 291–294, Madrid, Spain, 1995).

## Contexte applicatif

Le champ applicatif des systèmes de capture des mouvements faciaux, de parole ou d'expressions faciales, depuis la vidéo est vaste : applications télécom pour la communication (personnages virtuels interactifs — avatar, agent conversationnel animé —, normalisation MPEG-4/SNHC) ; production d'effet spéciaux par transposition des mouvements capturés sur un autre modèle d'animation ; applications pour les nouvelles technologies éducatives (orthophonie, correction phonétique, apprentissage des langues étrangères) ; applications des technologies vocales (synthèse et reconnaissance audiovisuelle, reconnaissance ou vérification du locuteur, biométrie) ; enfin, applications utilitaires d'acquisition de mouvements pour des études théoriques d'analyse et de perception du geste.

Cette thèse s'est pleinement inscrite dans le cadre d'un projet RNRT ; comme des parties du cahier des charges de ce projet ont pu influer sur certains choix effectués dans la thèse — notamment la cadence de fonctionnement du système

—, il est utile de décrire de manière plus détaillée ce qui constitue une application technologique *réalisée* de notre étude.

## Du projet *TempoValse*

Dans la perspective des réseaux mobiles 3G et au-delà, le projet RNRT précompétitif *TempoValse* a visé la réalisation d'une maquette de terminal multimédia portable basé sur la normalisation MPEG-4. L'application retenue était la visiophonie « *scalable* » ; le terminal développé est composé principalement d'un module de capture des signaux audiovisuels fournissant un signal acoustique et une image vidéo centrée sur le visage du locuteur ou de la locutrice, et d'un module de traitement des signaux audiovisuels assurant le codage et le décodage des signaux audiovisuels issus du module de capture et envoyés vers l'interlocuteur ou l'interlocutrice, ainsi que ceux provenant de cet interlocuteur ou cette interlocutrice. L'utilisation du terminal en situation de déplacement est garantie par un dispositif de type casque et oreillette où les capteurs sont fixes et solidaires de la tête ; plusieurs terminaux développés au cours du projet sont représentés la figure 4. Des expérimentations menées par France Télécom [3] ont permis de valider une utilisation mobile du terminal et de préconiser que le regard soit dirigé vers le capteur vidéo, comme dans le cas d'un rendu effectué sur des lunettes « *see-through* ».

Au sein de ce projet, les participants de l'ICP avaient notamment en charge : la réalisation de la maquette physique du terminal ; la modélisation des locuteurs et des locutrices pour la création de clones ; les études algorithmiques et le développement logiciel pour l'analyse et l'animation faciale ; les modèles du visage pour la synthèse ; le moteur d'animation ; l'interface avec la chaîne de transmission des paramètres ; et, les études sur le gain perceptif de l'apport visuel.

Les travaux effectués dans cette thèse peuvent être mis en correspondance avec presque tous les aspects ci-dessus. Des collaborateurs de l'ICP ont participé, parfois de manière importante, à certaines parties de ces travaux ; cela est noté au bas de la première page des chapitres concernés.

3. Bailly (G.), F. Elisei, M. Odisio, D. Pelé, D. Caillière et K. Grein-Cochard. Talking faces for MPEG-4 compliant scalable face-to-face telecommunication. *In Proc. of the Smart Objects Conf.*, pages 204–207, Grenoble, France, mai 2003.

FIGURE 4 – Terminaux portables pour le projet *TempoValse* : le casque de la société Ganymédia (à gauche) ; les capteurs audiovisuels et le module de rendu de l'ICP (au centre) ; un dispositif de France Télécom où le monocle « *see-through* » aligné avec le capteur vidéo permet une communication face-à-face (à droite).

## *Deux types de conditions expérimentales*

Nous avons retenu deux types de conditions expérimentales, et deux corpora audiovisuels sont enregistrés : un corpus, appelé « billes », est destiné à la construction du modèle géométrique de la forme du visage, où le visage est marqué par des billes permettant de repérer de manière précise un réseau dense de points de chair ; un corpus, appelé « téléconférence », correspond à des conditions expérimentales compatibles avec le projet *TempoValse*, où le locuteur ou la locutrice est filmé par une caméra fixée à un casque relié rigidement à la tête et où son visage ne comporte plus de marqueurs — plus exactement, les seuls marqueurs présents sont utilisés pour déterminer les mouvements de référence mais ils sont gommés lors des analyses vidéo pour ne pas introduire de biais. Cela est illustré sur la figure 5.

Trois locuteurs et une locutrice ont été étudiées dans cette thèse ; les résultats principaux pour chacune de ces personnes seront donnés, mais pour une plus grande clarté nous faisons le choix dans ce mémoire d'illustrer avec la locutrice, française, la méthodologie développée. Notons également que pour des raisons techniques le matériau de parole à disposition pour les corpora « téléconférence » est malheureusement plus limité et les expériences les plus poussées ont été menées avec les corpora « billes ».

FIGURE 5 – Deux types de conditions expérimentales : la locutrice pour les corpora « billes » (à gauche) et « téléconférence » (à droite).

## PLAN DU MÉMOIRE

Le mémoire est organisé selon le plan suivant. Le chapitre premier est consacré à la construction du modèle tridimensionnel de la forme du visage. Le chapitre 2 présente les deux types de modèles de l'apparence du visage, un modèle de la texture du visage et un modèle de l'apparence locale d'une sélection automatique de points 3D ; ces deux modèles sont construits pour chacune des deux conditions expérimentales considérées. Le chapitre 3 présente l'architecture du notre système d'estimation des mouvements ; il consiste principalement en une boucle d'analyse par la synthèse qui cherche à ajuster à l'image des modèles du visage construits au préalable. Les deux derniers chapitres sont consacrés à l'évaluation de notre système, un point sur lequel nous nous sommes particulièrement focalisés. Les résultats d'évaluations objectives des mouvements estimés sont repris au chapitre 4, tandis que le chapitre 5 présente plusieurs évaluations perceptives de ces mouvements.

# 1

# Modélisation tridimensionnelle de la forme du visage

## 1.1. INTRODUCTION

La nature tridimensionnelle du conduit vocal, les applications des têtes parlantes virtuelles incitent à préférer un modèle 3D de la géométrie du visage. Par ailleurs, en analyse d'image on sait que les modèles 3D permettent de simplifier la tâche en rendant compte — intrinsèquement ou plus simplement — de variations parfois complexes et non-linéaires en 2D.

Des dizaines de muscles de la face, des lèvres, du cou et de la langue sont impliqués lorsque le visage est en mouvement [1] ; ces muscles ne sont pas contrôlés indépendamment : p. ex., la description de mouvements expressifs observés est faite classiquement selon 46 composantes faciales [2]. Lorsqu'une personne parle, le nombre des degrés de liberté effectivement observés géométriquement est encore plus petit. Cette redondance, ces corrélations, naturelle est à la base de la méthodologie que nous avons adoptée pour guider l'émergence, à partir d'une collection de données 3D spécifique au locuteur ou à la locutrice, de modèles

L'acquisition des données audiovisuelles doit beaucoup à C. Savariaux et à A. Arnal. L'étiquetage et la méthodologie d'analyse de ces données sont le fruit d'un travail commun, fait avec (par ordre alphabétique) : G. Bailly, F. Elisei et B. Holm. H. Lœvenbruck a accepté d'être sujet de cette étude.

1. Parke (F. I.) et K. Waters. *Computer facial animation.* A K Peters Ltd, Wellesley, USA, 1996.

2. Ekman (P.) et W. V. Friesen. *Facial Action Coding System (Investigator's Guide).* Consulting Psychologists Press, Inc., 1978.

de la forme du visage pilotés linéairement par un petit nombre de paramètres articulatoires. Avant de présenter ce paradigme de construction, voyons d'abord quels autres modèles sont employés pour l'animation faciale de parole.

## 1.1.1. Modèles 3D pour l'animation faciale

### *1.1.1.1. Les modèles génériques : paramétriques ou physiologiques*

On peut distinguer en deux types les modèles génériques, suivant qu'ils sont paramétriques (ou encore géométriques) ou bien physiologiques. Ces modèles *a priori* doivent être ajustés à la morphologie du locuteur ou de la locutrice avant d'être utilisés dans une tâche d'analyse de séquence vidéo. Cette adaptation est faite à partir de vues où des points caractéristiques de la géométrie du visage *cible* sont repérés, soit manuellement soit automatiquement. Après adaptation, il n'est pas clair que ces modèles puissent reproduire de manière correcte toutes les postures faciales.

Pour les modèles paramétriques, les positions des sommets du maillage, qui représente la surface du visage, sont calculées localement à partir des positions et de l'influence de points de contrôle ; ce sont les déplacements de ces points qui permettent au modèle de rendre compte des mouvements faciaux [3].

3. P. ex., Eisert (P.) et B. Girod. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5):70–78, septembre 1998 ; Ahlberg (J.). *Model-based coding — Extraction, coding, and evaluation of face model parameters*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, septembre 2002 ; Cohen (M. M.), D. W. Massaro et R. Clark. Training a talking head. *In Proc. of the IEEE Internat. Conf. on Multimodal Interfaces*, pages 499–504, Pittsburgh, USA, octobre 2002 ; Tao (H.) et T. S. Huang. Visual estimation and compression of facial motion parameters — Elements of a 3D model-based video coding system. *Internat. J. of Computer Vision*, 50(2):111–125, novembre 2002 ; Pelachaud (C.). Visual text-to-speech. *In* Pandzic (I. S.) et R. Forchheimer, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 8, pages 125–140. Wiley, 2002 ; Kalberer (G. A.), P. Mueller et L. V. Gool. Visual speech, a trajectory in viseme space. *Internat. J. of Imaging Systems and Technology*, 13(1):74–84, juin 2003 ; Basu (S.), N. Oliver et A. Pentland. 3D lip shapes from video : A combined physical-statistical model. *Speech Communication*, 26:131–148, 1998 ; DeCarlo (D.) et D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Internat. J. of Computer Vision*, 38(2):99–127, juillet 2000. Certains de ces modèles sont des améliorations des modèles *princeps*, de Parke : Parke (F. I.). Parameterized models for facial animation. *IEEE Computer Graphics & Applications*, 2:61–68, novembre 1982, et du modèle Candide : Rydfalk (M.). CANDIDE, a parameterized face. Rapport technique LiTH-ISY-I-866, Dept. of Electrical Engineering, Linköping University, 1987.

Les modèles physiologiques ou biomécaniques veulent reproduire explicitement les caractéristiques physiologiques réelles du visage comme : la structure et l'activité musculaires ; certaines lois de la physique appliquées aux composants du modèle ; des contraintes biomécaniques. Ils sont pilotés par des commandes musculaires [4]. Notons que ces modèles — pourtant plus proches de la nature du visage — n'ont pas atteint encore la maturité nécessaire pour une tâche d'analyse vidéo : les calculs numériques associés peuvent être instables pour des modèles simples de type réseau de masses–ressorts ; certaines propriétés manquent encore, même aux modèles les plus évolués de type éléments finis 3D — pour lesquels le temps de calcul est important.

### *1.1.1.2. Des modèles spécifiques pour chaque personne*

Plutôt que d'acquérir des données puis d'adapter à ces données un modèle générique, il est possible de construire un modèle direct ; c'est le choix que nous avons fait. Puisqu'ils sont basés sur des données *ad hoc* ces modèles, *a posteriori* et spécifiques à la personne étudiée, sont capables de couvrir l'espace d'apprentissage [5] ; pour aboutir à un modèle linéaire qui exploite au mieux la redon-

4. Terzopoulos (D.) et K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):569–579, juin 1993 ; King (S. A.) et R. E. Parent. Creating speech-synchronized animation. *IEEE Trans. on Visualization and Computer Graphics*, 11(3):341–352, 2005 ; Lucero (J. C.) et K. G. Munhall. A model of facial biomechanics for speech production. *J. of the Acoustical Society of America*, 106 (5):2834–2848, novembre 1999 ; Choe (B.), H. Lee et H.-S. Ko. Performance-driven muscle-based facial animation. *J. of Visualization and Computer Animation*, 12:67–79, 2001 ; Gomi (H.), J. Nozoe, J. Dang et K. Honda. Physiologically based lip model for generating speech articulation. *In Proc. of the Internat. Seminar on Speech Production*, pages 79–84, Sydney, Australia, décembre 2003.

5. P. ex., Guenter (B.), C. Grimm, D. Wood, H. Malvar et F. Pighin. Making faces. *In Proc. of SIGGRAPH*, Computer Graphics Proceedings, Annual Conference Series, pages 55–66. ACM SIGGRAPH / Addison Wesley, juillet 1998 ; Pighin (F.), R. Szeliski et D. H. Salesin. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002 ; Hong (P.), Z. Wen et T. S. Huang. Real-time speech-driven face animation. *In* Pandzic (I. S.) et R. Forchheimer, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 7, pages 115–124. Wiley, 2002 ; Theobald (B.-J.), J. A. Bangham, I. Matthews et G. C. Cawley. Visual speech synthesis using statistical models of shape and appearance. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 78–83, 2001 ; Blanz (V.), C. Basso, T. Poggio et T. Vetter. Reanimating faces in images and video. *In Proc. of the Annual Conf. of the European Association for Computer Graphics*, Granada, Spain, 2003 ; Kuratate (T.). *Talking head animation system driven by facial motion mapping and a 3D face database.* PhD thesis, Department of Information Processing, Nara Institute of Science and Technology, Nara, Japan, juin 2004.

dance des données, ces dernières font l'objet de traitements statistiques relevant de l'analyse factorielle : l'analyse en composantes principales est très utilisée, nous aurons l'occasion d'y revenir.

### 1.1.2. Présentation de notre méthode de modélisation

Comme il est classique, nous modélisons le visage comme un réseau polygonal tridimensionnel basé par construction sur la morphologie du locuteur ou de la locutrice. Ce maillage 3D est particulièrement dense au niveau de la zone cruciale des lèvres où un modèle de type *spline* est utilisé pour reproduire la surface labiale ; pour le reste du visage, les sommets du maillage correspondent à des points de chair marqués — lors de l'apprentissage — par des billes collées sur le visage. Les mouvements du maillage sont appris à partir d'analyses statistiques d'un petit ensemble de mouvements soigneusement capturés sur la personne étudiée ; ces analyses statistiques sont supervisées en fonction des connaissances phonétiques de telle sorte que les paramètres de contrôle correspondent à des degrés de liberté pertinents pour la parole et de manière à éviter les corrélations fortuites qui seraient issues d'un ensemble d'apprentissage réduit (p. ex., entre les sourcils et les lèvres).

Voyons maintenant plus en détail la construction d'un tel modèle.

## 1.2. CORPUS RETENU POUR LA MODÉLISATION DE LA PAROLE

L'approche que nous avons adoptée pour la construction d'un modèle tridimensionnel est basée sur l'analyse, et donc sur leur collecte préalable, de données spécifiques recueillies sur le locuteur ou la locutrice.

À partir de mesures géométriques des lèvres obtenues grâce à un système de *Chromakey*, une classification statistique d'un corpus pour le français a dégagé empiriquement un ensemble de 23 postures représentatives du corpus total : les auteurs ont appelé *visèmes* ces 23 postures [6]. Des visèmes ont aussi été définis à

6. Benoît (C.), T. Lallouache, T. Mohamadi et C. Abry. A set of French visemes for visual speech synthesis. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines : Theories, Models and Designs*, pages 485–501. Elsevier B.V., 1992.

partir de mesures perceptives [7]. Dans les deux cas, il s'agit de définitions expérimentales *a posteriori*.

La méthodologie d'analyse statistique que nous avons utilisée — elle sera détaillée ci-après — a déjà été appliquée pour la modélisation articulatoire de la langue ; il a été montré qu'une sélection composée des voyelles et des cibles consonantiques du corpus initial (20 configurations sélectionnées parmi 1 222) permettait de construire un modèle presque aussi précis [8] qu'un modèle construit à partir de l'ensemble du corpus [9].

Nous avons donc fait le choix pour le corpus d'un ensemble restreint de configurations articulatoires ; ces configurations sont censées couvrir l'espace des déformations faciales visibles au cours de l'activité de parole et nous les appellerons, *a priori*, visèmes. Ces visèmes sont définis comme les centres des réalisations des phonèmes suivant :

- les voyelles orales et nasales du français { i, e, ɛ, a, ɔ, o, u, y, ø, œ, ɑ̃, ɛ̃, ɔ̃, œ̃ } ;
- les consonnes sous forme de triphones en contexte vocalique symétrique *VCV* , où *V* est l'une des voyelles { i, e, a, o, u, œ } et *C* est l'une des consonnes { p, t, k, b, d, g, f, v, s, z, ʃ, ʒ, ʀ, l, m, n, ð, θ } [10] ;
- deux postures *silencieuses* : *rest* (bouche fermée au repos) et *preph* (bouche entrouverte en position préphonatoire).

7. Fisher (C. G.). Confusions among visually perceived consonants. *J. of Speech and Hearing Research*, 15:474–482, 1968 ; Walden (B. E.), R. A. Prosek, A. A. Mongomery, C. K. Scherr et C. J. Jones. Effects of training on the visual recognition of consonants. *J. of Speech and Hearing Research*, 20:130–145, 1977. Rappelons aussi que la norme MPEG-4 propose une liste de 14 visèmes pour l'anglais.

8. C.-à-d. autour d'un millimètre, toujours de l'ordre de grandeur de la précision des données initiales.

9. Badin (P.), G. Bailly, M. Raybaudi et C. Segebarth. A three-dimensional linear model articulatory model based on MRI data. *In Proc. of the Third ESCA/COCOSDA Internat. Workshop on Speech Synthesis*, pages 249–254, Jenolan Caves, Australia, 1998.

10. Les fricatives dentales [ð] et [θ] ont été incluses pour tester l'éventuelle application du modèle à l'anglais. La locutrice, de langue maternelle française, est bilingue.

## 1.3. DISPOSITIF POUR LA CAPTURE DE MOUVEMENTS

Les gestes de parole résultent parfois de mouvements du visage très fins : p. ex., seules des différences de quelques mm² sur l'ouverture des lèvres distinguent la voyelle [u], la fricative [f] en contexte vocalique arrondi [ufu] ou une fermeture complète comme [p] dans [upu]. L'acquisition des données doit être au moins aussi précise. Les systèmes de capture de la géométrie et des mouvements sont maintenant facilement disponibles. Leur utilisation n'est pourtant pas simple ou souhaitable pour ce que nous voulons :

- les scanners 3D produisent des données qui doivent ensuite être mises en correspondance avec un modèle géométrique du visage où les points 3D seront ancrés dans la chair ; de plus la durée de l'acquisition est trop longue pour que le sujet reste parfaitement immobile, ce qui semble empêcher le système d'atteindre la précision requise en parole [11] ;
- les systèmes à base de capteurs actifs ou passifs délivrent des mouvements de plusieurs dizaines de points de chair à des fréquences suffisantes pour l'étude de la parole ; peu invasif, le système Vicon [12] est censé avoir une précision nominale de 0,1 mm ; mais, les auteurs qui l'utilisent ont constaté des erreurs ponctuelles beaucoup plus importantes et qu'il faut essayer de corriger par des post-traitements *ad hoc* des signaux [13].

11. Dans l'étude de Kuratate, la durée d'un scan avec le système Cyberware est de 17 s ; les données sont adaptées ensuite grâce à points de chair marqués par des senseurs d'un système Optotrack ; les données dans la région labiale sont trop bruitées et sont remplacées par l'adaptation d'un modèle géométrique : Kuratate (T.). *Talking head animation system driven by facial motion mapping and a 3D face database.* PhD thesis, Department of Information Processing, Nara Institute of Science and Technology, Nara, Japan, juin 2004.

12. http://www.vicon.com

13. P. ex., alors que la fréquence d'acquisition est de 120 Hz, un filtrage passe-bas dont la fréquence de coupure est 10 Hz est appliqué ; les trames où plus de 9 % des points ne sont pas détectés sont éliminées dans Maeda (S.), M. Toda, A. J. Carlen et L. Maftahi. Functional modeling of the face during speech production. *In Actes des Journées d'études sur la parole*, pages 341–344, Nancy, France, juin 2002. Gibert et coll. rapportent aussi des lissages temporels des trajectoires et des interventions manuelles fastidieuses : Gibert (G.), G. Bailly, D. Beautemps, F. Elisei et R. Brun. Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *J. of the Acoustical Society of America*, 118(2):1144–1153, 2005 ; Lucero (J. C.), S. T. R. Maciel, D. A. Johns et K. G. Munhall. Empirical modeling of human face kinematics during speech using motion clustering. *J. of the Acoustical Society of America*, 118

Ces deux types de système requièrent de plus des interventions manuelles pour corriger partie des données. En outre, les données obtenues pour les lèvres posent problème : les deux types de système sont inaptes à fournir la forme de l'ouverture labiale — pour cela des capteurs devraient être placés sur la partie humide des lèvres.

Nous avons fait le choix d'un système qui utilise des caméras vidéos 50 Hz où les données seront étiquetées de manière contrôlée, manuelle et directe. Un grand nombre de points de chair (plus de 200) sont marqués par des billes collées sur le visage du locuteur ou de la locutrice. Les billes sont placées en accord avec la structure osseuse et musculaire sous-jacente (p. ex. sur le sillon nasogénien [14]). La densité est maximale pour le bas du visage : c'est lui qui varie le plus en parole neutre (voir la figure 1.1). Il ne serait pas possible de coller des billes sur les lèvres, c'est pourquoi un modèle géométrique spécifique sera utilisé.

## 1.3.1. Dispositif expérimental

Les conditions d'enregistrement sont contrôlées autant que possible. L'enregistrement a lieu dans une chambre sourde ; l'éclairage, artificiel, est assuré par deux ballasts de quatre néons de type « lumière du jour ». Le capteur vidéo consiste en deux caméras de haute qualité et deux miroirs ; ce capteur permet l'acquisition simultanée de quatre vues — des deux profils et deux vues de face — du locuteur ou de la locutrice ; le cadrage est centré sur le visage, en légère contre-plongée pour mieux visualiser le dessous du menton et le larynx (voir la figure 1.1). Les deux caméras sont synchronisées entre elles par *genlock* ; l'enregistrement se fait sur *Betacam SP* ; un petit montage électronique à base de LEDs produisant un flash pendant 20 ms — la durée d'une trame — permet de caler les deux séquences d'images lors de la numérisation [15]. Les deux miroirs n'ont pas de couche de verre protectrice [16] afin d'éviter une deuxième réflexion parasite sur la vitre. Enfin, le signal audio est enregistré, de manière synchrone, avec la vidéo.

(1):405–409, 2005.

14. « ride du sourire » située en arrière de l'aile du nez qui limite la lèvre supérieure sur le côté de la joue.

15. Depuis, le laboratoire est équipé d'un système grâce auquel l'acquisition peut se faire directement en numérique.

16. Ils sont donc très fragiles et doivent être protégés de tout contact.

FIGURE 1.1 – Image issue du dispositif pour l'acquisition des données. Les billes collées sur le visage de la locutrice permettent de reconstruire la position 3D de 245 points de chair. Ici, le visème [iki].

## 1.3.2. Calibrage du capteur vidéo

Le calibrage d'une caméra [17] consiste à instancier un modèle géométrique qui rend compte du processus de formation des images, de la manière dont le monde tridimensionnel se projette dans le plan image.

### *1.3.2.1. Modèle sténopé*

Nous avons choisi pour les caméras le modèle sténopé : ce modèle permet de reproduire l'effet de perspective [18] et, du point de vue mathématique, les cal-

17. Dans cette section, le mot « caméra » est utilisé pour désigner indifféremment une caméra ou un miroir ; les miroirs sont considérés comme des caméras indépendantes. Certains calculs, comme la reconstruction métrique, peuvent différer *légèrement* — la différence n'est pas significative en ce qui nous concerne — de ceux résultant d'un cadre mathématique qui intégrerait les propriétés physiques de la réflexion : LIN (I.-C.), J.-S. YEH et M. OUHYOUNG. Extracting 3D facial animation parameters from multiview video clips. *IEEE Computer Graphics & Applications*, 22 (6):72–80, novembre-décembre 2002.

18. Chacun a déjà pu remarquer qu'une fois *dans* la photo ou l'image deux droites parallèles dans le monde 3D se rejoignent à l'infini ; dans certaines de nos prises de vue nous avons constaté que le parallélisme n'était pas préservé par la projection, ce qui nous a conduit à rejeter le modèle de projection affine — dans les descriptions ci-après, adopter ce modèle plus simple reviendrait

culs associés sont simples puisque ce modèle est une transformation *linéaire* de l'espace projectif vers le plan projectif[19].

Le modèle sténopé est défini par des paramètres extrinsèques, qui situent la caméra dans le repère du monde, et des paramètres intrinsèques qui représentent la projection perspective du repère de la caméra vers le repère lié à l'image.

Les paramètres intrinsèques sont :

– la distance focale en pixel dans les directions horizontale $\alpha_u$ et verticale $\alpha_v$ ;

– les coordonnées du point principal dans le repère pixels, $u_0$ et $v_0$.

Ces paramètres forment la matrice **C** :

$$\mathbf{C} = \begin{pmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.1}$$

Soient $\boldsymbol{t}$ la position de la caméra et **R** la matrice de rotation qui oriente la caméra. **R** est construite à partir des angles d'Euler, notés ici $r_x$, $r_y$ et $r_z$ :

$$\mathbf{R} = \begin{pmatrix} \cos r_z & -\sin r_z & 0 \\ \sin r_z & \cos r_z & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos r_y & 0 & \sin r_y \\ 0 & 1 & 0 \\ -\sin r_y & 0 & \cos r_y \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos r_x & -\sin r_x \\ 0 & \sin r_x & \cos r_x \end{pmatrix} \tag{1.2}$$

La matrice de projection $3\times4$ **P** de la caméra est enfin[20] :

$$\mathbf{P} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{R} & -\mathbf{R}\boldsymbol{t} \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{1.3}$$

La projection perspective d'un point 3D s'écrit :

$$\begin{pmatrix} su \\ sv \\ s \end{pmatrix} = \begin{pmatrix} P_{1,1} & P_{1,2} & P_{1,3} & P_{1,4} \\ P_{2,1} & P_{2,2} & P_{2,3} & P_{2,4} \\ P_{3,1} & P_{3,2} & P_{3,3} & P_{3,4} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{1.4}$$

à forcer à $\begin{pmatrix} \mathbf{0}^T & 1 \end{pmatrix}$ la troisième ligne de la matrice de projection **P**.

19. Les calculs sont donc conduits en coordonnées homogènes. Pour une présentation complète, voir en français Horaud (R.) et O. Monga. *Vision par ordinateur : outils fondamentaux.* Hermès, Paris, France, deuxième édition, 1995 ; ou en anglais Faugeras (O.). *Three-Dimensional Computer Vision — A Geometric Viewpoint.* MIT Press, Cambridge, USA, 1993.

20. à un facteur scalaire près

Les coordonnées $(u;v)$ dans le plan image sont alors :

$$\begin{cases} u = \dfrac{P_{1,1}X+P_{1,2}Y+P_{1,3}Z+P_{1,4}}{P_{3,1}X+P_{3,2}Y+P_{3,3}Z+P_{3,4}} \\ v = \dfrac{P_{2,1}X+P_{2,2}Y+P_{2,3}Z+P_{2,4}}{P_{3,1}X+P_{3,2}Y+P_{3,3}Z+P_{3,4}} \end{cases} \tag{1.5}$$

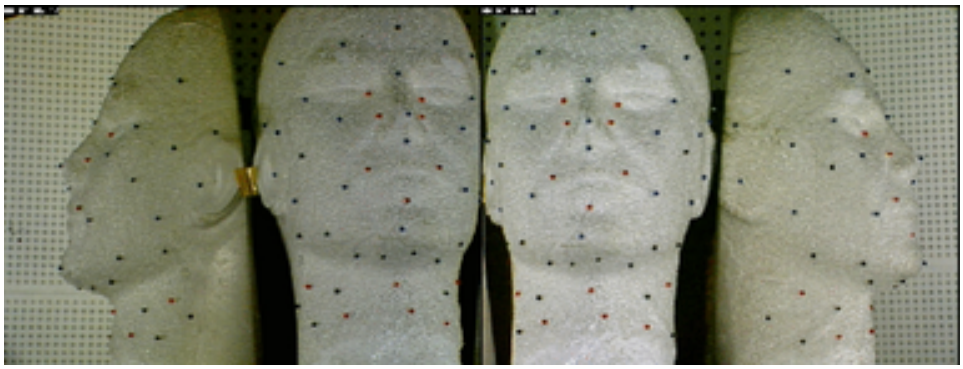


FIGURE 1.2 – Vues de l'objet de calibrage.

### *1.3.2.2. Calcul de la matrice de projection*

Afin de pouvoir déterminer la projection dans le plan image de points 3D, il faut connaître la matrice **P**. Pour cela, on introduit dans la scène un objet 3D connu : c'est Lazlo, la tête de polystyrène [21] représentée sur la figure 1.2 ; nous avons défini la structure tridimensionnelle de Lazlo à partir des mesures avec un pied à coulisse de plus de 200 distances entre les billes [22]. Utiliser comme mire de calibrage un objet de la forme d'une tête présente l'intérêt de bien calibrer la portion de l'espace 3D où sera ensuite le visage du locuteur ou de la locutrice.

21. Sa peau de polystyrène a brûlé ; d'où son nom de patient anglais…

22. Un scan 3D de Lazlo permettrait d'obtenir une meilleure définition de sa géométrie ; cette forme 3D serait utilisable en synthèse pour des expériences d'évaluations de méthodes de calibrage automatique, etc.

Après étiquetage manuel des positions dans l'image $\{(u_i; v_i)\}$ des points 3D $\{\mathbf{X}_i\}$ de Lazlo, on aboutit à partir de l'équation (1.5) au système linéaire suivant :

$$\begin{pmatrix} & & & & & & & \vdots & & & \\ \mathrm{X}_i & \mathrm{Y}_i & \mathrm{Z}_i & 1 & 0 & 0 & 0 & 0 & -u_i\mathrm{X}_i & -u_i\mathrm{Y}_i & -u_i\mathrm{Z}_i \\ 0 & 0 & 0 & 0 & \mathrm{X}_i & \mathrm{Y}_i & \mathrm{Z}_i & 1 & -v_i\mathrm{X}_i & -v_i\mathrm{Y}_i & -v_i\mathrm{Z}_i \\ & & & & & & & \vdots & & & \end{pmatrix} \begin{pmatrix} \mathrm{P}_{1,1} \\ \mathrm{P}_{1,2} \\ \mathrm{P}_{1,3} \\ \mathrm{P}_{1,4} \\ \mathrm{P}_{2,1} \\ \mathrm{P}_{2,2} \\ \mathrm{P}_{2,3} \\ \mathrm{P}_{2,4} \\ \mathrm{P}_{3,1} \\ \mathrm{P}_{3,2} \\ \mathrm{P}_{3,3} \end{pmatrix} = \begin{pmatrix} \vdots \\ u_i\mathrm{P}_{3,4} \\ v_i\mathrm{P}_{3,4} \\ \vdots \end{pmatrix} \tag{1.6}$$

Comme $\mathbf{P}$ est définie à un facteur près et comme la solution $\mathbf{P} = \mathbf{0}$ n'est pas intéressante, nous imposons $\mathrm{P}_{3,4} = 1$. Le système est résolu ensuite au sens des moindres carrés.

Il est possible de déterminer les paramètres intrinsèques et extrinsèques à partir de $\mathbf{P}$. Les calculs que nous serons amenés à faire par la suite ne nécessitent pas de connaître explicitement ces paramètres : nous avons choisi de conserver pour $\mathbf{P}$ l'estimation ci-dessus pour disposer d'une matrice de projection qui rende compte au mieux des observations, au détriment de la sémantique « physico-géométrique » de ses paramètres.

Cette procédure est répétée pour les quatre caméras du dispositif expérimental ; dans le plan coronal « moyen » du visage, la résolution calculée de ce dispositif est de 2,3 pixels par mm ; la résolution de l'étiquetage manuel est subpixellique.

## 1.4. COLLECTION DES POINTS 3D

### 1.4.1. Géométrie 3D associée à l'image d'une posture

D'un point de vue géométrique, une posture articulatoire est représentée par un ensemble de postions 3D composé des points marqués par les billes sur le

visage et de 30 points sur les lèvres qui contrôlent un modèle de la géométrie de ces dernières.

### *1.4.1.1. Des points de chair marqués par les billes*

Une personne experte étiquette sur les quatre vues les positions dans l'image de toutes les billes collées sur le visage [23].

Ensuite, les positions dans l'espace 3D de chaque point sont déterminées par reconstruction stéréoscopique. Soient $n$ $(n > 1)$ le nombre de vues où le point est visible et $\{(u_i; v_i),\ 1 \leq i \leq n\}$ les positions dans l'image d'un point 3D dont on recherche les coordonnées $\begin{pmatrix} \mathrm{X} & \mathrm{Y} & \mathrm{Z} \end{pmatrix}^{\mathrm{T}}$. À partir de l'équation de projection (1.5) il vient facilement :

$$\begin{pmatrix} \vdots & & \\ \mathrm{P}_{i\ 1,1} - u_i \mathrm{P}_{i\ 3,1} & \mathrm{P}_{i\ 1,2} - u_i \mathrm{P}_{i\ 3,2} & \mathrm{P}_{i\ 1,3} - u_i \mathrm{P}_{i\ 3,3} \\ \mathrm{P}_{i\ 2,1} - v_i \mathrm{P}_{i\ 3,1} & \mathrm{P}_{i\ 2,2} - v_i \mathrm{P}_{i\ 3,2} & \mathrm{P}_{i\ 2,3} - v_i \mathrm{P}_{i\ 3,3} \\ & \vdots & \end{pmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \end{pmatrix} = \begin{pmatrix} \vdots \\ u_i \mathrm{P}_{i\ 3,4} - \mathrm{P}_{i\ 1,4} \\ v_i \mathrm{P}_{i\ 3,4} - \mathrm{P}_{i\ 2,4} \\ \vdots \end{pmatrix} \tag{1.7}$$

Ce système linéaire est résolu au sens des moindres carrés.

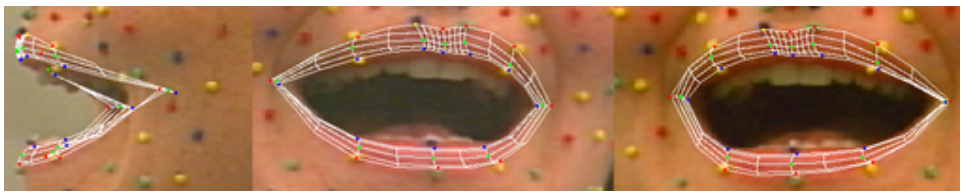


FIGURE 1.3 – Adaptation du modèle géométrique des lèvres pour un [a].

23. Cet étiquetage tout manuel est assisté par un logiciel qui montre des renseignements sur la position probable du point en cours de saisie : p. ex., les droites épipolaires ; la reconstruction 3D estimée à partir des vues où l'étiquetage est déjà fait.

#### 1.4.1.2. Des trente points de lèvres

Des billes collées sur les lèvres perturberaient l'articulation : la géométrie des lèvres est modélisée par un modèle paramétrique 3D [24]. Ce modèle consiste en un maillage basé sur un ensemble de contours polynomiaux de degrés deux et trois ; ces contours sont contrôlés par les positions de trente points situés sur les lèvres. Comme représenté sur la figure 1.3, les points de contrôle sont positionnés manuellement de telle sorte que la surface 3D générée s'ajuste à l'image des lèvres ; des contraintes *ad hoc* pendant cet ajustement manuel garantissent la sémantique physique de ces 30 points.

### 1.4.2. Expression dans un repère crânien de référence

Dans un référentiel lié aux caméras, les positions des points 3D extraites ci-dessus sont le résultat de mouvements faciaux et des mouvements rigides de la tête. Si les mouvements rigides sont dus à l'activité de parole, ils ne présentent pas ici d'intérêt puisque le modèle articulatoire est basé sur des données « statiques » [25]. Les données qui nous importent sont celles liées aux seuls mouvements faciaux : les positions 3D sont donc exprimées dans un repère du monde lié au crâne et défini sur une posture articulatoire de référence.

Un ensemble de points rigides $\{\mathbf{X_i},\ i = 1 \ldots 6\}$ est montré sur la figure 1.4 ; ces points sont utilisés pour déterminer le mouvement de tête de la posture articulatoire vers le repère de la posture articulatoire de référence.

Pour cela nous cherchons la translation $\boldsymbol{t}$ et les trois angles d'Euler $\boldsymbol{r}$ qui permettent d'ajuster les projections de cet ensemble de points dans l'image — sur les quatre vues — à leurs positions étiquetées $\boldsymbol{x_i}$. La matrice de rotation est calculée comme dans l'équation (1.2). Après roto-translation, les points rigides sont :

$$\begin{pmatrix} \mathbf{X}_{\boldsymbol{i}}^{'} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X}_{\boldsymbol{i}} \\ 1 \end{pmatrix} \tag{1.8}$$

24. Revéret (L.) et C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 207–212, Terrigal, Australia, décembre 1998 ; pour une description détaillée, voir Revéret (L.). *Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, mai 1999.

25. Le modèle articulatoire est censé couvrir l'espace des postures du régime dynamique.

FIGURE 1.4 – Points du visage utilisés pour le calcul du mouvement rigide de la tête ; en plus de points positionnés sur l'os nasal et les tempes, cet ensemble inclut des points positionnés sur le haut du visage, quasi-immobiles dans le repère crânien au cours de la parole sans contexte expressif. Cette caractéristique est vérifiée *a posteriori*.

On note $\boldsymbol{u_i'} = \begin{pmatrix} u_i' & v_i' \end{pmatrix}^{\mathrm{T}}$ les projections des points $\mathbf{X_i'}$ selon l'équation (1.5). Les paramètres du mouvement rigide recherché sont alors :

$$\arg\min_{\boldsymbol{t},\boldsymbol{r}} \sum_i \left\| \boldsymbol{x_i} - \boldsymbol{u_i'} \right\|^2 \tag{1.9}$$

Cette minimisation non-linéaire est résolue numériquement.

Le repère crânien utilisé *in fine*, dans lequel sont exprimés les points rigides de référence, est défini ainsi :

- l'origine est la jonction basse de l'extérieur des incisives supérieures ;
- l'axe $x$ (resp. $y$, $z$) est perpendiculaire au plan sagittal (resp. axial, coronal) ;
- le repère est direct, l'axe des $x$ est orienté vers la gauche du visage, l'axe des $y$ est orienté vers le haut, l'axe des $z$ orienté vers l'extérieur du visage.

## 1.5. ÉMERGENCE DU MODÈLE ARTICULATOIRE 3D PAR ANALYSE FACTORIELLE

C'est par analyse factorielle qu'est construit le modèle de la géométrie du visage à partir de la collection des données 3D. Cette méthode d'analyse linéaire conduit à un modèle où les coordonnées 3D $\mathbf{X}$ sont calculées par combinaison linéaire de facteurs de déformation $\boldsymbol{\alpha}$ associés à des vecteurs de déplacement 3D $\mathbf{M_X}$. Soit, en notant $\mathbf{X_0}$ la posture neutre, N le nombre de points 3D et $\mathbf{X}^j$ les données du $j$-ème visème :

$$\mathbf{X}^j = \begin{pmatrix} X_1^j & Y_1^j & Z_1^j & \dots & X_N^j & Y_N^j & Z_N^j \end{pmatrix}^T = \mathbf{X_0} + \mathbf{M_X} \cdot \boldsymbol{\alpha}^j \tag{1.10}$$

Les facteurs de déformation sont orthogonaux — c.-à-d. indépendants — si leur corrélation est nulle sur la collection de données. Plusieurs techniques existent pour déterminer ces facteurs de déformation ; l'analyse en composantes principales (ACP) est une technique non-supervisée, optimale au sens de la variance expliquée, très utilisée [26]. Reprenant des travaux antérieurs en parole sur la modélisation articulatoire [27] nous avons fait le choix d'une technique supervisée — plutôt, guidée — basée sur la régression linéaire. Le processus est supervisé en choisissant la nature des facteurs de déformation et donc en choisissant la répartition de la variance des données. Ainsi, les facteurs de déformations et leurs vecteurs de déplacement associés seront interprétables fonctionnellement

26. Voir p. ex. pour les mouvements faciaux THEOBALD (B.-J.), J. A. BANGHAM, I. MATTHEWS et G. C. CAWLEY. Visual speech synthesis using statistical models of shape and appearance. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 78–83, 2001 ; HONG (P.), Z. WEN et T. S. HUANG. Real-time speech-driven face animation. *In* PANDZIC (I. S.) et R. FORCHHEIMER, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 7, pages 115–124. Wiley, 2002 ; AHLBERG (J.). *Model-based coding — Extraction, coding, and evaluation of face model parameters.* PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, septembre 2002.

27. MAEDA (S.). On articulatory and acoustic variabilities. *J. of Phonetics*, 19:321–331, 1991 ; BEAUTEMPS (D.), P. BADIN et G. BAILLY. Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *J. of the Acoustical Society of America*, 109(5):2165–2180, mai 2001 ; BADIN (P.), G. BAILLY, L. REVÉRET, M. BACIU, C. SEGEBARTH et C. SAVARIAUX. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *J. of Phonetics*, 30(3):533–553, 2002 ; MAEDA (S.), M. TODA, A. J. CARLEN et L. MAFTAHI. Functional modeling of the face during speech production. *In Actes des Journées d'études sur la parole*, pages 341–344, Nancy, France, juin 2002.

en tant que degrés de liberté des articulateurs de la parole — c.-à-d., du point de vue du contrôle du conduit vocal.

Les vecteurs de déplacement sont déterminés l'un après l'autre de la manière suivante.

1. La collection des données 3D est centrée autour de la posture moyenne.
2. (a) Une ACP est faite sur les coordonnées $y$ des points de l'arc mandibulaire et de LT (point à la jonction supérieure des incisives inférieures). La première composante de l'ACP est retenue comme le facteur $\alpha_1$.
   (b) Le vecteur de déplacement $\mathbf{M_{X1}}$ est déterminé en effectuant une régression linéaire des données (pour tous les points) sur le facteur $\alpha_1$.
   (c) La contribution de $\alpha_1$ est calculée comme $\mathbf{M_{X1}} \cdot \alpha_1$ ; cette contribution est soustraite des données pour la suite.
3. (a) Une ACP est faite sur les coordonnées $xyz$ des points des lèvres et des points correspondant aux billes du contour autour des lèvres. La première composante de l'ACP est retenue comme le facteur $\alpha_2$.
   (b) Le vecteur de déplacement $\mathbf{M_{X2}}$ est déterminé en effectuant une régression linéaire des données (pour tous les points sauf LT) sur le facteur $\alpha_2$.
   (c) La contribution de $\alpha_2$ est calculée comme $\mathbf{M_{X2}} \cdot \alpha_2$ ; cette contribution est soustraite des données pour la suite.
4. (a) Une ACP est faite sur les coordonnées $y$ des points de la lèvre inférieure et des points correspondant aux billes autour de la lèvre inférieure. La première composante de l'ACP est retenue comme le facteur $\alpha_3$.
   (b) Le vecteur de déplacement $\mathbf{M_{X3}}$ est déterminé en effectuant une régression linéaire des données (pour tous les points sauf LT) sur le facteur $\alpha_3$.
   (c) La contribution de $\alpha_3$ est calculée comme $\mathbf{M_{X3}} \cdot \alpha_3$ ; cette contribution est soustraite des données pour la suite.
5. (a) Une ACP est faite sur les coordonnées $y$ des points de la lèvre supérieure et des points correspondant aux billes autour de la lèvre supérieure. La première composante de l'ACP est retenue comme le facteur $\alpha_4$.

(b) Le vecteur de déplacement $\mathbf{M}_{\mathbf{X}4}$ est déterminé en effectuant une régression linéaire des données (pour tous les points sauf LT) sur le facteur $\alpha_4$.

(c) La contribution de $\alpha_4$ est calculée comme $\mathbf{M}_{\mathbf{X}4} \cdot \alpha_4$ ; cette contribution est soustraite des données pour la suite.

6. (a) Une ACP est faite sur les coordonnées $y$ des points des lèvres et des points correspondant aux billes du contour autour des lèvres. La première composante de l'ACP est retenue comme le facteur $\alpha_5$.

(b) Le vecteur de déplacement $\mathbf{M}_{\mathbf{X}5}$ est déterminé en effectuant une régression linéaire des données (pour tous les points sauf LT) sur le facteur $\alpha_5$.

(c) La contribution de $\alpha_5$ est calculée comme $\mathbf{M}_{\mathbf{X}5} \cdot \alpha_5$ ; cette contribution est soustraite des données pour la suite.

7. (a) Une ACP est faite sur les coordonnées $z$ des points de l'arc mandibulaire et de LT. La première composante de l'ACP est retenue comme le facteur $\alpha_6$.

(b) Le vecteur de déplacement $\mathbf{M}_{\mathbf{X}6}$ est déterminé en effectuant une régression linéaire des données (pour tous les points) sur le facteur $\alpha_6$.

(c) La contribution de $\alpha_6$ est calculée comme $\mathbf{M}_{\mathbf{X}6} \cdot \alpha_6$ ; cette contribution est soustraite des données pour la suite.

8. (a) Une ACP est faite sur les coordonnées $xyz$ de tous les points sur la gorge. La première composante de l'ACP est retenue comme le facteur $\alpha_7$.

(b) Le vecteur de déplacement $\mathbf{M}_{\mathbf{X}7}$ est déterminé en effectuant une régression linéaire des données (pour tous les points sauf les lèvres et LT) sur le facteur $\alpha_7$.

(c) La contribution de $\alpha_7$ est calculée comme $\mathbf{M}_{\mathbf{X}7} \cdot \alpha_7$ ; cette contribution est soustraite des données pour la suite.

Il est temps d'expliquer les choix de notre supervision. Ce sont des adaptations *ad hoc* de choix déjà faits dans les études précédentes [28] et auxquelles nous

28. Beautemps (D.), P. Badin et G. Bailly. Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *J. of the Acoustical Society of America*, 109(5):2165–2180, mai 2001 ; Badin (P.), G. Bailly, L. Revéret, M. Baciu, C. Segebarth et C. Savariaux. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *J. of Phonetics*, 30(3):533–553, 2002.

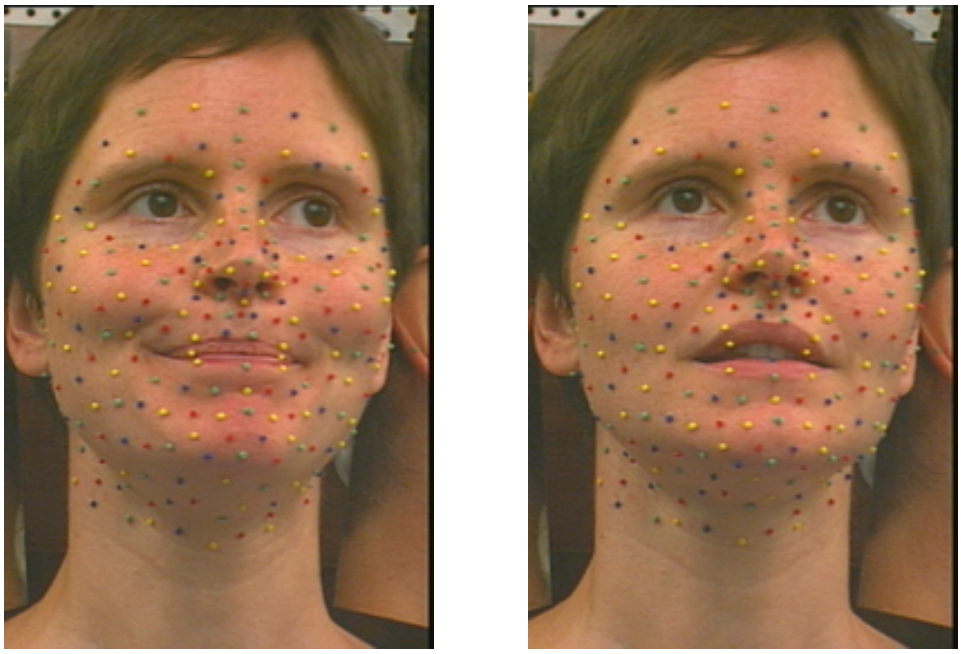


FIGURE 1.5 – Deux visèmes typiques illustrant la nécessité du paramètre lips4 : entre [epe] (à gauche) et [asa] (à droite), les différences géométriques ne sont pas dues qu'à l'articulation de la parole dite « neutre » mais aussi à l'esquisse d'un sourire. Ces deux visèmes sont le maximum (à gauche) et le minimum (à droite) du paramètre lips4 sur la base d'apprentissage.

renvoyons pour une justification complète. Simplement, ces choix sont inspirés des nombreuses études de la littérature consacrées aux degrés de liberté observés des articulateurs :

– la mâchoire est un objet rigide qui semble n'avoir que deux degrés de liberté *fonctionnels* en parole [29] ; le geste d'abaissement de la mâchoire est un geste porteur des lèvres (et de la langue), c'est pourquoi il vient en premier dans la modélisation ; ce n'est pas le cas du deuxième degré de liberté,

29. P. ex., deux paramètres expliquent 97 % de la variance de LT dans BEAUTEMPS (D.), P. BADIN et G. BAILLY. Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *J. of the Acoustical Society of America*, 109(5):2165–2180, mai 2001.

qui vient après les mouvements des lèvres ;
- pour les langues et les locuteurs que nous avons étudiées (français, allemand et arabe) les lèvres ont trois degrés de liberté [30] ; mais, comme la locutrice a parfois souri durant la session d'enregistrement (voir la figure 1.5) un quatrième paramètre — $\alpha_5$ — sans signification autre en parole a été introduit pour prendre en compte le mouvement vertical des commissures des lèvres et des joues propre au sourire ;
- un paramètre de contrôle supplémentaire doit enfin être introduit pour la gorge afin de reproduire un mouvement (*résiduel* vis à vis de la mâchoire et des lèvres) lié au gonflement du dessous du menton.

Ces choix faits nous permettent de *nommer* les facteurs $\alpha_i$. Nous les appelons paramètres *articulatoires*, de $\alpha_1$ à $\alpha_7$ : jaw1 ; lips1 ; lips2 ; lips3 ; lips4 ; jaw2 ; et lar1.

Il est à noter que seule une partie du corpus a été considérée pour la construction du modèle : nous avons conservé seulement les 14 voyelles et 54 consonnes, principalement en contexte { i, a, u }. Le modèle est donc bâti à partir d'une collection de 275 points 3D pour 68 visèmes.

Faisons enfin un résumé pratique : le modèle de la forme du visage de la personne étudiée est 3D, linéaire et piloté par un jeu de sept paramètres articulatoires $\boldsymbol{\alpha}$ indépendants ; de plus, la position et l'orientation de la tête sont définies par une translation $\boldsymbol{t}$ et une rotation exprimée avec les trois angles d'Euler $\boldsymbol{r}$. Le modèle du visage est donc entièrement contrôlé par le jeu de paramètres $\boldsymbol{p}$ :

$$\mathbf{X} = \begin{pmatrix} X_1 & Y_1 & Z_1 & \dots & X_N & Y_N & Z_N \end{pmatrix}^T = \mathbf{X_0} + \mathbf{M_X} \cdot \boldsymbol{\alpha} \tag{1.11}$$

$$\boldsymbol{p} = \begin{pmatrix} \boldsymbol{\alpha}^T & t_x & t_y & t_z & r_x & r_y & r_z \end{pmatrix}^T \tag{1.12}$$

## 1.6. RÉSULTATS

### 1.6.1. Précision des modèles

Le tableau 1.1 détaille les contributions de chaque paramètre articulatoire à la réduction de variance des données 3D ; est présenté le modèle de la locutrice

30. Revéret (L.). *Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, mai 1999.

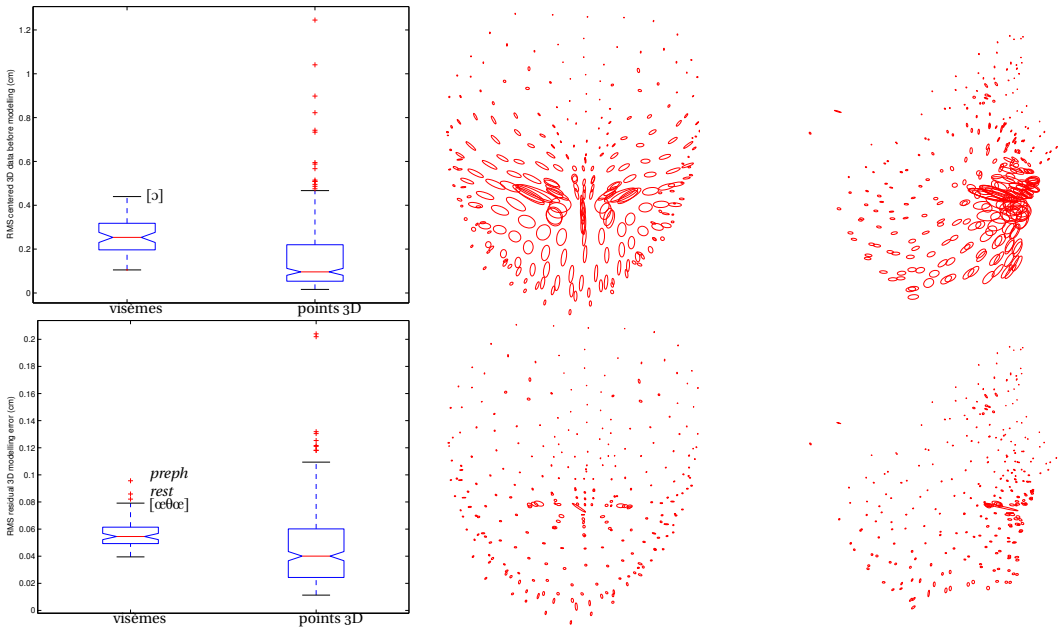



FIGURE 1.6 – Données 3D d'apprentissage (en haut) et résidu après modélisation articulatoire (en bas) ; à gauche : boîtes à moustaches des déviances RMS (en cm) — les deux échelles sont différentes en haut et en bas ; à droite : ellipses de dispersion correspondantes.

française *hl* qui est utilisé comme illustration dans toute cette thèse, mais aussi trois autres modèles individuels que nous avons construits suivant la même méthode. À chaque fois, le modèle explique plus de 93 % de la variance des données — pour *hl*, 96 %. Le détail des contributions montre des différences qui reflètent celles entre les langues concernées — français, allemand et arabe — et donc entre les corpora, entre les morphologies des locuteurs et de la locutrice, et entre leurs stratégies articulatoires.

Comme on le voit sur la figure 1.6, le modèle articulatoire peut reproduire fidèlement la géométrie des visèmes. Les ellipses de dispersion montrent les deux premiers modes de variation de tous les points ; les variations avant modélisation sont de bien plus grande ampleur que les variations résiduelles. À gauche sur la figure, les mêmes données sont représentées avec des boîtes à moustaches. L'erreur médiane de modélisation est d'environ 0,5 mm ; cette erreur est inférieure

TABLEAU 1.1 – Contribution de chaque paramètre articulatoire à la réduction de variance des données articulatoires 3D (en %). Les modèles individuels d'une locutrice (noté *hl*) et de trois locuteurs (pour le français, l'arabe et l'allemand) sont présentés, ainsi que le nombre de points compris dans chaque modèle.

| Langue | N | Paramètre articulatoire | | | | | | | Cumul |
|---|---|---|---|---|---|---|---|---|---|
| | | jaw1 | lips1 | lips2 | lips3 | lips4 | jaw2 | lar1 | |
| *hl* | 275 | 16,1 | 50,1 | 10,2 | 9,6 | 6,7 | 1,6 | 2,0 | 96,3 |
| français | 198 | 30,3 | 57,1 | 4,5 | 3,6 | | 0,4 | 0,8 | 96,9 |
| arabe | 231 | 14,8 | 69,4 | 4,1 | 2,7 | | 1,8 | 1,6 | 94,4 |
| allemand | 283 | 32,0 | 31,3 | 13,2 | 10,9 | | 4,7 | 1,3 | 93,3 |

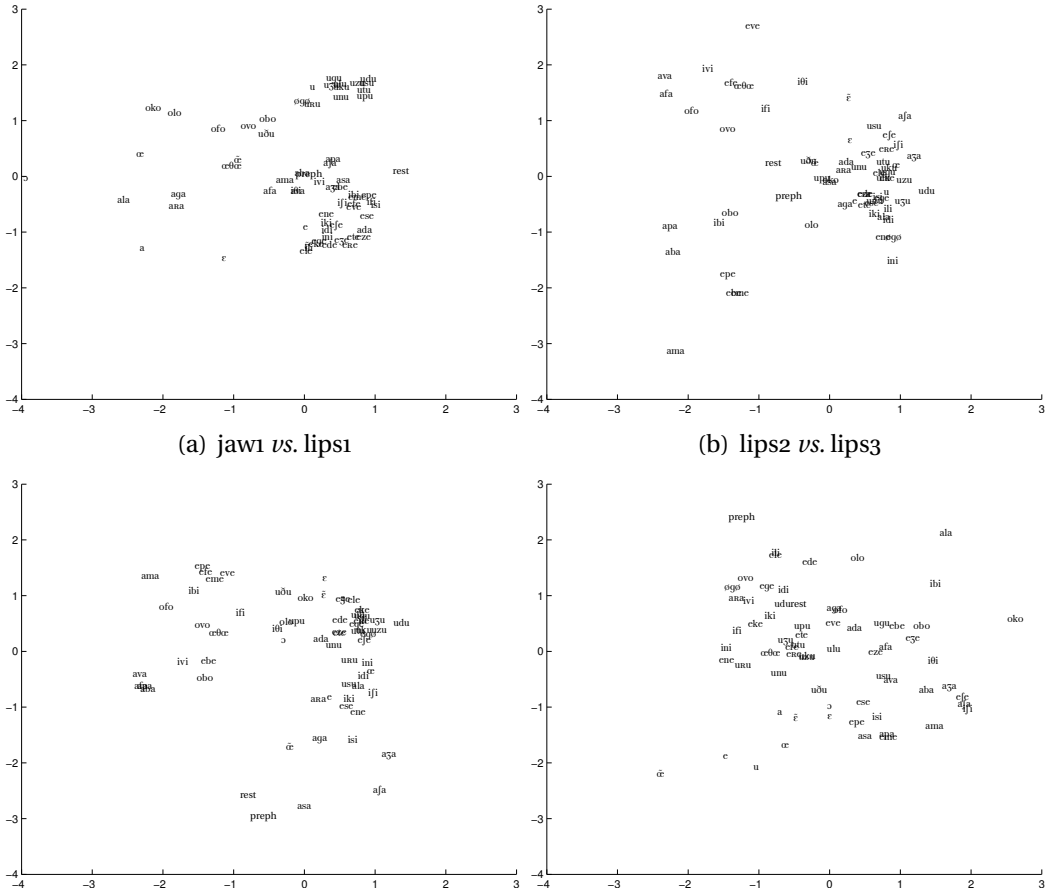
à 0,6 mm pour 75 % des visèmes. Les visèmes pour qui la modélisation est la moins bonne sont *preph*, *rest* et [œθœ] ; il s'agit des deux postures *silencieuses* et d'un visème « composé » d'une voyelle ([œ]) et d'une consonne ([θ]) représentées seulement deux fois dans la partie du corpus utilisé pour la construction. Même pour ces trois visèmes l'erreur de modélisation est inférieure à 1 mm. Les points pour qui la modélisation est la moins bonne appartiennent au contour interne des lèvres : comme ce contour n'est pas défini physiquement — il peut dépendre de la prise de vue — ce sont les points pour lesquels l'expert qui ajuste le modèle de surface des lèvres peut éprouver des difficultés à ancrer avec précision des points de chair.

## 1.6.2. Interprétations articulatoires

La figure 1.7 montre les projections des visèmes sur quatre plans factoriels — les facteurs ne sont pas corrélés [31]. Étant donnés les effets des paramètres articulatoires, qui sont détaillés ci-après, les observations sont en accord avec les connaissances phonétiques : p. ex., pour les traits articulatoires principaux, on distingue dans le plan jaw1–lips1 des classes pour les contextes vocaliques {a (consonnes en lieu arrière), i–e, u, o} ; dans le plan lips2–lips3, pour les labiodentales, les bilabiales, les fricatives labialisées.

La figure 1.8 représente les effets de chaque paramètre articulatoire pour les

31. Cette figure est plus lisible dans la version électronique de ce document qui permet le zoom.

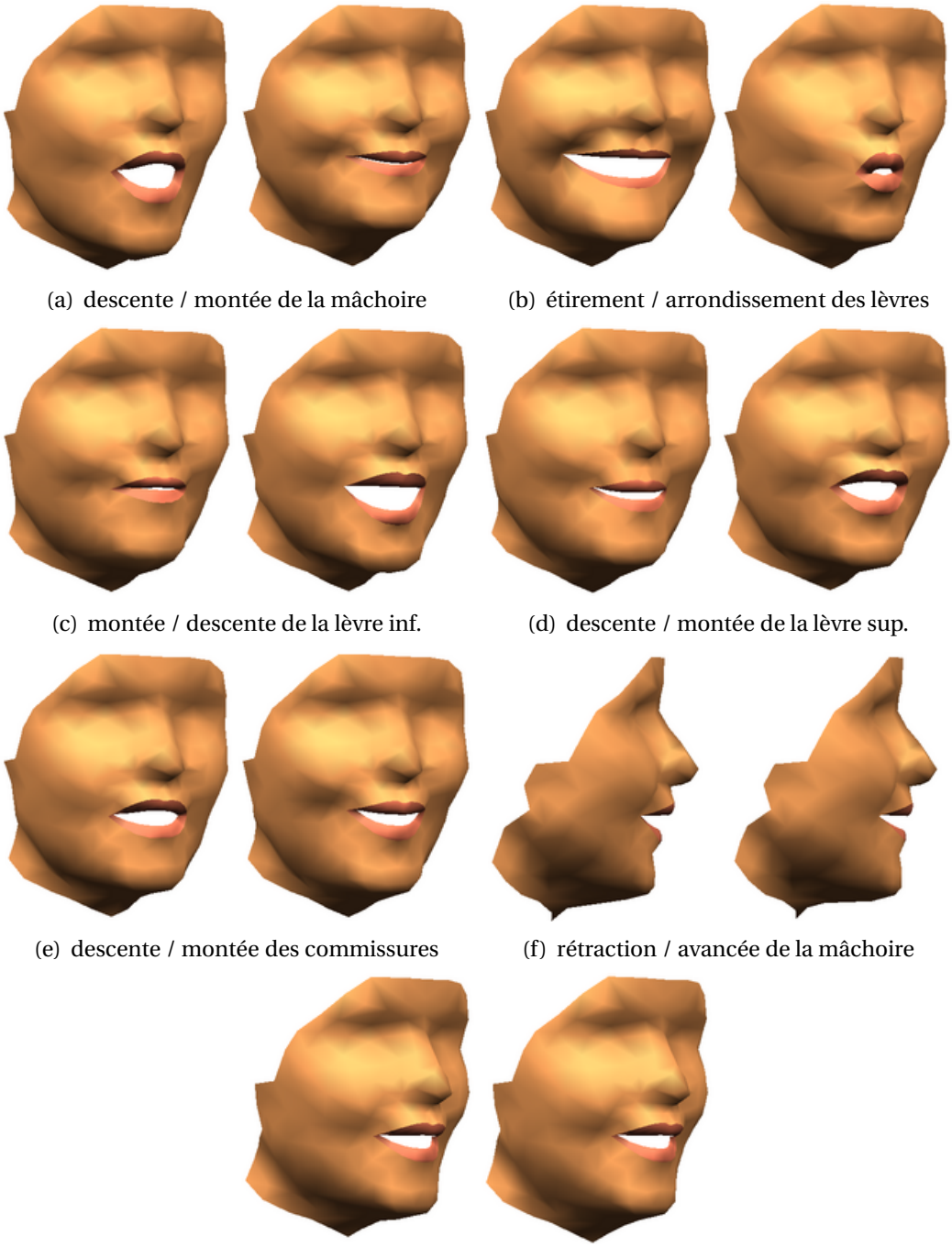


(a) jaw1 *vs.* lips1 (b) lips2 *vs.* lips3

(c) lips2 *vs.* lips4 (d) jaw2 *vs.* lar1

FIGURE 1.7 – Projection des visèmes sur des plans factoriels du modèle articulatoire.

valeurs ±2 (*Nota* : les paramètres sont normalisés par leurs écart-types) ; *a posteriori* on constate que ces mouvements élémentaires ont une interprétation phonétique claire : les nomogrammes pour jaw1 n'appellent pas de commentaire car ces gestes étaient attendus que la montée et descente de la mâchoire ; lips1 pilote l'étirement et l'arrondissement des lèvres et sera recruté pour la gamme des contextes vocaliques /i-u/ ; lips2 pilote la montée et descente de la lèvre inférieure et sera recruté notamment pour les fricatives labiodentales (/f/, /v/) ; lips3 pilote la montée et descente de la lèvre supérieure et sera recruté notamment pour les fricatives labialisées (/ʃ/, /ʒ/) ; comme cela était désiré lips4 n'a lui



(a) descente / montée de la mâchoire (b) étirement / arrondissement des lèvres

(c) montée / descente de la lèvre inf. (d) descente / montée de la lèvre sup.

(e) descente / montée des commissures (f) rétraction / avancée de la mâchoire

(g) montée / descente du larynx

FIGURE 1.8 – Les mouvements élémentaires du modèle articulatoire. De (a) à (g) : nomogrammes ($\pm 2\sigma$) de jaw1, lips1, lips2, lips3, lips4, jaw2 et lar1.

aucune interprétation phonétique et pourrait traduire l'effet d'un sourire (voir aussi la figure 1.5) ; jaw2 pilote la rétraction et l'avancée de la mâchoire et sera recruté notamment pour les fricatives labiodentales afin de positionner la lèvre inférieure sous les incisives supérieures ; enfin lar1 pilote la montée et descente du larynx et agit surtout sur le dessous du menton. Des interprétations claires ont aussi été constatées pour les modèles des trois autres locuteurs.

Par combinaison linéaire pondérée de ces postures articulatoires élémentaires, le modèle 3D devra être suffisamment généralisable pour pouvoir reproduire l'ensemble des formes 3D que peut prendre le visage de la locutrice en parole dynamique. En inversant le modèle sur des postures articulatoires hors de la base d'apprentissage, nous avons observé deux groupes de formes que le modèle ne sait pas reproduire : ces formes sont produites en parole dynamique où, d'une part, parfois la locutrice descend la mâchoire selon un axe qui n'est pas exactement vertical — ou du moins différent de celui de jaw1 — ; d'autre part après une occlusion bilabiale les lèvres peuvent se décoller de manière asymétrique. Le premier cas pourrait être réglé en rajoutant les visèmes *ad hoc* lors de l'apprentissage et en introduisant un troisième degré de liberté pour la mâchoire [32] ; le deuxième par l'utilisation d'un modèle biomécanique des lèvres [33]. En l'état, le modèle *régularise* ces formes non-canoniques.

32. C'est du reste la démarche adoptée par Maeda et coll. qui ont aussi fait de telles observations : Maeda (S.), M. Toda, A. J. Carlen et L. Maftahi. Functional modeling of the face during speech production. *In Actes des Journées d'études sur la parole*, pages 341–344, Nancy, France, juin 2002.

33. Kuratate mentionne une adaptation, encore en cours de développement, du modèle géométrique que nous avons utilisé qui permettrait selon leur auteur d'améliorer le rendu du contour interne des lèvres en prolongeant le modèle de surface à l'intérieur de la bouche : Kuratate (T.). *Talking head animation system driven by facial motion mapping and a 3D face database*. PhD thesis, Department of Information Processing, Nara Institute of Science and Technology, Nara, Japan, juin 2004. Avec une démarche similaire nous étions arrivé à des conclusions similaires lors de notre D.E.A. sans parvenir finalement à un modèle qui nous satisfît, notamment à cause de notre difficulté à obtenir des contours fiables sur les coupes IRM. Cependant un tel modèle pourrait mieux reproduire les phénomènes de collision ce qui permettrait de résoudre le problème initial avec l'ajout de visèmes *ad hoc* lors de l'apprentissage.

## 1.7. CONCLUSION

Nous avons présenté une méthodologie de construction de modèles tridimensionnels du visage capables de reproduire les gestes faciaux de parole d'un locuteur ou d'une locutrice étudié. De tels modèles sont pilotés linéairement par un petit nombre — 6 ou 7 — de paramètres indépendants, *articulatoires* parce qu'ils correspondent fonctionnellement à la parole. Ces paramètres sont déterminés en fonction des connaissances sur les degrés de liberté des articulateurs du conduit vocal. Les effets de ces paramètres — les mouvements élémentaires associés — émergent d'une analyse factorielle des corrélations observées sur des données capturées de la personne modélisée. Les modèles construits expliquent plus de 93 % de la variance initiale des données et les mouvements élémentaires sont pertinents sur le plan phonétique.

Si nous allons dans la suite utiliser ces modèles pour l'analyse de séquences vidéo, ils peuvent bien sûr être au cœur d'autres applications, notamment en synthèse audiovisuelle de la parole [34].

Pour améliorer le réalisme et l'attractivité de ces modèles, il serait intéressant d'étendre leur domaine de validité, aujourd'hui limité à la parole en contexte expressif neutre. Couvrir les autres expressions pourrait se faire suivant une démarche similaire mais en introduisant un mode supplémentaire pour les expressions dans la collecte des données et les analyses statistiques [35].

Le coût expérimental pour la construction des modèles reste important ; pouvoir conformer — tout en conservant finesse et précision — une famille de modèles existants, c.-à-d. un modèle pluri-locuteurs, à la morphologie et à l'articulation d'une nouvelle personne accroîtrait encore le domaine des applications possibles pour ces têtes parlantes [36].

34. Revéret (L.), G. Bailly et P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *In Proc. of the Internat. Conf. on Spoken Language Processing*, volume 2, pages 755–758, Beijing, China, octobre 2000 ; Attina (V.), D. Beautemps, M.-A. Cathiard et M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables : Rules for a Cued Speech synthesizer. *Speech Communication*, 44(1–4):197–214, octobre 2004 ; Gibert (G.), G. Bailly, D. Beautemps, F. Elisei et R. Brun. Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *J. of the Acoustical Society of America*, 118(2):1144–1153, 2005.

35. Kroonenberg (P. M.). *Three-mode principal component analysis. Theory and applications*. DSWO press, Leiden University, Netherlands, 1983.

36. Bérar (M.), G. Bailly, M. Chabanas, F. Elisei, M. Odisio et Y. Payan. Towards a generic talking head. *In Proc. of the Internat. Seminar on Speech Production*, pages 7–12, Sydney, Australia, décembre 2003.

# 2

# Modélisation de l'apparence du visage

## 2.1. INTRODUCTION

L'apparence d'un visage est fonction de nombreux facteurs, qui peuvent interagir : des mouvements faciaux d'abord ; des propriétés photométriques intrinsèques de la peau varient en fonction de *stimuli externes*, de l'environnement, comme la transpiration ou le hâle ; d'*états internes*, propres à la personne, comme la gêne ou des pleurs. Comme signal, l'image d'un visage dépend aussi des conditions de la prise de vues, notamment de la caméra, de sa position et de son orientation dans le repère facial — ou, le mouvement rigide de la tête — ainsi que des conditions d'illumination.

Nous présentons dans ce chapitre deux modèles de l'apparence du visage liés aux mouvements faciaux, c.-à-d. aux paramètres articulatoires du modèle de la forme du visage développé au chapitre précédent. Une modélisation directe entre forme et apparence est retenue ici. Le premier modèle est un modèle, classique de la texture du visage. Ensuite sera décrit un modèle de l'apparence locale (basé sur la notion de champs réceptifs couleur) d'un sous-ensemble pertinent des points 3D du modèle de forme.

L'acquisition des données audiovisuelles doit beaucoup à C. Savariaux et à A. Arnal. B. Holm, G. Gibert et F. Elisei ont participé à l'étiquetage de ces données. H. Lœvenbruck a accepté d'être sujet de cette étude.

## 2.2. MODÈLES DE LA TEXTURE DU VISAGE

### 2.2.1. *Eigenvisemes* : les modes principaux de variation de l'apparence

Les descriptions statistiques des images de visages ont d'abord été proposées pour la reconnaissance et la détection [1] ; les principaux modes de variation des images de visages ont été appelés *eigenfaces* ; ils sont déterminés par ACP sur les images en niveaux de gris et forment une base de l'espace des visages : par sa décomposition sur cette base il est possible de déterminer si une image appartient à l'espace des visages et desquels elle est la plus proche. Cette technique a ensuite été reprise pour la parole, avec les *eigenlips*, les principaux modes de variation d'images des lèvres ; en reconnaissance [2] comme en identification [3], il a été montré qu'une dizaine d'*eigenlips* peuvent représenter efficacement les images des lèvres.

Par combinaison linéaire, *eigenfaces* et *eigenlips* permettent de synthétiser des nouvelles images de visages ou de lèvres ; mais, comme les espaces image ne sont pas convexes, la synthèse peut aussi produire des *monstres*. Une manière d'éviter ce problème consiste à opérer sur les images une normalisation géométrique [4]. Ces images normalisées définissent alors des textures qui peuvent être

1. Turk (M.) et A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–86, 1991.

2. Bregler (C.) et Y. Konig. "eigenlips" for robust speech recognition. *In Proc. of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, pages 669–672, Adelaide, Australia, avril 1994 ; Revéret (L.). From raw images of the lips to articulatory parameters : a viseme-based prediction. *In Proc. of the European Conf. on Speech Communication and Technology*, volume 4, pages 2011–2014, Rhodes, Greece, septembre 1997.

3. Brooke (N. M.) et S. D. Scott. PCA image coding schemes and visual speech intelligibility. *In Proc. of the Institute of Acoustics, Autumn Meeting*, pages 123–129, Windermere, UK, 1994.

4. P. ex., Cootes (T. F.), G. J. Edwards et C. J. Taylor. Active appearance models. *In* Burkhardt (H.) et B. Neumann, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 1407 de *Lecture Notes in Computer Science*, pages 484–498, Freiburg, Germany, juin 1998. Springer-Verlag ; Ström (J.). *Model-based head tracking and coding*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2002 ; Theobald (B.-J.), J. A. Bangham, I. Matthews et G. C. Cawley. Visual speech synthesis using statistical models of shape and appearance. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 78–83, 2001 ; Blanz (V.) et T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, septembre 2003.

appliquées au modèle de forme. Nous allons maintenant vérifier que les images de postures articulatoires peuvent être normalisées vers un repère pour la texture et que dans ce référentiel les redondances sur les valeurs des texels [5] peuvent être recodées avec un petit nombre de paramètres.

### 2.2.1.1. *Normalisation géométrique 3D des images*

Un espace commun pour les textures est préalablement défini ; il correspond à une normalisation de chaque texture par rapport à une forme 3D de référence, souvent prise comme la posture neutre. Pour rapporter chaque image dans ce référentiel commun, un algorithme de déformation 3D [6] est utilisé ; en spécifiant les géométries — c.-à-d. les sommets du maillage 3D — de départ et d'arrivée, la technique du *texture mapping* permet de calculer les transformations de l'image d'origine [7]. En général, cette transformation n'est pas bijective ; c.-à-d., à un pixel destination il ne correspond pas un seul pixel source, mais plusieurs pixels — p. ex., cas d'une zone plus petite dans l'image destination — ou aucun : une opération de filtrage doit donc être effectuée. OpenGL propose pour cela deux possibilités : choisir le plus proche pixel ou faire un mélange pondéré linéaire des 2×2 pixels les plus proches. Nous avons retenu la première méthode de filtrage, plus simple et plus rapide, mais qui peut entraîner l'apparition d'*aliasing* [8]. Des transformations de l'image plus élaborées peuvent être appliquées [9].

La figure 2.1 montre un exemple de cette transformation. En plus de ramener les données RGB dans un espace où l'hypothèse de convexité est plus à même d'être vraie [10], cette normalisation permet de considérer des vecteur de données de taille constante, $3m$ si $m$ est le nombre de texels. La texture du visage $\mathbf{T}$ est :

$$\mathbf{T} = \begin{pmatrix} \mathbf{T_1}^{\mathrm{T}} & \mathbf{T_2}^{\mathrm{T}} & \cdots & \mathbf{T_m}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}} \tag{2.1}$$

où $\mathbf{T_i} = \begin{pmatrix} \mathrm{R}_i & \mathrm{G}_i & \mathrm{B}_i \end{pmatrix}^{\mathrm{T}}$ est le $i$-ème texel de la texture.

5. Ici, un texel est défini comme l'équivalent pour la texture d'un pixel pour une image.

6. ou *warping 3D*

7. Les calculs ont été effectués directement par la carte graphique 3D grâce à la bibliothèque graphique OpenGL ; nous ne rentrons pas ici dans le détail des opérations — notamment la gestion des cas particuliers.

8. Woo (M.), J. Neider et T. Davis. *OpenGL programming guide : The official guide to learning OpenGL, version 1.1.* Addison-Wesley, second édition, 1997.

9. P. ex., en utilisant des fonctions à base radiale.

10. Nous avons discuté ce problème, équivalent, pour la modèle articulatoire de la géométrie.

FIGURE 2.1 – Normalisation géométrique pour l'acquisition des données RGB d'apprentissage de texture. Le visème [a] est déformé vers la posture de référence.

TABLEAU 2.1 – Contribution des dix premiers *eigenvisemes* (et somme cumulée) à la réduction de variance des données RGB de texture (en %).

| *Corpus* | Eigenvisemes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *« billes »* | 38,2 (38,2 ) | 7,4 (45,6 ) | 5,3 (51,0 ) | 4,2 (55,1 ) | 3,6 (58,7 ) | 3,1 (61,8 ) | 3,0 (64,9 ) | 2,1 (67,0 ) | 1,9 (68,8 ) | 1,7 (70,5 ) |
| *« téléconférence »* | 28,4 (28,4 ) | 12,0 (40,4 ) | 5,9 (46,3 ) | 5,0 (51,3 ) | 4,3 (55,6 ) | 3,7 (59,3 ) | 3,3 (62,6 ) | 2,9 (65,5 ) | 2,8 (68,3 ) | 2,4 (70,7 ) |

### *2.2.1.2. Résultats*

Pour chacune des deux collections de textures, correspondant aux conditions expérimentales « billes » et « téléconférence », une ACP a été réalisée. Pour « billes » le corpus consiste en les 68 visèmes utilisés pour la construction du modèle articulatoire. Pour « téléconférence », 80 images ont été sélectionnées selon un algorithme *k-means* appliqué à l'ensemble des paramètres articulatoires échantillonnés aux centres des réalisations acoustiques des phonèmes du corpus — une

phrase a été écartée pour tester le suivi — ; une vérification *a posteriori* montre que pour les postures sélectionnées les paramètres articulatoires sont distribués de manière adéquate, sauf pour lips4 (la locutrice n'a pas « souri » sur ce corpus).

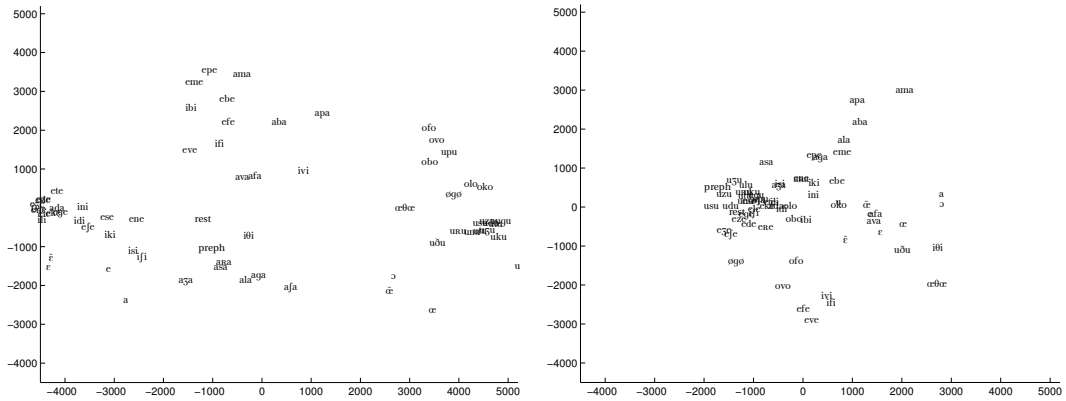


FIGURE 2.2 – Projection des visèmes pour « billes » sur les deux premiers plans factoriels issus de l'ACP. À gauche, 1 *vs.* 2 ; à droite, 3 *vs.* 4. Ces quatre facteurs expliquent plus de 55 % de la variance initiale des données. La figure de gauche montre que les données ne sont pas distribuées normalement (voir les deux classes « extrêmes » de part et d'autre du premier axe factoriel).

Le tableau 2.1 reprend les contributions des dix premiers *eigenvisemes* à la réduction de variance des données de textures. Pour les deux corpora les dix premiers *eigenvisemes* permettent d'expliquer plus de 70 % de la variance des données. Les contributions peuvent être jugées peu importantes ; nous croyons que cela est dû aux conditions controlées d'acquisition des données : les sources d'éclairage ne varient pas et la normalisation géométrique a notamment éliminé la zone à l'intérieur de la bouche. On peut voir sur la figure 2.5 (page 55) la variance des données avant et après soustraction des contributions des *eigenvisemes* ; le contraste a été rehaussé pour faciliter la visualisation — le résidu n'est sinon pas visible à l'œil sauf pour certaines frontières où l'on observe de l'*aliasing*. Pour « billes », les zones de variations les plus importantes sont : les sillons nasogéniens (cette ride disparaît pour les articulations où les lèvres sont arrondies) ; le dessous de la lèvre inférieure (une ombre portée apparaît pour les articulations où les lèvres sont arrondies) ; le dessous du menton (dont l'apparence varie lors

FIGURE 2.3 – *Eigenvisemes* pour 68 images du corpus « billes » résultant de l'analyse en composantes principales de la texture du visage après normalisation géométrique. De gauche à droite, les dix premiers modes de variation de l'apparence expliquent plus de 70 % de la variance initiale des données. Pour la visualisation, les images ont été normalisées vectoriellement.

des montées et descentes de la mâchoire) ; les billes (les zones les plus bruitées lors de la déformation). Sauf les billes, retirées des images pour l'évaluation du suivi, pour « téléconférence » les zones majeures de variation sont similaires.

Les *eigenvisemes* sont représentés sur les figures 2.3 et 2.4. Pour « billes », le premier vecteur est lié à l'apparition de l'ombre sous la lèvre inférieure conjointement avec la disparition des rides du sillon naso-génien, c.-à-d. au paramètre articulatoire lips1 qui expliquait la moitié de la variance des données 3D ; le deuxième à la disparition de l'ombre sous le menton conjointement à l'apparition de zones ombrées sous les pommettes, c.-à-d. au paramètre articulatoire jaw1 ; le troisième reflète un changement global du niveau d'illumination du visage ; le quatrième est lié à l'apparition d'une ombre au dessus de la lèvre supérieure conjointement à une augmentation de la luminance des lèvres, peut-être typique des consonnes fricatives labialisées et des labiodentales ; l'interprétation des autres *eigenvisemes* est plus délicate, les quatre derniers donnent l'impression de coder principalement les erreurs d'alignement du modèle 3D . La figure 2.2 conforte ces interprétations ; elle montre les projections des visèmes

FIGURE 2.4 – *Eigenvisemes* pour 80 images du corpus « téléconférence » résultant de l'analyse en composantes principales de la texture du visage après normalisation géométrique. De gauche à droite, les dix premiers modes de variation de l'apparence expliquent plus de 70 % de la variance initiale des données. Les trous dans la texture correspondent aux zones du visage où quelques billes étaient collées ; ils n'ont pas été pris en compte dans le calcul. Pour la visualisation, les images ont été normalisées vectoriellement.

sur les deux premiers plans factoriels [11]. Les classes vocaliques {i, e} et {u, o} sont projetées de part et d'autre du premier axe factoriel, la classe vocalique {a} étant au centre ; le deuxième facteur distingue les consonnes frontales des autres ; le quatrième facteur distingue les bilabiales des labiodentales, en contexte non-arrondi. On note également le fait que de larges zones des deux plans ne sont pas atteintes par les visèmes ; cela pourrait vouloir dire que malgré la normalisation géométrique effectuée, l'espace de l'apparence du visage n'est pas convexe ; les zones non-atteintes correspondraient alors aux *monstres* que nous avons évoqués précédemment [12]. Ce problème serait résolu en considérant chacune des classes observées comme un sous-espace dans lequel serait faite une ACP [13].

Pour « téléconférence », les résultats sont plus en demi-teinte : alors que le traitement est vectoriel et contrairement au cas précédent où ce phénomène n'était que marginal, on observe des variations intra-*eigenviseme* des trois canaux couleur distinctes. Cela est probablement dû à la saturation des images

11. Cette figure est plus lisible dans la version électronique de ce document qui permet le zoom.

12. Nous n'avons cependant pas vérifié perceptivement cette hypothèse.

13. MOGHADDAM (B.) et A. PENTLAND. Probabilistic visual learning for object representation. *In* NAYAR (S. K.) et T. POGGIO, éditeurs. *Early visual learning*, pages 99–130. Oxford University, 1996.

d'origine. On retient également qu'à cette résolution le calcul rend compte des variations d'apparence des plis et replis des lèvres, alors que ces différences sont trop subtiles géométriquement pour que le modèle géométrique utilisé pour représenter la surface des lèvres soit capable de les traiter.

Ces résultats d'ACP mettent en évidence que les variations majeures de l'apparence du visage peuvent être représentées par un petit nombre de paramètres et que, bien que les images soient normalisées géométriquement, ces paramètres sont principalement liés d'une part aux paramètres articulatoires et d'autre part à des changements uniformes de luminance. Dans le cadre classique des *Active Appearance Models*[14] la modélisation est poursuivie en effectuant une ACP entre les modèles de forme et d'apparence afin d'éliminer les corrélations entre les deux modèles ; cela impose de pondérer les modèles de manière *ad hoc* pour les rendre plus homogènes. Il est aussi possible de conserver les deux modèles mais au prix d'un nombre accru des paramètres de contrôle du modèle de visage — ce qui alourdirait la tâche du système de suivi.

Nous avons vu que les variations de l'apparence sont dans notre cas une conséquence simple de changements articulatoires, c'est pourquoi les modèles de l'apparence que nous présentons dans le reste de ce chapitre ne sont pilotés que par les seuls paramètres articulatoires. Les effets principaux de l'illumination seront gommés lors du suivi en effectuant une normalisation en luminance des images[15].

14. Cootes (T. F.), G. J. Edwards et C. J. Taylor. Active appearance models. *In* Burkhardt (H.) et B. Neumann, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 1407 de *Lecture Notes in Computer Science*, pages 484–498, Freiburg, Germany, juin 1998. Springer-Verlag.

15. La méthode employée, simple, sera vue au chapitre 3. Des études s'attaquent à l'estimation pour l'analyse vidéo des effets importants de l'illumination — des variations fortes des conditions d'illuminations peuvent avoir un impact important sur l'image. Si l'on dispose de la géométrie 3D du visage, deux approches sont possibles, soit à partir d'un modèle d'éclairage *explicite*, comme des adaptations du modèle de Phong : Eisert (P.) et B. Girod. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5):70–78, septembre 1998 ; Blanz (V.) et T. Vetter. A morphable model for the synthesis of 3D faces. *In Proc. of SIGGRAPH*, Computer Graphics Proceedings, Annual Conference Series, pages 187–194. Addison Wesley, août 1999 ; Garcia (E.) et J.-L. Dugelay. Applications and specificities of synthetic/synthetic projective registration. *In Proc. of the IEEE Internat. Conf. on Multimedia and Exposition*, 2004, soit avec un modèle d'illumination *implicite*, sous la forme d'une carte pour le shading (résultant d'un mélange pondéré de cartes d'éclairage apprises). Cette carte de shading est multipliée avec la texture du visage : Gross (R.), I. Matthews et S. Baker. Fisher light-fields for face recognition across pose and illumination. *In DAGM*, volume 2449 de *Lecture Notes in Computer Science*, pages 481–489, Zürich, Switzerland, 2002. Springer-Verlag.

FIGURE 2.5 – Variance des données RGB de la texture du visage d'apprentissage. En haut, le corpus « billes » : variance des données avant modélisation, variance du résidu après modélisation articulatoire directe et variance du résidu après modélisation par les dix premières composantes de l'ACP ; en bas, le corpus « téléconférence » : variance avant et après modélisation articulatoire directe — le résidu après ACP, qui nécessite une quantité de mémoire virtuelle importante, n'est pas montré. Pour la visualisation, les images ont subi une même transformation affine vectorielle.

### 2.2.2. Un modèle articulatoire de la texture

Comme pour la forme du visage, la construction du modèle de texture est basée sur une analyse factorielle supervisée des données. Les facteurs de variation de la texture sont choisis comme étant les paramètres articulatoires, c.-à-d. les paramètres de contrôle du modèle de la forme du visage. Le modèle de la texture

s'écrit donc :

$$\mathbf{T} = \begin{pmatrix} \mathbf{T_1} \\ \mathbf{T_2} \\ \vdots \\ \mathbf{T_m} \end{pmatrix} = \mathbf{T_0} + \mathbf{M_T} \cdot \boldsymbol{\alpha} \tag{2.2}$$

où comme précédemment $\mathbf{T_i} = \begin{pmatrix} \mathrm{R}_i & \mathrm{G}_i & \mathrm{B}_i \end{pmatrix}^{\mathrm{T}}$ est le $i$-ème texel de la texture.

La construction du modèle $\mathbf{M_T}$ fait appel à la même procédure que précédemment pour collecter les données de texture ; puis, dans l'espace associé à la forme 3D de référence une régression multi-linéaire robuste [16] est effectuée des données sur les paramètres articulatoires. Un principe équivalent a déjà été employé avec succès pour relier des différences entre deux textures et une variation des paramètres de forme de visage [17].

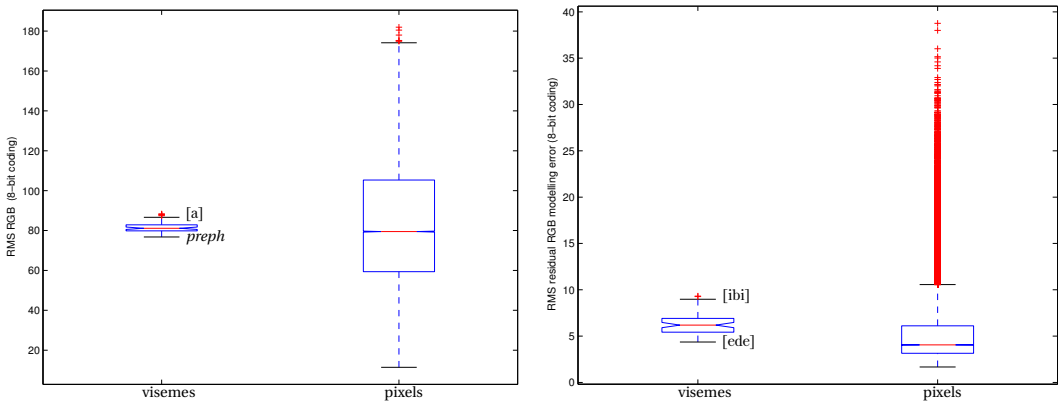



FIGURE 2.6 – Boîtes à moustaches des déviances RMS des données RGB d'apprentissage (à gauche) et résidu après modélisation articulatoire (à droite) — les deux échelles sont différentes. Voir aussi la figure 2.5.

16. Pour chaque composante RGB de chaque pixel une méthode itérative pondère les erreurs d'estimation afin de donner moins d'influence aux points aberrants.

17. COOTES (T. F.) et P. KITTIPANYA-NGAM. Comparing variations on the active appearance model algorithm. *In Proc. of the British Machine Vision Conf.*, volume 2, pages 837–846, Cardiff, UK, 2002.

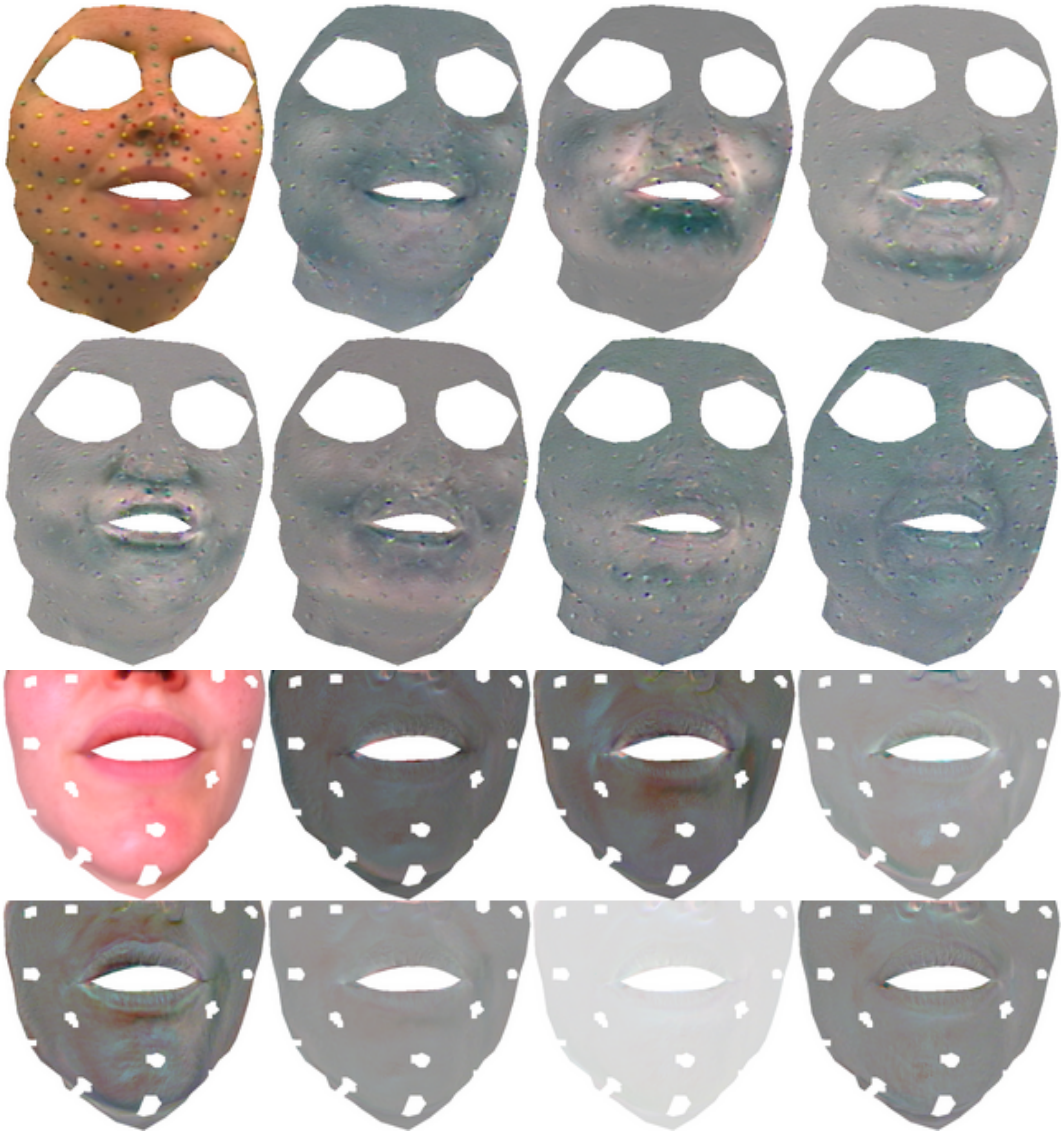


FIGURE 2.7 – Modèles articulatoires directs de l'apparence du visage résultant d'une régression linéaire multiple de la texture du visage après normalisation géométrique. En haut, le corpus « billes » où les sept paramètres articulatoires expliquent plus de 56 % de la variance initiale de 68 visèmes. En bas, le corpus « téléconférence » où les paramètres articulatoires expliquent plus de 40 % de la variance initiale de 80 postures articulatoires. Dans l'ordre de lecture : valeur moyenne, jaw1, lips1, lips2, lips3, lips4, jaw2, lar1. Pour la visualisation, les images ont été normalisées vectoriellement.

TABLEAU 2.2 – Contribution de chaque paramètre articulatoire à la réduction de variance des données d'apparence (en %). Quand cela a été possible, le calcul a aussi été effectué sur des données de test (absentes de la modélisation) ; le résultat ne présentait pas de différences majeures. Comme les paramètres articulatoires utilisés comme prédicteurs peuvent être corrélés, le cumul global n'est pas égal à la somme des contributions.

| *Corpus* | Apparence | Paramètre articulatoire | | | | | | | Cumul |
|---|---|---|---|---|---|---|---|---|---|
| | | jaw1 | lips1 | lips2 | lips3 | lips4 | jaw2 | lar1 | |
| *« billes »* | RGB | 5,5 | 36,0 | 5,8 | 3,6 | 3,7 | 1,7 | 2,3 | 56,8 |
| | LA | 6,6 | 26,0 | 7,8 | 4,0 | 4,3 | 2,1 | 2,6 | 50,0 |
| *« téléconférence »* | RGB | 6,4 | 10,8 | 9,1 | 8,2 | 9,5 | 8,5 | 2,8 | 40,2 |
| | LA | 15,5 | 8,5 | 17,2 | 8,2 | 14,4 | 5,3 | 1,6 | 49,5 |

#### *2.2.2.1. Résultats*

On peut voir sur la figure 2.6 les déviances RMS des données RGB avant et après modélisation articulatoire pour « billes ». Les visèmes sont bien représentés par la médiane ; après modélisation l'erreur résiduelle est comprise entre 4 et 10 (pour une valeur maximale théorique de 255). Pour 75 % des pixels, l'erreur résiduelle est inférieure à 6 : cette différence est à peine perceptible à l'œil, c'est pourquoi sur la figure 2.5 le contraste des images des variances ont été rehaussés. Cela permet de voir que les pixels les moins bien modélisés correspondent à des zones où le bruit dû à la déformation est le plus fort, à savoir les régions des billes et les commissures des lèvres — pour ce dernier cas, ce bruit de *warping* est principalement dû à des défauts d'alignement du modèle 3D (comme rapporté au chapitre précédent). Pour « téléconférence », en plus des remarques précédentes, on voit que les régions les moins bien modélisées sont l'intérieur des narines et les plis ou replis des lèvres ; le premier cas n'est pas important ; le deuxième montre que les variations les plus subtiles ne sont pas prises en compte par le modèle.

Le tableau 2.2 reprend les contributions des paramètres articulatoires à la réduction de variance des données. Les sept paramètres articulatoires permettent d'expliquer plus de 56 % de la variance pour « billes » soit l'équivalent de la contribution des cinq premiers *eigenvisemes* ; pour « téléconférence », 40 %, l'équivalent des deux premiers *eigenvisemes*.

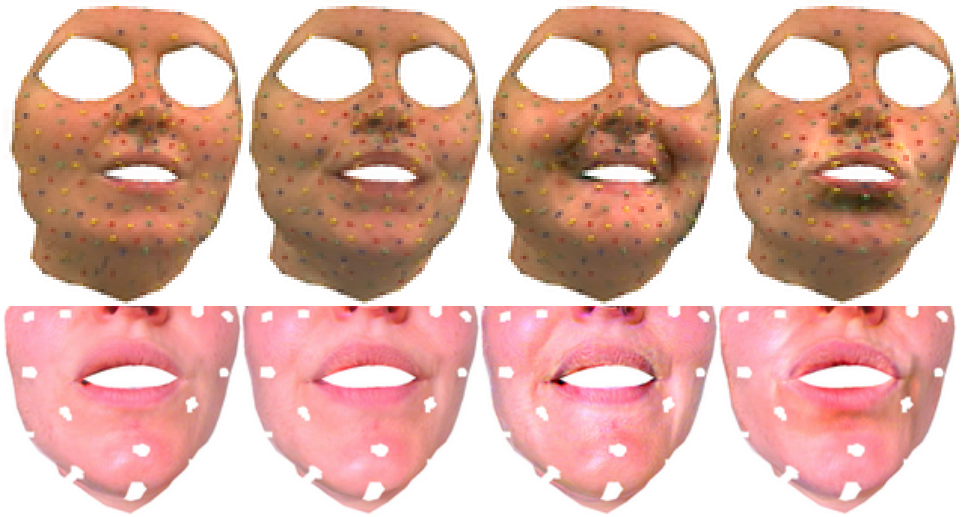


FIGURE 2.8 – Exemples de synthèse de textures sur la forme de référence 3D. Nomogrammes (±2σ) de jaw1 (paire de gauche) et de lips1 (paire de droite). Les mouvements articulatoires correspondants sont représentés sur la figure 1.8.

Si seule une partie de la variance des données est capturée par le modèle, on peut voir sur la figure 2.7 que l'effet des paramètres articulatoires est pertinent : pour « billes » on remarque les similarités déjà évoquées avec les *eigenvisemes* (voir les effets de jaw1 et lips1 en regard des deux premiers *eigenvisemes*) ; lips4 rend bien l'assombrissement du bas des pommettes lors du sourire ; les effets de jaw2 et lar1 sont plus délicats à interpréter. Pour « téléconférence », hors jaw1 et lips1, le modèle n'est pas interprétable ; p. ex. l'effet de lips2 semble lié à une variation uniforme de luminance. Rappelons que sur ce corpus les erreurs d'alignement du modèle 3D sont plus importantes [18] et que les postures articulatoires sont des postures articulées normalement — et non très clairement, comme pour « billes ». Il est possible que pour ces plus petites variations en parole normale les variations de l'apparence possèdent deux degrés de liberté pertinents.

18. Le modèle de la caméra rend moins compte de ses défauts et plus souvent la locutrice a produit des articulations non-atteignables par le modèle de forme — p. ex., la moitié droite des lèvres, notamment pour la lèvre supérieure, est plus « raide » ici ; la dissymétrie qui en résulte n'est pas modélisée.

Enfin, ces modèles de texture peuvent être utilisés pour un rendu fidèle et vidéo-réaliste en animation graphique. Comme le montrent les nomogrammes de la figure 2.8, les variations de l'apparence comme celles des sillons nasogéniens sont bien reproduites suivant que les lèvres sont plus ou moins étirées ou arrondies.

## 2.3. MODÉLISATION PAR DESCRIPTION LOCALE

Un visage sans marqueurs contient de larges régions où la texture n'apporte pas beaucoup d'information sur les mouvements articulatoires sous-jacents — p. ex., les joues —; certaines variations, comme dans la région des yeux, sont importantes mais ne sont pas dues à l'articulation.

Pour une application à l'analyse d'image, des gains en rapidité et peut-être en robustesse sont à attendre d'une modélisation de l'apparence limitée aux régions les plus informatives. Notre approche est basée sur une modélisation de l'apparence locale d'un ensemble sélectionné de points 3D : ici c'est un vecteur de caractéristiques qui pour une vue donnée décrit localement l'image là où se projette un point 3D. C'est sur le principe similaire en 3D aux ASM en 2D, utilisés pour l'ajustement sur des images de visages [19] ou pour la reconnaissance audiovisuelle de la parole [20] ; dans ce domaine et pour la reconnaissance de posture manuelle, il a déjà été proposé des représentations locales qui incorporent la notion d'échelle [21].

### 2.3.1. Famille des dérivées gaussiennes

Pour la description locale, notre choix s'est porté sur les premiers termes de la décomposition de Taylor du signal image ; les dérivées de ce signal sont calculées

19. Cootes (T. F.), D. Cooper, C. J. Taylor et J. Graham. Active Shape Models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

20. Luettin (J.) et N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

21. Matthews (I.), T. F. Cootes et J. A. Bangham. Extraction of visual features for lipreading. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):198–213, février 2002 ; parmi les travaux utilisant les filtres de Gabor : Nölker (C.) et H. Ritter. Visual recognition of continuous hand postures. *IEEE Trans. on Neural Networks*, 13(4):983–994, juillet 2002.

par convolution avec le noyau des dérivées gaussiennes [22]. Cette représentation locale a été utilisée pour la reconnaissance d'objets [23] et la détection de visage [24].

### 2.3.1.1. *Dérivées gaussiennes 1D*

Les filtres gaussiens sont séparables : par commodité nous pouvons donc travailler en 1D. Les formules discrètes pour les dérivées d'ordre 0 à 2 sont :

$$g_0(n) = \mathrm{A}_0 \mathrm{e}^{\frac{-n^2}{2\sigma^2}} \quad \text{avec} \quad \sum_{i=-\infty}^{+\infty} g_0(i) = 1 \tag{2.3}$$

$$g_1(n) = \mathrm{A}_1 \frac{-n}{\sigma^2} g_0(n) \quad \text{avec} \quad \sum_{i=-\infty}^{+\infty} i\, g_1(i) = 1 \tag{2.4}$$

$$g_2(n) = \mathrm{A}_2 \frac{n^2 - \sigma^2}{\sigma^4} g_0(n) \quad \text{avec} \quad \sum_{i=-\infty}^{+\infty} \frac{i^2}{2} g_2(i) = 1 \tag{2.5}$$

Les coefficients de normalisation $\mathrm{A}_i$ sont choisis de manière à assurer une réponse correcte si les filtres sont appliqués à des signaux polynomiaux [25].

22. D'abord introduit par Marr (D.) et E. Hildreth. Theory of edge detection. *Proc. of the Royal Society of London : Biological Sciences*, 207:187–217, 1980, il a été prouvé que ce noyau possédait des propriétés remarquables, permettant p. ex. de définir un espace d'échelle : Koenderink (J. J.) et A. J. van Doorn. The structure of images. *Biological Cybernetics*, 50:363–370, 1984 ; Lindeberg (T.). Feature detection with automatic scale selection. *Internat. J. of Computer Vision*, 30(2):77–116, 1998.Les fonctions gaussiennes permettent une caractérisation locale en résolution et en orientation de l'image ; c'est aussi le cas des ondelettes bidimensionnelles, comme les filtres de Gabor. Des bancs de filtres de Gabor (à diverses échelles et orientations) ont été couramment utilisés pour la reconnaissance faciale ; il serait donc intéressant d'évaluer aussi les performances de ces filtres dans notre système. L'emploi aisé des dérivées gaussiennes nous a paru propice à des expérimentations « exploratoires » et dans la perspective d'un suivi capable de s'appuyer suivant l'articulation — et la prise de vue — sur d'autres points d'intérêt ou d'analyser localement l'image à une échelle *intrinsèque articulatoire* ; nous n'avons toutefois pas mené ces expériences.

23. Schmid (C.), R. Mohr et C. Bauckhage. Evaluation of interest point detectors. *Internat. J. of Computer Vision*, 37(2):151–172, 2000 ; Hall (D.), V. Colin de Verdière et J. L. Crowley. Object recognition using coloured receptive fields. *In Proc. of the European Conf. on Computer Vision*, pages 164–177, Dublin, Ireland, juin 2000.

24. Vogelhuber (V.) et C. Schmid. Face detection based on generic local descriptors and spatial constraints. *In Proc. of the Internat. Conf. on Pattern Recognition*, volume 1, pages 1084–1087, Barcelona, Spain, septembre 2000.

25. Horaud (R.) et O. Monga. *Vision par ordinateur : outils fondamentaux*. Hermès, Paris, France, deuxième édition, 1995.

La figure 2.9 montre les premiers filtres dérivateurs gaussiens 2D jusqu'à l'ordre deux.

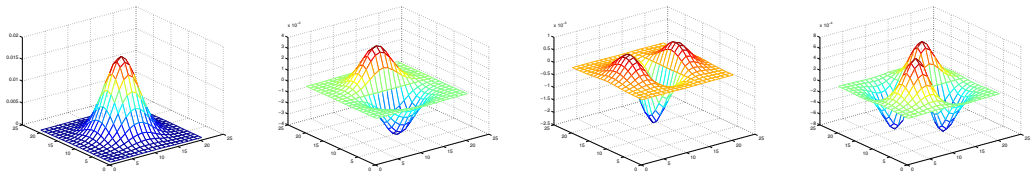


FIGURE 2.9 – Visualisation des filtres gaussiens jusqu'à l'ordre 2 appliqués sur les canaux couleurs pour décrire l'apparence locale : de gauche à droite, les coefficients des masques de convolutions des filtres $G_{00}$, $G_{01}$, $G_{02}$, $G_{11}$. Les filtres $G_{10}$ et $G_{20}$ sont identiques aux filtres $G_{01}$ et $G_{02}$ à une rotation près. Les filtres correspondent à une échelle $\sigma = 3$ et à une erreur de troncature $e = 0{,}000\ 01$. Rappelons que la transformée de Fourier d'une gaussienne est une gaussienne (de variance différente).

### *2.3.1.2. Filtres gaussiens chromatiques*

Ces filtres sont appliqués sur des images couleur ; plutôt que de travailler dans l'espace RGB nous avons retenu l'espace $YC_bC_r$ ; cet espace permet de distinguer plus aisément luminance et chrominance ; la meilleure interprétabilité des composantes couleur pourra être exploitée [26]. Les formules de passage de l'espace RGB vers l'espace $YC_bC_r$ sont :

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} a_R & a_G & a_B \\ \frac{-a_R}{2(1-a_B)} & \frac{-a_G}{2(1-a_B)} & \frac{1}{2} \\ \frac{1}{2} & \frac{-a_G}{2(1-a_R)} & \frac{-a_B}{2(1-a_R)} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.6}$$

où $a_R = 0{,}222\ 0$, $a_G = 0{,}706\ 7$ et $a_B = 0{,}071\ 3$. Soit finalement :

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} 0{,}222\ 0 & 0{,}706\ 7 & 0{,}071\ 3 \\ -0{,}119\ 5 & -0{,}380\ 5 & 0{,}500\ 0 \\ 0{,}500\ 0 & -0{,}454\ 2 & -0{,}045\ 8 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.7}$$

26. Il aurait été plus rigoureux scientifiquement d'utiliser aussi cet espace couleur $YC_bC_r$ pour les *eigenvisemes* ; il était plus rapide de travailler dans l'espace RGB pour notre implémentation logicielle appuyée sur la bibliothèque graphique OpenGL.

### 2.3.1.3. Taille des filtres

La troncature du filtre *idéal* introduit des erreurs dans le calcul de la convolution. La taille des filtres — des masques de convolution — est déterminée par une erreur de troncature fixée *a priori*[27]. Privilégiant la précision au détriment de la vitesse[28] des calculs intermédiaires, nous avons utilisé une erreur de troncature $e = 0{,}000\,01$.

### 2.3.1.4. Définition du vecteur décrivant l'apparence locale

Comme cela est représenté sur la figure 2.10, seules les dérivées gaussiennes jusqu'à l'ordre 1 sont retenues dans le vecteur **D** pour décrire l'apparence locale d'un point 3D — rappelons que les convolutions sont calculées à la projection dans le plan image du point. Nous avons constaté empiriquement que dans notre cas les dérivées d'ordres supérieurs étaient trop bruitées. Enfin, la dérivée d'ordre 0 de la luminance Y est écartée : c'est une manière de rendre **D** moins sensible aux changements d'illumination, et plus précisément ici invariant aux changements produisant un décalage en luminance ; il reste donc huit composantes pour **D** :

$$\begin{aligned}\mathbf{D} &= \begin{pmatrix} \mathrm{Y} \\ \mathrm{C_b} \\ \mathrm{C_r} \end{pmatrix} \star \begin{pmatrix} \mathrm{G}_0(\sigma_0) & \mathrm{G}_1^x(\sigma_0) & \mathrm{G}_1^y(\sigma_0) \end{pmatrix} \qu\setminus \quad \left\{ \mathrm{Y}_{\mathrm{G}_0(\sigma_0)} \right\} \\ &= \begin{pmatrix} \mathrm{Y}^x_{\mathrm{G}_1} \mathrm{Y}^y_{\mathrm{G}_1} \mathrm{C}_{\mathrm{b}_{\mathrm{G}_0}} \mathrm{C}^x_{\mathrm{b}_{\mathrm{G}_1}} \mathrm{C}^y_{\mathrm{b}_{\mathrm{G}_1}} \mathrm{C}_{\mathrm{r}_{\mathrm{G}_0}} \mathrm{C}^x_{\mathrm{r}_{\mathrm{G}_1}} \mathrm{C}^y_{\mathrm{r}_{\mathrm{G}_1}} \end{pmatrix}^{\mathrm{T}}\end{aligned} \tag{2.8}$$

où $\star$ signifie « convolution ». Dans nos expériences la valeur de l'échelle $\sigma_0$ est constante et égale à la valeur attendue pour les points du visage[29].

27. Étant donné un seuil d'erreur pour la troncature $e$, la taille du filtre $n$ est égale à $2m-1$ où $m$ est le plus petit entier supérieur à deux tel que $\mathrm{erfc}(\frac{m-1}{\sigma}) \leq 1 - \sqrt{1-e}$. Ce calcul ne tient pas compte de l'erreur due à la discrétisation. Pour une étude des conséquences théoriques de la discrétisation, voir Lindeberg (T.). Principles for automatic scale selection. *In Handbook on Computer Vision and Applications*, volume 2, pages 239–274. Academic Press, Boston, USA, 1999.

28. Étant donnés les nombreuses études sur les implémentations rapides d'approximations des filtres gaussiens, ce choix très conservatif semble exagéré.

29. Pour le corpus « billes », l'échelle retenue est de l'ordre de grandeur de la taille des marqueurs ; pour « téléconférence » nous avons multiplié cette valeur par le rapport d'échelle entre les deux prises de vue ($\sigma_0 = 5{,}417\,4$). Il serait possible d'introduire une plus grande sémantique dans le choix de l'échelle, en travaillant à l'échelle intrinsèque des points du visage ; voir Linde-

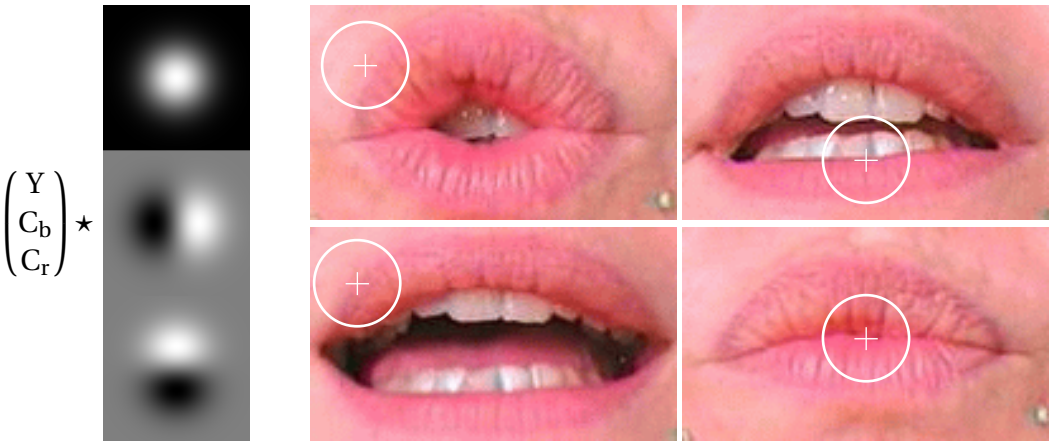



FIGURE 2.10 – Calcul de l'apparence locale. À gauche, la famille des filtres gaussiens appliqués à chacun des canaux Y, $C_b$ et $C_r$ — la dérivée d'ordre 0 de Y est écartée. À droite, des articulations différentes induisent des changements d'apparence.

## 2.3.2. Un contrôle articulatoire des descripteurs couleur locaux

Comme pour le modèle de texture, la construction du modèle de l'apparence locale est basée sur une analyse factorielle supervisée des données. Les facteurs de variation sont choisis comme étant les paramètres articulatoires, c.-à-d. les paramètres de contrôle du modèle de la forme du visage. Le modèle de l'apparence locale s'écrit donc :

$$\mathbf{LA} = \begin{pmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \vdots \\ \mathbf{D_l} \end{pmatrix} = \mathbf{LA_0} + \mathbf{M_{LA}} \cdot \boldsymbol{\alpha} \tag{2.9}$$

où $l$ est le nombre de points 3D inclus dans le modèle (voir ci-après) et $\mathbf{D_i} = \begin{pmatrix} Y^x_{G_1,i} & \cdots & C^y_{r_{G_1},i} \end{pmatrix}^T$ est le vecteur des descripteurs de l'apparence pour le $i$-ème point.

BERG (T.). Feature detection with automatic scale selection. *Internat. J. of Computer Vision*, 30 (2):77–116, 1998.

Comme précédemment le modèle est déterminé par une régression multilinéaire robuste des données sur les paramètres articulatoires.

Notons tout de suite qu'une relation linéaire n'est pas la plus adaptée pour représenter ces données ; l'emploi de méthodes non-linéaires requerrait une base d'apprentissage plus importante. Dans ce cas moins simple, les gradients devraient aussi être exprimés plus naturellement, en angle et amplitude : plutôt que le couple $(G^x; G^y)$, il serait plus efficace de considérer les descripteurs $\left(\theta = \arctan \frac{G^x}{G^y}; \|(G^x; G^y)\|_2\right)$.

### 2.3.2.1. *Sélection d'un sous-ensemble de points 3D*

Pour que le système gagne en vitesse, et peut-être en robustesse, il reste maintenant à ne conserver plus que les points les plus « informatifs ». La sélection est faite automatiquement, de la manière suivante : un critère permet pour chaque paramètre articulatoire de trier les points 3D ; seuls les 20 premiers points sont conservés. Le critère est la somme des variances des descripteurs de l'apparence locale expliquées par le paramètre articulatoire — d'autres choix sont bien sûr possibles. Le jeu des $l$ points 3D $\mathbf{X_{LA}}$ automatiquement sélectionnés consiste en la réunion des sous-ensembles pour chaque paramètre articulatoire.

### 2.3.2.2. *Résultats*

La figure 2.11 montre sur une vue de face le résultat de la sélection pour « téléconférence » (les points situés à proximité des quelques billes collées sur le visage ne prennent pas part au processus de sélection) ; les $l = 51$ points retenus semblent pertinents : ils sont principalement distribués sur les lèvres, la gorge et l'arc mandibulaire — il y a aussi un point sis au sommet du sillon naso-génien [30]. Sur ce corpus, la répartition des points retenus est dissymétrique [31] ; il semble alors intéressant dans le cas général d'introduire une étape d'affinage pour symétriser la sélection des points retenus, ou bien d'intégrer la symétrie comme contrainte lors du choix des points.

30. Le résultat pour « billes » n'est pas montré car la présence des billes biaise le processus de sélection : les petites erreurs d'ajustement et le bruit de *warping* ôte toute sémantique au résultat.

31. Cela est en partie dû à un moins bon ajustement de la moitié gauche du visage — qui est le reflet d'une « asymétrie physico-musculaire »— sur ce corpus.

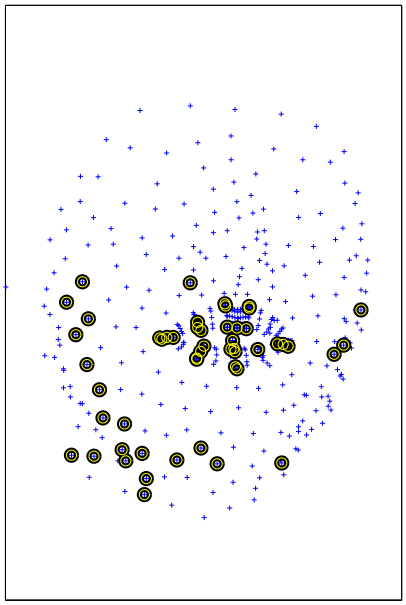


FIGURE 2.11 – Vue de face des points 3D ; les grosses pastilles indiquent les points sélectionnés automatiquement pour le modèle de l'apparence locale pour « téléconférence ».

Le tableau 2.2 (page 58) reprend les contributions des paramètres articulatoires à la réduction de variance des données. Les sept paramètres articulatoires permettent d'expliquer la moitié de la variance pour « billes » et pour « téléconférence » ; pour « téléconférence », ce modèle de l'apparence locale explique mieux la variance des données que le modèle de texture — cela peut être vu comme une vérification *a posteriori* du fait que sans marqueurs un petit groupe de régions permet de rendre compte des variations de l'apparence du visage.

## 2.4. LES MODÈLES DE L'APPARENCE UTILISÉS PAR LA SUITE

Pour chacune des conditions expérimentales « billes » et « téléconférence », un modèle de texture et un modèle de l'apparence locale ont été construits ; parce qu'ils varient linéairement en fonction des paramètres articulatoires ces deux modèles seront appelés par la suite *tex_lin* et *la_lin*. Pour comparaison, nous considérerons aussi deux modèles constants — c.-à-d. pour lesquels $\mathbf{M_T} = \mathbf{0}$ et $\mathbf{M_{LA}} = \mathbf{0}$ — qui seront appelés *tex_cst* et *la_cst*. Le modèle *tex_cst* correspond à la texture extraite sur une image de référence, comme il est classique, et non à la moyenne des textures sur la base d'apprentissage ; de même pour *la_cst* sur le corpus « billes » — sur le corpus « téléconférence », *la_cst* correspond à $\mathbf{LA} = \mathbf{LA_0}$ [32].

## 2.5. DISCUSSION

Nous avons présenté deux représentations de l'apparence du visage ; en se plaçant dans des conditions expérimentales contrôlées les variations de l'apparence résultent des changements articulatoires : les modèles de l'apparence construits sont contrôlés par les paramètres articulatoires, c.-à-d. par les paramètres du modèle de forme du visage, et ils permettent de rendre compte des principaux modes de variation des données d'apparence.

Plusieurs différences importantes existent entre les deux types de modèles d'apparence. Le modèle de texture (du domaine de l'informatique graphique) est un modèle complet de l'apparence du visage qui peut donc être utilisé pour le rendu vidéo-réaliste de séquences d'animation. L'utilisation du modèle d'apparence locale (du domaine de la vision par ordinateur) ne permet pas de synthétiser de nouvelles apparences : il est restreint aux traitements liés à l'analyse d'image.

Alors que le modèle de texture fait intervenir des déformations de l'image vers un espace de référence qui peuvent être importantes, le modèle de l'apparence locale ne fait pas intervenir de normalisation géométrique : plus dépendant de la

32. On verra au chapitre consacré à l'évaluation objective que ce modèle n'a pas eu à être utilisé pour le suivi ; cette différence de définition entre les deux corpora n'est donc pas importante ici.

prise de vue, il reflète localement la structure de l'image et est capable, p. ex., de faire la différence entre les lèvres et l'intérieur de la bouche suivant que les lèvres sont ouvertes ou fermées. Le choix de relations non-linéaires (réseau de modèles linéaires, réseau de neurones) des paramètres articulatoires aux descripteurs de l'apparence, qui nécessitera plus de données d'apprentissage, rendra mieux compte des observations. Comme le modèle de texture, le modèle d'apparence locale est utilisable à des résolutions différentes de celle d'apprentissage : il suffit pour cela de multiplier l'échelle à laquelle s'effectuent les convolutions par le rapport de la résolution de l'image analysée et de la résolution du corpus d'apprentissage. Il serait aussi possible d'utiliser des échelles différentes suivant les points considérés et leurs structures locales dans l'image [33].

Les modèles de l'apparence restent dans le cas général spécifiques à chaque condition d'enregistrement ; une modélisation plus détaillée de la scène qui inclurait des modèles explicites des sources d'illumination accroîtrait la robustesse aux variations d'éclairement ; pour la prise en compte des grands mouvements rigides de la tête, ou des grandes différences de la position et de l'orientation de la caméra, la construction d'un réseau, *multi-vues*, de modèles d'apparence, *mono-vues*, nous semble une piste intéressante [34] parce qu'elle conserverait les qualités intrinsèques de chacun des modèles ; en revanche, il faut noter que la taille d'un tel réseau croîtrait exponentiellement en fonction de nombre de degrés de liberté non-linéaires : cela constitue une limite de cette approche.

L'apprentissage automatique, non-supervisé, des modèles d'apparence serait bien sûr appréciable ; pour cela il serait possible d'adapter des méthodes « tout image » récemment proposées [35] en tirant profit de la connaissance *a priori* du modèle 3D ; le modèle d'apparence locale pourrait être basé sur — ou utilisé

33. Lindeberg (T.). Feature detection with automatic scale selection. *Internat. J. of Computer Vision*, 30(2):77–116, 1998.

34. Cootes (T. F.), K. N. Walker et C. J. Taylor. View-based active appearance models. *In Proc. of the Internat. Conf. on Face and Gesture Recognition*, pages 227–232, Grenoble, France, 2000 ; Fillbrandt (H.), S. Akyol et K.-F. Kraiss. Extraction of 3D hand shape and posture from image sequences for sign language recognition. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 181–186, Nice, France, octobre 2003.

35. Cootes (T. F.), S. Marsland, C. J. Twining, K. Smith et C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. *In* Pajdla (T.) et J. Matas, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 3024 de *Lecture Notes in Computer Science*, pages 316–327, Prague, Czech Republic, mai 2004. Springer-Verlag ; Baker (S.), I. Matthews et J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(10):1380–1384, 2004.

pour — le suivi dans l'image de quelques points d'ancrage sur le visage (p. ex., les contours et les commissures des lèvres) ; au fur et à mesure que le système gagne en précision et en robustesse cet ensemble de points pourrait s'enrichir de points additionnels sur le visage, pour conduire finalement à un modèle de texture linéaire [36].

36. WALKER (K. N.), T. F. COOTES et C. J. TAYLOR. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(5–6):435–440, avril 2002.

# 3

# Ajustement du modèle du visage aux images d'une séquence vidéo

## 3.1. INTRODUCTION

Dans les deux chapitres précédents, nous avons construit un modèle du visage, constitué d'un modèle de forme, articulé et en 3D, puis de deux types de modèles de l'apparence. Ce modèle du visage est au cœur du système qui fait l'objet de ce chapitre, un système qui vise à estimer, à suivre, les mouvements 3D depuis chaque image d'une séquence vidéo. Cette problématique d'ajustement d'un modèle 3D à une image est classique en vision par ordinateur, notamment pour l'estimation des mouvements faciaux ou la reconnaissance de visages au repos. Si c'est moins le cas dans le domaine de recherche « parole », il peut s'apparenter au classique problème de l'estimation de la géométrie du conduit vocal à partir du son, appelé inversion articulatoire — il s'agit alors de retrouver les gestes ayant produit un son donné.

L'introduction par les éditeurs d'un numéro spécial de *IEEE Computer Graphics and Applications* consacré au suivi débute ainsi : « *All we want is fast, « accurate, low-latency tracking of our head, hands, elbows, knees, or feet in a « cockpit, office, lab, or warehouse. Oh, and it should predict where we'll be a short « time in the future so that we compensate for delays in the rest of the system*[1]. » À cette liste d'exigences pléthorique, ajoutons sans hésiter la robustesse[2].

1. JULIER (S.) et G. BISHOP. Guest editors' introduction : Tracking : How hard can it be ? *IEEE Computer Graphics & Applications*, 22:22–23, novembre-décembre 2002.

2. au bruit, gaussien ou non ; à l'occultation, partielle ou complète, etc.

L'utilisation de modèles dans un système de suivi (audio-) visuel permet de contraindre et donc simplifier la tâche ; cela, sous l'hypothèse que chaque image de la séquence vidéo contient une projection 2D du modèle 3D. L'estimation des mouvements faciaux et des mouvements de tête revient à ajuster le modèle 3D de telle sorte que sa projection soit conforme aux informations de l'image.

## 3.1.1. Des ajustements de modèles à une image

### *3.1.1.1. Approches basées sur des points caractéristiques de l'image*

La posture 3D peut être estimée à partir des positions et de leurs mouvements dans l'image des projections d'un ensemble de points 3D du modèle. Ces points 3D sont soit des nœuds du modèle 3D — sommets du maillage, points de contrôle —, soit sont sélectionnés d'après leurs positions, connues, sur l'image initiale selon des critères qui jugent de leur capacité à « être suivis »[3]. Cet ensemble doit contenir suffisamment de points pour que l'inversion 2D vers les paramètres de contrôle du modèle 3D soit sur-déterminée : les contraintes liées au *flot optique* entre deux images de la séquence aboutissent à un problème linéaire fonction des paramètres du modèle 3D ; ces paramètres sont alors estimés par inversion au sens des moindres carrés. L'emploi d'une boucle fermée d'analyse par la synthèse permet d'éviter une trop grosse accumulation de l'erreur au cours de la séquence[4]. Une autre manière de boucler pour rendre le système plus robuste consiste à réaliser l'ajustement de manière successive sur une décomposition pyramidale, à résolution croissante, de l'image[5] ou en utilisant un filtre de

3. Ström et coll. (1999) définissent un critère basé sur le déterminant du hessien : Ström (J.), T. Jebara, S. Basu et A. Pentland. Real time tracking and modeling of faces : an EKF-based analysis by synthesis approach. *In Proc. of the Internat. Conf. on Computer Vision*, Corfu, Greece, septembre 1999. Voir aussi les travaux sur les points d'intérêt, p. ex. : Shi (J.) et C. Tomasi. Good features to track. *In Proc. of the Internat. Conf. on Computer Vision and Pattern Recognition*, pages 593–600, 1994 ; Schmid (C.), R. Mohr et C. Bauckhage. Evaluation of interest point detectors. *Internat. J. of Computer Vision*, 37(2):151–172, 2000 ; Bretzner (L.). *Multi-Scale Feature Tracking and Motion Estimation*. PhD thesis, Computational Vision and Active Perception Laboratory, KTH, Stockholm, Sweden, octobre 1999.

4. Li (H.), P. Roivainen et R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):545–555, juin 1993.

5. Eisert (P.) et B. Girod. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5):70–78, septembre 1998 ; Tao (H.) et T. S. Huang. Visual estimation and compression of facial motion parameters — Elements of a 3D model-based video coding

Kalman [6]. Des contraintes additionnelles sont parfois intégrées : DeCarlo et Metaxas combinent dans leur système des forces physiques liées à l'ajustement des arêtes du modèle 3D [7]. Enfin, ce type d'approches est présenté plus formellement dans (Li et Forchheimer, 2002) [8].

### *3.1.1.2. Approches basées sur des comparaisons d'image*

Au prix d'un accroissement des calculs requis, une autre approche, très répandue, consiste à utiliser cette fois tous les pixels pour l'ajustement d'un modèle de visage texturé. L'ajustement du modèle est effectué de manière itérative grâce à une boucle d'analyse par la synthèse qui nécessite — en plus du modèle — la définition d'une méthode de mesure de l'écart entre la synthèse et l'image analysée, et un algorithme de minimisation de cet écart. Cette méthode est employée pour estimer les mouvements de tête [9] et de parole [10], les expressions [11] et

system. *Internat. J. of Computer Vision*, 50(2):111–125, novembre 2002 ; Kroos (C.), T. Kuratate et E. Vatikiotis-Bateson. Video-based face motion measurement. *J. of Phonetics*, 30(3):569–590, juillet 2002 ; Cootes (T. F.), C. J. Taylor et A. Lanitis. Active Shape Models : Evaluation of a multi-resolution method for improving image search. *In Proc. of the British Machine Vision Conf.*, pages 327–336, 1994.

6. Ström et coll. (1999), *op. cit.*

7. DeCarlo (D.) et D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Internat. J. of Computer Vision*, 38(2):99–127, juillet 2000.

8. Li (H.) et R. Forchheimer. Model-based coding : The complete system. *In* Pandzic (I. S.) et R. Forchheimer, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 11, pages 187–218. Wiley, 2002.

9. La Cascia (M.), S. Sclaroff et V. Athitsos. Fast, reliable head tracking under varying illumination : an approach based on registration of texture-mapped 3D models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(4):322–336, avril 2000.

10. Revéret (L.), G. Bailly et P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *In Proc. of the Internat. Conf. on Spoken Language Processing*, volume 2, pages 755–758, Beijing, China, octobre 2000.

11. Pighin (F.), R. Szeliski et D. H. Salesin. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002. L'on trouve ici les méthodes dérivées des *AAM* (Cootes (T. F.), G. J. Edwards et C. J. Taylor. Active appearance models. *In* Burkhardt (H.) et B. Neumann, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 1407 de *Lecture Notes in Computer Science*, pages 484–498, Freiburg, Germany, juin 1998. Springer-Verlag) : Matthews (I.) et S. Baker. Active appearance models revisited. *Internat. J. of Computer Vision*, 60(2):135–164, 2004 ; Abboud (B.), F. Davoine et M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing : Image Communication*, 19:723–740, 2004 ; Yan (S.), C. Liu, S. Li, H. Zhang, H. Shum et Q. Cheng. Texture-constrained active shape

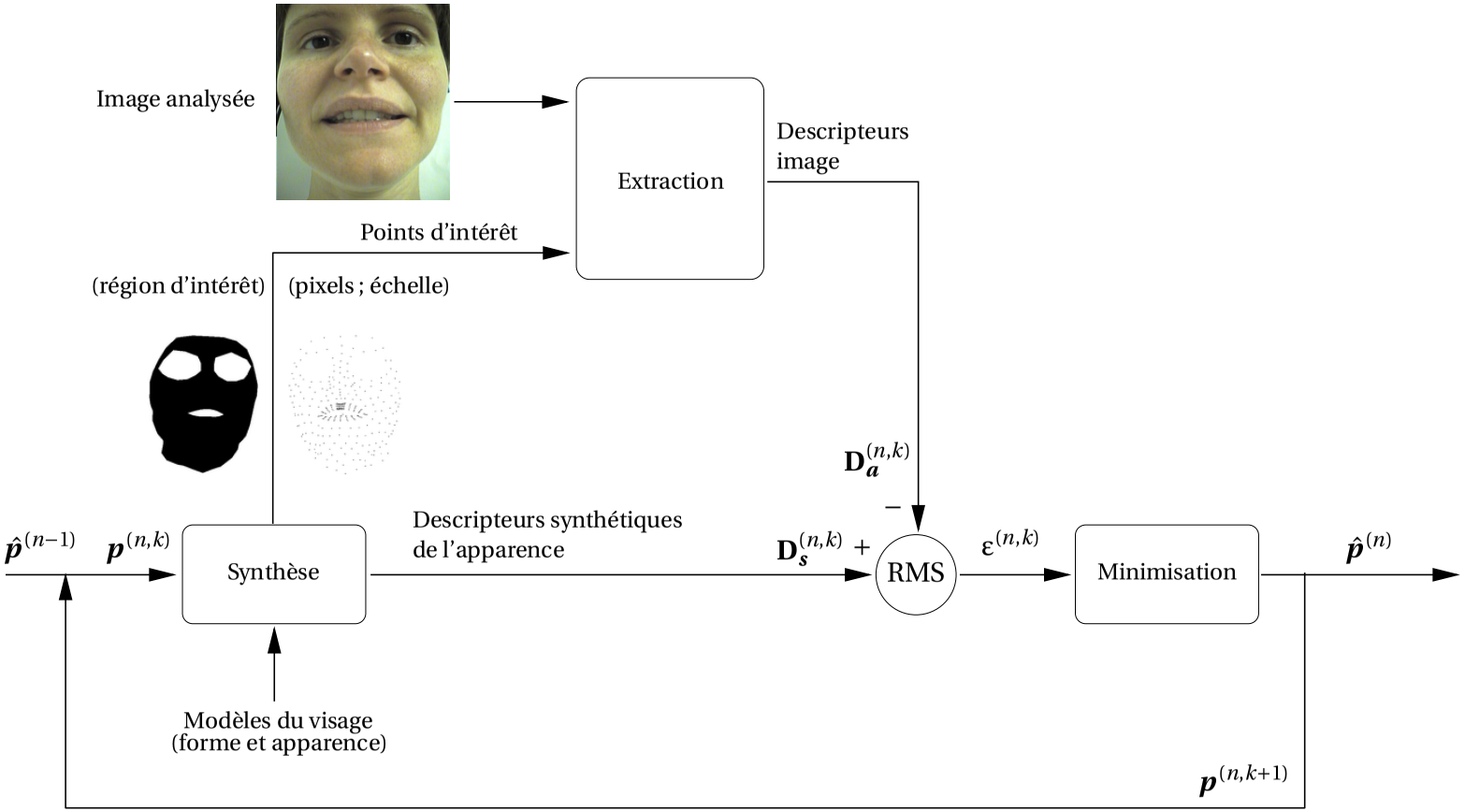



FIGURE 3.1 – Architecture du suivi utilisant une boucle d'analyse par la synthèse pour la $n$-ième image d'une séquence. À la $k$-ième itération, l'estimation courante des paramètres $\boldsymbol{p}$ permet de synthétiser un ensemble $\mathbf{D}_{\boldsymbol{s}}$ de descripteurs de l'apparence du visage, ainsi que de calculer l'ensemble $\mathbf{D}_{\boldsymbol{a}}$ des descripteurs correspondant sur l'image analysée. L'écart $\varepsilon$ entre l'analyse et la synthèse est fourni à l'algorithme de minimisation qui, si l'écart minimum n'est pas encore atteint, délivre une nouvelle estimation des paramètres $\boldsymbol{p}$ pour la $(k+1)$-ième itération.

pour la reconnaissance faciale [12].

C'est dans cette catégorie que se situe notre système ; la description qui suit introduit, motive et détaille les choix effectués pour la fonction d'écart et la méthode d'optimisation. Quand il le faudra nous reviendrons sur les travaux cités ici pour expliquer plus en détail certaines caractéristiques et différences pertinentes.

## 3.2. ÉCART ENTRE LA VIDÉO ET LE VISAGE SYNTHÉTIQUE

Comme représentée sur la figure 3.1, l'estimation (ou ajustement) des paramètres de contrôle est basée sur une boucle d'analyse par la synthèse. La différence entre l'analyse et la synthèse est utilisée pour guider la recherche de minima. Les modèles de la forme et de l'apparence du visage étant appris et les caméras étant calibrées [13], cette fonction d'écart $\varepsilon$ dépend des paramètres de contrôle ; nous détaillons ci-après comment s'exprime cette dépendance dans les deux cas correspondant aux deux modèles d'apparence du chapitre 2.

models. *In Proc. of The First Internat. Workshop on Generative-Model-Based Vision*, Copenhagen, Denmark, mai 2002 ; DORNAIKA (F.) et J. AHLBERG. Efficient active appearance model for real-time head and facial feature tracking. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, Nice, France, octobre 2003.

12. BLANZ (V.) et T. VETTER. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, septembre 2003 — il est à noter que pour accélérer la durée des calculs la fonction d'écart de ce système est basée seulement sur une sélection de 40 points définie aléatoirement à chaque itération —; ROMDHANI (S.), V. BLANZ et T. VETTER. Face identification by fitting a 3D morphable model using linear shape and texture error functions. *In Proc. of the European Conf. on Computer Vision*, volume 2353 de *Lecture Notes in Computer Science*, pages 3–19, Copenhagen, Denmark, mai 2002. Springer-Verlag.

13. En fait, notre système gère aussi l'intégration des paramètres du modèle des caméras au jeu des paramètres estimés.

### 3.2.1. Calcul de la fonction d'écart avec le modèle de texture

Les paramètres articulatoires sont notés $\boldsymbol{\alpha}$ ; les paramètres du mouvement rigide — translation et rotation — de la tête sont notés $\boldsymbol{t}$ et $\boldsymbol{r}$. Ces trois vecteurs sont regroupés dans le vecteur de contrôle du modèle de visage, $\boldsymbol{p}$. Les paramètres des modèles des caméras sont notés $\boldsymbol{cam}$. L'image analysée est notée $\mathbf{I}_{\boldsymbol{a}}$.

Dans l'espace texture, normalisé par rapport à une forme 3D de référence, la texture du visage $\mathbf{T}$ est (voir l'équation (2.2), page 56)

$$\mathbf{T} = \begin{pmatrix} \mathbf{T_1} \\ \mathbf{T_2} \\ \vdots \\ \mathbf{T}_{\boldsymbol{m}} \end{pmatrix} = \mathbf{T_0} + \mathbf{M_T} \cdot \boldsymbol{\alpha} \tag{3.1}$$

$\mathscr{W}$ est l'opérateur qui applique la texture sur la forme 3D et calcule la projection de cette forme texturée sur le plan image. L'image de synthèse est

$$\mathbf{I}_{\boldsymbol{s}}(\boldsymbol{p}) = \mathscr{W}\left(\mathbf{T}(\boldsymbol{\alpha}); (\boldsymbol{t}; \boldsymbol{r}; \boldsymbol{cam})\right) \tag{3.2}$$

Le support de projection de la synthèse — de taille variable — est

$$\mathscr{S}(\boldsymbol{p}) = \left\{(u, v) \in \mathbb{N}^2 \,\middle|\, \mathbf{I}_{\boldsymbol{s}}(\boldsymbol{p})(u, v) \text{ existe}\right\} \tag{3.3}$$

Afin de limiter les effets d'éventuelles différences dans les conditions d'éclairement, les images analysée et synthétique sont normalisées en intensité avant d'être comparées : pour les pixels non-noirs les valeurs RGB sont divisées par $\mathrm{L} = \frac{\mathrm{R+G+B}}{3}$ (nous ne décrivons pas plus formellement cette opération pour ne pas alourdir inutilement les notations). La différence entre analyse et synthèse $\boldsymbol{e}$ est alors

$$\boldsymbol{e}(\boldsymbol{p}) = \mathbf{I}_{\boldsymbol{a}}\left(\mathscr{S}(\boldsymbol{p})\right) - \mathbf{I}_{\boldsymbol{s}}(\boldsymbol{p}) \tag{3.4}$$

La taille du support de $\boldsymbol{e}$ n'est pas fixe, elle dépend de $\boldsymbol{p}$.

## 3.2.2. Calcul de la fonction d'écart avec le modèle de l'apparence locale

L'apparence locale $\mathbf{LA}_{\boldsymbol{s}}$ des $l$ points 3D $\mathbf{X}_{\mathbf{LA}}$ automatiquement sélectionnés pour l'estimation des mouvements est (voir l'équation (2.9), page 64)

$$\mathbf{LA}_{\boldsymbol{s}} = \begin{pmatrix} \mathbf{D_1} \\ \mathbf{D_2} \\ \vdots \\ \mathbf{D}_{\boldsymbol{l}} \end{pmatrix} = \mathbf{LA_0} + \mathbf{M_{LA}} \cdot \boldsymbol{\alpha} \tag{3.5}$$

$\mathscr{P}$ est l'opérateur qui projette un point 3D dans le plan image. Le support de projection de la synthèse est

$$\mathscr{S}(\boldsymbol{p}) = \bigcup_{i=1}^{l} \mathscr{P}\left(\mathbf{X}_{\mathbf{LA}_i}(\boldsymbol{\alpha}); (\boldsymbol{t}; \boldsymbol{r}; \boldsymbol{cam})\right) \tag{3.6}$$

$\mathscr{A}$ est l'opérateur qui calcule les descripteurs de l'apparence pour des points de l'image. Il vient

$$\mathbf{LA}_{\boldsymbol{a}}(\boldsymbol{p}) = \mathscr{A}\left(\mathbf{I}_{\boldsymbol{a}}; \mathscr{S}(\boldsymbol{p})\right) \tag{3.7}$$

La différence entre analyse et synthèse $\boldsymbol{e}$ est alors

$$\boldsymbol{e}(\boldsymbol{p}) = \mathbf{LA}_{\boldsymbol{a}}(\boldsymbol{p}) - \mathbf{LA}_{\boldsymbol{s}}(\boldsymbol{p}) \tag{3.8}$$

La taille du vecteur $\boldsymbol{e}$ ne dépend pas de $\boldsymbol{p}$ mais uniquement du nombre $l$ de points 3D retenus — et donc constante dans le système.

### *3.2.2.1. Approximation du jacobien*

Cette dernière propriété permet d'établir une relation linéaire entre $\boldsymbol{e}$ — ou une variation $\boldsymbol{\delta_e}$ — et une mise à jour des paramètres $\boldsymbol{\delta_p}$ ; une telle matrice est en fait le jacobien. Certains travaux[14] font l'approximation que le jacobien est

14. Cootes (T. F.), G. J. Edwards et C. J. Taylor. Active appearance models. *In* Burkhardt (H.) et B. Neumann, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 1407 de *Lecture Notes in Computer Science*, pages 484–498, Freiburg, Germany, juin 1998. Springer-Verlag ; Dornaika (F.) et J. Ahlberg. Efficient active appearance model for real-time head and facial feature tracking. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, Nice, France, octobre 2003 ; Jurie (F.) et M. Dhome. Real time tracking of 3D objects : an efficient and robust approach. *Pattern Recognition*, 35:317–328, 2002.

constant ; il peut alors être appris au préalable et, si les dérivées de la fonction d'écart doivent être utilisées, cela permet des calculs rapides. Cependant le calcul de la fonction d'écart fait intervenir des étapes où l'influence des paramètres de contrôle n'est pas linéaire : le jacobien n'est donc pas constant. Quitte à perdre en rapidité [15], le jacobien est recalculé à chaque fois dans les applications qui privilégient la précision de l'ajustement [16].

### 3.2.3. Expression de la fonction d'écart

Pour les deux cas correspondant aux deux modèles d'apparence, un vecteur différence entre analyse et synthèse $\boldsymbol{e}$ peut être calculé. Dans les deux cas, le vecteur différence, de taille $\mathrm{N_e}$, est composé par concaténation de plusieurs vecteurs constitués d'informations sur l'apparence du visage : en généralisant, l'analyse et la synthèse sont comparées via deux ensembles de descripteurs de l'apparence du visage calculés à partir de la forme du visage. C'est ce qui est représenté sur la figure 3.1.

Enfin, la fonction d'écart $\varepsilon$ répond à un problème dit des moindres carrés non-linéaires ; une fonction de pénalisation pour des paramètres articulatoires statistiquement improbables $K$ est ajoutée ; $\varepsilon$ prend la forme

$$\varepsilon(\boldsymbol{p}) = \frac{1}{\mathrm{N_e}} \|\boldsymbol{e}\|_2 + K(\boldsymbol{\alpha}) \tag{3.9}$$

$$\text{avec} \qquad K(\boldsymbol{\alpha}) = \sum_i \rho(\boldsymbol{\alpha}_i) \qquad \text{où} \qquad \rho(x) = \begin{cases} \lambda\left(\mathrm{e}^{(|x|-\gamma)^2} - 1\right) & \text{si} \quad |x| \geq \gamma \\ 0 & \text{si} \quad |x| < \gamma \end{cases} \tag{3.10}$$

et où $\| \ \|_2$ désigne la norme euclidienne [17] de $\mathbb{R}^{\mathrm{N_e}}$. $\lambda$ et $\gamma$ sont des valeurs positives ;

15. Il est reporté pour (Blanz et Vetter, 1999) 5 minutes de calcul par une machine tournant un processeur Pentium IV cadencé à 2 GHz pour réaliser l'ajustement du modèle à une image. Bien sûr le temps de calcul ne se réduit pas au seul calcul du jacobien.

16. Blanz (V.) et T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, septembre 2003 ; Pighin (F.), R. Szeliski et D. H. Salesin. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002.

17. Il est possible d'intégrer un modèle de l'erreur en utilisant d'autres distances, comme la distance de Mahalanobis ou la norme « robuste » proposée dans Geman (S.) et D. E. McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistic Institute*, LII-4:5–21, 1987.

par défaut [18] le système utilise $\lambda = \frac{1}{5}$ et $\gamma = 3$. Les raisons et propriétés suivantes ont guidé le choix de $\rho$ :

- les paramètres articulatoires sont issus d'analyses en composantes principales et ils sont distribués normalement et normalisés par leur écart-type ; une déviation supérieure à trois correspond à moins de 3 ‰ des valeurs [19]. D'autres formes de pénalisation plus ou moins *ad hoc* ont été proposées [20].
- $\rho$ et donc $K$ sont de classe $\mathscr{C}^1$ : la fonction de pénalisation est suffisamment lisse pour ne pas perturber le choix et le comportement de l'algorithme d'optimisation ; cela sera vu ci-après.

Notons enfin qu'une formulation statistique des termes de l'équation (3.9) peut permettre de définir la fonction d'écart comme l'opposé du logarithme de la probabilité *a posteriori* des paramètres de contrôle étant donnés l'image analysée et le modèle du visage [21].

## 3.3. MESURES DES VARIATIONS DE LA FONCTION D'ÉCART

Le système présenté à la figure 3.1 est composé de plusieurs modules ; sans connaissances particulières *a priori*, il faut supposer que les représentations choisies et les traitements associés à ces données influent et interagissent sur les performances globales du système : la fonction d'écart et la méthode employée pour minimiser cette dernière sont un exemple des interdépendances entre modules.

Pour évaluer, caractériser la fonction d'écart *per se* il est possible de collecter des échantillons de la fonction d'écart [22]. Une telle topographie de $\varepsilon$ a été réali-

18. Les valeurs de $\lambda$ et de $\gamma$ peuvent être spécifiées par passage d'arguments à notre programme.

19. Remarquons que si les données statiques utilisées pour l'apprentissage du modèle sont distribuées normalement, ce n'est *a priori* pas le cas pour les postures dynamiques, intermédiaires.

20. P. ex., une forme quadratique : Pighin et coll. (2002), *op. cit.* ; une forme exponentielle liée à la probabilité des paramètres : Blanz et Vetter (2003), *op. cit.*.

21. Voir p. ex. Blanz et Vetter (2003), *op. cit.* ; Basu (S.), N. Oliver et A. Pentland. 3D lip shapes from video : A combined physical-statistical model. *Speech Communication*, 26:131–148, 1998.

22. Perret (Y.). *Suivi de paramètres de modèle géométriques à partir de séquences vidéo multivues*. Thèse de doctorat, Université Claude Bernard, Lyon, France, décembre 2001 ; Gérard (P.) et A. Gagalowicz. Three dimensional model-based tracking using texture learning and matching. *Pattern Recognition Letters*, 21:1095–1103, 2000 ; Sminchisescu (C.). *Estimation algorithms*

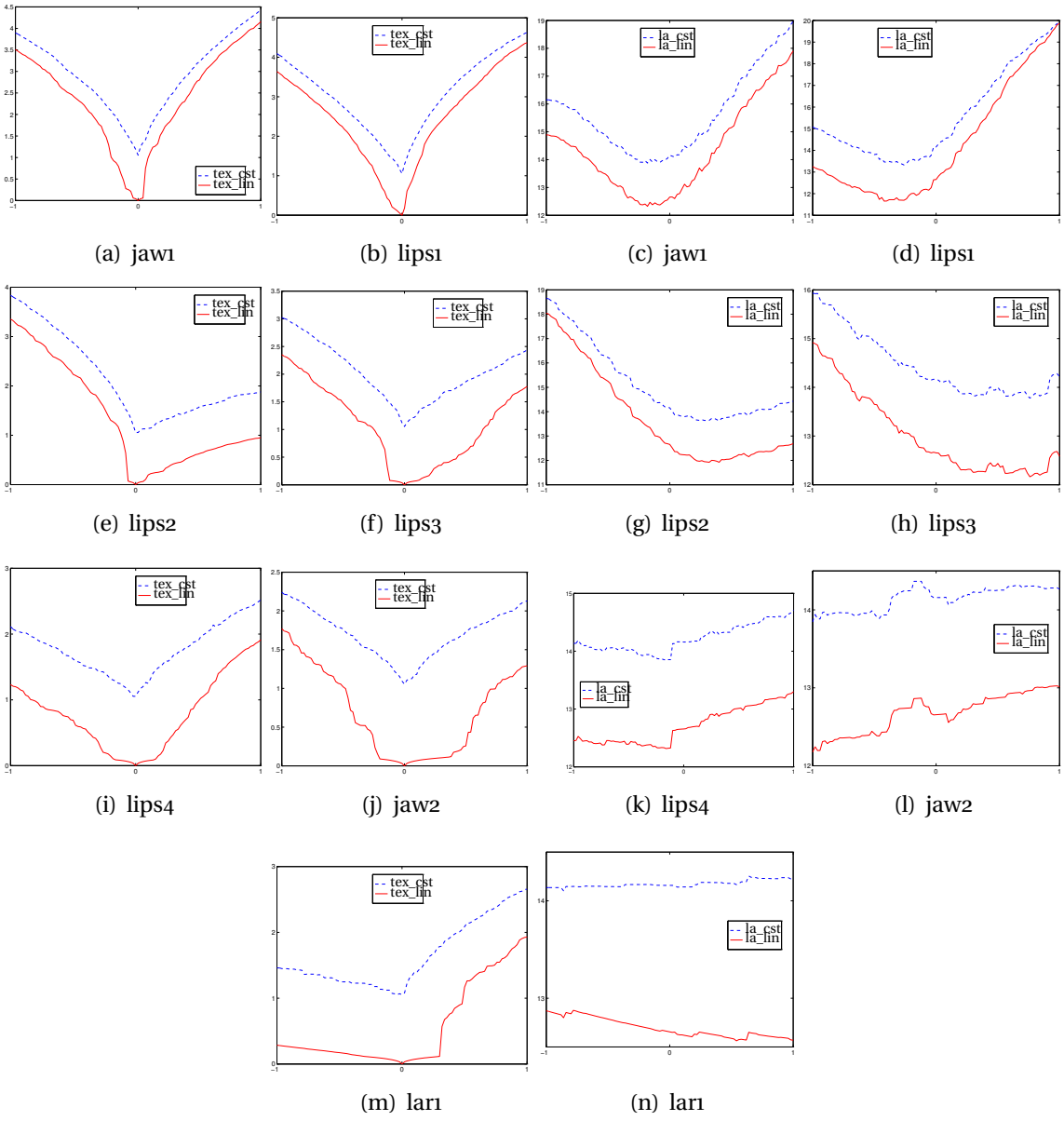
FIGURE 3.2 – Deux rendus synthétiques d'un modèle articulatoire au repos : à gauche, un rendu avec une texture riche correspondant au corpus « billes » ; à droite, un rendu plus écologique avec une texture correspondant au corpus « téléconférence ». Les textures sont calculées à partir des modèles de texture correspondants.

sée pour deux conditions expérimentales ; des mesures ont été faites autour de la position neutre en utilisant comme images analysées la forme 3D de référence synthétisée avec deux textures correspondant aux modèles de texture linéaire des corpora « billes » et « téléconférence ». Ces images analysées sont représentées sur la figure 3.2 ; une seule vue, de face, est utilisée. Les mesures des figures 3.3, 3.4, 3.5 et 3.6 correspondent aux quatre modèles d'apparence *tex_lin*, *la_lin*, *tex_cst* et *tex_cst*. Cela permet de discuter les résultats en testant les conditions expérimentales, les modèles d'apparence et les paramètres de contrôle du visage, articulatoires et rigides.

Constatons d'abord que pour tous les modèles la fonction d'écart présente des minima locaux.

INFLUENCE DU CORPUS Dans les conditions où la texture du visage est enrichie, les pentes de la fonction d'écart sont marquées et le nombre de minima locaux

*for ambiguous visual models — Three dimensional human modeling and motion reconstruction in monocular video sequences.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, juillet 2002.



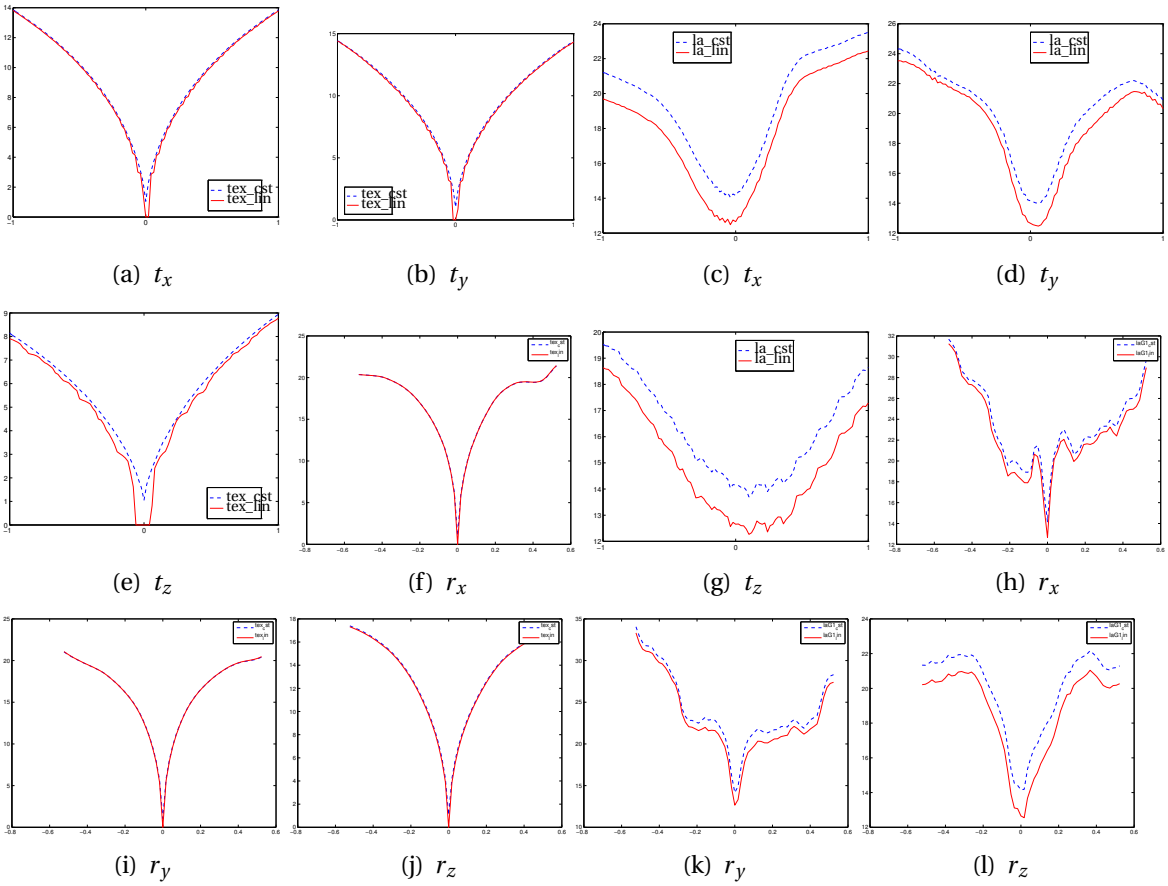

(a) jaw1 (b) lips1 (c) jaw1 (d) lips1

(e) lips2 (f) lips3 (g) lips2 (h) lips3

(i) lips4 (j) jaw2 (k) lips4 (l) jaw2

(m) lar1 (n) lar1

FIGURE 3.3 – Échantillonnage de la fonction d'écart pour les paramètres articulatoires pour « billes » : à gauche, les modèles de texture ; à droite, les modèles de l'apparence locale. Les pointillés correspondent aux modèles qui ne varient pas avec l'articulation ; les traits pleins, à ceux qui varient. Les échelles verticales diffèrent ; voir aussi la figure 3.2.

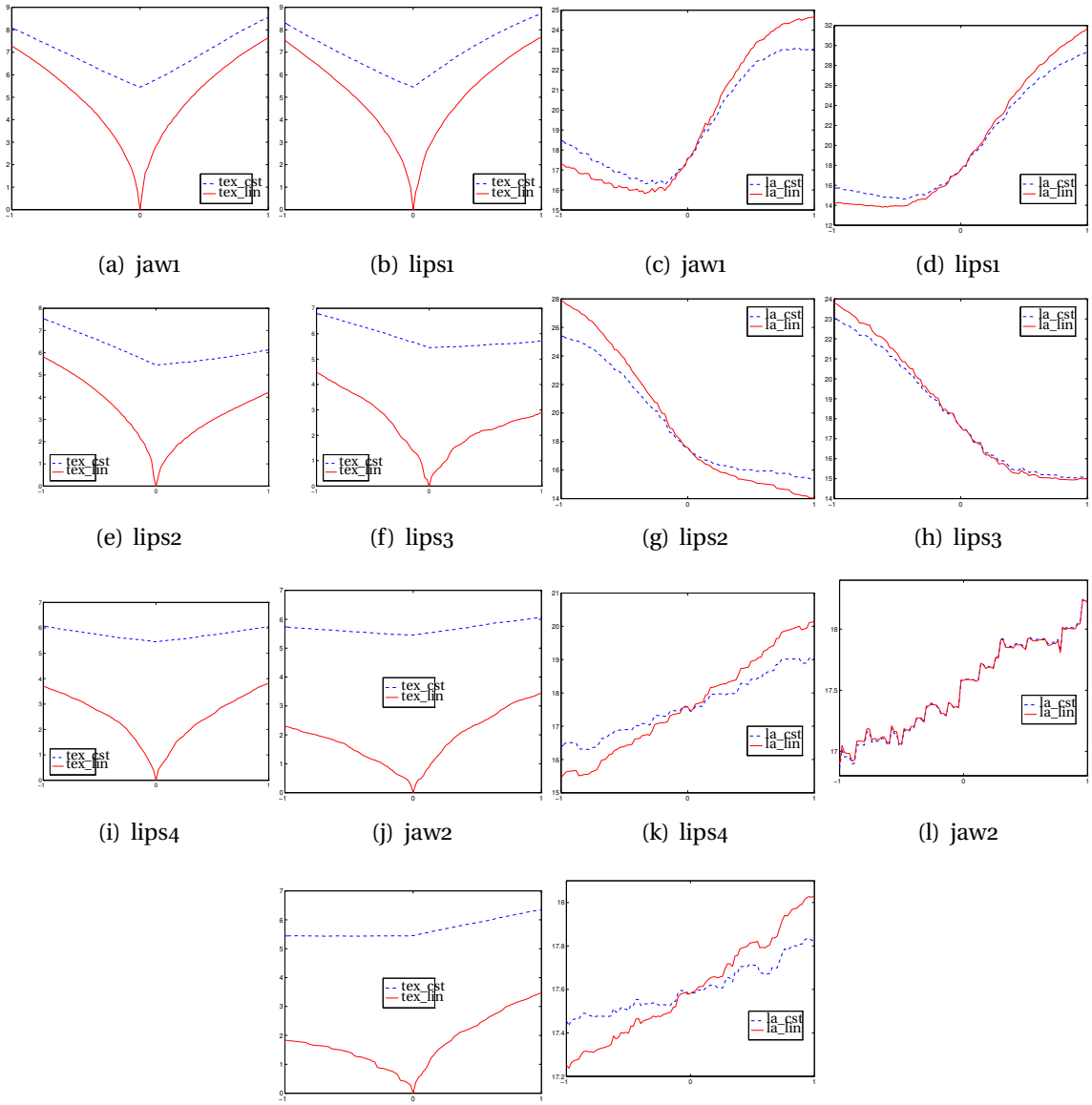



(a) $t_x$ (b) $t_y$ (c) $t_x$ (d) $t_y$

(e) $t_z$ (f) $r_x$ (g) $t_z$ (h) $r_x$

(i) $r_y$ (j) $r_z$ (k) $r_y$ (l) $r_z$

FIGURE 3.4 – Échantillonnage de la fonction d'écart pour les paramètres de mouvement rigide pour « billes » : à gauche, les modèles de texture ; à droite, les modèles de l'apparence locale. Les pointillés correspondent aux modèles qui ne varient pas avec l'articulation ; les traits pleins, à ceux qui varient. Les angles de rotation, exprimés en rad, varient entre $-\frac{\pi}{6}$ et $\frac{\pi}{6}$ ; les translations sont exprimées en cm. Les échelles verticales diffèrent ; voir aussi la figure 3.2.



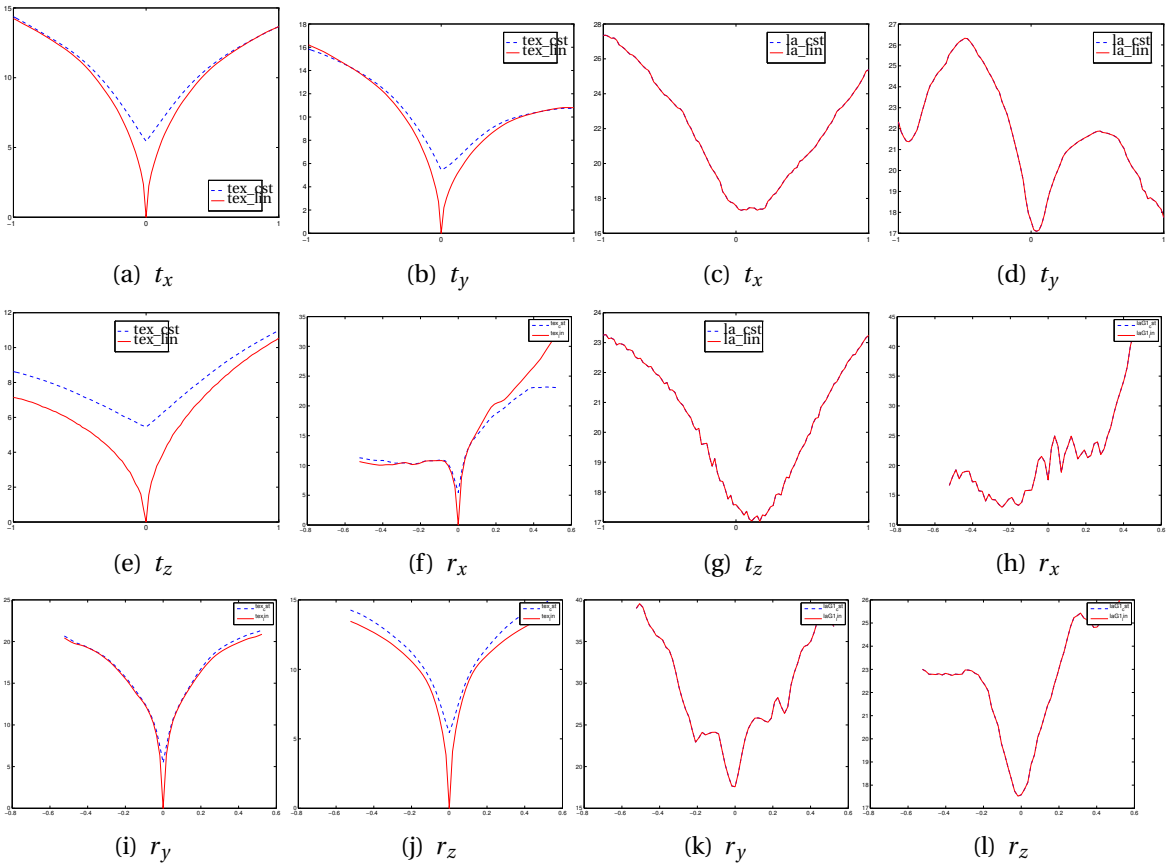

(a) jaw1 (b) lips1 (c) jaw1 (d) lips1

(e) lips2 (f) lips3 (g) lips2 (h) lips3

(i) lips4 (j) jaw2 (k) lips4 (l) jaw2

(m) lar1 (n) lar1

FIGURE 3.5 – Échantillonnage de la fonction d'écart pour les paramètres articulatoires pour « téléconférence » : à gauche, les modèles de texture ; à droite, les modèles de l'apparence locale. Les pointillés correspondent aux modèles qui ne varient pas avec l'articulation ; les traits pleins, à ceux qui varient. Les échelles verticales diffèrent ; voir aussi la figure 3.2.




(a) $t_x$ (b) $t_y$ (c) $t_x$ (d) $t_y$

(e) $t_z$ (f) $r_x$ (g) $t_z$ (h) $r_x$

(i) $r_y$ (j) $r_z$ (k) $r_y$ (l) $r_z$

FIGURE 3.6 – Échantillonnage de la fonction d'écart pour les paramètres de mouvement rigide pour « téléconférence » : à gauche, les modèles de texture ; à droite, les modèles de l'apparence locale. Les pointillés correspondent aux modèles qui ne varient pas avec l'articulation ; les traits pleins, à ceux qui varient. Sur ce corpus, les deux modèles de l'apparence locale ne diffèrent que par des changements articulatoires : leurs courbes sont dont confondues. Les angles de rotation, exprimés en rad, varient entre $-\frac{\pi}{6}$ et $\frac{\pi}{6}$ ; les translations sont exprimées en cm. Les échelles verticales diffèrent ; voir aussi la figure 3.2.

est restreint, laissant suggérer que le minimum global est atteignable de manière automatique. En revanche, dans les conditions d'un visage sans marqueurs les minima locaux sont parfois plus nombreux ; il n'existe pas toujours un minimum global (voir jaw2 et lar1). En pratique, cela peut aboutir à un problème mal posé ; dans certains cas spectaculaires, on sait que la tâche d'estimation de la posture 3D peut être rendue très ardue [23].

INFLUENCE DES PARAMÈTRES Il faut également noter les différences sur les variations de la fonction d'écart entre les paramètres : les paramètres qui influent peu sur la forme et l'apparence d'un visage vu de face — la translation en $z$ et les paramètres articulatoires jaw2 et lar1, voire lips4 — risquent d'être difficilement estimés avec précision. On notera la robustesse pour les rotations — jusqu'à 30 ° — des modèles de texture.

INFLUENCE DU TYPE DE MODÈLES D'APPARENCE Si ces dernières observations sont surtout des confirmations attendues, on observe que la fonction d'écart est globalement moins bien « conditionnée » pour les modèles de l'apparence locale par rapport aux modèles de texture : les minima locaux sont plus nombreux et la valeur minimale de la fonction d'écart ne correspond pas toujours aux valeurs exactes des paramètres (à savoir 0 sur les figures) — ce décalage pourrait se retrouver dans les paramètres estimés par le système. Les images servant de support à cette étude sont issues du modèle de texture ce qui peut constituer un biais pour les comparaisons entre les types de modèles.

INFLUENCE DE LA MODÉLISATION ARTICULATOIRE DE L'APPARENCE On remarque qu'en ce qui concerne les modèles d'apparence, les topographies des fonctions d'écart ont des minima globaux plus marqués — les bassins d'attraction sont plus grands — et seraient donc plus propres à la minimisation dans le cas des modèles d'apparence qui varient avec l'articulation que pour les modèles constants. Ce point est positif puisque la construction des modèles d'apparence n'inclut pas explicitement de contraintes liées à leur utilisation pour l'ajustement sur une image. Les valeurs des fonctions d'écart sont plus petites pour les modèles variant avec l'articulation. On notera également que pour le seul mouvement rigide

23. SMINCHISESCU (C.). *Estimation algorithms for ambiguous visual models — Three dimensional human modeling and motion reconstruction in monocular video sequences.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, juillet 2002.

l'apport de la modélisation semble insignifiant [24]. Cela pourrait faciliter l'intégration de modèles d'apparence multi-vues.

RÉSUMÉ Ces mesures de la fonction d'écart ne donnent qu'un aperçu de sa topographie, pour un cas ; nous n'avons pas présenté les effet possibles des interactions entre facteurs ou du bruit. Il serait difficile d'exploiter systématiquement ce jeu restreint d'abaques ; cependant, ces mesures mettent en évidence la complexité de cette topographie et vont permettre de guider le choix de la méthode de minimisation qui sera utilisée.

## 3.4. RECHERCHE DU JEU OPTIMAL DES PARAMÈTRES DE CONTRÔLE

Des échantillons prélevés dans plusieurs conditions expérimentales tendent à conforter cette hypothèse intuitive que, même en appliquant certains pré-traitements normatifs, la fonction d'écart n'est pas convexe [25] et que des calculs approchés des dérivées par différences finies seront très bruités [26]. Concrètement, ces deux caractéristiques de la fonction d'écart guident le choix de son processus de minimisation. L'algorithme retenu doit pouvoir atteindre une bonne solution — voire le minimum global — en présence de minima locaux et sans utiliser les dérivées.

24. En tout cas au premier ordre ; des interactions entre paramètres pourraient exister. Rappelons que par définition, cet apport est nul dans le cadre des modèles de l'apparence locale… et pourtant ce n'est pas le cas pour « billes » : le modèle de texture constant est en fait une texture prise sur une posture articulatoire — [afa] — proche de la posture neutre ; de même pour le modèle de l'apparence locale constant (ce n'est plus le cas pour « téléconférence », ce qui explique que les courbes de ce type de modèle soient confondues).

25. La fonction d'écart est positive, et nulle seulement si les descripteurs de synthèse et ceux de l'image sont identiques : elle n'est bien sûr pas concave.

26. Il n'est pas possible d'obtenir une expression analytique complète des dérivées de la fonction d'écart. En développant le calcul de la dérivée première p. ex., il est possible d'aboutir à une somme d'une expression analytique et d'une différentielle correspondant aux descripteurs de l'image qu'il faudra estimer par différences finies ; l'estimation résultante du gradient n'est *a fortiori* pas plus performante : PIGHIN (F.), R. SZELISKI et D. H. SALESIN. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002.

De nombreuses méthodes existent pour minimiser sans contraintes une fonction non-linéaire à valeur scalaire de plusieurs variables réelles. De façon générale, on distingue d'une part les méthodes locales qui recherchent un minimum voisin d'un point de départ donné et les méthodes globales qui recherchent un minimum dans tout un domaine donné. On trouve dans cette dernière catégorie, par exemple, les méthodes tabous ; à base de recuit-simulé ; d'algorithme génétique. Si elles conviennent bien à certaines classes de problèmes où la recherche d'un « gain » fort prime sur le temps de calcul, nous avons choisi de ne pas les appliquer ici pour maintenir un fonctionnement du système en temps interactif (toutefois, leur emploi serait adapté pour initialiser le système lorsque le point de départ est éloigné de la solution). [27] La première catégorie, les méthodes locales, comprend les méthodes qui nécessitent de calculer d'une manière ou d'une autre les dérivées de la fonctionnelle, parmi lesquelles les variantes de la méthode de quasi-Newton généralisée [28] sont considérées comme les plus performantes, et les méthodes de recherche directe qui n'utilisent pas les valeurs *numériques* de la fonctionnelle mais seulement un classement relatif d'un nombre finis de points d'évaluation de la fonctionnelle.

Les critères de choix que nous avons retenus nous ont amené à utiliser une méthode de recherche directe. Nous avons mis en œuvre des tests pilotes pour des implémentations de plusieurs de ces méthodes, notamment la méthode des variations locales, la méthode de Powell mais aussi une variante de la méthode de quasi-Newton généralisée et la méthode de Levenberg-Marquardt [29] ; de manière empirique, ces tests nous ont conduit à retenir l'algorithme de Nelder-Mead.

27. Une stratégie de type « force brute » n'est bien sûr pas envisageable : en considérant que les paramètres articulatoires varient entre −3 et 3 et que la deuxième décimale est pertinente, on obtient 600 valeurs possibles pour chaque paramètre ; $600^{13}$ évaluations de la fonctionnelle sont alors nécessaires ; à la cadence approximative de mille évaluations par seconde, la durée du calcul est… inhumaine.

28. Ces méthodes approximent localement au deuxième ordre la fonctionnelle.

29. Marquardt (D.). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441, 1963 ; la méthode de Levenberg-Marquardt est faite *sur mesure* pour les fonctions d'écart de type « moindres carrés », c.-à-d. qui peuvent s'écrire sous la forme $f(\boldsymbol{p}) = \|\boldsymbol{r}(\boldsymbol{p})\|_2^2$, ce qui est le cas de la fonction d'écart $\varepsilon$. Cette méthode combine une optimisation par descente de gradient et par approximation au deuxième ordre : il n'est pas surprenant *a fortiori* que dans notre cas ses performances soient moindres que celles de méthodes de recherche directe.

### 3.4.1. Algorithme de Nelder-Mead

Cette méthode pour l'optimisation non-linéaire sans contraintes a été proposée par Nelder et Mead [30]. L'algorithme de Nelder-Mead est simple à mettre en œuvre et, lorsqu'il fonctionne correctement, requiert à chaque itération peu d'évaluations de la fonctionnelle. Il est à noter que d'un point de vue théorique cet algorithme ne converge pas forcément vers un minimum ; par exemple, cela a été démontré pour une famille de fonctions strictement convexes en dimension deux [31]. Son utilisation très répandue dans de nombreux domaines montre que d'un point de vue pratique l'algorithme, ou des adaptations *ad hoc* [32], donne satisfaction. Dans leur article où la méthode de Nelder-Mead est décrite de manière algorithmique — la description ci-après reprend largement ce travail —, Lagarias et coll. concluent ainsi : « *Our general conclusion about the Nelder-Mead « algorithm is that the main mystery to be solved is not wether it ultimately « converges to a minimizer—for general (nonconvex) functions, it does not—but « rather why it tends to works so well in practice by producing a rapid initial « decrease in function values.* [33] »

La méthode de Nelder-Mead maintient à chaque itération un simplexe non-dégénéré, c.-à-d. un objet géométrique de dimension $n$ — $n$ est le nombre de variables de la fonctionnelle $\varepsilon$ —, de volume non-nul et qui est l'enveloppe convexe des $n+1$ sommets $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n+1}$ [34]. Le simplexe est utilisé pour explorer l'espace de recherche. Des règles permettent de déformer le simplexe pour l'adapter à la géométrie de la fonctionnelle au cours de la minimisation.

L'algorithme utilise quatre paramètres ; il s'agit des coefficients de réflection

30. Nelder (J. A.) et R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965 ; Press (W. H.), S. A. Teukolsky, W. T. Vetterling et B. P. Flannery. Downhill simplex method in multidimensions. *In Numerical Recipes in C*, chapitre 10-4, pages 408–412. Cambridge University Press, 1992.

31. McKinnon (K. I. M.). Convergence of the Nelder-Mead simplex method to a nonstationary point. *SIAM Journal of Optimization*, 9(1):148–158, 1998.

32. P. ex., Perret (Y.). *Suivi de paramètres de modèle géométriques à partir de séquences vidéo multi-vues.* Thèse de doctorat, Université Claude Bernard, Lyon, France, décembre 2001 ; Perret introduit des perturbations aléatoires qui peuvent s'ajouter aux coordonnées des moins bons sommets ; cela peut permettre au simplexe de *sortir* d'un minimum local et de continuer l'optimisation. Les propriétés mathématiques du simplexe ne sont alors plus garanties.

33. Lagarias (J. C.), J. A. Reeds, M. H. Wright et P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

34. P. ex., pour une fonctionnelle de deux variables, le simplexe sera un triangle.

(ρ), d'expansion (χ), de contraction (γ), et de rétrécissement (σ). Ces paramètres doivent satisfaire $\rho > 0$, $\chi > 1$, $\chi > \rho$, $0 < \gamma < 1$, et $0 < \sigma < 1$. Leurs valeurs classiques, que nous avons utilisées, sont $\rho = 1$, $\chi = 2$, $\gamma = \frac{1}{2}$ et $\sigma = \frac{1}{2}$.

Une itération de l'algorithme consiste en les opérations présentées ci-après ; une itération aboutit soit à déterminer et accepter un nouveau sommet qui remplace le *pire* sommet $\boldsymbol{x}_{n+1}$ pour l'itération suivante, soit, pour une opération de rétrécissement, à déterminer $n$ nouveaux sommets qui composent avec $\boldsymbol{x}_1$ le simplexe à l'itération suivante.

1. *Classement*
   Classer les $n+1$ sommets de telle sorte que $\varepsilon(\boldsymbol{x}_1) \leq \varepsilon(\boldsymbol{x}_2) \leq \cdots \leq \varepsilon(\boldsymbol{x}_{n+1})$.
2. *Décision de terminer ou non l'algorithme*
   Cette étape ne fait pas l'objet de règles précises dans la description classique de l'algorithme proprement dit ; les choix que nous avons retenus sont repris à la suite de cette description.
3. *Réflection*
   Calculer le point de réflexion $\boldsymbol{x}_\mathrm{r}$ d'après
   $$\boldsymbol{x}_\mathrm{r} = \bar{\boldsymbol{x}} + \rho(\bar{\boldsymbol{x}} - \boldsymbol{x}_{n+1}) = (1+\rho)\bar{\boldsymbol{x}} - \rho\boldsymbol{x}_{n+1}, \tag{3.11}$$
   où $\bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i}{n}$ est l'isobarycentre des $n$ meilleurs points (c.-à-d. de tous les sommets sauf $\boldsymbol{x}_{n+1}$). Évaluer $\varepsilon_\mathrm{r} = \varepsilon(\boldsymbol{x}_\mathrm{r})$. Si $\varepsilon_1 \leq \varepsilon_\mathrm{r} < \varepsilon_n$, accepter le point de réflection $\boldsymbol{x}_\mathrm{r}$ et terminer l'itération.
4. *Expansion*
   Si $\varepsilon_\mathrm{r} < \varepsilon_1$, calculer le point d'expansion $\boldsymbol{x}_\mathrm{e}$ d'après
   $$\boldsymbol{x}_\mathrm{e} = \bar{\boldsymbol{x}} + \chi(\boldsymbol{x}_\mathrm{r} - \bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} + \rho\chi(\bar{\boldsymbol{x}} - \boldsymbol{x}_{n+1}) = (1+\rho\chi)\bar{\boldsymbol{x}} - \rho\chi\boldsymbol{x}_{n+1}, \tag{3.12}$$
   et évaluer $\varepsilon_\mathrm{e} = \varepsilon(\boldsymbol{x}_\mathrm{e})$. Si $\varepsilon_\mathrm{e} < \varepsilon_\mathrm{r}$, accepter $\boldsymbol{x}_\mathrm{e}$ et terminer l'itération ; sinon (si $\varepsilon_\mathrm{e} \geq \varepsilon_\mathrm{r}$), accepter $\boldsymbol{x}_\mathrm{r}$ et terminer l'itération.
5. *Contraction*
   Si $\varepsilon_\mathrm{r} \geq \varepsilon_n$, faire une contraction entre $\bar{\boldsymbol{x}}$ et le meilleur de $\boldsymbol{x}_{n+1}$ et $\boldsymbol{x}_\mathrm{r}$.
   (a) *extérieure.* Si $\varepsilon_n \leq \varepsilon_\mathrm{r} < \varepsilon_{n+1}$ (c.-à-d. $\boldsymbol{x}_\mathrm{r}$ est strictement meilleur que $\boldsymbol{x}_{n+1}$), faire une contraction extérieure : calculer
   $$\boldsymbol{x}_\mathrm{c} = \bar{\boldsymbol{x}} + \gamma(\boldsymbol{x}_\mathrm{r} - \bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} + \gamma\rho(\bar{\boldsymbol{x}} - \boldsymbol{x}_{n+1}) = (1+\gamma\rho)\bar{\boldsymbol{x}} - \gamma\rho\boldsymbol{x}_{n+1}, \tag{3.13}$$
   et évaluer $\varepsilon_\mathrm{c} = \varepsilon(\boldsymbol{x}_\mathrm{c})$. Si $\varepsilon_\mathrm{c} \leq \varepsilon_\mathrm{r}$, accepter $\boldsymbol{x}_\mathrm{c}$ et terminer l'itération ; sinon, aller à l'étape 6 (faire un rétrécissement).

(b) *intérieure.* Si $\varepsilon_r \leq \varepsilon_{n+1}$, faire une contraction intérieure : calculer

$$\boldsymbol{x}_{cc} = \bar{\boldsymbol{x}} - \gamma(\bar{\boldsymbol{x}} - \boldsymbol{x}_{n+1}) = (1-\gamma)\bar{\boldsymbol{x}} + \gamma \boldsymbol{x}_{n+1}, \tag{3.14}$$

et évaluer $\varepsilon_{cc} = \varepsilon(\boldsymbol{x}_{cc})$. Si $\varepsilon_{cc} < \varepsilon_{n+1}$, accepter $\boldsymbol{x}_{cc}$ et terminer l'itération ; sinon, aller à l'étape 6 (faire un rétrécissement).

6. *Rétrécissement*
Évaluer $\varepsilon$ aux $n$ points $\boldsymbol{v}_i = \boldsymbol{x}_1 + \sigma(\boldsymbol{x}_i - \boldsymbol{x}_1)$, $i = 2, \ldots, n+1$. Les sommets (non-ordonnés) du simplexe à la prochaine itération sont $\boldsymbol{x}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n+1}$.

Il est à noter la possibilité de spécifier des règles de départage pour l'ordonnancement des points dans le cas où des valeurs de $\varepsilon$ sont égales en plusieurs points du simplexe [35].

La décision d'arrêter l'algorithme est prise après l'étape 1 (classement). Les critères d'arrêt sont :

- la différence absolue entre les valeurs de la fonctionnelle aux meilleur et pire sommets, $\frac{\varepsilon_{n+1}-\varepsilon_1}{\varepsilon_{n+1}+\varepsilon_1} \leq \delta_a$ ;
- la différence relative entre ces mêmes valeurs, $\varepsilon_{n+1} - \varepsilon_1 \leq \delta_r$ ;
- la distance *normalisée* entre ces deux sommets, $\frac{1}{n}\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\|_2 \leq \delta_d$ ;
- le nombre d'évaluations de la fonctionnelle qui doit être inférieur à un nombre maximal N.

Les valeurs des seuils $\delta$ et N peuvent être spécifiées par passage d'arguments à notre programme ; par défaut nous avons choisi les valeurs conservatrices [36] :

- $\delta_a = 0{,}000\,1$ ;
- $\delta_r = 1{,}0$, le critère correspondant est alors toujours vérifié ;
- $\delta_d = 0{,}05$ ;
- N = 1 000.

L'algorithme est terminé seulement si les trois premières inégalités sont vraies ou si le nombre maximal d'évaluations de la fonctionnelle est atteint. Ce dernier critère permet d'une part d'éviter à notre programme de rester bloqué si l'on se trouve dans un cas *pathologique* (l'on a vu plus haut qu'il en existe) et d'autre part de contrôler facilement la durée de l'optimisation, si l'on veut par exemple un fonctionnement plus rapide du système (et *a priori* moins précis).

35. Lagarias (J. C.), J. A. Reeds, M. H. Wright et P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

36. Les choix par défaut privilégient la précision du résultat plutôt que la petitesse du nombre d'itérations.

# 3.5. CONCLUSION

## 3.5.1. Récapitulatif

Dans ce chapitre ont été présentées l'architecture et les méthodes employées de notre système d'estimation des mouvements tridimensionnels du visage depuis la vidéo. Ce système consiste en une boucle d'analyse par la synthèse qui cherche à ajuster à l'image des modèles du visage — forme 3D et apparence — construits au préalable. À partir d'un jeu de paramètres de contrôle ces modèles permettent de synthétiser des descripteurs de l'apparence du visage et de déterminer un ensemble de pixels de l'image analysée. Ces pixels d'intérêt permettent de calculer sur l'image analysée des descripteurs de l'apparence qui sont comparés à leurs homologues synthétiques : c'est ainsi qu'est définie une fonction d'écart entre l'image analysée et les paramètres de contrôle. La fonction d'écart comprend également un terme qui rend compte de la vraisemblance statistique des paramètres considérés. Des mesures par échantillonnage des variations de cette fonction d'écart ont montré notamment que cette dernière n'est pas convexe et comporte des minima locaux : si l'obtention d'un *bon* jeu de paramètres de contrôle est possible, la recherche automatique de ce jeu de paramètres doit utiliser un algorithme de minimisation qui tient compte de ces caractéristiques. Notre choix s'est porté sur l'algorithme de Nelder-Mead ; cette méthode de recherche directe a été décrite de manière algorithmique. Pour faciliter l'implémentation du système, nous avons veillé à donner les valeurs numériques de ses réglages obligés.

Dans sa version actuelle ce système gère de manière simple certains aspects qui dans d'autres applications pourraient faire l'objet d'attentions supplémentaires ; ce sont autant de perspectives pour le système que nous discutons ci-après.

## 3.5.2. Discussion

L'étape d'initialisation, plus difficile que le suivi au sein de la séquence parce que le système est en général éloigné de la solution, ne fait pas l'objet d'un traitement spécifique (les paramètres sont initialisés à zéro) ; pour cette tâche, incluant éventuellement un affinement du calibrage des caméras, l'emploi d'une

méthode d'optimisation globale serait profitable. Un tel mode pourrait aussi servir à « récupérer » le système en cas de décrochage manifeste (forte et soudaine hausse de la fonction d'écart) [37]. Dans des cas plus difficiles, il serait possible aussi de s'appuyer sur un système de détection de visage.

Comment une prédiction temporelle, moins triviale que celle utilisée, au sein d'une séquence pourrait permettre d'améliorer de manière forte l'estimation des paramètres de contrôle ? Des essais préliminaires simples ne nous ont pas permis d'aboutir à une précision supérieure ; l'emploi de méthodes plus sophistiquées serait sûrement plus bénéfique [38] ; il serait également possible d'utiliser un modèle de coarticulation : l'analyse d'un corpus vidéo par le système de suivi permet d'apprendre un tel modèle [39] et l'utilisation dans le suivi constituerait une méthode pour raffiner de manière itérative le modèle de coarticulation. De plus, cette prédiction temporelle nous semble intéressante pour diminuer le nombre de passages dans la boucle d'optimisation, et donc accélérer, sans perte de qualité, la cadence de traitement du flux vidéo. Introduire à ce point un module basé sur le signal audio serait une façon intéressante d'utiliser la complémentarité et la redondance intrinsèques des signaux de la parole audiovisuelle, multimodale.

Suivre quelques unes de ces pistes semble un pré-requis avant d'utiliser le système pour une analyse de scène plus complexe, notamment où le visage serait partiellement caché ou, de manière plus générale, difficilement localisable.

37. Perret (Y.). *Suivi de paramètres de modèle géométriques à partir de séquences vidéo multivues.* Thèse de doctorat, Université Claude Bernard, Lyon, France, décembre 2001.

38. Hamlaoui (S.) et F. Davoine. Facial action tracking using particle filters and active appearance models. *In Proc. of the Smart Objects Conf.*, pages 165–169, Grenoble, France, 2005.

39. Revéret (L.), G. Bailly et P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *In Proc. of the Internat. Conf. on Spoken Language Processing*, volume 2, pages 755–758, Beijing, China, octobre 2000.

# 4

# Évaluation objective

## 4.1. INTRODUCTION

Nous avons présenté dans le chapitre précédent un système d'estimation des mouvements du visage depuis des séquences vidéo. En les construisant nous nous sommes déjà efforcés de caractériser chacun de ses différents composants. Il est temps de « convertir en expériences » ce système ; celles que nous présentons dans ce chapitre ont pour but d'évaluer et de comparer de manière objective les performances du système pour chacun des quatre modèles d'apparence décrits au chapitre 2.

Les angles d'attaque de ce problème d'évaluation quantitative ne manquent pas : pour l'informatique ce peut être la complexité algorithmique, la description du flot des données au sein de l'architecture et les durées d'exécution (exprimées en nombre d'instructions, de cycles, de temps de calcul, de fréquence nominale de fonctionnement), la consommation ; pour le codage basé modèle, le résidu de l'image, le PSNR ; pour la capture de mouvements, la précision des mouvements estimés, la robustesse au bruit ; etc.

Les critères d'évaluation que nous avons retenus sont la précision — en 3D — des mouvements capturés et leur pertinence phonétique. D'abord nous présentons les paradigmes d'évaluations que l'on rencontre, en omettant les indices « informatiques » vus ci-dessus [1].

L'acquisition des données audiovisuelles doit beaucoup à C. Savariaux et à A. Arnal. B. Holm, G. Gibert et F. Elisei ont participé à l'étiquetage de ces données. H. Lœvenbruck a accepté d'être sujet de cette étude.

1. Mêlée à un projet regroupant des partenaires industriels intéressés par une certaine rapi-

## 4.1.1. Paradigmes d'évaluations

### *4.1.1.1. Regards (peu) qualitatifs*

Nous commençons ce tour d'horizon par un constat d'insuffisance : dans la plupart des travaux consacrés à l'extraction de mouvements faciaux, les résultats expérimentaux sont souvent réduits à la portion congrue [2].

Cependant, toujours des travaux ont démontré qu'il est souhaitable et possible d'aller au delà de ces évaluations simples. Cela est détaillé ci-après.

### *4.1.1.2. Confrontation avec des connaissances*

Sans disposer des mouvements effectivement produits, certains auteurs font une véritable inspection des résultats en confrontant les paramètres estimés à des connaissances disponibles par ailleurs : des faits phonétiques [3] ; voire, plus simplement, des adéquations avec des observations visuelles [4].

---

dité de la cadence de fonctionnement, cette thèse s'est focalisée sur des méthodes qui assurent un fonctionnement en temps interactif. Aujourd'hui, dans le cas le plus favorable, le système fonctionne à environ 2 Hz sur des machines de bureau très performantes ; le système fonctionnerait *grosso modo* en temps réel si le temps global de calcul était diminué par un facteur 30. Comme de plus les méthodes utilisées sont implémentées au niveau logiciel de manière sous-optimale pour la durée d'exécution, nous pensons qu'il est réaliste de dire que l'évolution des matériels informatiques permettrait bientôt d'atteindre le temps réel.

2. Nous convenons toutefois que lors des congrès des démonstrations en direct peuvent montrer la robustesse des méthodes, ou du moins y contribuer.

3. Revéret (L.), G. Bailly et P. Badin. MOTHER : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *In Proc. of the Internat. Conf. on Spoken Language Processing*, volume 2, pages 755–758, Beijing, China, octobre 2000 ; Revéret (L.) et I. Essa. Visual coding and tracking of speech related facial motion. *In IEEE International Workshop on Cues in Communication*, Hawaii, USA, décembre 2001 ; Kroos et coll. comparent la pertinence phonétique de deux jeux de vecteurs propres issus d'ACP, l'un sur le corpus de test et l'autre sur le même corpus mais enregistré en présence — très « invasive » — de senseurs 3D attachés au visage : Kroos (C.), T. Kuratate et E. Vatikiotis-Bateson. Video-based face motion measurement. *J. of Phonetics*, 30(3):569–590, juillet 2002.

4. Basu (S.), I. Essa et A. Pentland. Motion regularization for model-based head tracking. *In Proc. of the Internat. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

### 4.1.1.3. Images synthétiques

Générer des séries d'images synthétiques permet de disposer des mouvements de référence ; mais quels paramètres utiliser ? Des valeurs arbitraires présentent des variations temporelles irréalistes mais elles peuvent être efficaces pour le diagnostic [5]. Pour obtenir des trajectoires moins artificielles, certains travaux estiment d'abord les paramètres à partir de l'analyse d'une séquence réelle, puis les utilisent pour la synthèse des séquences de test [6]. Toutefois cette méthode ne produit pas des trajectoires naturelles ; de plus les résultats peuvent être biaisés par l'utilisation de valeurs peut-être privilégiées par le système de suivi.

En charnière avec la présentation des utilisations d'images réelles, nous distinguons les évaluations pour nous remarquables menées par Eisert [7]. La méthodologie adoptée est systématique et d'abord atomique : l'estimation du mouvement rigide est évaluée avec une tête artificielle, en synthèse puis avec des images réelles ; dans sa version synthétique cette tête sert aussi à évaluer l'estimation des paramètres photométriques ; l'estimation des expressions faciales est évaluée à partir de séquences synthétisées d'après des valeurs arbitraires des paramètres de contrôle d'une version adaptée de Candide. À chaque fois, l'effet de deux types de bruit est mesuré : un bruit gaussien ajouté aux pixels de l'image ou aux coordonnées 3D. Enfin, pour les séquences réelles le critère retenu est le PSNR.

5. Li (H.), P. Roivainen et R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):545–555, juin 1993 ; Prêteux (F.) et M. Malciu. Model-based head tracking and 3D pose estimation. *In Proceedings of SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision*, pages 94–110, San Diego, USA, juillet 1998 ; Perret (Y.). *Suivi de paramètres de modèle géométriques à partir de séquences vidéo multi-vues.* Thèse de doctorat, Université Claude Bernard, Lyon, France, décembre 2001 ; Pighin (F.), R. Szeliski et D. H. Salesin. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002 ; Zhang (Y.) et C. Kambhamettu. 3D head tracking under partial occlusion. *Pattern Recognition*, 35:1545–1557, 2002 ; Matthews (I.) et S. Baker. Active appearance models revisited. *Internat. J. of Computer Vision*, 60 (2):135–164, 2004.

6. Basu et coll. (1996), *op. cit.* ; Ström (J.). *Model-based head tracking and coding.* PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2002 ; Dornaika (F.) et J. Ahlberg. Efficient active appearance model for real-time head and facial feature tracking. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, Nice, France, octobre 2003

7. Eisert (P.). *Very low bit-rate video coding using 3-D models.* PhD thesis, University of Erlangen-Nuremberg, octobre 2000.

### *4.1.1.4. Images réelles*

Dans une optique de codage basé modèle [8], les indices portent sur la qualité de la reconstruction de l'image [9] : ce peut être un résidu pixel à pixel mesuré avec l'erreur RMS ou sa formulation logarithmique classique, le PSNR [10].

En utilisant de manière annexe des systèmes de capture de la géométrie, des mouvements ou de la position — maintenant plus accessibles —, des données de références fiables peuvent être obtenues [11]. Notons que même sans disposer de senseurs, une configuration particulière du dispositif expérimental peut fournir de telles données [12].

Enfin, il existe la possibilité de mesurer la distance des projections — dans le cas des modèles 2D, les positions — de points du modèle à leurs positions

8. Pour un codage de l'image sans pertes, le codeur doit encore encoder la différence entre l'image analysée et l'image synthétique. Plus grande cette différence, moins utile le modèle.

9. Li (H.), P. Roivainen et R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):545–555, juin 1993 ; Antoszczyszyn (P. M.), J. M. Hannah et P. M. Grant. Reliable tracking of facial features in semantic-based video coding. *IEE Proceedings — Vision Image and Signal Processing*, 145(4):257–263, août 1998 ; Eisert (P.). *Very low bit-rate video coding using 3-D models*. PhD thesis, University of Erlangen-Nuremberg, octobre 2000 ; Pighin (F.), R. Szeliski et D. H. Salesin. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002 ; Cootes (T. F.) et P. Kittipanya-ngam. Comparing variations on the active appearance model algorithm. *In Proc. of the British Machine Vision Conf.*, volume 2, pages 837–846, Cardiff, UK, 2002.

10. En notant M la valeur théorique maximale de l'erreur RMSE, la formule est : $\mathrm{PSNR(dB)} = 20\log\frac{M^2}{\mathrm{RMSE}^2}$.

11. DeCarlo (D.) et D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Internat. J. of Computer Vision*, 38(2):99–127, juillet 2000 ; La Cascia (M.), S. Sclaroff et V. Athitsos. Fast, reliable head tracking under varying illumination : an approach based on registration of texture-mapped 3D models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(4):322–336, avril 2000 ; Zhang (Y.) et C. Kambhamettu. 3D head tracking under partial occlusion. *Pattern Recognition*, 35:1545–1557, 2002 ; Blanz (V.) et T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, septembre 2003 ; Sherrah (J.) et S. Gong. Fusion of perceptual cues for robust tracking of head pose and position. *Pattern Recognition*, 34:1565–1572, 2001 ; Morency (L.-P.), P. Sundberg et T. Darell. Pose estimation using 3D view-based eigenspaces. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 45–52, Nice, France, octobre 2003.

12. P. ex., sur la base CMU l'angle entre deux vues, acquises simultanément, est connaissable : Sim (T.), S. Baker et M. Bsat. The CMU pose, illumination, and expression (PIE) database. *In Proc. of the IEEE Internat. Conf. on Automatic Face and Gesture Recognition*, pages 53–58, 2002 (d'après Blanz et Vetter (2003), *op. cit.*).

exactes, étiquetées manuellement dans chaque image [13]. Des points qui ne sont pas repérables — comme le milieu d'un menton « lisse » — peuvent être intégrés grâce à des marqueurs collés sur le visage ; les régions autour des marqueurs devenant inutilisables par le système de suivi — sauf à biaiser fortement les résultats —, ces marqueurs ne sont pas placés dans des zones « critiques » [14].

On ne saurait trop promouvoir l'existence de bases de données *publiques*, qui permettent de comparer les systèmes entre eux et de mutualiser l'effort d'étiquetage [15].

### 4.1.2. Présentation des expériences réalisées

Notre système a pour but de capturer des postures, des mouvements tridimensionnels de la parole depuis l'image : les critères d'évaluation que nous avons retenus sont la précision — au sens 3D — des mouvements capturés ainsi que leur pertinence phonétique. Les évaluations ont porté sur des images réelles de la locutrice *hl* ; avec une seule vue, de face, du visage ; sans variations importantes des conditions d'illumination. Il s'agit là d'un paradigme, d'un scénario applicatif, classique et où la tâche d'estimation est encore difficile. La première série d'expériences s'est faite sur des images bien connues, tirées du corpus « billes » ;

13. Antoszczyszyn et coll. (1998), *op. cit.* ; Cootes et Kittipanya-ngam (2002), *op. cit.* ; Yan (S.), C. Liu, S. Li, H. Zhang, H. Shum et Q. Cheng. Texture-constrained active shape models. *In Proc. of The First Internat. Workshop on Generative-Model-Based Vision*, Copenhagen, Denmark, mai 2002 ; Kroos (C.), S. Masuda, T. Kuratate et E. Vatikiotis-Bateson. Towards the facecoder : Dynamic face synthesis based on image motion estimation in speech. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 24–29, Aalborg, Denmark, septembre 2001 ; DeCarlo et Metaxas (2000), *op. cit.* ; Daubias introduit de plus un critère de recouvrement entre les surfaces de lèvres estimées et celles étiquetées : Daubias (P.). *Modèles* a posteriori *de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle.* Thèse de doctorat, Université du Maine, Le Mans, France, décembre 2002 ; Huang (X.), S. Zhang, Y. Wang, D. Metaxas et D. Samaras. A hierarchical framework for high resolution facial expression tracking. *In Proc. of the IEEE Internat. Workshop on Articulated and Nonrigid Motion*, Washington D.C., USA, juin 2004 ; Revéret (L.). *Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, mai 1999 ; Eveno (N.), A. Caplier et P.-Y. Coulon. Accurate and quasi-automatic lip tracking. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(5):706–715, mai 2004.

14. DeCarlo et Metaxas (2000), *op. cit.* ; Huang et coll. (2004) ; *op. cit.*

15. P. ex., Messer (K.), J. Matas, J. Kittler, J. Luettin et G. Maître. XM2VTSDB : The extended M2VTS database. *In Proceedings of the Conference on Audio- and Video-based Biometric Personal Authentification*, pages 72–77, Washington, DC, USA, mars 22–23 1999.

ces expériences de validation caractérisent notre système — décliné en quatre versions correspondant aux quatre modèles d'apparence — dans des conditions expérimentales contrôlées. Le corpus « téléconférence » a été utilisé pour la seconde série d'expérience ; il s'agit là de conditions difficiles : les défauts de la caméra font intervenir des déformations dans l'image que le modèle projectif ne peut pas prendre en compte ; le visage, en gros plan, est proche du centre de projection ; si l'éclairage varie peu, les images sont saturées. Mais, il n'est pas besoin d'estimer les mouvements rigides puisque la caméra est solidaire de la tête.

Les résultats des différents systèmes d'estimation des mouvements seront comparés aux mouvements réels. La section suivante présente la manière dont ces mouvements « vrais » sont déterminés.

#### *4.1.2.1. Obtention des données de référence, dites « vérité terrain »*

Les mouvements de référence sur une séquence sont obtenus de manière semi-automatique en quatre étapes :

1. un algorithme de suivi de points par corrélation d'informations de couleur détermine les positions dans l'image (ou les images, si plusieurs vues sont disponibles) d'un ensemble de billes collées sur le visage ;
2. une personne experte inspecte chaque image, procède le cas échéant aux corrections nécessaires puis étiquette les positions des points repérables (comme les commissures des lèvres) ;
3. un jeu de paramètres de contrôle du modèle 3D est estimé par résolution numérique d'une minimisation non-linéaire à partir des positions dans l'image (selon une méthode similaire à celle utilisée pour déterminer les mouvements rigides, page 33) ;
4. une fois toute la séquence traitée, un filtre temporel lisse les trajectoires des paramètres de contrôle.

Ces mouvements de référence ne correspondent donc pas précisément à la réalité puisqu'ils sont « filtrés » par le modèle de forme 3D ; il a été vu que ce modèle ne peut pas reproduire *exactement toutes* les postures articulatoires. Ces réserves faites, ces mouvements de référence seront assimilés dans la suite de ce document à la « vérité terrain »[16].

16. En reprenant ce terme répandu, peut-être mal traduit de l'anglais *ground-truth data*, nous cédons avec regret à l'emploi du jargon scientifique.

## 4.2. DEUX FAMILLES D'EXPÉRIENCES CONTRÔLÉES

### 4.2.1. Visèmes de test

Un sous-ensemble — 20 % — des visèmes utilisés pour la construction du modèle articulatoire a été écarté de l'apprentissage des modèles d'apparence. Les mouvements de tête ont été précisément déterminés durant la construction du modèle de forme ; l'estimation n'a porté que sur les seuls paramètres articulatoires ; la posture neutre est utilisée comme posture de départ de l'algorithme.

La figure 4.1 montre l'erreur résiduelle 3D RMS, calculée avec tous les points du modèle géométrique. Avec les modèles d'apparence *tex_lin*, *tex_cst* et *la_lin* le système retrouve bien la géométrie des visèmes : pour une large part — la moitié des visèmes pour *tex_lin* et *la_lin* — l'erreur résiduelle 3D est inférieure à la précision de la modélisation géométrique. Les résultats pour *tex_lin* et *la_lin* sont meilleurs que ceux avec *tex_cst* et *la_cst* : cela montre les gains procurés par la modélisation articulatoire de l'apparence. Toutefois des échecs ont eu lieu, comme pour [asa]. Le système était initialisé loin de la solution ; en choisissant comme point de départ une posture articulatoire plus proche, le système parvient à une estimation correcte. Au cours d'une séquence, seules des faibles variations entre deux trames successives doivent être estimées [17].

La figure 4.2 représente les erreurs RMS d'estimation des paramètres articulatoires et de quatre paramètres utilisés classiquement en phonétique pour décrire la géométrie des lèvres (la largeur A' ; l'ouverture B' ; la protrusion de la lèvre inférieure Pi ; la protrusion de la lèvre supérieure Ps) [18]. En plus de retrouver les constats précédents, on voit que les paramètres articulatoires jaw2, lar1 et dans une moindre mesure lips4 sont les moins bien retrouvés. On notera que la précision obtenue pour les paramètres descriptifs de la géométrie labiale est de l'ordre du millimètre pour *tex_lin* et *la_lin*.

Les expériences suivantes s'intéressent au comportement dynamique du système, lorsqu'il s'agit de suivre sur toute une séquence les mouvements de parole. Avec une base importante de données de parole naturelle, elles vont permettre d'affiner les remarques faites ici.

17. Une fréquence d'acquisition de 50 Hz est suffisante pour les mouvements faciaux de parole.

18. Ces paramètres sont calculés facilement à partir des trente points 3D qui contrôlent la géométrie des lèvres ; voir ABRY (C.), L.-J. BOË, P. CORSI, R. DESCOUT, M. GENTIL et P. GRAILLOT. *Labialité et phonétique.* Université des langues et lettres de Grenoble, 1980.

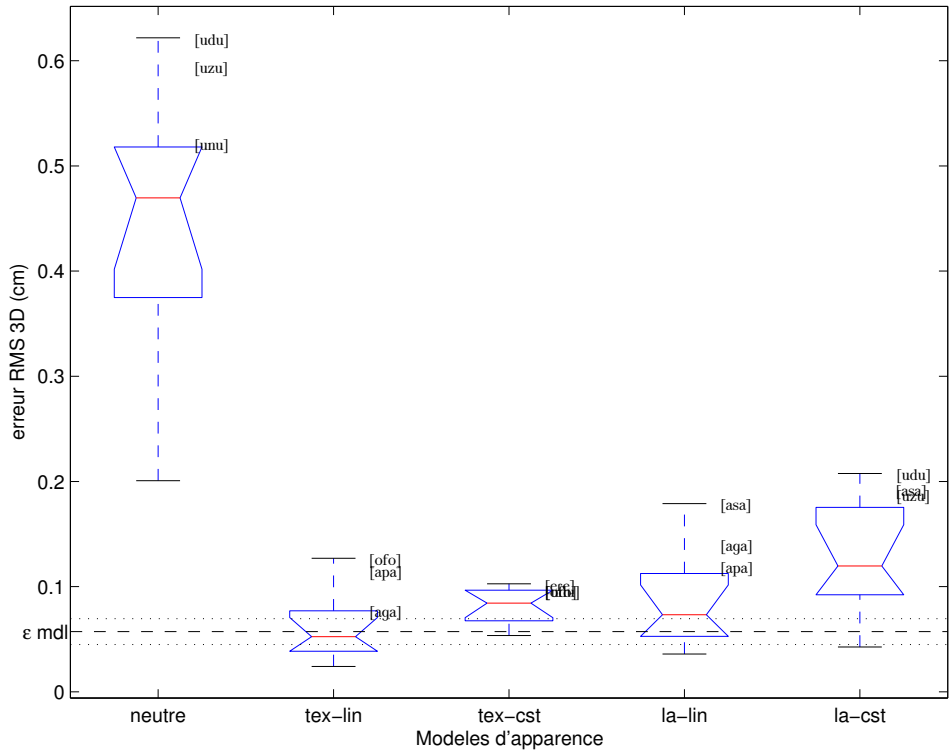



FIGURE 4.1 – Boîtes à moustaches de l'erreur RMS d'estimation de la géométrie 3D sur les visèmes de tests, pour chaque modèle d'apparence. Les trois visèmes les plus mal estimés sont indiqués dans chaque groupe. À titre de comparaison, la distribution des erreurs si l'on considérait la posture neutre — qui est aussi la posture de départ de l'algorithme — comme solution est représentée par la boîte *neutre*. L'erreur de modélisation 3D et son écart-type sont également indiqués.

*tex_lin*

*tex_cst*

*la_lin*

*la_cst*

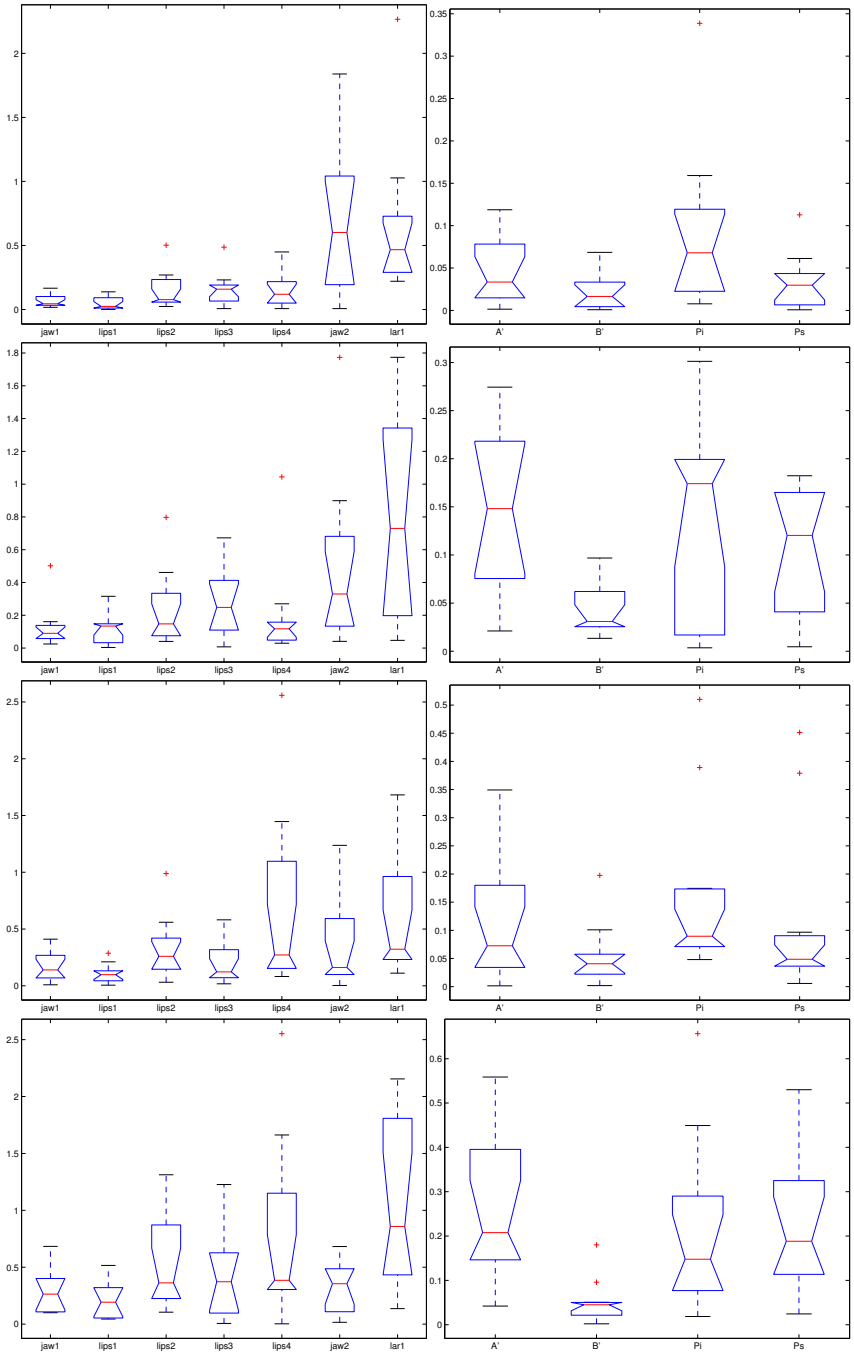



FIGURE 4.2 – Erreurs RMS d'estimation des paramètres articulatoires et de paramètres descriptifs de la géométrie des lèvres (largeur, ouverture, protrusions inf. et sup., exprimés en cm.) sur les visèmes de tests. Les échelles diffèrent.

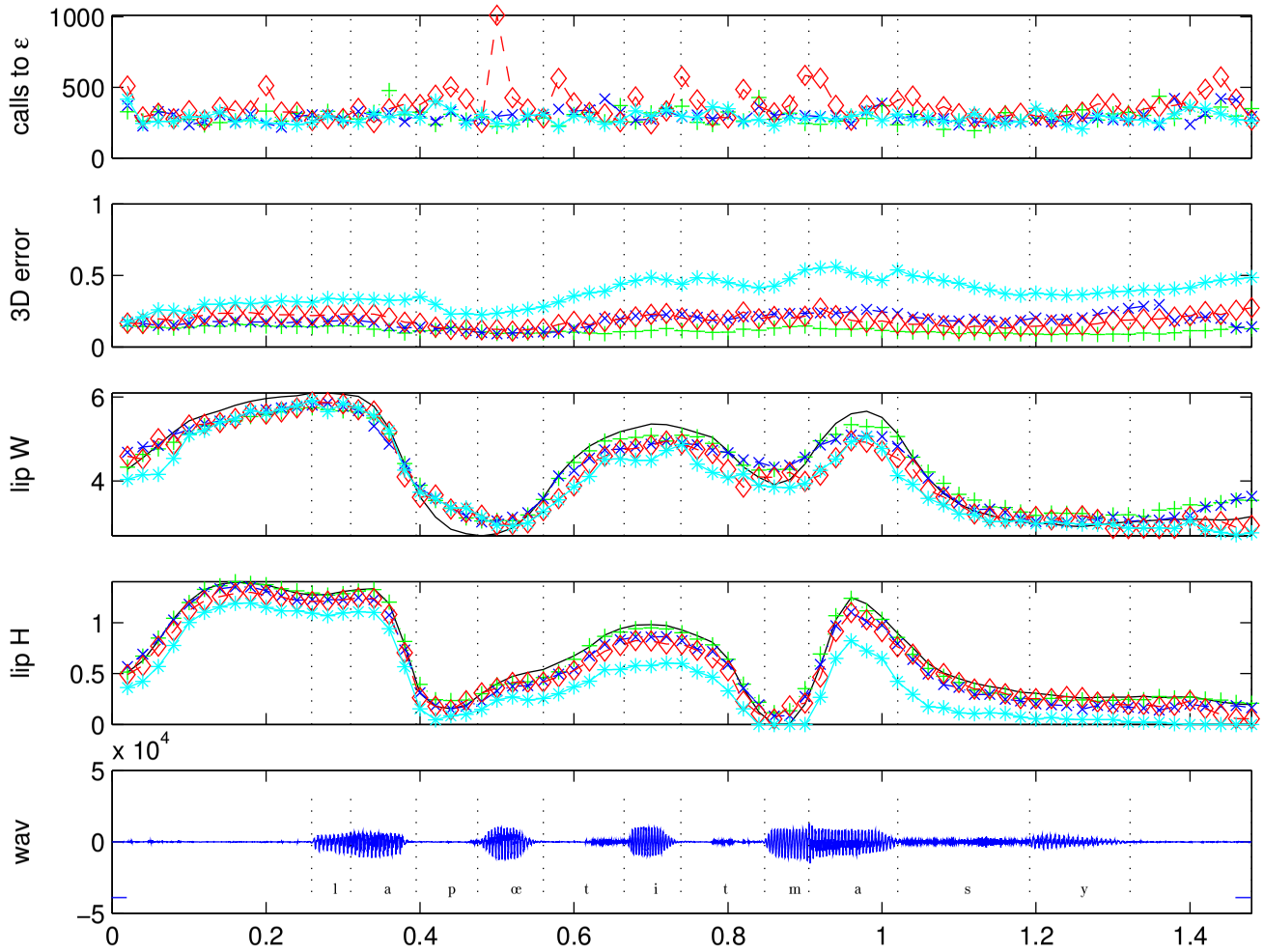



FIGURE 4.3 – Dans les conditions « billes », suivi de la séquence « la petite massue » avec différents modèles d'apparence : *tex_lin*('+'), *tex_cst*('x'), *la_lin*('◇'), *la_cst* ('*'). Le trait plein indique les données de référence. De haut en bas : nombre d'évaluations de la fonction d'écart $\varepsilon$ ; erreur RMS 3D (cm) ; largeur des lèvres (cm) ; ouverture des lèvres (cm) ; signal audio étiqueté.

## 4.2.2. Soixante-dix sept phrases

Pour cette deuxième expérience, le corpus étudié (voir en appendice, page 161) est composé de 77 phrases :

- 68 phrases courtes constituant le corpus articulatoirement riche [19] établi par Bothorel et coll. [20] ;
- parmi les vingt listes proposées par Combescure [21], une liste de 10 phrases phonétiquement équilibrées [22] ; pour des raisons techniques une des dix phrases n'a pas pu être exploitée.

Les 7 546 trames de ces 77 phrases ont été analysées pour estimer les mouvements rigide de tête et les mouvements faciaux [23].

### *4.2.2.1. Gros plan sur « la petite massue »*

Les résultats du suivi pour la phrase « la petite massue » sont illustrés sur la figure 4.3. Les meilleurs résultats sont obtenus avec *tex_lin* ; *tex_cst* et *la_lin* sont équivalents ; *la_cst* est clairement moins bien : les mouvements obtenus avec ce modèle ont moins d'amplitude, comme si, à la différence de *tex_cst*, ce modèle n'était plus valide sur les formes éloignées de la posture où il a été défini.

Les mesures géométriques de largeur et d'ouverture des lèvres montrent que les paramètres articulatoires estimés reproduisent les gestes d'anticipation et atteignent — parfois imparfaitement — les cibles *phonétiques* comme la fermeture des lèvres pour les occlusions bilabiales [p] et [m]. Notons enfin qu'en moyenne la fonction d'écart est appelée environ 300 fois par trame jusqu'à convergence ; c'est principalement dû au fait que nous avons conservé les valeurs conservatrices — utilisées pour les visèmes — pour décider de la fin de l'optimisation.

19. Les phonèmes sont prononcés dans des contextes phonétiques divers qui reflètent la variabilité due à la coarticulation.

20. Bothorel (A.), P. Simon, F. Wioland et J.-P. Zerling. *Cinéradiographie des voyelles et consonnes du français.* Institut de Phonétique de Strasbourg, 1986.

21. Combescure (P.). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56:34–38, 1981.

22. Combescure (1981), *op. cit.* : « Les fréquences relatives pour chaque phonème reflètent celles qui existent dans la langue française. »

23. Seuls 68 visèmes sont utilisés pour la construction des modèles.

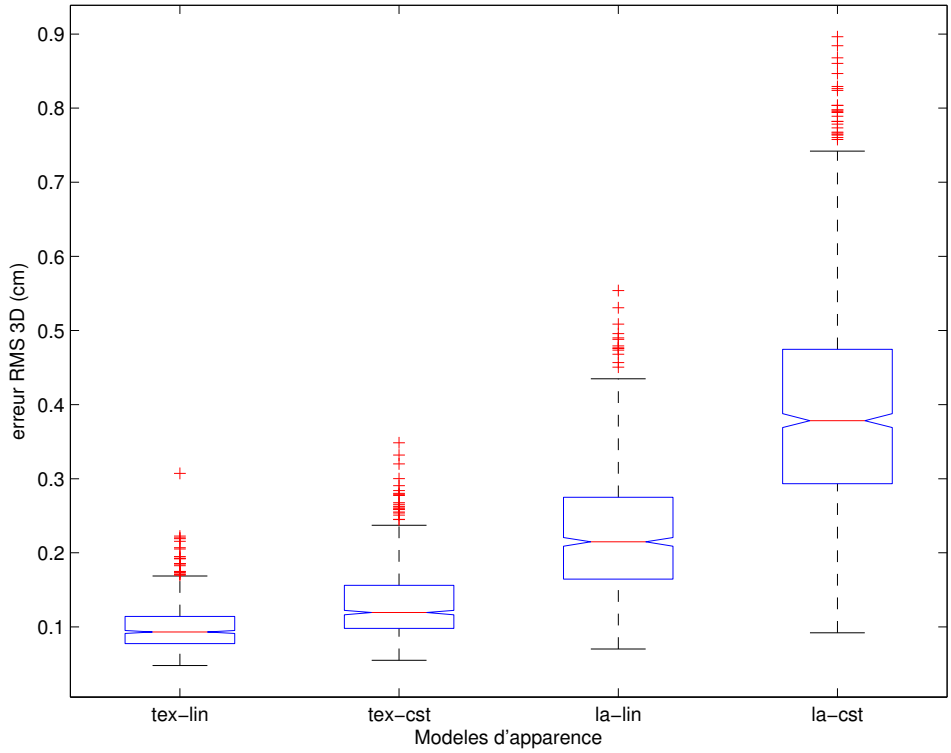
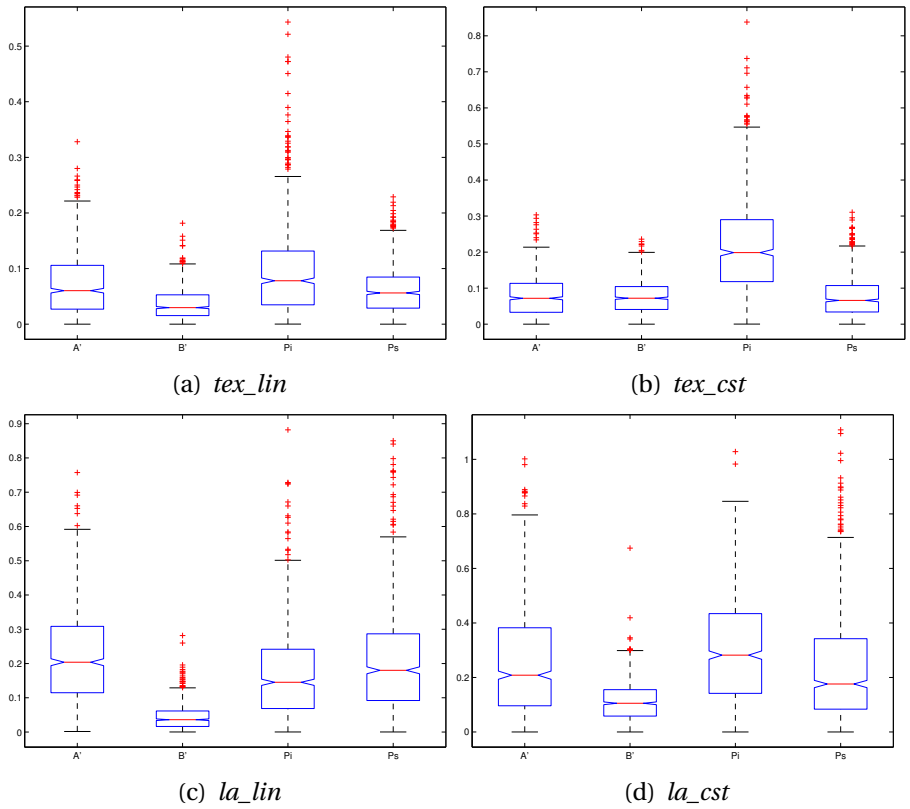



FIGURE 4.4 – Boîtes à moustaches de l'erreur RMS d'estimation de la géométrie 3D aux centres des réalisations acoustiques sur le corpus « 77 phrases », pour chaque modèle d'apparence.

#### *4.2.2.2. Résultats aux centres des réalisations acoustiques*

Sur l'ensemble des 77 phrases les postures articulatoires 3D — les mouvements de tête n'ont pas été considérés — ont été échantillonnées aux centres des réalisations acoustiques des voyelles (385 occurrences en tout) et des consonnes (456 occurrences en tout).

La figure 4.4 montre l'erreur résiduelle 3D RMS, calculée avec tous les points du modèle géométrique. L'erreur ici est plus importante que celle obtenue pour les visèmes ; cela peut s'expliquer par le fait que les images analysées correspondent à des postures articulatoires intermédiaires que le modèle géométrique




(a) *tex_lin* (b) *tex_cst*

(c) *la_lin* (d) *la_cst*

FIGURE 4.5 – Erreurs RMS d'estimation des paramètres descriptifs de la géométrie des lèvres (largeur, ouverture, protrusions inf. et sup.) aux centres des réalisations acoustiques sur le corpus « 77 phrases ». Les échelles diffèrent.

peut avoir des difficultés à reproduire [24]. Les résultats permettent de classer distinctement les modèles d'apparence : le meilleur est *tex_lin* (l'erreur médiane est inférieure à 1 mm) ; vient ensuite *tex_cst* ; puis *la_lin* (l'erreur médiane est environ 2 mm) ; enfin *la_cst* où l'on constate une chute qualitative. Ce classement est confirmé par la figure 4.5 où sont représentées les erreurs RMS d'estimation des quatre paramètres qui décrivent la géométrie labiale. On notera que la précision obtenue pour *tex_lin* est encore de l'ordre du millimètre.

24. Voir la section 1.6.2. (page 44) : le modèle articulatoire ne peut pas s'adapter aux formes associées aux « décollements » asymétriques des lèvres après les occlusions bilabiales.

### *4.2.2.3. Arbres de confusion aux centres des réalisations acoustiques*

Les deux ensembles — voyelles et consonnes — de configurations géométriques 3D de la section précédente ont été classifiés hiérarchiquement afin de construire des arbres de confusions, ou dendrogrammes.

La construction de ces arbres de confusion requiert un tableau donnant les distances pour toutes les paires de groupes. Le calcul de ces distances suppose de pouvoir calculer la distance entre deux postures 3D, $\boldsymbol{a}$ et $\boldsymbol{b}$. Ces deux postures 3D peuvent être représentées sous forme de vecteurs de longueur 3N, où N est le nombre de points 3D du modèle géométrique. La distance euclidienne peut alors être utilisée pour $\boldsymbol{a}$ et $\boldsymbol{b}$ :

$$d(\boldsymbol{a},\boldsymbol{b})^2 = (\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{a}-\boldsymbol{b})$$

Entre deux ensembles $\mathscr{A}$ et $\mathscr{B}$ de postures 3D (par exemple entre [b] et [v]), on peut calculer la distance de Hausdorff :

$$\mathrm{D_H}(\mathscr{A},\mathscr{B}) = \max\{h(\mathscr{A},\mathscr{B}),\ h(\mathscr{B},\mathscr{A})\} \quad \text{avec} \tag{4.1}$$

$$h(\mathscr{U},\mathscr{V}) = \max_{\boldsymbol{u}\in\mathscr{U}} \min_{\boldsymbol{v}\in\mathscr{V}} d(\boldsymbol{u},\boldsymbol{v})$$

Une mesure plus robuste est :

$$h(\mathscr{U},\mathscr{V}) = \mathrm{median}\left\{\min_{\boldsymbol{v}\in\mathscr{V}} d(\boldsymbol{u},\boldsymbol{v}), \quad \boldsymbol{u}\in\mathscr{U}\right\} \tag{4.2}$$

C'est cette dernière forme, appelée distance modifiée de Hausdorff [25], que nous avons retenue.

Enfin, les dendrogrammes sont construits à partir du critère d'agrégation de Ward [26].

La figure 4.6 montre les arbres de confusion géométrique pour les consonnes et pour les voyelles pour chacun des systèmes de suivi et pour la « vérité terrain ».

25. L'inégalité triangulaire n'est pas vraie : mathématiquement il ne s'agit plus d'une distance. Cela reste toutefois une estimation robuste de la dissimilarité entre deux ensembles.

26. Ce critère est basé sur l'inertie : à chaque itération les deux groupes qui seront agrégés sont déterminés de façon à minimiser la variance intra-groupes ; Lebart (L.), A. Morineau et M. Piron. *Statistique exploratoire multidimensionnelle.* Dunod, 1995, (page 167).

Pour les voyelles on retrouve pour tous les systèmes des résultats similaires à ceux obtenus — à partir de données « terrain » — par Robert-Ribes et coll. [27] :

- les voyelles arrondies sont très distinctes des voyelles non-arrondies ;
- les voyelles non-arrondies peuvent être séparées en deux classes, {i, e, ɛ} et {a}.

Pour les consonnes les résultats diffèrent suivant les systèmes. Si l'on choisit l'arbre *vérité terrain* comme référence, on peut se donner la valeur 0,08 comme seuil « d'élagage ». Cette valeur permet de dégager pour la « vérité terrain » sept classes de consonnes, en accord avec les définitions des visèmes de la littérature :

- les bilabiales {p, b, m} ;
- les labiodentales {f, v} ;
- les dentales {t, d, n} ;
- l'uvulaire {ʁ} ;
- l'alvéolaire latérale {l} ;
- une classe composite {s, z, k, g} ;
- les consonnes post-alvéolaires — et arrondies — {ʃ, ʒ}.

Notons que les postures 3D n'incluent pas de points spécifiques à l'articulation linguale ; or si la géométrie du visage permet d'obtenir des informations sur la géométrie de la langue, Bailly et Badin [28] ont montré que cela n'est pas suffisant pour déterminer le lieu de constriction linguale. Cela pourrait expliquer le regroupement dans une seule classe des deux groupes {s, z} et {k, g}.

À cette même hauteur, on retrouve les sept mêmes classes avec *tex_cst*. Les classes obtenues pour *tex_lin* sont elles aussi très pertinentes : on retrouve les classes {p, b, m}, {f, v} et {ʃ, ʒ} ; une classe {ʁ, l} regroupe ces deux consonnes qui diffèrent surtout par la position de la langue ; une classe pour les constrictives alvéolaires {s, z} ; une classe pour les occlusives vélaires {k, g, d}, où toutefois d était plus attendu avec les autres dentales dans la classe {t, n}. Pour les deux systèmes basées sur l'apparence locale — *la_lin* et *la_cst* — la classification obtenue est moins bonne : seule ressort vraiment la distinction entre les consonnes bilabiales et les autres consonnes. Cela peut traduire des performances moindres du suivi aux instants d'échantillonnage.

27. Robert-Ribes (J.), J.-L. Schwartz, T. Lallouache et P. Escudier. Complementarity and synergy in bimodal speech : Auditory, visual and audio-visual identification of French oral vowels in noise. *J. of the Acoustical Society of America*, 103(6):3677–3689, juin 1998.

28. Bailly (G.) et P. Badin. Seeing tongue movements from outside. *In Proc. of the Internat. Conf. on Spoken Language Processing*, pages 1913–1916, Boulder, USA, 2002.

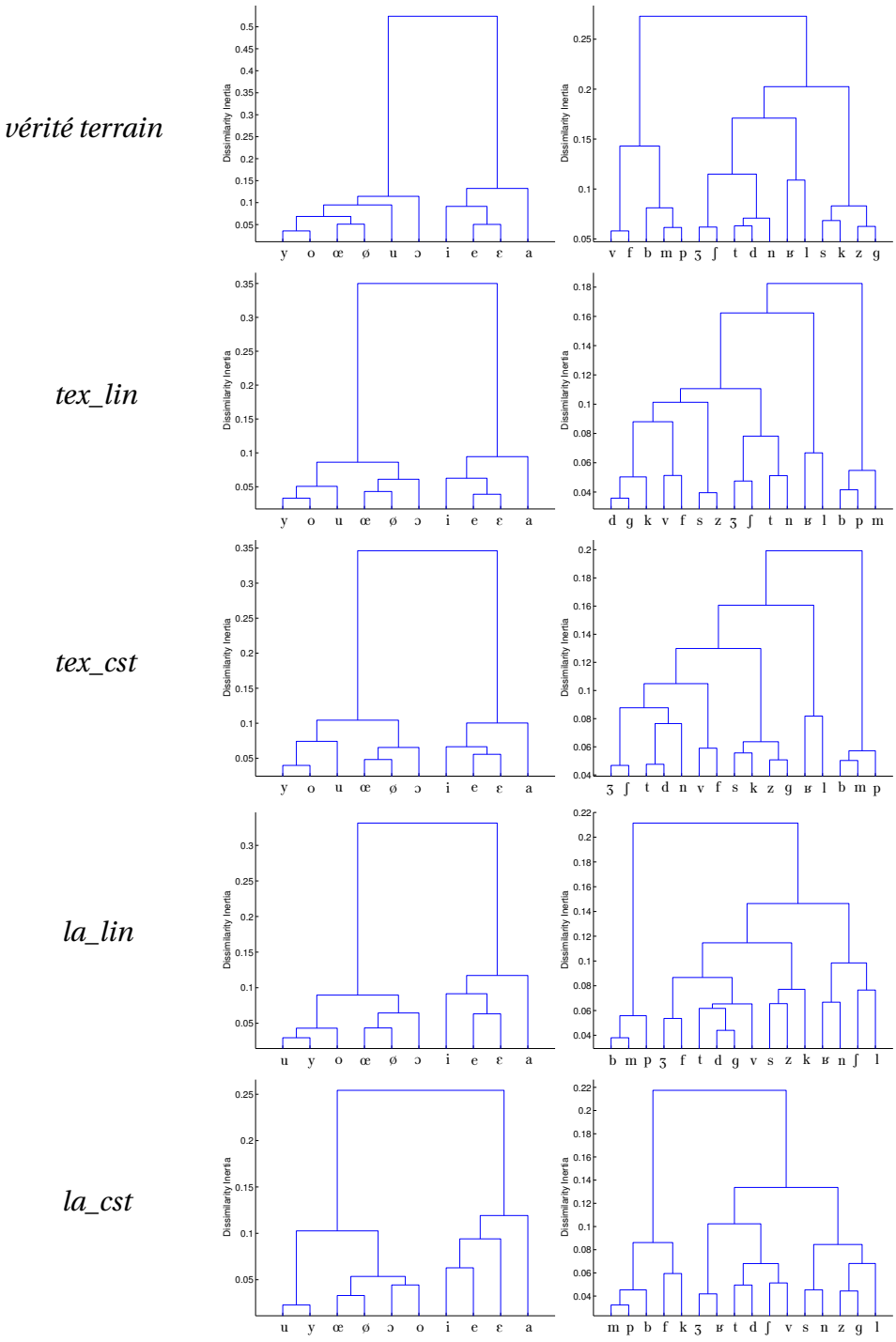



FIGURE 4.6 – Arbres de confusion 3D pour les voyelles et pour les consonnes sur un corpus de 77 phrases. Les classifications hiérarchiques sont représentées pour chaque système de suivi.

## 4.3. DEUX EXPÉRIENCES PLUS « ÉCOLOGIQUES »

Appliqué sur le corpus « téléconférence », le système de suivi se trouve dans des conditions expérimentales offrant aussi une variabilité applicative. Même si le corpus est moins fourni [29], comme pour « billes » les modèles d'apparence ont été évalués sur des visèmes puis sur des phrases. Enfin, le modèle d'apparence *la_cst* qui s'était déjà révélé insuffisant pour « billes » n'a pas été évalué.

### 4.3.1. Huit visèmes

La caméra est solidaire de la tête : le mouvement de tête a été estimé au préalable ; l'estimation n'a porté que sur les seuls paramètres articulatoires ; la posture neutre est utilisée comme posture de départ de l'algorithme. Les visèmes avaient été utilisés pour l'apprentissage des modèles d'apparence.

La figure 4.7 montre l'erreur résiduelle 3D RMS, calculée avec tous les points du modèle géométrique. Les erreurs médianes sont plus grandes que celles obtenues sur les visèmes « billes » ; en revanche pour *tex_lin* et *la_lin* elles sont équivalentes à celles obtenues sur les 77 phrases « billes » ; les résultats pour *tex_cst* sont eux inférieurs. Comme précédemment des échecs — p. ex. [i] — ont eu lieu ; on peut regarder des exemples d'ajustements réussi et raté sur la figure 4.9. Comme précédemment le système était initialisé loin de la solution et le choix pour le point de départ d'une posture articulatoire plus proche conduit à une estimation correcte.

La figure 4.8 représente les erreurs RMS d'estimation des paramètres articulatoires et de quatre paramètres descriptifs de la géométrie labiale. Les paramètres articulatoires lips4, jaw2 et lar1 ne sont pas retrouvés. Cette dégradation se retrouve pour les paramètres labiaux.

Rappelons que le modèle projectif de la caméra ne permet pas de prendre en compte certains défauts de la caméra utilisée : cela nuit à l'estimation de la vérité terrain (d'autant plus qu'il n'y a qu'une seule vue à disposition), notamment pour les paramètres qui influent surtout sur les coordonnées en $z$ comme jaw2.

29. Nous avons déjà eu l'occasion de dire (page 19) les limites quantitatives du matériau de ce corpus.

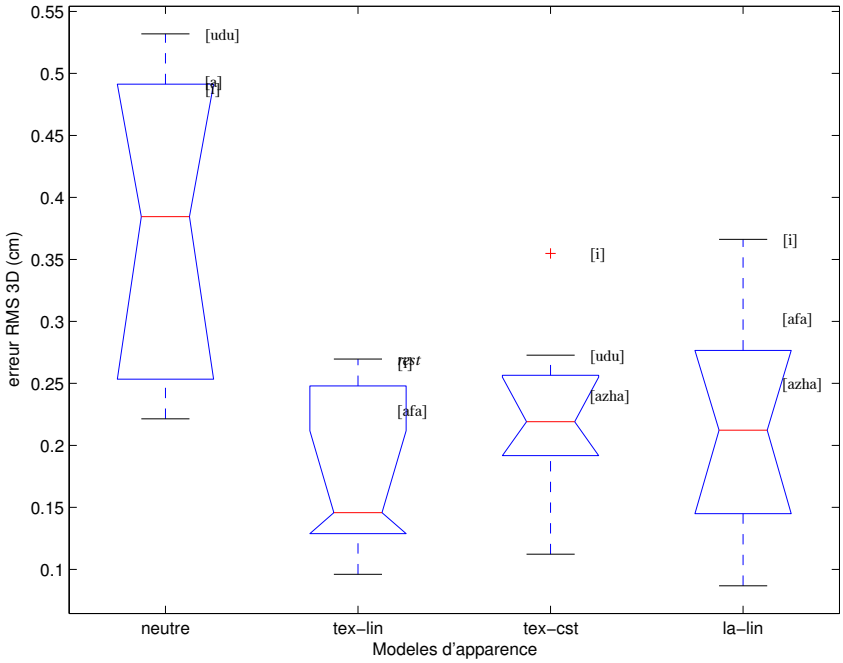



FIGURE 4.7 – Boîtes à moustaches de l'erreur RMS d'estimation de la géométrie 3D sur les visèmes « téléconférence », pour chaque modèle d'apparence. Les trois visèmes les plus mal estimés sont indiqués dans chaque groupe. À titre de comparaison, la distribution des erreurs si l'on considérait la posture neutre — qui est aussi la posture de départ de l'algorithme — comme solution est représentée par la boîte *neutre*.

*tex_lin*

*tex_cst*

*la_lin*

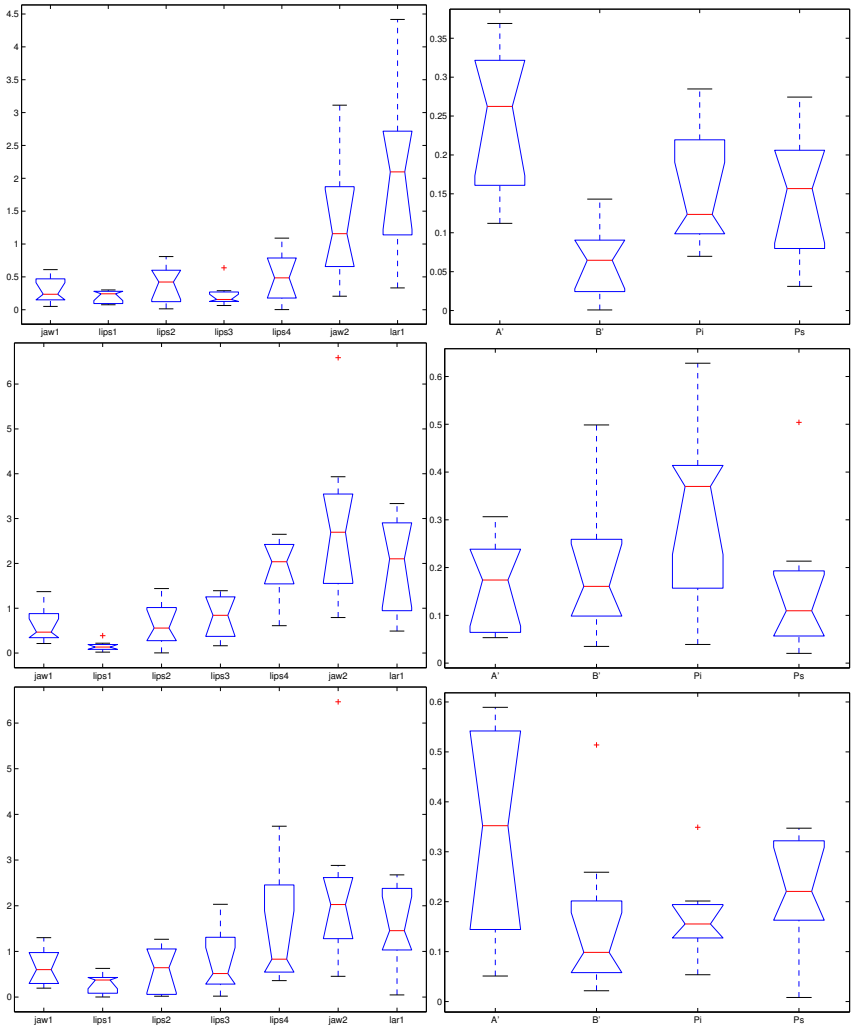



FIGURE 4.8 – Erreurs RMS d'estimation des paramètres articulatoires et de paramètres descriptifs de la géométrie des lèvres (largeur, ouverture, protrusions inf. et sup.) sur les visèmes « téléconférence ». Les échelles diffèrent.

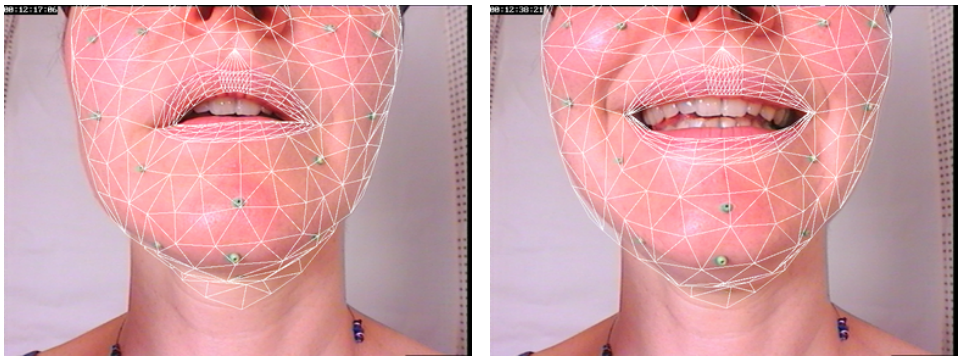


FIGURE 4.9 – Meilleur et pire ajustements obtenus avec le modèle d'apparence *la_lin* pour les visèmes « téléconférence » : à gauche, le visème *préphonatoire* ; à droite, [i]. Des pré-traitements appliqués aux images utilisées pour la construction des modèles d'apparence empêchent le système de tirer profit des billes collées sur le visage (voir page 52).

### 4.3.2. Gros plan sur « la vaisselle propre »

Seule, la séquence « la vaisselle propre est mise sur l'évier » a été écartée de la construction des modèles d'apparence. Comme on le voit sur la figure 4.10, dans cette expérience les résultats pour *tex_lin* et *la_lin* sont bien meilleurs que ceux pour *tex_cst* : notamment, avec ce dernier les occlusions bilabiales [p], [b] et [m] ne sont pas reproduites — les lèvres ne ferment pas complètement.

En fin de séquence, quand il n'y a plus de signal audio et que les lèvres retournent à la position de repos, on observe des oscillations, principalement avec le modèle *la_lin*. Ces oscillations ne sont pas présentes en début de séquence, également en absence de signal audio. Cela pourrait s'expliquer ainsi : d'une part par le fait que le visème *rest* n'était pas bien retrouvé tandis que le visème *prephonatoire*, si ; d'autre part les trames paire et impaire délivrées par la caméra ont des propriétés photométrique dont les différences sont visibles à l'œil nu [30], ce qui a un impact sur le système.

30. Sur un écran d'ordinateur ; ce n'est pas le cas une fois les images imprimées en niveau de gris… Le système traite successivement les deux trames d'une image.

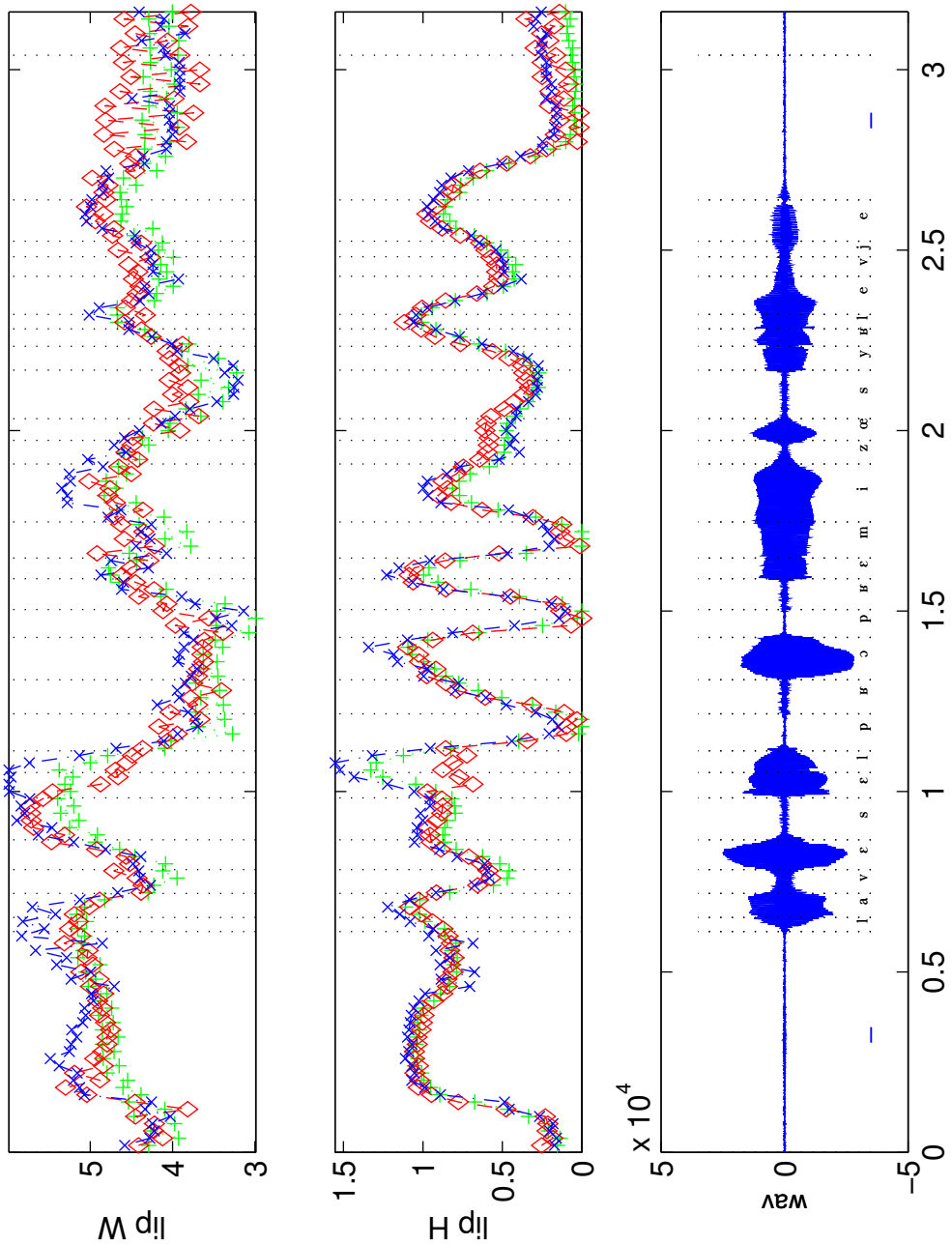



FIGURE 4.10 – Dans les conditions « téléconférence », suivi de la séquence « la vaisselle propre est mise sur l'évier » pour différents modèles d'apparence : *tex_lin* ('+'), *tex_cst* ('x'), *la_lin* ('◇'). De haut en bas : largeur des lèvres (cm) ; ouverture des lèvres (cm) ; signal audio étiqueté.

## 4.4. CONCLUSION

Alors qu'en général — dans les travaux publiés — la place consacrée à l'évaluation des méthodes d'analyse des mouvements faciaux mises en œuvre est restreinte, nous nous sommes attachés à quantifier la qualité de l'estimation sur des séquences réelles, par rapport aux mouvements effectivement produits et par rapport à des choses connues en phonétique. Nous avons choisi de tester notre système dans un paradigme monoculaire, le visage face à la caméra ; ce système fonctionne aussi — et bien sûr d'autant mieux — avec plusieurs vues synchrones du visage.

La première série d'expériences a porté sur des images acquises dans les conditions bien contrôlées de la construction du modèle articulatoire ; un corpus riche à disposition, la bonne qualité de la caméra, la prise de vue en gros plan sur le visage, l'éclairage constant, le peu d'amplitude des mouvements rigides, la disposition aisée des mouvements réels et enfin la présence de nombreuses billes collées sur le visage ont permis de bien poser le problème de l'évaluation, comparée ou non, des modèles d'apparence. Cette série d'expérience a permis de constater que les modèles d'apparence qui varient avec l'articulation ont des performances supérieures à celles des modèles constants ; plus exactement, le modèle *tex_lin* est meilleur, mais les modèles *tex_cst* et *la_lin* — dans une moindre mesure — permettent eux aussi de retrouver précisément les mouvements faciaux produits. Le modèle constant de l'apparence locale *la_cst* n'a lui clairement pas la qualité requise.

La deuxième série d'expériences a porté sur des images acquises dans des conditions réalistes d'une application de type téléconférence virtuelle ; le bas du visage, cette fois sans marqueurs interférants, filmé en gros plan par une caméra (solidaire de la tête), les images saturées par l'éclairage ambiant ont contribué à créer des conditions où l'estimation était difficile. Les expériences, trop restreintes pour que l'on puisse conclure, confirment que les meilleures performances sont obtenues avec *tex_lin* ; c'est cette fois le modèle *la_lin* qui est un peu mieux que *tex_cst*.

Afin d'aller plus loin dans ces évaluations quantitatives des résultats, de pouvoir comparer plus facilement plusieurs méthodes d'analyse des mouvements faciaux, nous trouverions utile de pouvoir accéder à une base de données publique, dans la continuité des bases construites pour tester les systèmes de reconnaissance d'identité ou de reconnaissance de la parole. Cette base compor-

terait notamment des données de références 3D pour un corpus de parole important et varié ; la mise à disposition aux autres chercheurs de la communauté des données de chacun pourrait déjà permettre d'amorcer le processus.

# 5

# Évaluation perceptive

## 5.1. INTRODUCTION

Lors des évaluations présentées dans le chapitre précédent, nous avons comparé d'un point de vue géométrique les mouvements et les postures articulatoires estimés avec la vérité terrain et les connaissances phonétiques. Les scénarios mettant en jeu de la synthèse de mouvements par des têtes parlantes virtuelles constituent une des grandes familles d'applications de notre système d'analyse de la vidéo. Les problématiques liées à la caractérisation de l'interaction entre l'humain et la machine constituent un champ très vaste de recherche ; même s'il est plus restreint, un enjeu important pour « connaître » la qualité d'une animation synthétique de parole est la connaissance de la perception (au sens « état cognitif ») que peut avoir une personne à l'écoute de cette animation.

Ainsi, l'animation faciale (re-) synthétisée à partir des mouvements estimés est-elle effectivement perçue comme un visage en train de parler ? Ou, plus concrètement, reproduit-elle pour la perception les propriétés de signaux de parole naturels ?

En complément de l'étude de leurs performances intrinsèques, les têtes parlantes de synthèse sont aussi évaluées de manière perceptive comparativement à la parole réelle, c.-à-d. produite par l'humain. De même que pour les systèmes

L'acquisition des données audiovisuelles doit beaucoup à C. Savariaux et à A. Arnal. C. Bulfone et N. Medves ont configuré et adapté le matériel informatique pour qu'il puisse être utilisé lors des expériences de perception. Le test de l'acuité visuelle nous a été fourni et expliqué par S. Bechon. C. Lavergne, J.-L. Schwartz et F. Berthommier nous ont donné divers conseils et références. Des étudiants et des étudiantes de l'ENSERG ont participé aux expériences de perception.

de synthèse audio [1], on peut distinguer les évaluations perceptives des têtes parlantes virtuelles en deux catégories complémentaires :

- les évaluations qualitatives, basées sur une appréciation directe des têtes parlantes vues comme un partenaire de la communication — comment est explicitement perçu le médium, ou canal d'encodage et de transmission des signaux multimodaux ?
- les évaluations quantitatives, portant sur la transmission de l'information de parole — quelle est l'efficacité, la capacité de ce canal de transmission ?

Nous présenterons dans ce chapitre l'évaluation de nos résultats dans ces deux catégories.

### 5.1.1. Regards et appréciations : évaluations qualitatives

Theobald et coll. ont évalué leur synthèse de parole visuelle — couplée avec le signal audio d'origine — en faisant noter la naturalité [2] de sa dynamique pour des phrases de test [3].

Leur tête parlante est basée sur un *Active Appearance Model* ; pour la synthèse d'un phonème, les paramètres de contrôle résultent du mélange de plusieurs exemplaires, sélectionnés dans un dictionnaire d'après leur similarité selon un contexte triphonique ; la série temporelle de ces paramètres de contrôle est ensuite filtrée par une *spline* pour adoucir les transitions.

L'expérience 1 évaluait l'effet de ce filtrage par *spline* en notant conjointement des séquences vidéos régularisées par le modèle statistique du visage du même locuteur, lissées ou non. La dynamique des séquences non lissées a été jugée plus naturelle que celle des séquences lissées. Ces séquences lissées servaient de référence *a priori* haute dans l'expérience 2 qui évaluaient l'effet du nombre d'exemplaires utilisés pour la synthèse d'un phonème donné. La référence *a priori* basse (une sélection aléatoire dans le dictionnaire) est jugée significativement moins naturelle que les séquences de synthèse — quelque soit le nombre d'exemplaires —, elles-mêmes significativement moins naturelles que

1. Benoît (C.) et L. C. W. Pols. On the assessment of synthetic speech. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines : Theories, Models and Designs*, pages 435–441. Elsevier B.V., 1992.

2. un néologisme pour traduire l'anglais *naturality*

3. Theobald (B.-J.), J. A. Bangham, I. Matthews et G. C. Cawley. Evaluation of a talking head based on appearance models. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 187–192, St Jorioz, France, septembre 2003.

les séquences « d'origine » (régularisées par le modèle statistique et lissées).

En résumé, la synthèse visuelle a été jugée moins naturelle que les séquences vidéos d'origine. Notons également que la présentation du même locuteur en synthèse et en naturel permet de s'affranchir de la variabilité interpersonnelle.

PANDZIC ET COLL. ont étudié la perception de plusieurs têtes parlantes virtuelles pour un scénario applicatif donné [4]. Selon les expériences, 145 ou 190 personnes ont participé ; 39 % d'entre elles n'étaient pas de de langue maternelle anglaise.

L'expérience 1 comparait l'intelligibilité de deux têtes parlantes de synthèse : un modèle 3D contrôlé d'après (Cohen et Massaro, 1993) [5] rendu « à la Gouraud », c.-à-d. avec un aspect « cireux » (à l'image des nomogrammes de la figure 1.8, les polygones ne sont pas texturés, ils ont des propriétés photométriques beaucoup plus simples que celles d'un vrai visage), et un système de synthèse par patches 2D superposés pour composer l'image finale. Nous reviendrons plus loin sur les gains d'intelligibilité *per se* ; par ailleurs, la durée totale de l'expérience était plus grande en présence de bruit *vs.* parole claire, et plus grande pour la synthèse par patches 2D *vs.* tête « à la Gouraud » ou audio seul. Même si une vidéo naturelle manquait comme référence, ces mesures indirectes semblent montrer que l'effort cognitif de traitement de l'information audiovisuelle était plus grand pour ce synthétiseur visuel par patches 2D, traitement qui aurait donc pu être de nature différente.

Dans l'expérience 2, un questionnaire concernant les impressions ressenties après utilisation du service — agent d'accueil sur un portail web — était rempli. Comme le disent les auteurs, le système d'animation faciale n'était pour ce service qu'un gadget ; les auteurs trouvent remarquable que dans ces conditions la tête parlante ait été jugée en moyenne légèrement utile, mais cela n'apprend rien sur la qualité intrinsèque de l'animation ou sur la perception par l'humain.

Nous citons brièvement une troisième expérience qui comparait l'attrait de ces deux têtes parlantes de synthèse et d'une version texturée de la tête 3D, après présentation d'une phrase d'accueil. Les résultats nous paraissent difficilement interprétables : qu'est-ce qui a plu ? déplu ?

4. PANDZIC (I. S.), J. OSTERMANN et D. MILLEN. User evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15:330–340, 1999.

5. COHEN (M. M.) et D. W. MASSARO. Modeling coarticulation in synthetic visual speech. *In Proceedings of Computer Animation*. Springer-Verlag, 1993.

GEIGER ET COLL. ont confronté dans des tests de Turing leur synthèse de parole visuelle *Mary 101* aux séquences vidéos d'origine de la même locutrice [6]. *Mary 101* consiste en une image de fond sur laquelle est superposée au niveau du visage une image de synthèse, construite en fonction du contexte phonétique par mélange de plusieurs visèmes.

Pour ces tests, les participantes et les participants devaient catégoriser les séquences présentées comme étant des séquences naturelles ou de synthèse. Ces tests étaient déclinés sous trois versions, à chaque fois pour des phrases et pour des mots : dans l'expérience 1, les séquences naturelles et synthétiques étaient présentées couplées avec le son d'origine et en ordre aléatoire ; pour l'expérience 2, la principale différence était la contrainte pour les participants et les participantes de répondre rapidement (dans un intervalle de temps de 1 s) ; enfin, dans l'expérience 3 pour chaque séquence la synthèse et le naturel étaient présentés conjointement (paire S–N ou paire N–S), et sans le son.

Dans ces trois déclinaisons du test de Turing, la synthèse n'a pas été différenciée de la vidéo naturelle. Ce résultat impressionnant est toutefois remis en perspective par leur quatrième expérience — reprise ci-dessous — de mesure de l'intelligibilité des deux présentations.

UNE CONCLUSION SUR CES TROIS ÉTUDES Même lorsqu'elles sont correctement appréhendées, ces expériences où l'on demande un jugement explicite sur l'animation présentée ne permettent pas d'aller vraiment au delà des résultats immédiats : si une différence entre naturel et synthèse est observée, où se situe-t-elle ? qu'est-ce qui est apprécié ? quels sont les composantes de l'animation qu'il faudra améliorer ?

### 5.1.2. L'intelligibilité : une évaluation quantitative

Dans les applications qui nous intéressent — celles où la parole constitue le cœur du système —, il est crucial que la tête parlante soit intelligible.

6. GEIGER (G.), T. EZZAT et T. POGGIO. Perceptual evaluation of video-realistic speech. AI Memo 2003-003, Massachusetts Institute of Technology, Cambridge, USA, février 2003.

### 5.1.2.1. Arguments pour la détermination du corpus

Les signaux émis par la tête parlante sont interprétés en tant que tels, avec leurs qualités et leurs défauts, et aussi d'après les connaissances et acquis linguistiques propres à l'individu qui l'écoute, l'observe [7]. Ainsi, des imperfections éventuelles du signal de parole peuvent être supplées par des processus linguistiques de haut niveau mettant en jeu chez l'observateur ou l'observatrice : le lexique, le contexte au niveau du mot, au niveau de la phrase, la sémantique, etc.

Plusieurs protocoles de mesure d'intelligibilité, qui « se placent » à des niveaux différents (phrases ou mots avec ou sans signification, test de rime, etc.) ont été proposés [8].

Pour continuer de disséquer notre tête parlante de synthèse, nous avons fait le choix d'étudier son intelligibilité au niveau consonantique sur un corpus de logatomes de type voyelle–consonne–voyelle (*VCV*). Les dix consonnes que nous avons retenues — elles seront énumérées à la section 5.4.1.1. — permettent de considérer la perception de tous les traits phonétiques, à l'exception du contraste voisement *vs.* non-voisement.

### 5.1.2.2. Évaluations de têtes parlantes

Pour une fois, cette section commence par une conclusion : toutes les évaluations des têtes parlantes de synthèse ont montré que ces dernières apportaient un gain d'intelligibilité par rapport au seul signal audio en présence de bruit, et

7. Lindblom (B.). Role of articulation in speech perception : Clues from production. *J. of the Acoustical Society of America*, 99(3):1683–1692, mars 1996.

8. Voir p. ex. Voiers (W. D.). Performance evaluation of speech processing devices : Diagnostic evaluation of speech intelligibility. AF Cambridge Research Laboratories Final Report AFCLR-67-0101, Contract AF19(628)-4987, 1967 ; Peckels (J. P.) et M. Rossi. Le test de diagnostic par paires minimales, adapatation au français du *Diagnostic Rhyme Test* de Voiers. *Revue d'Acoustique*, 27:245–262, 1973 ; House (A. S.), C. E. Williams, M. H. L. Hecker et K. D. Kryter. Articulation-testing method : Consonantal differentiation with a closed response set. *J. of the Acoustical Society of America*, 37:158–166, 1965 ; Falaschi (A.). Segmental quality assessment by pseudo-words. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines : Theories, Models and Designs*, pages 455–472. Elsevier B.V., 1992 ; et, pour plus une présentation générale, Benoît (C.) et L. C. W. Pols. On the assessment of synthetic speech. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines : Theories, Models and Designs*, pages 435–441. Elsevier B.V., 1992 ; Dutoit (T.). *An introduction to text-to-speech synthesis*, chapitre 7.3, pages 195–200. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1997.

souvent même en milieu non-bruité, mais que ce gain était inférieur à celui d'une vidéo naturelle d'une personne.

Puisqu'en parole claire l'acoustique suffit à la compréhension générale, classiquement le signal audio est bruité afin d'assurer une plus grande importance de la modalité visuelle.

Nous commençons une énumération en citant Cohen et coll., qui ont mesuré en lecture labiale pour des mots monosyllabiques l'intelligibilité de Baldi, une des toutes premières têtes parlantes de synthèse [9] ; Baldi est un fils du modèle de Parke [10], mais avec une langue ; il est contrôlé d'après (Cohen et Massaro, 1993) [11]. Dans une étude plus récente où Baldi était conformé à l'anatomie et à l'articulation d'un locuteur, son intelligibilité a été évaluée sur des phrases en présence de bruit blanc [12].

L'intelligibilité des têtes parlantes de l'ICP de la première génération a été évaluée sur des logatomes *VCVCV* pour plusieurs niveaux de bruit blanc [13]. Une évaluation similaire pour le japonais a été réalisée avec des logatomes *CV* par Yamamoto et coll. [14].

Les expériences au laboratoire KTH ont porté sur des logatomes *VCV* [15], incluant aussi des phrases [16] ; suivant l'application visée, s'il est présenté le signal

9. Cohen (M. M.), R. L. Walker et D. W. Massaro. Perception of synthetic visual speech. *In* Stork (D. G.) et M. E. Hennecke, éditeurs. *Speechreading by Humans and Machines*, volume 150 de *Computer and Systems Sciences*, pages 153–168. Springer, 1996.

10. Parke (F. I.). Parameterized models for facial animation. *IEEE Computer Graphics & Applications*, 2:61–68, novembre 1982.

11. Cohen (M. M.) et D. W. Massaro. Modeling coarticulation in synthetic visual speech. *In Proceedings of Computer Animation*. Springer-Verlag, 1993.

12. Cohen (M. M.), D. W. Massaro et R. Clark. Training a talking head. *In Proc. of the IEEE Internat. Conf. on Multimodal Interfaces*, pages 499–504, Pittsburgh, USA, octobre 2002.

13. Guiard-Marigny (T.), D. Ostry et C. Benoît. Speech intelligibility of synthetic lips and jaw. *In Proc. of the Internat. Congress of Phonetic Sciences*, volume 3, pages 222–225, Stockholm, Sweden, août 1995.

14. Yamamoto (E.), S. Nakamura et K. Shikano. Lip movement synthesis from speech based on Hidden Markov models. *Speech Communication*, 26(1-2):105–115, octobre 1998.

15. Voir p. ex. Beskow (J.). Animation of talking agents. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 149–152, Rhodes, Greece, septembre 1997.

16. Voir p. ex. Agelfors (E.), J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens et T. Öhman. Synthetic faces as a lipreading support. *In Proc. of the Internat. Conf. on Spoken Language Processing*, pages 3047–3050, Sydney, Australia, 1998 ; Siciliano (C.), G. Williams, J. Beskow et A. Faulkner. Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. *In Proc. of the Internat. Congress of Phonetic Sciences*, pages 131–134, Barcelona, Spain, août 2003.

audio est filtré pour simuler une transmission téléphonique ou l'effet de pathologies auditives.

Olivès et coll. [17] ont évalué leur tête parlante finnoise sur des logatomes *VCV* pour plusieurs niveaux de bruit rose. Williams et Katsaggelos [18] ont évalué un synthétiseur audio-vers-visuel (d'images de lèvres et bouche, superposées sur une image de fond) sur des mots et phrases pour plusieurs niveaux de bruit blanc. Fagel et Clemens [19] ont évalué le synthétiseur visuel de leur tête parlante au moyen d'un test de rime en allemand pour un niveau de bruit blanc de −6 dB.

Revenons à l'étude de Pandzic et coll. [20]. L'expérience 1 comparait l'intelligibilité de deux têtes parlantes de synthèse pour des séries de cinq chiffres. Il est à noter que ce vocabulaire n'exploite qu'un sous-espace articulatoire : en anglais — et en français — les chiffres (ou les nombres jusqu'à 999) sont articulés sans occlusions bilabiales. En parole claire, le score de l'audio seul était au niveau de celui des deux têtes parlantes ; en présence de bruit, l'audio seul était significativement moins intelligible que les deux têtes parlantes.

Enfin, dans l'expérience 4 de (Geiger et coll., 2003) [21], les auteurs ont comparé pour la lecture labiale de phrases et de mots l'intelligibilité de *Mary 101* et celle de la locutrice réelle. Les performances de la synthèse étaient inférieures à celles du naturel. *Si les participants et les participantes ne pouvaient pas différencier explicitement la vidéo de synthèse de la vidéo naturelle, ils l'ont donc fait implicitement en essayant de comprendre le message.* Cela illustre la complémentarité entre les deux points de vue de l'évaluation évoqués précédemment : la qualité générale et l'intelligibilité.

17. Olivès (J.-L.), J. Kulju, R. Möttönen et M. Sams. Audio-visual speech synthesis for Finnish. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 157–162, Santa Cruz, USA, 1999.

18. Williams (J. J.) et A. K. Katsaggelos. An HMM-based speech-to-video synthesizer. *IEEE Trans. on Neural Networks*, 13(4):900–915, juillet 2002.

19. Fagel (S.) et C. Clemens. An articulation model for audiovisual speech synthesis — determination, adjustment, evaluation. *Speech Communication*, 44(1–4):141–154, octobre 2004.

20. Pandzic (I. S.), J. Ostermann et D. Millen. User evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15:330–340, 1999.

21. Geiger (G.), T. Ezzat et T. Poggio. Perceptual evaluation of video-realistic speech. AI Memo 2003-003, Massachusetts Institute of Technology, Cambridge, USA, février 2003.

## 5.2. LE PARADIGME DES POINTS LUMINEUX

### 5.2.1. Quelques points sur la perception du mouvement biologique

Nos travaux peuvent en être un exemple, le domaine de la vision par ordinateur s'intéresse beaucoup au problème de la détermination de la forme 3D à partir d'images 2D. Une approche classique est la *forme d'après le mouvement*[22], dans laquelle la forme 3D, ou structure spatiale, est déterminée à partir de positions dans le plan image (ou les plans images dans le cas de dispositifs stéréoscopiques par exemple) de primitives 2D — p. ex. points, contours ; ces primitives, qui correspondent aux projections de l'objet 3D, sont supposées avoir été extraites au préalable par des pré-traitements d'image.

En vision par ordinateur, l'inférence de la structure 3D à partir de ces primitives 2D est computationnelle ; même si les algorithmes utilisés peuvent inclure des contraintes sur la forme 3D, la détermination de cette forme 3D (ainsi que les mouvements entre les vues) résulte de calculs purement mathématiques. En plus des travaux cités au chapitre 3 (page 72), nous renvoyons à Jebara et coll. [23] pour une présentation générale [24].

En revanche, d'un point de vue perceptif, pour déterminer la structure 3D l'humain ne semble pas traiter l'information de mouvement suivant de telles analyses mathématiques, mais plutôt en utilisant des heuristiques particulières [25].

22. une traduction (inusitée !) de l'anglais *structure-from-motion*

23. Jebara (T.), A. Azarbayejani et A. Pentland. 3D structure from 2D motion. *IEEE Signal Processing Magazine*, 16(3):66–84, 1999.

24. Nous faisons cependant une exception et citons les travaux précurseurs d'Ullman ; p. ex. (Ullman, 1976) prouve que trois vues orthographiques de quatre points non-coplanaires permettent de déterminer les distances entre ces points à un facteur d'échelle près ; Ullman (S.). The interpretation of structure from motion. AI Memo 476, Massachusetts Institute of Technology, Cambridge, USA, octobre 1976.

25. Voir Braunstein (M. L.). Structure from motion. *In* Smith (A. T.) et R. J. Snowden, éditeurs. *Visual Detection of Motion*, pages 367–393. Academic Press, 1994 ; Domini (F.) et C. Caudek. 3-D structure perceived from dynamic information : a new theory. *TRENDS in Cognitive Sciences*, 7 (10):444–449, octobre 2003.

### 5.2.1.1. *Expériences* princeps

Afin de comprendre la perception visuelle de ce qu'il a appelé le mouvement biologique, Johansonn a étudié l'interprétation d'affichages de points lumineux en mouvement [26] — ces points lumineux étaient situés sur les principales articulations (c.-à-d. chevilles, genoux, hanches, poignets, coudes, épaules) du corps de personnes, au cours d'activités diverses. Les résultats montrent que la cinématique de ces points lumineux — dont les mouvements sont cohérents — suffit pour identifier des activités comme marcher, danser, courir ou faire des pompes, alors qu'une configuration statique de ces points reste indéchiffrable.

Kozlowski et Cutting [27] ont montré que des stimuli cinématiques de cette nature permettent de distinguer un marcheur d'une marcheuse ; d'autre part, des stimuli « moins biologiques » (en changeant la fréquence de la marche ou l'amplitude de la course des bras) rendent l'interprétation plus difficile.

Ces stimuli permettent de déterminer des paramètres dynamiques : dans l'expérience de Runeson et Frykholm les observateurs et les observatrices peuvent estimer la masse d'une boîte portée par un mime [28].

En ce qui concerne la perception du visage, Bassili [29] a montré qu'un rendu sous forme de points lumineux permet d'identifier des émotions — grâce au mouvement ; en statique, le visage n'est même pas identifié.

Depuis ces premières expériences et grâce à la disponibilité des systèmes de capture de mouvement optiques, les affichages sous forme de points lumineux sont devenus un outil méthodologique classique pour étudier la perception du mouvement.

26. Johansson (G.). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973 ; Jansson (G.), S. S. Bergström et W. Epstein, éditeurs. *Perceiving events and objects*. Lawrence Erlbaum Associates, 1994.

27. Kozlowski (L. T.) et J. E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

28. Runeson (S.) et G. Frykholm. Visual perception of lifted weight. *J. of Experimental Psychology : Human Perception and Performance*, 7:733–740, 1981.

29. Bassili (J. N.). Facial motion in the perception of faces and of emotional expressions. *J. of Experimental Psychology : Human Perception and Performance*, 4:373–379, 1978.

### *5.2.1.2. Perception de la parole sous forme de points lumineux*

Ce paradigme d'étude de la perception s'est développé pour la parole depuis les travaux de Summerfield [30], où quatre points situés sur le contour des lèvres n'avaient pourtant pas permis de procurer un gain d'intelligibilité significatif.

Avec plus de points lumineux (voir la colonne centrale de la figure 5.1), Rosenblum et coll. ont observé un gain d'intelligibilité [31], toutefois inférieur à celui procuré par la vidéo naturelle. Une étude plus récente de Bergeson et coll. confirme que ces stimuli visuels peuvent être intégrés avec le signal audio pour rehausser l'intelligibilité [32].

Dans une autre étude [33], Rosenblum et Saldaña ont montré que des stimuli de parole de ce type permettent en outre de reproduire dans une moindre mesure l'effet McGurk [34].

De plus, dans une étude récente, Santi et coll. ont étudié en imagerie fonctionnelle la catégorisation de stimuli de parole sous forme de points lumineux pour une tâche de lecture labiale. Ces stimuli « points lumineux » semblent mettre en jeu les mêmes aires cérébrales que des signaux de parole naturels [35].

---

30. Summerfield (A. Q.). Use of visual information in phonetic perception. *Phonetica*, 36:314–331, 1979.

31. Rosenblum (L. D.), J. A. Johnson et H. M. Saldaña. Visual kinematic information for embellishing speech in noise. *J. of Speech and Hearing Research*, 39(6):1159–1170, 1996.

32. Bergeson (T. R.), D. B. Pisoni et J. T. Reynolds. Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 55–60, St Jorioz, France, septembre 2003.

33. Rosenblum (L. D.) et H. M. Saldaña. An audiovisual test of kinematic primitives for visual speech perception. *J. of Experimental Psychology : Human Perception and Performance*, 22(2):318–331, 1996.

34. L'effet McGurk est une illusion robuste qui illustre l'intégration, la fusion, perceptive des deux modalités audio et vidéo du signal de parole : dans la version canonique de l'effet, suite à des présentations audio-visuelles incongruentes [ba–ga], le percept généralement perçu est [da] ; McGurk (H.) et J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, décembre 1976. Dans l'expérience de Rosenblum et Saldaña, il s'agit dans cette étude d'une présentation audio-visuelle [ba–va] ; si le percept entendu n'est pas [ba], on considère que la modalité visuelle a un effet.

35. Santi (A.), P. Servos, E. Vatikiotis-Bateson, T. Kuratate et K. G. Munhall. Perceiving biological motion : Dissociating talking from walking. *J. of Cognitive Neuroscience*, 15(6):800–809, 2003.

### 5.2.2. Les points lumineux : une technique de rendu visuel

La plupart des expériences d'évaluation que nous avons présentées au début de ce chapitre mettent en jeu l'ensemble des modules composant le moteur de synthèse : le module responsable de la génération des paramètres articulatoires (fournis soit par l'analyse de séquences naturelles, soit par la synthèse à partir du texte ou du son), le modèle de forme qui est chargé de synthétiser la géométrie du visage sur l'écran, et le modèle d'apparence qui a la charge du rendu final de chaque pixel du visage.

Les résultats en partie controversés obtenus dans (Pandzic et coll., 1999) et (Geiger et coll., 2003) [36], où les faibles scores d'intelligibilité semblent en contradiction avec l'excellente acceptabilité de l'animation, pourraient s'expliquer par le fait que les scores d'évaluation sont le produit complexe des comportements potentiellement déficients de ces trois composants essentiels. Ainsi, dans une tâche de jugement global de la qualité portant sur le réalisme ou l'adéquation avec l'acoustique, des mouvements corrects peuvent être jugés inacceptables s'ils sont rendus par un modèle d'apparence inadéquat (p. ex. une texture unique, trop grossière) ; de même un modèle d'apparence très précis peut compenser des paramètres de contrôle du mouvement inappropriés.

Pour se focaliser sur l'évaluation des mouvements (c.-à-d. les paramètres de contrôle et leur dynamique), on peut substituer un rendu en points lumineux au modèle d'apparence ; cela ne permet pas de faire abstraction du modèle de forme puisqu'il doit être possible pour le rendu de déterminer quels points sont cachés par le visage en mouvement.

Cohen et coll. [37] ont utilisé ce principe pour Baldi afin d'étudier la perception visuelle de la parole dans une tâche de lecture labiale de mots monosyllabiques. Pour tous les visèmes, consonantiques et vocaliques, l'identification était inférieure pour la condition points lumineux comparée à la condition visage « à la

36. Pandzic (I. S.), J. Ostermann et D. Millen. User evaluation : synthetic talking faces for interactive services. *The Visual Computer*, 15:330–340, 1999 ; Geiger (G.), T. Ezzat et T. Poggio. Perceptual evaluation of video-realistic speech. AI Memo 2003-003, Massachusetts Institute of Technology, Cambridge, USA, février 2003.

37. Cohen (M. M.), R. L. Walker et D. W. Massaro. Perception of synthetic visual speech. *In* Stork (D. G.) et M. E. Hennecke, éditeurs. *Speechreading by Humans and Machines*, volume 150 de *Computer and Systems Sciences*, pages 153–168. Springer, 1996.

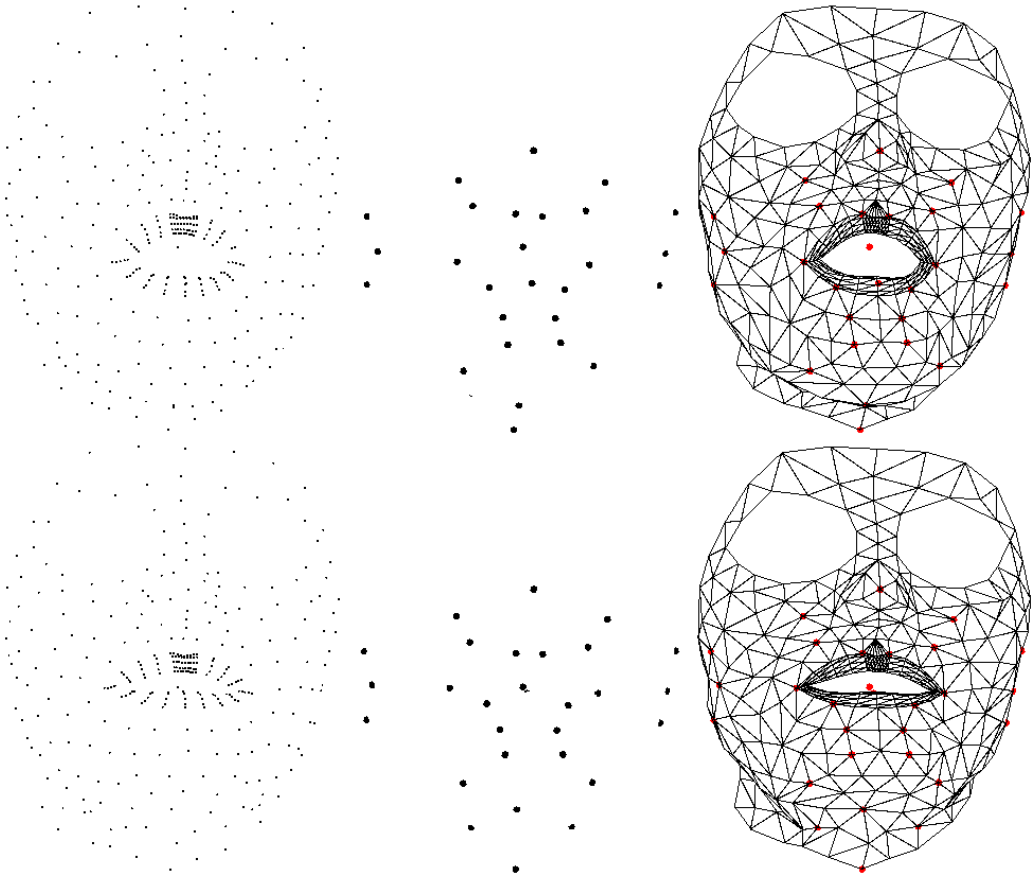


FIGURE 5.1 – Rendus du visage pour deux articulations sous forme de points lumineux, suivant deux techniques : à gauche le dispositif retenu avec une forte densité de points « pixelliques », au milieu le dispositif classique avec peu de pastilles, pastilles identifiables en rouge sur le réseau fil de fer de droite. Dans les deux cas, le visage en points lumineux comprend deux points pour les incisives supérieures et inférieures. Le visage est rendu en deux passes : une première passe initialise le Z-buffer et la seconde fait effectivement le rendu des points qui peuvent éventuellement être masqués suite à la première passe. C'est le cas pour les deux points situés de part et d'autre du larynx sur l'articulation du haut, et pour le point des incisives inférieures qui est partiellement visible sur l'articulation du bas.

Gouraud » [38]. Mais, dans une première expérience, Baldi « à la Gouraud » produisant les mêmes mouvements était significativement moins intelligible qu'une vidéo de parole naturelle. Pour nous, cela remet en cause le caractère biologique des mouvements de Baldi ; sous cette hypothèse d'*artificialité*, Baldi en points lumineux est *naturellement* moins intelligible, jetant ainsi un doute sur la portée des résultats quant au but initial (l'étude de la perception visuelle de parole). Nous rejoignons donc Cohen et coll. quand ils concèdent (page 167) : « *Some « might argue that this performance difference* [Baldi « à la Gouraud » *vs.* Baldi « en points lumineux] *ocurred because of the overall inferiority of the synthetic « face relative to a natural face.* »

### 5.2.3. Notre instanciation du rendu sous forme de points lumineux

Dans les études de perception de parole sous forme de points lumineux que nous avons mentionnées, le locuteur ou la locutrice est typiquement enregistré sous un éclairage faible, le visage noirci avec du maquillage, et avec de grosses pastilles réfléchissantes [39] collées sur la partie inférieure de son visage.

Certains de ces dispositifs induisent vraisemblablement des informations cinématiques 3D supplémentaires car :

- un chromakey imparfait des vidéos d'origine peut laisser des traces des mouvements de la tête et de la peau [40] ;
- la surface du visage et les normales à cette surface changent, et donc la géométrie apparente des pastilles changent également (p. ex., une pastille située à la proximité d'une commissure des lèvres apparaîtra comme un cercle pendant les articulations étirées et comme une ellipse pendant les articulations arrondies).

Nous avons fait le choix d'une technique où chaque point lumineux est un carré de 2×2 pixels, équivalent à 1 mm de diamètre dans le monde 3D. Une pre-

38. voir *supra* pour notre définition de « à la Gouraud »

39. 3 mm de diamètre dans Rosenblum (L. D.) et H. M. Saldaña. An audiovisual test of kinematic primitives for visual speech perception. *J. of Experimental Psychology : Human Perception and Performance*, 22(2):318–331, 1996.

40. Bergeson (T. R.), D. B. Pisoni et J. T. Reynolds. Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 55–60, St Jorioz, France, septembre 2003.

mière passe fait le rendu de tout le maillage polygonal du visage et calcule le Z-buffer ; les points lumineux effectivement affichés lors de la seconde passe correspondent à l'illumination de points de chair qui font face à la caméra et qui ne sont pas masqués par la tête ; deux points lumineux sont considérés pour les incisives inférieures et supérieures.

Nos points lumineux sont donc des *vrais* points 2D en mouvement sur l'écran. Bien que cela ne constitue pas non plus un affichage canonique de points lumineux [41], nous avons choisi un affichage à haute densité de points qui, même statiquement, est identifié aisément comme un visage (donc il « contient » aussi certaines informations structurelles). Cela permet de saisir immédiatement la finesse des déplacements sur l'ensemble du visage ; toutefois, la sur-représentation au niveau des lèvres tend à « noyer » les points pour les dents.

La figure 5.1 illustre ces deux possibilités de rendu en points lumineux : au milieu le rendu d'un petit nombre de pastilles bien choisies, à gauche le rendu de points 2D, dense et sans ambiguïté, que nous avons retenu.

#### 5.2.3.1. *Évaluation qualitative de notre tête parlante et validation du rendu sous forme de points lumineux*

Nous avons déjà employé cette technique de rendu lors d'une expérience de perception de systèmes de synthèse de parole visuelle à partir du texte [42].

Les 23 participantes et participants devaient juger sur une échelle MOS [43] à cinq valeurs de l'adéquation des mouvements joués avec le signal audio. Le corpus de l'étude comportait dix phrases de tests. En plus de cinq systèmes de synthèse de mouvements à partir de la chaîne phonétique, les mouvements correspondant à la vérité terrain (voir section 4.1.2.1.) et les mouvements « opposés [44] » étaient évalués.

Comme on le voit sur le tableau 5.1, ces deux dernières familles de mouvements ont obtenues des notes respectivement très bonnes et très mauvaises ;

41. On l'a vu, dans un paradigme *structure-from-motion*, un affichage en points lumineux ne devrait procurer aucune clef sur la structure 3D sous-jacente en l'absence de mouvement.

42. Nous ne présentons ici qu'un résumé court de cette expérience dont la thématique est en marge de nos travaux de thèse ; Bailly (G.), G. Gibert et M. Odisio. Evaluation of movement generation systems using the point-light technique. *In Proc. of the IEEE Workshop on Speech Synthesis*, pages 27–30, Santa Monica, USA, septembre 2002.

43. « *Mean Opinion Score* »

44. selon l'expression « *on the other side of mean* »

même si ce n'est qu'une confirmation attendue dans cette tâche comparative, cette évaluation qualitative de la tête parlante de synthèse constitue une validation de notre approche. Aussi, les participants et les participantes ont tous indiqué avoir vu une tête parlante, notamment grâce aux lèvres.

TABLEAU 5.1 – 23 personnes ont noté l'adéquation avec le son de dix phrases de tests pour plusieurs systèmes de synthèse de mouvements (Bailly et coll., 2002, *op. cit.*). L'échelle MOS comporte cinq niveaux (de 5 pour « très bien » à 1 pour « très insuffisant »). *org* correspond à la vérité terrain, *orginv* aux mouvements « opposés » ; les autres systèmes représentent cinq synthétiseurs de mouvements. *org* et *orginv* sont respectivement le mieux et le plus mal notés.

| *Systèmes* | *org* | *synl* | *reg* | *syn* | *mlphrtst* | *mlapp* | *orginv* |
|---|---|---|---|---|---|---|---|
| Moyenne des notes | 4,3 | 4,2 | 3,7 | 3,7 | 2,5 | 2,1 | 1,9 |
| Écart-type des notes | 0,8 | 0,8 | 1,0 | 1,0 | 1,1 | 0,8 | 1,0 |

## 5.3. CADRE DES TESTS RETENUS POUR ÉVALUER LE SUIVI

Nous présentons maintenant un protocole expérimental constitué de deux tests d'intelligibilité qui permet de caractériser des systèmes de capture de mouvements par rapport à la vérité terrain. La première expérience permet d'étalonner la vérité terrain en points lumineux en fournissant les scores d'identification audiovisuelle des séquences d'origine pour trois types de condition visuelle : audio seul, enregistrement vidéo original et points lumineux représentant le visage, animés d'après la vérité terrain. La seconde expérience fournit les scores d'identification obtenus par le système de suivi utilisant plusieurs modèles d'apparence pour l'analyse de ces vidéos et rendu sous forme de points lumineux.

L'ensemble des matrices de confusion est obtenu à l'issue d'un premier traitement des résultats de ces expériences. Une partie de ces matrices sont reprises dans le texte pour éclairer les analyses. Les principaux résultats présentés ci-dessous sont issus d'analyses basées sur des indices calculés à partir des matrices de confusion.

# 5.4. EXPÉRIENCE I — INTELLIGIBILITÉ D'UN VISAGE RENDU EN POINTS LUMINEUX

## 5.4.1. Méthode

### *5.4.1.1. Les stimuli* VCV

Le jeu de stimuli se compose de logatomes *VCV* clairement articulés, sans phrase porteuse : *V* est l'une des voyelles { a, i, u } et *C* est l'une des dix consonnes voisées { b, d, g, v, z, ʒ, ʀ, l, m, n }. La locutrice est enregistrée dans les mêmes conditions que lors de la construction du modèle géométrique du visage (comme cela est illustré sur la figure 1.1). En ce qui concerne l'acoustique, les signaux audio d'origine sont combinés avec un bruit blanc, pour chacun des rapports signal sur bruit (SNR) dans { −24, −18, −12, −6, 0, 6 } dB. Trois conditions visuelles sont testées : audio seul, enregistrement vidéo d'origine et le visage en points lumineux, animé d'après la vérité terrain. Ces trois conditions seront dénommées dans la suite *nil*, *natural* et *plf*.

Par « vérité terrain » nous entendons les paramètres articulatoires estimés à partir de l'inversion du modèle 3D depuis un étiquetage semi-automatique, tel que cela est décrit à la section 4.1.2.1.. De plus, le mouvement rigide de tête variant très peu pendant ces séquences, lors de la re-synthèse le visage en points lumineux est animé à partir des seuls paramètres articulatoires.

De plus, pour une étude préliminaire, le jeu de stimuli comprend également des stimuli de type McGurk [45] : une vidéo [aga] doublée avec des stimuli audio [aba] soigneusement alignés.

Pour chaque condition visuelle, il y a donc 186 stimuli ($3 \times 10 \times 6 + 1 \times 1 \times 6$).

### *5.4.1.2. Procédure*

L'expérience est divisée en trois sessions successives, qui correspondent aux conditions visuelles *nil*, *natural* et enfin *plf*. Dans chaque session, tous les 186 stimuli sont joués dans un ordre aléatoire. En guise de familiarisation et d'entraînement, quelques stimuli supplémentaires sont joués au début de chaque ses-

45. McGurk (H.) et J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, décembre 1976.

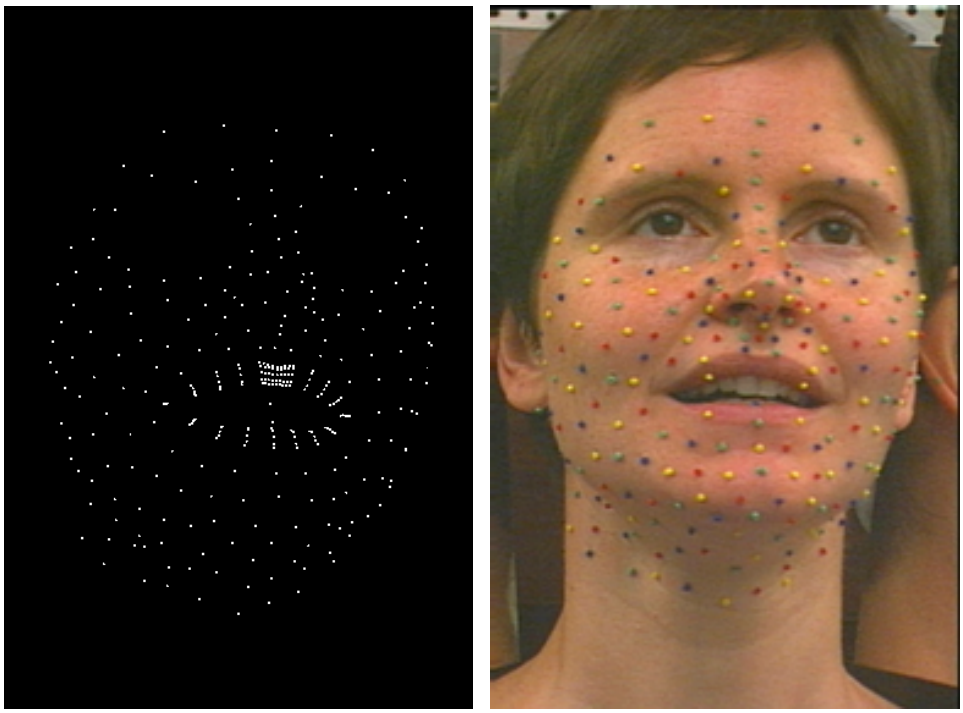


FIGURE 5.2 – Les conditions visuelles *plf* (à gauche) et *natural* (à droite).

sion. En outre, avant la session *plf*, les participants et les participantes sont prévenus qu'ils vont voir une tête parlante représentée sous forme d'un ensemble de points lumineux en mouvement ; une vidéo illustratrice qui montre un fondu enchaîné entre les conditions visuelles *natural* et *plf* au cours d'une même séquence est alors jouée. Les deux conditions visuelles sont représentées sur la figure 5.2.

Les stimuli sont joués sur l'écran 15 pouces d'un ordinateur portable équipé d'écouteurs ; la taille de la fenêtre montrant les animations est de 384×576 pixels. Le visage affiché est légèrement orienté (rotation de $-10^{\circ}$ autour de l'axe des $y$) et correspond à une taille sur l'écran de 12 cm de haut.
Les vidéos *natural* sont produites avec le logiciel *Adobe Premiere* ; leur fréquence vidéo est 25 Hz. Les animations *plf* sont rendues en blanc sur fond noir. Elles sont générées à la volée et en temps réel grâce aux fonctionnalités graphiques offertes par une carte accélératrice 3D standard.

Les participants et les participantes utilisent une interface graphique. Ils indiquent la consonne qu'ils ont identifiée par un clic sur le bouton qui lui correspond ; parce que le choix est forcé, il s'agit bien d'une tâche de catégorisation et non d'une tâche d'identification [46]. Toutefois, il nous arrivera d'employer par la suite le mot « identification » ou l'un de ses dérivés pour désigner la tâche ou les résultats : hormis les consonnes non-voisées visuellement très semblables (p. ex., /p/ pour /b/, ou /t/ pour /d/) que nous avons choisies délibérément de ne pas considérer, les dix consonnes proposées couvrent bien l'ensemble des réponses possibles. Un indice de progression indique la part de la session déjà effectuée. Juste après qu'un choix est donné, le prochain stimulus est joué.

Les participantes et les participants ont reçu l'instruction de faire leur choix relativement rapidement. Cela permet d'une part de limiter la durée totale de l'expérience aux alentours de 35 minutes ; d'autre part, cela augmente la possibilité d'apparition d'un *effet d'ordre* [47], notamment quand un stimulus très bruité suit un stimulus non bruité ; toutefois, nous espérons qu'un tel biais est compensé en moyenne par les ordres aléatoires qui sont différents pour chaque participant et chaque participante.

### *5.4.1.3. Participantes et participants*

Dix-sept personnes ont pris part à l'expérience (treize participants et quatre participantes ; âges de 20 à 26 ans, âge moyen 22,8 ans ± 1,7 ans). Leur langue maternelle est le français ; elles sont naïves au regard de l'expérience ; elles ne connaissent pas la locutrice. Leur acuité visuelle (en tenant compte de leur correction le cas échéant) est estimée par le test optométrique de Parinaud [48]. Tous les participants et toutes les participantes lisent de manière fluide au niveau 3, ce qui indique une bonne acuité visuelle pour la vision de près. Leur acuité auditive n'est pas testée mais ils ne se connaissent pas de problème auditif ; leur capacité auditive pour la tâche d'identification est évaluée avec la condition visuelle *nil*.

46. dite aussi « à choix ouvert »

47. Le problème de l'*effet d'ordre* est classique en psychologie expérimentale ; la catégorisation perceptive d'un stimulus peut dépendre des stimuli précédents et de leurs catégorisations. Ici, la fréquence importante à laquelle les stimuli sont enchaînés constitue peut-être un facteur supplémentaire d'interférence.

48. Pour ce test de la vision de près, on demande à la personne de lire sur une plaquette située à 33 cm d'elle. Cette plaquette contient des textes de tailles de caractères décroissantes. Le résultat du test correspond à la limite de lecture fluide. Nous nous sommes procuré pour quelques jours une telle plaquette chez une orthoptiste.

## 5.4.2. Résultats

### *5.4.2.1. Identification et confusions*

Les pourcentages d'identification consonantique correcte sont représentés sur la figure 5.3 et dans le tableau 5.2. Globalement, les scores avec *plf* sont inférieurs à ceux avec *natural*. Une analyse de la variance (ANOVA) à deux facteurs avec mesures répétées met en évidence des effets principaux pour les conditions visuelles ($F(2,32) = 289{,}72$, $p < 0{,}001$) et pour le niveau de bruit audio ($F(5,80) = 395{,}97$, $p < 0{,}001$), ainsi qu'une interaction significative ($F(10,160) = 13{,}224$, $p < 0{,}001$). Des comparaisons multiples en ajustant les valeurs de $p$ avec la procédure de Holm [49] montre à chaque niveau de bruit des différences significatives entre toutes les conditions visuelles.

Les propriétés remarquables des résultats comparés de *plf* et *natural* sont :

- l'identification encore plus faible de [z] ;
- l'identification incorrecte de [ʒ] en contexte arrondi ;
- l'identification correcte à tous les niveaux de bruit pour *natural* et *plf* des consonnes nasales ([m] et [n] ne sont pas confondues avec [b] et [d] — en accord avec les résultats audio alors que c'est le cas pour les arbres construits d'après la géométrie, page 108) ;
- l'asymétrie dans les confusions [l]–[ʀ] est certainement due au fait que la locutrice française a produit la trille [ʀ], alors qu'en français plus souvent c'est la fricative [ʁ] qui est utilisée ;
- les identifications correctes sont plus élevées en contexte vocalique [a] qu'en contexte vocalique [i] ou [u] — la vision de l'intérieur des lèvres explique en partie ce gain pour *natural* ; nous discuterons pour *plf* par la suite.

### *5.4.2.2. Distribution des réponses incorrectes*

L'intelligibilité, définie comme la proportion de réponses correctes, traite du taux d'erreur ; cela ne fournit pas d'information sur la manière dont les réponses

49. Holm (S.). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979 ; cette procédure d'ajustement est une variante en plusieurs étapes de la méthode de Bonferroni. Comme cette dernière, elle ne repose pas sur des hypothèses spécifiques et elle conserve la probabilité que l'ensemble des conclusions contienne au moins une erreur de première espèce sous le seuil de $\alpha$ ; en revanche, sa puissance est supérieure : Howell (D. C.). *Méthodes statistiques en sciences humaines.* De Boeck Université, 1998.

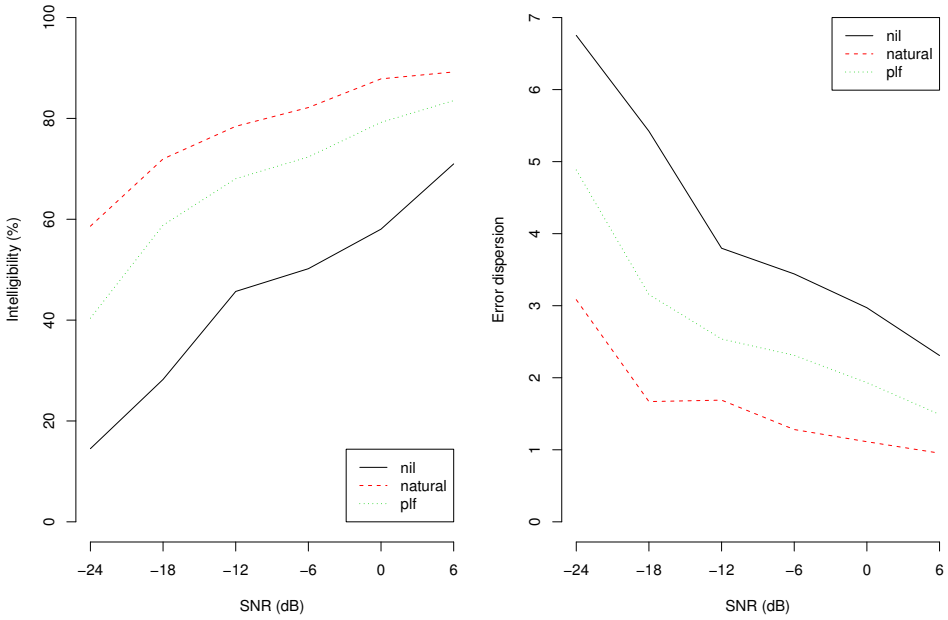



FIGURE 5.3 – Taux d'identification (à gauche) et dispersion de l'erreur, exprimée comme le nombre de catégories d'erreur par consonne (à droite), en fonction du niveau de bruit audio, pour les conditions visuelles de l'expérience 1 (*nil*, *natural* et *plf*).

TABLEAU 5.2 – Taux d'identification et dispersion de l'erreur, exprimée comme le nombre de catégories d'erreur par consonne, en fonction du niveau de bruit audio, pour les conditions visuelles de l'expérience 1 (*nil*, *natural* et *plf*).

| | *Condition visuelle* | SNR (dB) | | | | | |
|---|---|---|---|---|---|---|---|
| | | −24 | −18 | −12 | −6 | 0 | 6 |
| Taux id. (%) | *nil* | 15 | 28 | 46 | 50 | 58 | 71 |
| | *natural* | 59 | 72 | 78 | 82 | 88 | 89 |
| | *plf* | 40 | 59 | 68 | 72 | 79 | 84 |
| Disp. err. | *nil* | 6,75 | 5,42 | 3,80 | 3,44 | 2,97 | 2,31 |
| | *natural* | 3,08 | 1,67 | 1,69 | 1,28 | 1,11 | 0,95 |
| | *plf* | 4,88 | 3,16 | 2,54 | 2,31 | 1,93 | 1,49 |

incorrectes sont distribuées. La figure 5.3 montre également une mesure de la dispersion de l'erreur, telle que la définit van Son [50]. La dispersion de l'erreur pour les stimuli représente « *the effective number of error categories per stimulus* » [51] ; une de ses propriétés intéressantes est que la dispersion de l'erreur est relativement peu sensible au taux de l'erreur.

Pour illustrer et faciliter une reproduction éventuelle de nos calculs, nous expliciterons les calculs des indices — sur lesquels reposent nos analyses —, ainsi qu'une application numérique pour la matrice *plf* du tableau 5.3.

Soit **M** une matrice de confusion et N le nombre total de stimuli :

$$\mathrm{N} = \sum_{i,j} \mathrm{M}_{i,j} \qquad \text{a.n. :} \quad \mathrm{N} = 3\ 060 \tag{5.1}$$

La probabilité d'occurrence d'un stimulus $s_i$ est connue et égale à : $p(s_i) = \frac{1}{\mathrm{N}} \sum_j \mathrm{M}_{j,i}$. La probabilité d'occurrence d'une réponse $r_i$ est estimée par sa fréquence observée : $p(r_i) = \frac{1}{\mathrm{N}} \sum_j \mathrm{M}_{i,j}$. La probabilité d'occurrence conjointe d'un stimulus $s_i$ et d'une réponse $r_j$ est estimée par sa fréquence observée : $p(s_i, r_j) = \frac{1}{\mathrm{N}} \mathrm{M}_{i,j}$. L'information contenue dans les stimuli est :

$$\mathrm{H}(s) = -\sum_i p(s_i) \log_2 p(s_i) \qquad \text{a.n. :} \quad \mathrm{H}(s) = 3{,}322 \text{ bit} \tag{5.2}$$

L'information contenue dans la matrice de confusion **M** est :

$$\mathrm{H}(\mathbf{M}) = -\sum_{i,j} p(s_i, r_j) \log_2 p(s_i, r_j) \qquad \text{a.n. :} \quad \mathrm{H}(\mathbf{M}) = 4{,}920 \text{ bit} \tag{5.3}$$

Le taux d'erreur ε est :

$$\varepsilon = \sum_{s_i \neq r_j} p(s_i, r_j) \qquad \text{a.n. :} \quad \varepsilon = 0{,}329 \tag{5.4}$$

L'entropie de l'erreur ε est :

$$\mathrm{H}_\varepsilon = -\varepsilon \log_2 \varepsilon - (1-\varepsilon) \log_2 (1-\varepsilon) \qquad \text{a.n. :} \quad \mathrm{H}_\varepsilon = 0{,}914 \text{ bit} \tag{5.5}$$

50. van Son (R. J. J. H.). A method to quantify the error distribution in confusion matrices. *In Proc. of the European Conf. on Speech Communication and Technology*, pages 2277–2280, Madrid, Spain, 1995.

51. le nombre effectif de catégories d'erreur par stimulus ; van Son (1995), *op. cit.*.

TABLEAU 5.3 – Matrices de confusion pour l'expérience 1 pour les trois conditions visuelles (de haut en bas) *nil*, *natural* et *plf*. Les réponses sont agrégées pour tous les participants et toutes les participantes, toutes les voyelles, et pour tous les niveaux de bruit. Les stimuli sont donnés en colonne. Les scores supérieurs à 10 % des choix sont mis en valeur.

| *nil* | b | m | v | d | n | l | z | ʒ | g | ʀ |
|---|---|---|---|---|---|---|---|---|---|---|
| b | **166** | 3 | **75** | **36** | 1 | 7 | 8 | 3 | **32** | 4 |
| m | | **87** | 3 | 2 | **39** | 11 | 4 | 5 | 1 | 4 |
| v | 18 | 10 | **81** | 22 | 18 | 16 | **46** | 29 | 17 | 12 |
| d | **50** | 5 | 24 | **135** | 5 | 11 | **39** | 6 | **38** | 7 |
| n | | **47** | 3 | 1 | **101** | 16 | 6 | 3 | 2 | 2 |
| l | | **97** | 2 | 2 | **82** | **160** | 5 | 11 | 5 | 12 |
| z | 11 | 9 | **33** | 18 | 10 | 13 | **83** | **45** | 14 | 15 |
| ʒ | 20 | 19 | **40** | **33** | 20 | 16 | **64** | **157** | 24 | 7 |
| g | **34** | 9 | 18 | **46** | 5 | 5 | **34** | 20 | **161** | 9 |
| ʀ | 7 | 20 | 27 | 11 | 25 | **51** | 17 | 27 | 12 | **234** |

| *nat* | b | m | v | d | n | l | z | ʒ | g | ʀ |
|---|---|---|---|---|---|---|---|---|---|---|
| b | **297** | 9 | 2 | 2 | | | | | | |
| m | 3 | **288** | | | 5 | 3 | | | 1 | 1 |
| v | 4 | | **293** | 1 | 2 | 1 | | 1 | 7 | 3 |
| d | 1 | 1 | 1 | **217** | 3 | | **38** | 3 | **32** | 1 |
| n | | 2 | | | **190** | 11 | 2 | | 3 | 1 |
| l | | 5 | | 1 | **83** | **220** | 1 | 4 | 6 | |
| z | | | 7 | 26 | 5 | 3 | **163** | **53** | 25 | 2 |
| ʒ | | | 2 | 11 | 4 | 5 | **70** | **227** | 23 | 5 |
| g | 1 | | | **41** | 1 | 2 | 22 | 5 | **202** | 2 |
| ʀ | | 1 | 1 | 7 | 13 | **61** | 10 | 13 | 7 | **291** |

| *plf* | b | m | v | d | n | l | z | ʒ | g | ʀ |
|---|---|---|---|---|---|---|---|---|---|---|
| b | **280** | 11 | **41** | 8 | | | 2 | 1 | 8 | |
| m | 3 | **252** | 2 | | 18 | 6 | 1 | | 2 | 2 |
| v | 14 | 5 | **221** | 14 | 10 | 4 | 16 | 26 | 7 | 6 |
| d | | | 7 | **185** | 1 | 4 | **55** | 6 | **32** | 3 |
| n | 2 | 20 | | | **152** | 16 | 2 | 1 | 2 | 4 |
| l | | 14 | 2 | 1 | **82** | **208** | 2 | 10 | 2 | 17 |
| z | 2 | 1 | 12 | 27 | 4 | 6 | **116** | **51** | 10 | 5 |
| ʒ | | | 11 | 16 | 16 | 8 | **46** | **172** | 20 | 7 |
| g | 4 | 2 | 7 | **48** | 4 | 5 | **57** | 15 | **213** | 9 |
| ʀ | 1 | 1 | 3 | 7 | 19 | **49** | 9 | 24 | 10 | **253** |

La dispersion de l'erreur pour les stimuli est alors :

$$d_s = 2^{\frac{\mathrm{H}(\mathbf{M})-\mathrm{H}(s)-\mathrm{H}_\varepsilon}{\varepsilon}} \qquad \text{a.n. :} \quad d_s = 4{,}2 \text{ catégories/stimulus} \tag{5.6}$$

Ces calculs faits et reportés dans le tableau 5.2, la dispersion des réponses incorrectes est plus importante avec *plf* qu'avec *natural*. Cependant, la structure reste la même ; le tableau 5.3 montre un aperçu de cette dégradation.

### 5.4.2.3. *Transmission de l'information phonétique*

Dans leur étude des confusions auditives des consonnes de l'anglais, Miller et Nicely [52] ont cherché à savoir dans quelle mesure étaient transmis plusieurs traits phonétiques : voisement, nasalité, friction, durée et lieux d'articulation. Dans ce but, ils ont défini l'information transmise relative, qui est calculée de la manière suivante.
L'information transmise des stimuli aux réponses est :

$$\mathrm{H}(s,r) = -\sum_{i,j} p(s_i,r_j)\log_2 \frac{p(s_i)\,p(r_j)}{p(s_i,r_j)} \qquad \text{a.n. :} \quad \mathrm{H}(s,r) = 1{,}701 \text{ bit} \tag{5.7}$$

L'information transmise relative est alors :

$$\mathrm{T} = \frac{\mathrm{H}(s,r)}{\mathrm{H}(s)} \qquad \text{a.n. :} \quad \mathrm{T} = 0{,}51 \tag{5.8}$$

À cause du faible nombre de réponses pour un triplet (condition visuelle ; SNR ; consonne) donné, nous avons dû regrouper pour les calculs les réponses de tous les participants et toutes les participantes. Notons que cette représentation de la performance moyenne du groupe dans son ensemble n'est vraisemblablement pas la plus précise [53].

L'information transmise relative reflète les relations entre les stimuli et les réponses dans leur ensemble. Cependant, dans cette expérience les caractéristiques du taux d'identification et de l'information transmise relative sont similaires.

52. Miller (G. A.) et P. E. Nicely. An analysis of perceptual confusions among some English consonants. *J. of the Acoustical Society of America*, 27(2):338–352, mars 1955.

53. J. Castelloe, communication personnelle (John.Castelloe@sas.com).

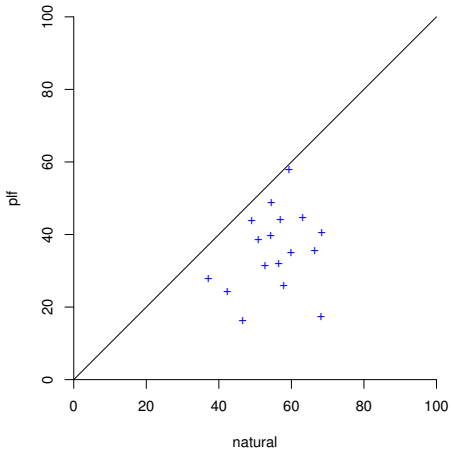



FIGURE 5.4 – Contributions relatives des conditions visuelles *natural* et *plf* des 17 participants et participantes à l'expérience I.

#### 5.4.2.4. *Contribution visuelle à la transmission de l'information phonétique*

Dans l'une des premières expériences [54] étudiant l'apport de la modalité visuelle à l'intelligibilité de la parole en présence de bruit, Sumby et Pollack se posent la question suivante : « *What is the visual informational contribution* «relative *to the* possible available *contribution in the absence of visual cues?* »

Ils répondent à cette question en définissant la contribution relative de l'information visuelle de la manière suivante [55] :

$$R_{vc} = \frac{T_{AV} - T_A}{1 - T_A} \tag{5.9}$$

54. Sumby et Pollack mentionnent notamment la thèse de doctorat présentée trois ans auparavant par J. J. O'Neill à l'Ohio State University : « *if visual factors supplementary to oral speech are utilized, we can tolerate higher noise interference levels than if visual factors are not utilized.* » ; SUMBY (W. H.) et I. POLLACK. Visual contribution to speech intelligibility in noise. *J. of the Acoustical Society of America*, 26(2):212–215, mars 1954

55. Rappelons que dans (Sumby et Pollack, 1954), *op. cit.*, cet indice est calculé à partir de l'information transmise relative, et non à partir du taux d'intelligibilité comme cela est pourtant souvent rapporté.

où $T_{AV}$ et $T_A$ désignent respectivement l'information transmise relative des conditions audiovisuelle et audio.

Les contributions relatives des conditions *natural* et *plf* sont représentées pour chaque participant et chaque participante sur la figure 5.4. En accord avec la littérature [56], on observe des variations interindividuelles des performances d'intégration audiovisuelle. Il est intéressant de noter que certains participants et certaines participantes ont des performances équivalentes avec les deux conditions visuelles.

En première approximation, un calcul global donne pour ces contributions relatives 58 % pour *natural* et 36 % pour *plf*. Elles varient, cependant ; elles sont notamment plus faibles au SNR −24 dB.

Cela induit les deux relations empiriques ci-dessous. Tout d'abord :

$$T_{AV} \approx R_{vc} + (1 - R_{vc})T_A \quad \text{avec } R_{vc} \text{ constant, } R_{vc} \in [0, 1] \tag{5.10}$$

Il est à noter que faute de disposer des réponses aux stimuli visuels seuls, il n'est pas possible de confronter nos données aux modèles d'intégration audiovisuelle de la littérature [57].

D'autre part, il vient :

$$T_{plf} \approx T_{natural} - 0{,}22(1 - T_{nil}) \tag{5.11}$$

En d'autres termes, l'information transmise par le visage en points lumineux est approximativement égale à l'information transmise par la vidéo d'origine diminuée de 22 % de l'information non transmise par le canal audio.

### 5.4.2.5. *Arbres de confusion*

Nous avons déjà construit des dendrogrammes au chapitre précédent pour hiérarchiser les consonnes selon la géométrie. De même, nous avons construit des dendrogrammes qui reflètent les confusions perceptives des consonnes.

56. P. ex. Grant (K. W.) et P. F. Seitz. Measures of auditory-visual integration in nonsense syllables and sentences. *J. of the Acoustical Society of America*, 104(4):2438–2450, octobre 1998 ; Lachs (L.) et D. B. Pisoni. Specification of cross-modal source information in isolated kinematic displays of speech. *J. of the Acoustical Society of America*, 116(1):507–518, 2004.

57. Grant (K. W.), B. E. Walden et P. F. Seitz. Auditory-visual speech recognition by hearing-impaired subjects : Consonant recognition, sentence recognition and auditory-visual integration. *J. of the Acoustical Society of America*, 103(5):2677–2690, mai 1998.

TABLEAU 5.4 – Matrice de similarité obtenue à partir de la matrice de confusion *plf* du tableau 5.3.

| | b | m | v | d | n | l | z | ʒ | g | ʀ |
|---|---|---|---|---|---|---|---|---|---|---|
| b | 306 | 25 | 64 | 29 | 22 | 16 | 26 | 23 | 26 | 17 |
| m | 25 | 306 | 24 | 18 | 61 | 44 | 16 | 21 | 23 | 29 |
| v | 64 | 24 | 306 | 63 | 37 | 34 | 61 | 68 | 57 | 35 |
| d | 29 | 18 | 63 | 306 | 43 | 35 | 170 | 87 | 129 | 38 |
| n | 22 | 61 | 37 | 43 | 306 | 144 | 49 | 65 | 48 | 64 |
| l | 16 | 44 | 34 | 35 | 144 | 306 | 41 | 62 | 43 | 96 |
| z | 26 | 16 | 61 | 170 | 49 | 41 | 306 | 147 | 142 | 44 |
| ʒ | 23 | 21 | 68 | 87 | 65 | 62 | 147 | 306 | 72 | 65 |
| g | 26 | 23 | 57 | 129 | 48 | 43 | 142 | 72 | 306 | 46 |
| ʀ | 17 | 29 | 35 | 38 | 64 | 96 | 44 | 65 | 46 | 306 |

Nous avons repris la méthode employée dans sa thèse par Robert-Ribes [58]. Soit **M** une matrice de confusion. La première étape consiste à déterminer une matrice symétrique de similarité **S** à partir de **M**. La similarité entre deux stimuli correspond au nombre d'occurrences où ces deux stimuli ont été catégorisés de manière identique :

$$S_{ij} = S_{ji} = \frac{1}{2}\sum_{k}(M_{ki} + M_{kj} - |M_{ki} - M_{kj}|) \tag{5.12}$$

Le tableau 5.4 donne la matrice de similarité de la matrice *plf* du tableau 5.3. Une matrice de dissimilarité **D** est ensuite calculée ainsi :

$$D_{ij} = 1 - \frac{S_{ij}}{\max \mathbf{S}} \tag{5.13}$$

Enfin, la classification hiérarchique est effectuée à partir de cette matrice **D** en utilisant le critère d'agrégation du saut minimal [59].

58. ROBERT-RIBES (J.). *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, février 1995, (page 127).

59. ou *single linkage* ; LEBART (L.), A. MORINEAU et M. PIRON. *Statistique exploratoire multidimensionnelle.* Dunod, 1995, (page 156). Nous avons effectué tous les calculs statistiques avec le logiciel libre *R* ; dans *R*, cela correspond à la méthode *single* de la fonction *hclust* ; R DEVELOPMENT CORE TEAM. *R : A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. URL http://www.R-project.org. ISBN 3-900051-07-0.

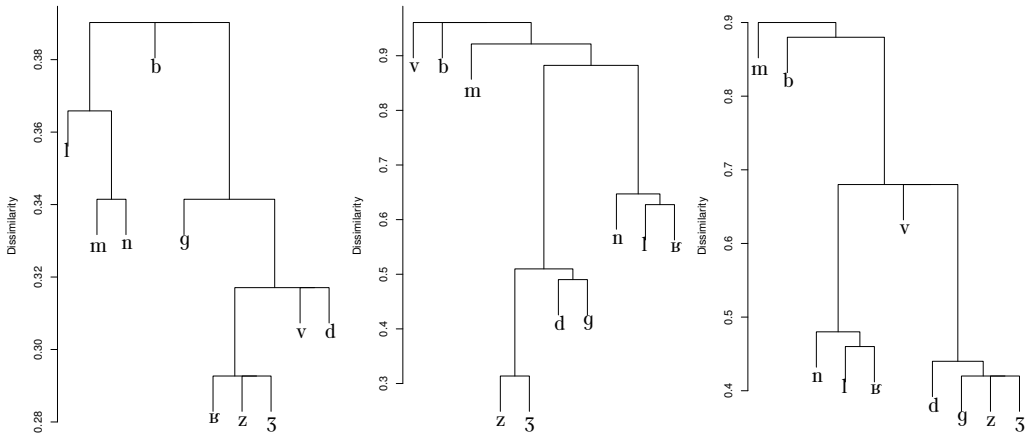



FIGURE 5.5 – Arbres de confusion perceptive au SNR −18 dB. De gauche à droite, les conditions *nil*, *natural* et *plf*.

La figure 5.5 montre les arbres de confusion pour les conditions *nil*, *natural* et *plf* au SNR −18 dB. À titre informatif, nous avons repris un arbre de confusion perceptive *visuelle* pour les consonnes de l'anglais sur la figure 5.6.

En choisissant *a posteriori* une valeur de dissimilarité « seuil », on peut définir des groupes au sein desquels les consonnes ne sont pas distinguées. Comme on le voit, un réglage astucieux de ce seuil permet facilement d'arriver à des interprétations « choisies » : nous prenons le parti de regarder ces arbres d'un peu plus loin.

Au SNR −18 dB les structures des arbres de confusion pour *nil*, *natural* et *plf* sont pertinentes. Pour *nil*, les dissimilarités sont faibles et resserrées, ce qui traduit une grande similarité des consonnes entre elles. Les dissimilarités des arbres pour *natural* et *plf* sont plus importantes et étalées ; ces arbres ont des structures équivalentes, mais les dissimilarités sont plus fortes pour *natural* : p. ex. de « l'altitude [60] » 0,5, on distingue huit classes pour *natural*, cinq pour *plf*, et une seule pour *nil*.

60. ou « à la profondeur », les arbres étant renversés…

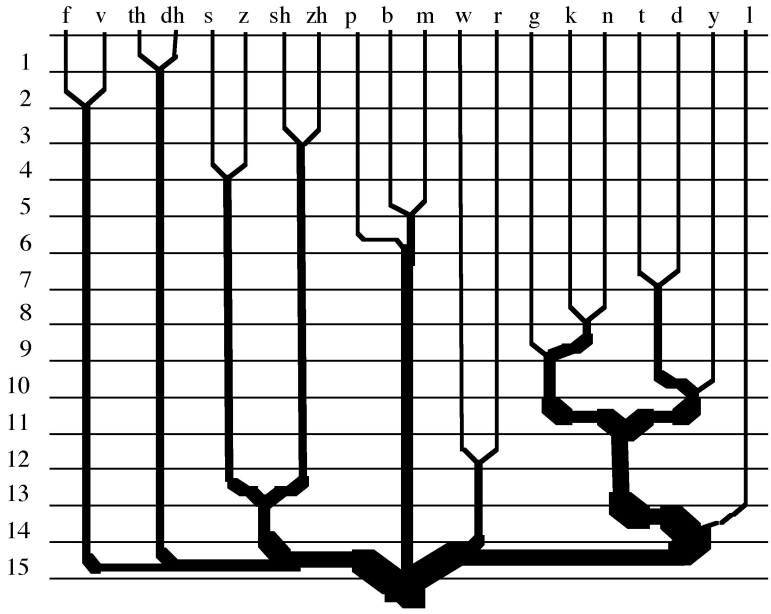



FIGURE 5.6 – Arbres de confusion perceptive visuelle des consonnes de l'anglais ; les participants et participantes présentent des déficiences auditives et sont entraînés pour la tâche d'identification de logatomes [Ca] (repris de WALDEN (B. E.), R. A. PROSEK, A. A. MONGOMERY, C. K. SCHERR et C. J. JONES. Effects of training on the visual recognition of consonants. *J. of Speech and Hearing Research*, 20:130–145, 1977).

### *5.4.2.6. Effet McGurk*

Les tableaux 5.5 et 5.6 présentent les réponses pour les stimuli McGurk. Le tableau 5.5 montre le pourcentage de réponses identiques à la consonne audio pour les stimuli McGurk. Ce tableau contient également les réponses pour [ava] : de manière imprévue, au fur et à mesure que le SNR augmente, la labiodentale [v] en contexte vocalique [a] est de plus en plus entendue comme la bilabiale [b], alors que par ailleurs la vidéo et les mouvements en points lumineux nous semblent parfaitement « canoniques ». Nous croyons que ces stimuli particuliers peuvent être considérés comme une combinaison McGurk (qui sera notée par la suite audio [ab$_v$a] – vidéo [ava]).

On observe un nombre plus faible de réponses de type « fusion » (sous l'hypothèse qu'une réponse différente de la modalité audio implique une intégration

TABLEAU 5.5 – Pourcentage de réponses « correctes » (basé sur l'audio) pour les stimuli *McGurk* de l'expérience 1. Pour la signification du stimulus [ab$_v$a], voir la section 5.4.2.6., page 144.

| *Condition visuelle* | Stimuli | SNR (dB) | | | | | |
|---|---|---|---|---|---|---|---|
| | | −24 | −18 | −12 | −6 | 0 | 6 |
| A–V incongruents | | | | | | | |
| *natural* | aba–aga | 6 | 6 | 12 | 24 | 29 | 41 |
| *plf* | aba–aga | 12 | 47 | 82 | 88 | 94 | 94 |
| *natural* | ab$_v$a–ava | 0 | 0 | 0 | 6 | 0 | 0 |
| *plf* | ab$_v$a–ava | 12 | 18 | 29 | 47 | 47 | 47 |
| Audio seul | | | | | | | |
| *nil* | aba | 29 | 94 | 100 | 100 | 100 | 100 |
| *nil* | aga | 71 | 71 | 88 | 100 | 100 | 100 |
| *nil* | ab$_v$a | 6 | 35 | 59 | 88 | 88 | 82 |
| A–V congruents | | | | | | | |
| *natural* | aba–aba | 100 | 100 | 100 | 100 | 100 | 100 |
| *plf* | aba–aba | 94 | 100 | 100 | 100 | 100 | 100 |
| *natural* | aga–aga | 47 | 82 | 94 | 100 | 100 | 100 |
| *plf* | aga–aga | 71 | 82 | 94 | 100 | 100 | 100 |

de la modalité visuelle, ici discordante) avec *plf* qu'avec *natural*. L'effet visuel dépend de l'intelligibilité audio : il est moins marqué quand le signal audio est plus intelligible, comme dans l'expérience — pour le japonais — de Sekiyama et Tohkura [61]. Comme on le voit sur le tableau 5.6, l'éventail des autres réponses est également conforme à la littérature.

Pour la combinaison [aba–aga], des tests du $\chi^2$ de McNemar [62] montrent des

61. SEKIYAMA (K.) et Y. TOHKURA. McGurk effect in non-English listeners : Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. of the Acoustical Society of America*, 90(4):1797–1805, octobre 1991.

62. Le test de McNemar mesure la concordance dans la catégorisation binaire d'un même échantillon de données entre deux « instruments ». Il est basé sur l'observation que si les deux « instruments » sont en accord dans leurs catégorisations, alors les données discordantes (c.-à-d. les données pour lesquelles les catégories diffèrent pour les deux « instruments ») doivent être réparties de manière égale entre les deux possibilités de discordance. Le calcul de la statistique est : $\chi_1 = \frac{(a-b)^2}{a+b}$, où $a$ et $b$ désignent les nombres de données dans les cases de discordance de la table de contingence — la table 2×2 est construite de telle sorte que chaque donnée n'y apparaît

TABLEAU 5.6 – Matrices de confusion des conditions visuelles *natural* et *plf* pour les stimuli *McGurk* de l'expérience 1. Voir également le tableau 5.5.

| | | *natural* | | | | | | | *plf* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SNR (dB) | b | v | d | l | z | g | ʀ | b | v | d | l | z | g | ʀ |
| aba–aga | −24 | 1 | 1 | 8 | | 6 | 1 | | 2 | 6 | 2 | 1 | 1 | 3 | 2 |
| | −18 | 1 | | 12 | | 4 | | | 8 | 5 | 3 | | 1 | | |
| | −12 | 2 | | 11 | | 3 | 1 | | 14 | 1 | 2 | | | | |
| | −6 | 4 | 1 | 9 | | 1 | 2 | | 15 | 2 | | | | | |
| | 0 | 5 | 2 | 8 | | 1 | 1 | | 16 | | 1 | | | | |
| | 6 | 7 | | 6 | | 1 | 3 | | 16 | 1 | | | | | |
| | SNR (dB) | b | v | d | z | ʒ | | | b | v | d | z | ʒ | | |
| $ab_{\upsilon}a$–ava | −24 | | 15 | 1 | | 1 | | | 2 | 14 | 1 | | | | |
| | −18 | | 17 | | | | | | 3 | 14 | | | | | |
| | −12 | | 16 | | 1 | | | | 5 | 10 | | 2 | | | |
| | −6 | 1 | 16 | | | | | | 8 | 9 | | | | | |
| | 0 | | 15 | | 2 | | | | 8 | 9 | | | | | |
| | 6 | | 17 | | | | | | 8 | 8 | | 1 | | | |

différences significatives, pour tous les SNR, entre *natural* et *nil* et entre *natural* et *plf*. Des différences significatives entre *plf* et *nil* sont mises en évidence à −24 dB ($\chi^2(1) = 5{,}00$, $p = 0{,}025$) et à −18 dB ($\chi^2(1) = 6{,}40$, $p = 0{,}011$).

Pour la combinaison [$ab_{\upsilon}a$–ava], les différences entre *natural* et *nil* sont significatives pour tous les SNR sauf à −24 dB. Des différences significatives sont trouvées entre *natural* et *plf* pour les SNR supérieurs à −18 dB. Des différences significatives sont trouvées entre *plf* et *nil* à −6, 0 et 6 dB (p. ex., à −6 dB : $\chi^2(1) = 5{,}44$, $p = 0{,}020$).

qu'une fois. Cette statistique est interprétée selon la distribution du $\chi^2$ à un degré de liberté ; le test est bilatéral.
Signalons au passage la page web (en anglais) de Dallal consacrée à la pratique statistique que nous avons trouvée très pédagogique ; DALLAL (G. E.). The little handbook of statistical practice. http://www.tufts.edu/~gdallal/LHSP.HTM. Last visited in September 2004.

## 5.4.3. Discussion

Comme on pouvait s'y attendre, l'apport perceptif procuré par les points lumineux est moins important que celui procuré par la vidéo d'origine. Les points lumineux semblent être de la parole visuelle globalement dégradée : par rapport à la vidéo, les points lumineux sont moins intelligibles (p. ex., [m] et [n] sont plus confondues) et ils ont des performances moindres pour l'intégration audiovisuelle.

Des idiosyncrasies expliquent certaines observations, comme l'asymétrie dans les confusions [l]–[ʀ]. Ce n'est pas le cas pour la distinction du trait de nasalité ([m] et [n] ne sont pas confondues avec [b] et [d]). Ces confusions géométriques mais pas auditives pourraient bien être exemplaires de la complémentarité perceptive des signaux audiovisuels ; la persistance de cette distinction pour les forts niveaux de bruit laisse pourtant ouverte la question de l'accord de cette explication avec des théories actuelles de la perception audiovisuelle de la parole [63] ; il est à noter que les confusions géométriques pourraient être dues à une insuffisance du modèle articulatoire à rendre certains mouvements « subtils » [64] ; enfin, les confusions géométriques ne sont pas *a priori* identiques aux confusions perceptives [65] : les phénomènes de coarticulation qui entrent en jeu pen-

63. Cette approche de l'analyse audiovisuelle de scène fait intervenir des interactions de bas niveau entre les modalités : p. ex., des traits visuels peuvent être « utilisés » pour débruiter ou guider l'analyse du signal acoustique ; Schwartz (J.-L.), F. Berthommier et C. Savariaux. Auditory syllabic identification enhanced by non-informative visible speech. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 19–24, St Jorioz, France, septembre 2003.

64. Une telle justification est présentée pour expliquer les meilleures performances en identification, et notamment des consonnes nasales, par les personnes présentant des déficiences auditives. Sur la figure 5.6 l'arbre des confusions perceptives visuelles montre pour l'anglais que des personnes entraînées ne confondent pas [d] et [n]. Cependant, une personne sourde, appareillée, a participé à cette expérience ; le tableau ci-dessous donne ses réponses au SNR −24 dB, c.-à-d., n'était la gêne due au bruit, dans des conditions de lecture labiale. Il semble intéressant de noter que : le taux d'erreur est important ; [m] est perçue comme [b] pour *natural*, et c'est l'inverse pour *plf*. Bien sûr la portée statistique de ce test est limitée.

| | *natural* | | | | | | | *plf* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | b | m | d | v | l | z | g | b | m | d | v | l | z | g |
| b | 3 | | | | | | | | 3 | | | | | |
| m | 2 | 1 | | | | | | | 3 | | | | | |
| d | | | 1 | | | 2 | | | | | | | 1 | 2 |
| n | | | 2 | | | 1 | | | | | 1 | 1 | 1 | |

65. Bernstein (L. E.), J. Jiang, A. Alwan et E. T. Auer. Similarity structure in visual phonetic

dant l'échantillonnage au centre des cibles acoustiques sont à distinguer des indices qui peuvent être extraits du décours temporel des consonnes en contexte lors de la tâche perceptive — informations temporelles alors effectivement capturées par le système de suivi ou le modèle articulatoire. Des expériences complémentaires, notamment un test en lecture labiale pure, sont ici nécessaires pour valider partie de ces hypothèses.

De manière intéressante, dans des conditions similaires (c.-à-d. pour les stimuli liés à la combinaison [ab$_v$a–ava], en parole claire) nous observons pour l'intégration les mêmes résultats avec notre tête parlante de synthèse que ceux observés avec un locuteur humain dans l'expérience de Rosenblum et Saldaña [66].

Revenons un instant sur le tableau 5.5, sur l'identification des stimuli [aga–aga] : de prime abord, il peut paraître surprenant, voire contradictoire avec les résultats sur les performances globales d'identification, que l'on observe à −24 dB un pourcentage de réponses correctes avec la vidéo inférieur à celui avec les points lumineux, et ce d'autant plus que l'audio est très robuste au bruit. Pour comprendre cela, nous avons réalisé une expérience de contrôle dans laquelle nous avons demandé à des participantes et des participants de catégoriser, à partir de la seule modalité visuelle, le stimulus [aga] pour les conditions visuelles *natural* et *plf*. Les résultats de cette expérience montrent que :

- pour ces deux conditions le taux d'identification n'est pas supérieur au hasard ;
- avec *natural* les réponses les plus fréquentes sont [d] et [z], ce qui correspond par ailleurs aux réponses incorrectes données à −24 dB ;
- avec *plf* les réponses les plus fréquentes sont [v] et [z] [67].

D'autre part, un test du $\chi^2$ de McNemar à −24 dB montre une différence significative entre *natural* et *plf* ($\chi^2(1) = 4{,}00$, $p = 0{,}046$).

Tout semble donc se passer comme si :

- en lecture labiale les deux conditions *natural* et *plf* avaient des comportements similaires ;
- au SNR −24 dB on observait pour *natural* une continuité par rapport aux réponses en lecture labiale (SNR $-\infty$ dB), alors que ce ne serait pas le cas

perception and optical phonetics. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 50–55, Aalborg, Denmark, septembre 2001.

66. Rosenblum (L. D.) et H. M. Saldaña. An audiovisual test of kinematic primitives for visual speech perception. *J. of Experimental Psychology : Human Perception and Performance*, 22(2):318–331, 1996.

67. Rappelons aussi que l'intérieur des lèvres varie pour *natural* tandis qu'il reste vide pour *plf*.

pour *plf*.

De telles observations pourraient s'expliquer en énonçant l'hypothèse qu'au SNR −24 dB l'intégration de l'audio et de la modalité visuelle *plf* se ferait au détriment complet de cette dernière, alors que la modalité visuelle *natural*, elle, serait plus « fiable » et donc prise en compte.

Pour ouvrir cette discussion, des études supplémentaires qui incluraient des conditions distinctes en points lumineux des articulateurs pourraient aider à comprendre comment la dégradation observée pourrait être due à la modalité visuelle « points lumineux » *per se*, à l'absence de la langue, ou à d'autres facteurs.

## 5.5. EXPÉRIENCE II — INTELLIGIBILITÉ DES MOUVEMENTS ESTIMÉS POUR PLUSIEURS MODÈLES D'APPARENCE

### 5.5.1. Méthode

Tous les participants et toutes les participantes de l'expérience I ont été recrutés pour l'expérience II une semaine plus tard. La tâche d'identification et l'interface étaient les mêmes. Les mouvements (qui correspondaient aux mêmes stimuli que ceux de l'expérience I) ont été estimés grâce à l'algorithme de suivi basé sur différents modèles d'apparence : *tex_lin*, *tex_cst*, et *la_lin*. Pour contrôler cette expérience, les stimuli incluaient également les mouvements *plf* de l'expérience I, qui sont renommés *ground_truth* dans cette section. Les mouvements résultant sont rendus avec des points lumineux et joués en synchronie avec les signaux audio d'origine, bruités aux SNR { −24, −18, 0 } dB [68]. Tous les stimuli sont présentés en ordre aléatoire, en une seule session ; les participantes et les participants sont prévenus qu'ils peuvent prendre une pause quand ils en ressentent le besoin.

68. Des mouvements correspondant à deux autres modèles d'apparence — non retenus ici — faisaient aussi partie de l'expérience II. Avec trois SNR et six conditions visuelles pour l'expérience II, le nombre total de stimuli est le même dans les deux expériences.

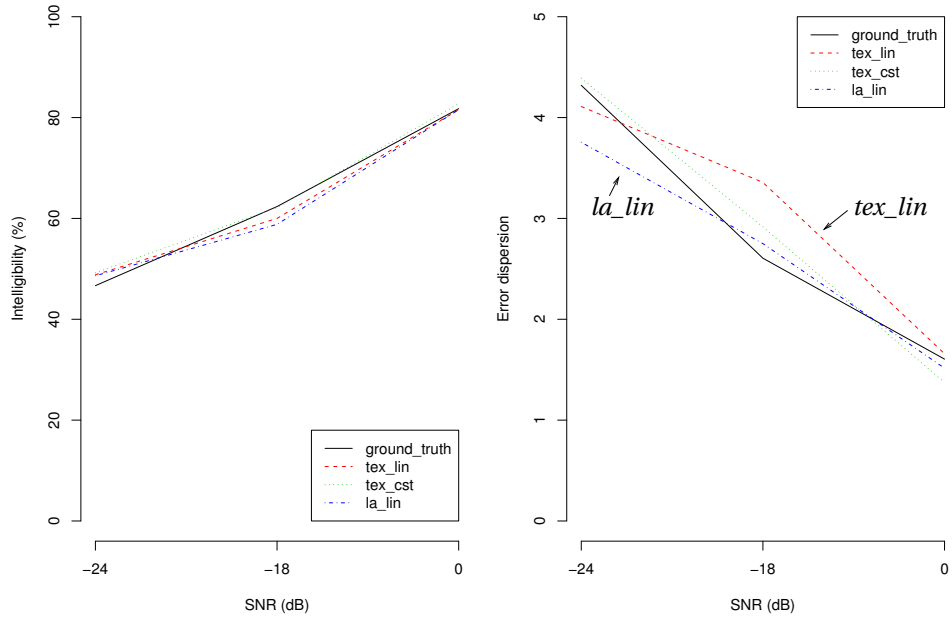



FIGURE 5.7 – Taux d'identification (à gauche) et dispersion de l'erreur, exprimée comme le nombre de catégories d'erreur par consonne (à droite), en fonction du niveau de bruit audio, pour chaque système de suivi.

## 5.5.2. Résultats

### 5.5.2.1. *Contrôle avec l'expérience* I

Une amélioration moyenne du taux d'identification correcte de 4 % est observée entre les deux expériences, c.-à-d. en considérant les résultats pour *plf* et *ground_truth* aux SNR $\{-24, -18, 0\}$ dB ; cela reflète les conversations post-expérimentales avec les participantes et les participants qui ont trouvé la tâche plus facile dans la seconde expérience.

Une ANOVA à deux facteurs avec mesures répétées montre des effets principaux significatifs pour la condition visuelle ($F(1,16) = 8{,}200\,4$, $p = 0{,}011$) et pour le niveau de bruit ($F(2,32) = 254{,}99$, $p < 0{,}001$), et une absence d'interaction ($F(2,32) = 0{,}878\,2$, valeur de $p$ supérieure au seuil de 0,05, sauf indication contraire). Des ANOVA à un facteur avec mesures répétées, faites pour chaque

niveau de bruit, montrent une différence significative à −24 dB ($F(1,16) = 7{,}427$, $p = 0{,}015$), et des différences non-significatives à −18 dB et à 0 dB.
Aussi, la dispersion de l'erreur pour les réponses incorrectes avec *ground_truth* est en moyenne plus petite de 0,48 catégories par rapport à *plf*.

Ces différences pourraient être dues au fait que les stimuli *ground_truth* pouvaient être présentés tout au long de l'expérience II, tandis que dans l'expérience I les stimuli *plf* n'étaient présentés qu'à la troisième session, et donc après la session « audio seul », jugée unanimement difficile et fatigante. Aussi, le plus grand nombre de niveaux de bruit dans l'expérience I a pu rendre la tâche plus difficile.

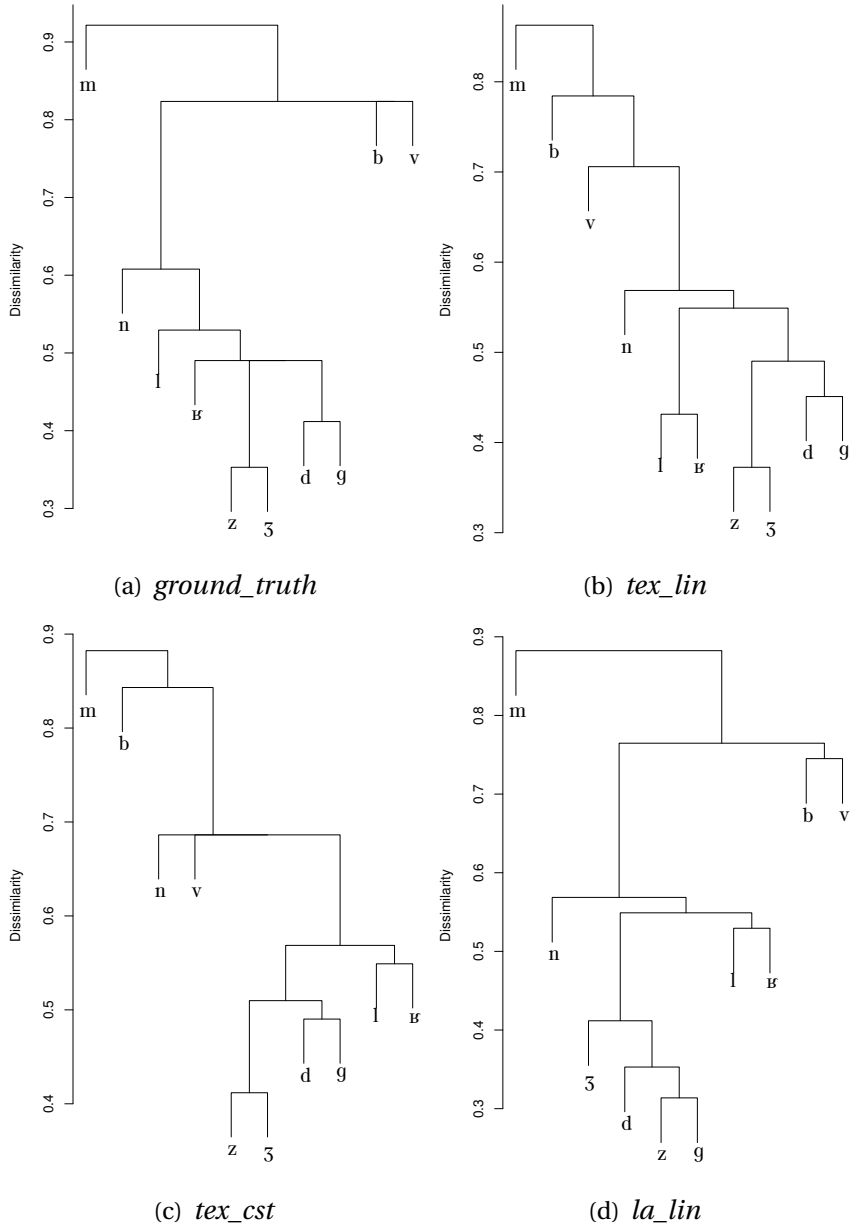
#### 5.5.2.2. *Identification et distribution des réponses incorrectes*

Le pourcentage d'identification consonantique correcte est représenté sur la figure 5.7. Globalement, les taux sont très proches les uns des autres. Une ANOVA à deux facteurs avec mesures répétées montre un effet significatif pour le niveau de bruit ($F(2,32) = 427{,}02$, $p < 0{,}001$) et un effet non significatif pour les conditions visuelles ($F(3,48) = 1{,}101$).
En revanche, les dispersions des réponses incorrectes sont différentes, particulièrement aux SNR les plus faibles. C'est pour le modèle d'apparence *la_lin* que la dispersion de l'erreur est la plus faible ; ce modèle d'apparence est donc plus « cohérent ». D'après les matrices de confusion, ces différences sur la dispersion de l'erreur ne semblent pas être dues à des stimuli particuliers mais plutôt refléter des tendances globales.

#### 5.5.2.3. *Arbres de confusion*

La figure 5.8 montre les arbres de confusion de chacune des conditions au SNR −18 dB. Vraisemblablement à cause de la présence de l'acoustique, ces arbres perceptifs sont différents des arbres géométriques de la figure 4.6. Comme les précédentes, cette représentation confirme que les différentes conditions visuelles ont un impact perceptif similaire. L'arbre pour *la_lin* est peut-être celui dont la structure est la plus proche de celle de l'arbre pour *ground_truth* qui correspond à la vérité terrain. Si l'on choisit d'utiliser le seuil classique de 75 % de confusion, pour toutes les conditions visuelles seules les articulations frontales sont bien distinguées, avec une seule classe pour les autres consonnes.

Comme dans l'expérience I, on remarque que [b] et [m] sont bien identifiées,




(a) *ground_truth* (b) *tex_lin*

(c) *tex_cst* (d) *la_lin*

FIGURE 5.8 – Arbres de confusion perceptive pour chaque système de suivi au SNR −18 dB. Les échelles sont différentes. L'arbre pour les réponses audio est représenté sur la figure 5.5.

TABLEAU 5.7 – Pourcentage de réponses « correctes » (basé sur l'audio) pour les stimuli *McGurk* de l'expérience II. Pour les réponses de type « audio seul », on peut se référer au tableau 5.5.

| *Modèle d'apparence* | Stimuli | SNR (dB) | | |
|---|---|---|---|---|
| | | −24 | −18 | 0 |
| A–V incongruents | | | | |
| *ground_truth* | aba–aga | 24 | 59 | 94 |
| *tex_lin* | aba–aga | 18 | 41 | 94 |
| *tex_cst* | aba–aga | 12 | 47 | 88 |
| *la_lin* | aba–aga | 29 | 53 | 100 |
| *ground_truth* | ab$_v$a–ava | 24 | 24 | 41 |
| *tex_lin* | ab$_v$a–ava | 24 | 29 | 41 |
| *tex_cst* | ab$_v$a–ava | 29 | 18 | 29 |
| *la_lin* | ab$_v$a–ava | 29 | 35 | 53 |
| A–V congruents | | | | |
| *ground_truth* | aba–aba | 94 | 88 | 100 |
| *tex_lin* | aba–aba | 88 | 94 | 100 |
| *tex_cst* | aba–aba | 100 | 94 | 100 |
| *la_lin* | aba–aba | 100 | 100 | 100 |
| *ground_truth* | aga–aga | 71 | 65 | 94 |
| *tex_lin* | aga–aga | 47 | 82 | 100 |
| *tex_cst* | aga–aga | 65 | 76 | 100 |
| *la_lin* | aga–aga | 53 | 71 | 100 |

alors que d'un point de vue géométrique nous avons montré (voir la figure 4.6) que ces deux consonnes formaient une seule classe, avec la consonne bilabiale non-voisée [p] ; et de même lors d'études sur la perception seulement visuelle [69]. Étant donné que pour l'audio (*nil*) ces deux consonnes sont bien distinctes l'une de l'autre mais peu identifiées, on peut dire que ce cas est un exemple de la complémentarité perceptive audiovisuelle procurée par les différentes conditions visuelles. Nous avons déjà discuté plus en profondeur ce point à la section 5.4.3.

69. WALDEN (B. E.), R. A. PROSEK, A. A. MONGOMERY, C. K. SCHERR et C. J. JONES. Effects of training on the visual recognition of consonants. *J. of Speech and Hearing Research*, 20:130–145, 1977.

#### 5.5.2.4. *Effet McGurk*

Le tableau 5.7 montre le pourcentage de réponses identiques à la consonne audio pour les stimuli McGurk. Bien que *tex_cst* semble être plus performant, les tests du $\chi^2$ de McNemar ne montrent pas de différences significatives. Notons que pour les stimuli [aga–aga], le score pour *ground_truth* est identique à celui pour *plf* dans l'expérience 1, mais les systèmes *tex_lin* et *la_lin* obtiennent eux des scores comparables à ceux dans l'expérience 1 de la vidéo *natural*.

### 5.5.3. Discussion

Cette expérience n'a pas permis de mettre en évidence des différences majeures quant à l'intelligibilité et à leur capacité à être intégrés avec le signal audio des mouvements estimés avec les modèles d'apparence *tex_lin*, *la_lin* et *tex_cst*.

Ces performances sont très satisfaisantes parce qu'elles sont similaires à celle des données terrain. Nous sommes dès lors tentés de conclure que le système d'estimation des mouvements fournit des résultats aussi intelligibles que la vérité terrain ; néanmoins, il faut rappeler que ce que nous appelons ici les mouvements de la vérité terrain sont en fait régularisés par le modèle de forme, qui ne peut pas reproduire *exactement chaque* posture faciale. Cela tend à infirmer l'hypothèse selon laquelle ces stimuli particuliers sont des *mouvements biologiques*. Une future expérience qui comparerait ces mouvements régularisés par le modèle de forme avec un affichage véritable et minutieusement contrôlé des points considérés permettrait de lever cette incertitude.

Dans les mêmes conditions vidéos très « contraintes », l'évaluation objective de ces modèles d'apparence réalisée au chapitre précédent avait abouti à la même conclusion. Dans des conditions moins contrôlées, les performances objectives de ces différents modèles de l'apparence variaient, cependant. Des tests perceptifs basés sur des mouvements extraits de vidéos plus difficiles mettraient vraisemblablement en lumière des traits discriminants.

## 5.6. CONCLUSION

Pour les équipes qui conçoivent des synthétiseurs de parole audiovisuelle, l'évaluation perceptive de leurs têtes parlantes virtuelles va de soi : le développement des modules d'animation (modèles de coarticulation, etc.) tient compte de connaissances géométriques, dynamiques en production de parole ; il prend aussi en compte la perception du système par la population qui pourrait être amenée à l'utiliser ; les expériences perceptives sont naturellement intégrées dans le cycle de mise au point des synthétiseurs.

Ce raisonnement centré sur l'humain n'est pas celui des nombreux travaux en vision par ordinateur — si l'on en juge d'après les publications dans ce domaine — consacrés à l'analyse et à la synthèse des mouvement faciaux.

Nous avons proposé une méthodologie pour l'évaluation perceptive de mouvements estimés d'après la vidéo et plus précisément, tout en ne négligeant pas l'évaluation de la qualité, pour juger de leur intelligibilité. Pour se focaliser sur l'évaluation des mouvements *per se* — en laissant autant que possible de côté la question de l'éventuel rendu synthétique —, nous avons utilisé le paradigme expérimental des « points lumineux ». Dans une première expérience de catégorisation de stimuli *VCV* avec l'audio bruité, les stimuli biologiques sont étalonnés : cela permet de situer la vérité terrain présentée sous forme de points lumineux par rapport à la vidéo d'origine et à l'absence de signal visuel. C'est dans la deuxième expérience — où la tâche est identique à celle de la première expérience — que les mouvements estimés d'après la vidéo sont comparés à la vérité terrain, sous une présentation unique en points lumineux.

La première expérience a permis de caractériser les points lumineux : ils sont effectivement intégrés au signal audio bruité pour rehausser la compréhension, mais leur apport est, de manière globale, inférieur à celui de la vidéo originale.

La deuxième expérience n'a pas mis en évidence de différences perceptives significatives entre les mouvements estimés à partir des différents modèles d'apparence. Manifestement les séquences vidéos que nous avons utilisées étaient trop faciles : le visage richement texturé de la locutrice, qui portait encore les billes servant à construire le modèle 3D, et l'hyper-articulation de logatomes courts ont pu masquer des différences de comportement entre les modèles d'apparence utilisés pour le suivi.

Une mise en parallèle a permis de constater que ces résultats perceptifs ne sont pas indépendants des résultats de l'évaluation objective (principalement

basée sur la géométrie) ; nous pensons toutefois qu'il serait imprudent de se passer du jugement humain — certes au fort coût expérimental — : par exemple, dans une tâche *a priori* plus difficile perceptivement mais équivalente du point de vue de l'évaluation géométrique, des tests d'identification (et non plus de catégorisation) de phrases courtes pourraient apporter une variabilité suffisante dans les conditions expérimentales pour mieux évaluer et distinguer les modèles d'apparence utilisés pour le suivi.

Enfin, ces expériences ont montré que la tête parlante virtuelle de notre locutrice, animée à partir des paramètres estimés d'après l'analyse de la vidéo, est capable de reproduire l'effet McGurk. Ce résultat, bien que préliminaire, est très encourageant : c'est à notre connaissance la première fois que cette illusion est répliquée avec de la parole visuelle synthétique [70].

Ce résultat permet de penser que notre système de synthèse de la parole audiovisuelle est utilisable pour permettre de mieux comprendre la parole naturelle. Le modèle 3D (partiellement) rendu en points lumineux pourrait devenir un outil méthodologique à disposition pour comprendre certains aspects du traitement du mouvement lors de l'intégration audiovisuelle de la parole.

70. Pour une tentative infructueuse récente, voir Cosker (D.), D. Marshall, P. L. Rosin, S. Paddock et S. Rushton. Towards perceptually realistic talking heads : models, methods and McGurk. *In Symposium on Applied Perception in Graphics and Visualization*, Los Angeles, USA, août 2004.

# Conclusion

## RÉCAPITULATIF DES CONTRIBUTIONS

Les travaux présentés dans cette thèse sont consacrés à l'élaboration et à l'évaluation d'un système de capture en trois dimensions des mouvements faciaux de parole à partir d'une séquence vidéo.

Les difficultés importantes auxquelles sont confrontés les systèmes d'analyse de la vidéo sont essentiellement dues à la grande variabilité des images à traiter ; dans le cas général de multiples facteurs — conditions d'enregistrement, différences inter-personnelles, déformations faciales — sont à considérer et contribuent à rendre complexe la tâche de l'estimation de la 3D. Nous avons choisi de nous restreindre au cas des mouvements du visage d'une personne connue engagée dans une communication parlée face à face dans un contexte expressif neutre.

L'approche adoptée pour l'estimation des mouvements aboutit à un système architecturé autour d'une boucle d'analyse par la synthèse qui cherche à ajuster à l'image un modèle de forme et d'apparence du visage. Cette modélisation du visage parlant est l'une de nos contributions principales. Ces modèles de la forme 3D et de l'apparence du visage sont construits au préalable à partir d'un ensemble choisi de données d'apprentissage représentatives et spécifiques au locuteur ou à la locutrice. La méthodologie de construction de modèles tridimensionnels du visage que nous avons proposée aboutit à des modèles pilotés linéairement par un petit nombre — six ou sept — de paramètres indépendants, *arti-*

*culatoires* parce que propres à la parole et déterminés en fonction des connaissances sur les degrés de liberté des articulateurs du conduit vocal ; il en est de même pour les deux types de modèles de l'apparence que nous avons présentés, directement pilotés par les mouvements faciaux.

Le deuxième point sur lequel nous nous sommes focalisés concerne l'évaluation du système et de ses composants. Par comparaison avec des données de référence, les évaluations objectives ont notamment porté sur l'erreur de recouvrement de la géométrie 3D et sur la pertinence, articulatoire, phonétique, des mouvements estimés. Ces mouvements ont également été soumis avec succès au jugement perceptif grâce à des expériences d'intelligibilité et d'intégration audiovisuelle. Les résultats obtenus ont permis d'illustrer le bon fonctionnement global du système quand il s'appuie sur des modèles du visage dont l'apparence dépend de l'articulation.

## PERSPECTIVES

Nous reprenons maintenant quelques points évoqués dans les conclusions de chaque chapitre.

Pour dépasser les limitations du modèle géométrique des lèvres mises en évidence au cours de nos expériences, son prolongement vers l'intérieur du conduit vocal devrait permettre de mieux représenter la zone interne des lèvres, visible surtout dans les configurations articulatoires arrondies, et assurer à la fois un meilleur ancrage de points de chair dans cette zone et un meilleur rendu visuel [1].

Si le contrôle *articulatoire* des modèles d'apparence s'est révélé pertinent, il s'est avéré que des variations d'apparence importantes ne sont pas explicables par un modèle linéaire : d'autres types de relations, non-linéaires, peuvent permettre de rendre compte de ces variations et ainsi améliorer encore les performances du système [2].

1. Kuratate (T.). *Talking head animation system driven by facial motion mapping and a 3D face database.* PhD thesis, Department of Information Processing, Nara Institute of Science and Technology, Nara, Japan, juin 2004 ; Mantel (C.). *Modélisation 3D non linéaire des lèvres et du visage à partir de données IRM par ACP à noyau.* Master SIPT, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

2. On pourra s'intéresser à Gacon (P.). *Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, En préparation.

Des études sur l'architecture du système pourraient aboutir à plusieurs profils du système d'estimation des mouvements, suivant que l'on souhaite p. ex. un fonctionnement plus proche du temps réel ou une gestion plus robuste de la phase d'initialisation.

Les autres perspectives naturelles de ce travail visent à intégrer des sources de variabilité pour l'instant « contrôlées ». Dans cette optique, la robustesse à des changements importants d'illumination passera par l'introduction de paramètres supplémentaires dans la modélisation. Par ailleurs, l'extension du modèle de contrôle à la parole en contexte expressif variable permettra au système de mieux gérer l'interaction avec l'environnement et d'améliorer le réalisme et l'attractivité des têtes parlantes. Concernant ce dernier point, il sera important de donner un regard [3] et une langue [4] au clone de synthèse : bien modélisés, ces deux organes auront des impacts perceptifs bénéfiques.

3. Itti (L.), N. Dhavale et F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. *In Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, pages 64–78, août 2003 ; Rossi (R.). Scrutation de scènes pour l'animation du regard d'un agent conversationnel. Rapport interne, Institut de la Communication Parlée, Grenoble, France, 2005.

4. Badin (P.), G. Bailly, L. Revéret, M. Baciu, C. Segebarth et C. Savariaux. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *J. of Phonetics*, 30(3):533–553, 2002.

# Appendice : le corpus « 77 phrases »

Ce corpus consiste en la réunion de deux corpora :

– un corpus composé de 68 courtes phrases, articulatoirement riche (c.-à-d., les phonèmes sont prononcés dans des contextes phonétiques divers qui reflètent la variabilité due à la coarticulation), établi à l'Institut de Phonétique de Strasbourg [1] ;
– un corpus composé de 9 phrases faisant partie d'une des vingt listes phonétiquement équilibrées (c.-à-d., d'après l'auteur, que « les fréquences relatives pour chaque phonème reflètent celles qui existent dans la langue française ») [2].

1. Bothorel (A.), P. Simon, F. Wioland et J.-P. Zerling. *Cinéradiographie des voyelles et consonnes du français.* Institut de Phonétique de Strasbourg, 1986.

2. Combescure (P.). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique,* 56:34–38, 1981.

# 68 PHRASES ARTICULATOIREMENT ÉQUILIBRÉES

Ma chemise est roussie.
Tiens-toi assis.
Elle m'étripa.
Deux machines à sous.
Ce mignon bout de chou.
La force du coup.
Six beaux tapis.
Ce plat de hachis.
La petite massue.
Il fume son tabac.
L'examen du cas.
Je vais chez l'abbé.
J'habite ici.
Une belle rascasse.
C'est la lune qui se couche.
Faire la nouba.
C'est Louis qui joue.
Pas plus de quatre rubis.
C'est lui qui me poussa.
Trop d'abus.
Le pied du gars.
Le vent mugit.
Tu ris beaucoup.

Voilà des bougies.
Il a du goût.
Une réponse ambigüe.
Un fourré touffu.
Voilà du filet cru.
Une pâte à choux.
Vous êtes exclue.
C'est une pièce crue.
C'est une vigie.
Un piège à poux.
Je suis à bout.
Deux jolis boubous.
J'apprends à naviguer.
Il fit tout pour l'attraper.
Un petit coucou.
Il supporta la secousse.
C'est ma tribu.
Une rocaille moussue.
La chaise du bout.
J'en ai assez.
Vous avez réussi.
Une autre roupie.
Mon sang se figea.

Donne un petit coup.
Comment l'as-tu su ?
Louis pense à ça.
Un tour de magie.
Mets tes beaux habits.
Prête-lui seize écus.
Il fait des achats.
Chevalier du gué.
Le jeune hibou.
Acheter du gui.
Il m'a donné une esquisse.
À qui est cette quiche ?
Elle a chu.
Achète ce fichu.
Il part pour Vichy.
Dix carrés de nougat.
Gilles m'attaqua.
Un pied fourchu.
Jean se coucha.
Jean est fâché.
Ils n'ont pas pu.
Deux beaux bijoux.

Combien de contrepèteries y trouverez-vous ?

## 10 PHRASES PHONÉTIQUEMENT ÉQUILIBRÉES

Il se garantira du froid avec ce bon capuchon.
Dès que le tambour bat, les gens accourent.
Les deux camions se sont heurtés de face.
Annie s'ennuie loin de mes parents.
La vaisselle propre est mise sur l'évier.
Ce petit canard apprend à nager.
Vous poussez des cris de colère ?
Mon père m'a donné l'autorisation.
Un loup s'est jeté immédiatement sur la petite chèvre.
La voiture s'est arrêtée au feu rouge.

Pour des raisons techniques la dixième phrase n'a pas pu être utilisée.

# Bibliographie


Abboud (B.), F. Davoine et M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19:723–740, 2004.

Abry (C.), L.-J. Boë, P. Corsi, R. Descout, M. Gentil et P. Graillot. *Labialité et phonétique*. Université des langues et lettres de Grenoble, 1980.

Agelfors (E.), J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens et T. Öhman. Synthetic faces as a lipreading support. *In Proc. of the Internat. Conf. on Spoken Language Processing*, pages 3047–3050, Sydney, Australia, 1998.

Ahlberg (J.). *Model-based coding — Extraction, coding, and evaluation of face model parameters*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, septembre 2002.

Antoszczyszyn (P. M.), J. M. Hannah et P. M. Grant. Reliable tracking of facial features in semantic-based video coding. *IEE Proceedings — Vision Image and Signal Processing*, 145(4):257–263, août 1998.

Attina (V.), D. Beautemps, M.-A. Cathiard et M. Odisio. A pilot study of temporal organization in Cued Speech production of French syllables: Rules for a Cued Speech synthesizer. *Speech Communication*, 44(1–4):197–214, octobre 2004.

BADIN (P.), G. BAILLY, M. RAYBAUDI et C. SEGEBARTH. A three-dimensional linear model articulatory model based on MRI data. *In Proc. of the Third ESCA/COCOSDA Internat. Workshop on Speech Synthesis*, pages 249–254, Jenolan Caves, Australia, 1998.

BADIN (P.), G. BAILLY, L. REVÉRET, M. BACIU, C. SEGEBARTH et C. SAVARIAUX. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *J. of Phonetics*, 30(3):533–553, 2002.

BAILLY (G.) et P. BADIN. Seeing tongue movements from outside. *In Proc. of the Internat. Conf. on Spoken Language Processing*, pages 1913–1916, Boulder, USA, 2002.

BAILLY (G.), F. ELISEI, M. ODISIO, D. PELÉ, D. CAILLIÈRE et K. GREIN-COCHARD. Talking faces for MPEG-4 compliant scalable face-to-face telecommunication. *In Proc. of the Smart Objects Conf.*, pages 204–207, Grenoble, France, mai 2003.

BAILLY (G.), G. GIBERT et M. ODISIO. Evaluation of movement generation systems using the point-light technique. *In Proc. of the IEEE Workshop on Speech Synthesis*, pages 27–30, Santa Monica, USA, septembre 2002.

BAKER (S.), I. MATTHEWS et J. SCHNEIDER. Automatic construction of active appearance models as an image coding problem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(10):1380–1384, 2004.

BASSILI (J. N.). Facial motion in the perception of faces and of emotional expressions. *J. of Experimental Psychology: Human Perception and Performance*, 4:373–379, 1978.

BASU (S.), I. ESSA et A. PENTLAND. Motion regularization for model-based head tracking. *In Proc. of the Internat. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

BASU (S.), N. OLIVER et A. PENTLAND. 3D lip shapes from video: A combined physical-statistical model. *Speech Communication*, 26:131–148, 1998.

BEAUTEMPS (D.), P. BADIN et G. BAILLY. Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *J. of the Acoustical Society of America*, 109(5):2165–2180, mai 2001.

Benoît (C.), T. Lallouache, T. Mohamadi et C. Abry. A set of French visemes for visual speech synthesis. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines: Theories, Models and Designs*, pages 485–501. Elsevier B.V., 1992.

Benoît (C.) et L. C. W. Pols. On the assessment of synthetic speech. *In* Bailly (G.) et C. Benoît, éditeurs. *Talking Machines: Theories, Models and Designs*, pages 435–441. Elsevier B.V., 1992.

Bérar (M.), G. Bailly, M. Chabanas, F. Elisei, M. Odisio et Y. Payan. Towards a generic talking head. *In Proc. of the Internat. Seminar on Speech Production*, pages 7–12, Sydney, Australia, décembre 2003.

Bergeson (T. R.), D. B. Pisoni et J. T. Reynolds. Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 55–60, St Jorioz, France, septembre 2003.

Bernstein (L. E.), J. Jiang, A. Alwan et E. T. Auer. Similarity structure in visual phonetic perception and optical phonetics. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 50–55, Aalborg, Denmark, septembre 2001.

Beskow (J.). Animation of talking agents. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 149–152, Rhodes, Greece, septembre 1997.

Blanz (V.), C. Basso, T. Poggio et T. Vetter. Reanimating faces in images and video. *In Proc. of the Annual Conf. of the European Association for Computer Graphics*, Granada, Spain, 2003.

Blanz (V.) et T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, septembre 2003.

Blanz (V.) et T. Vetter. A morphable model for the synthesis of 3D faces. *In Proc. of SIGGRAPH*, Computer Graphics Proceedings, Annual Conference Series, pages 187–194. Addison Wesley, août 1999.

Bothorel (A.), P. Simon, F. Wioland et J.-P. Zerling. *Cinéradiographie des voyelles et consonnes du français*. Institut de Phonétique de Strasbourg, 1986.

BRAUNSTEIN (M. L.). Structure from motion. *In* SMITH (A. T.) et R. J. SNOWDEN, éditeurs. *Visual Detection of Motion*, pages 367–393. Academic Press, 1994.

BREGLER (C.) et Y. KONIG. "eigenlips" for robust speech recognition. *In Proc. of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing*, pages 669–672, Adelaide, Australia, avril 1994.

BRETZNER (L.). *Multi-Scale Feature Tracking and Motion Estimation.* PhD thesis, Computational Vision and Active Perception Laboratory, KTH, Stockholm, Sweden, octobre 1999.

BROOKE (N. M.) et S. D. SCOTT. PCA image coding schemes and visual speech intelligibility. *In Proc. of the Institute of Acoustics, Autumn Meeting*, pages 123–129, Windermere, UK, 1994.

CHOE (B.), H. LEE et H.-S. KO. Performance-driven muscle-based facial animation. *J. of Visualization and Computer Animation*, 12:67–79, 2001.

COHEN (M. M.) et D. W. MASSARO. Modeling coarticulation in synthetic visual speech. *In Proceedings of Computer Animation.* Springer-Verlag, 1993.

COHEN (M. M.), R. L. WALKER et D. W. MASSARO. Perception of synthetic visual speech. *In* STORK (D. G.) et M. E. HENNECKE, éditeurs. *Speechreading by Humans and Machines*, volume 150 de *Computer and Systems Sciences*, pages 153–168. Springer, 1996.

COHEN (M. M.), D. W. MASSARO et R. CLARK. Training a talking head. *In Proc. of the IEEE Internat. Conf. on Multimodal Interfaces*, pages 499–504, Pittsburgh, USA, octobre 2002.

COMBESCURE (P.). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56:34–38, 1981.

COOTES (T. F.), D. COOPER, C. J. TAYLOR et J. GRAHAM. Active Shape Models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

COOTES (T. F.), G. J. EDWARDS et C. J. TAYLOR. Active appearance models. *In* BURKHARDT (H.) et B. NEUMANN, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 1407 de *Lecture Notes in Computer Science*, pages 484–498, Freiburg, Germany, juin 1998. Springer-Verlag.

COOTES (T. F.) et P. KITTIPANYA-NGAM. Comparing variations on the active appearance model algorithm. *In Proc. of the British Machine Vision Conf.*, volume 2, pages 837–846, Cardiff, UK, 2002.

COOTES (T. F.), S. MARSLAND, C. J. TWINING, K. SMITH et C. J. TAYLOR. Groupwise diffeomorphic non-rigid registration for automatic model building. *In* PAJDLA (T.) et J. MATAS, éditeurs. *Proc. of the European Conf. on Computer Vision*, volume 3024 de *Lecture Notes in Computer Science*, pages 316–327, Prague, Czech Republic, mai 2004. Springer-Verlag.

COOTES (T. F.), C. J. TAYLOR et A. LANITIS. Active Shape Models: Evaluation of a multi-resolution method for improving image search. *In Proc. of the British Machine Vision Conf.*, pages 327–336, 1994.

COOTES (T. F.), K. N. WALKER et C. J. TAYLOR. View-based active appearance models. *In Proc. of the Internat. Conf. on Face and Gesture Recognition*, pages 227–232, Grenoble, France, 2000.

COSKER (D.), D. MARSHALL, P. L. ROSIN, S. PADDOCK et S. RUSHTON. Towards perceptually realistic talking heads: models, methods and McGurk. *In Symposium on Applied Perception in Graphics and Visualization*, Los Angeles, USA, août 2004.

DALLAL (G. E.). The little handbook of statistical practice. http://www.tufts.edu/~gdallal/LHSP.HTM. Last visited in September 2004.

DAUBIAS (P.). *Modèles* a posteriori *de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle*. Thèse de doctorat, Université du Maine, Le Mans, France, décembre 2002.

DECARLO (D.) et D. METAXAS. Optical flow constraints on deformable models with applications to face tracking. *Internat. J. of Computer Vision*, 38(2):99–127, juillet 2000.

DOMINI (F.) et C. CAUDEK. 3-D structure perceived from dynamic information: a new theory. *TRENDS in Cognitive Sciences*, 7(10):444–449, octobre 2003.

DORNAIKA (F.) et J. AHLBERG. Efficient active appearance model for real-time head and facial feature tracking. *In Proc. of the IEEE ICCV Internat. Workshop on*

*Analysis and Modeling of Faces and Gestures*, pages 173–180, Nice, France, octobre 2003.

DUTOIT (T.). *An introduction to text-to-speech synthesis*, chapitre 7.3, pages 195–200. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1997.

EISERT (P.). *Very low bit-rate video coding using 3-D models*. PhD thesis, University of Erlangen-Nuremberg, octobre 2000.

EISERT (P.) et B. GIROD. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5):70–78, septembre 1998.

EKMAN (P.) et W. V. FRIESEN. *Facial Action Coding System (Investigator's Guide)*. Consulting Psychologists Press, Inc., 1978.

EVENO (N.), A. CAPLIER et P.-Y. COULON. Accurate and quasi-automatic lip tracking. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(5):706–715, mai 2004.

FAGEL (S.) et C. CLEMENS. An articulation model for audiovisual speech synthesis — determination, adjustment, evaluation. *Speech Communication*, 44(1–4): 141–154, octobre 2004.

FALASCHI (A.). Segmental quality assessment by pseudo-words. *In* BAILLY (G.) et C. BENOÎT, éditeurs. *Talking Machines: Theories, Models and Designs*, pages 455–472. Elsevier B.V., 1992.

FAUGERAS (O.). *Three-Dimensional Computer Vision — A Geometric Viewpoint*. MIT Press, Cambridge, USA, 1993.

FILLBRANDT (H.), S. AKYOL et K.-F. KRAISS. Extraction of 3D hand shape and posture from image sequences for sign language recognition. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 181–186, Nice, France, octobre 2003.

FISHER (C. G.). Confusions among visually perceived consonants. *J. of Speech and Hearing Research*, 15:474–482, 1968.

GACON (P.). *Analyse d'images et modèles de formes pour la détection et la reconnaissance. Application aux visages en multimédia*. Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, En préparation.

GARCIA (E.) et J.-L. DUGELAY. Applications and specificities of synthetic/synthetic projective registration. *In Proc. of the IEEE Internat. Conf. on Multimedia and Exposition*, 2004.

GEIGER (G.), T. EZZAT et T. POGGIO. Perceptual evaluation of video-realistic speech. AI Memo 2003-003, Massachusetts Institute of Technology, Cambridge, USA, février 2003.

GEMAN (S.) et D. E. MCCLURE. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistic Institute*, LII-4:5–21, 1987.

GÉRARD (P.) et A. GAGALOWICZ. Three dimensional model-based tracking using texture learning and matching. *Pattern Recognition Letters*, 21:1095–1103, 2000.

GIBERT (G.), G. BAILLY, D. BEAUTEMPS, F. ELISEI et R. BRUN. Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech. *J. of the Acoustical Society of America*, 118(2):1144–1153, 2005.

GOMI (H.), J. NOZOE, J. DANG et K. HONDA. Physiologically based lip model for generating speech articulation. *In Proc. of the Internat. Seminar on Speech Production*, pages 79–84, Sydney, Australia, décembre 2003.

GRANT (K. W.) et P. F. SEITZ. Measures of auditory-visual integration in nonsense syllables and sentences. *J. of the Acoustical Society of America*, 104(4):2438–2450, octobre 1998.

GRANT (K. W.), B. E. WALDEN et P. F. SEITZ. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition and auditory-visual integration. *J. of the Acoustical Society of America*, 103 (5):2677–2690, mai 1998.

GROSS (R.), I. MATTHEWS et S. BAKER. Fisher light-fields for face recognition across pose and illumination. *In DAGM*, volume 2449 de *Lecture Notes in Computer Science*, pages 481–489, Zürich, Switzerland, 2002. Springer-Verlag.

GUENTER (B.), C. GRIMM, D. WOOD, H. MALVAR et F. PIGHIN. Making faces. *In Proc. of SIGGRAPH*, Computer Graphics Proceedings, Annual Conference Series, pages 55–66. ACM SIGGRAPH / Addison Wesley, juillet 1998.

GUIARD-MARIGNY (T.), D. OSTRY et C. BENOÎT. Speech intelligibility of synthetic lips and jaw. *In Proc. of the Internat. Congress of Phonetic Sciences*, volume 3, pages 222–225, Stockholm, Sweden, août 1995.

HALL (D.), V. COLIN DE VERDIÈRE et J. L. CROWLEY. Object recognition using coloured receptive fields. *In Proc. of the European Conf. on Computer Vision*, pages 164–177, Dublin, Ireland, juin 2000.

HAMLAOUI (S.) et F. DAVOINE. Facial action tracking using particle filters and active appearance models. *In Proc. of the Smart Objects Conf.*, pages 165–169, Grenoble, France, 2005.

HOLM (S.). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

HONG (P.), Z. WEN et T. S. HUANG. Real-time speech-driven face animation. *In* PANDZIC (I. S.) et R. FORCHHEIMER, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 7, pages 115–124. Wiley, 2002.

HORAUD (R.) et O. MONGA. *Vision par ordinateur : outils fondamentaux.* Hermès, Paris, France, deuxième édition, 1995.

HOUSE (A. S.), C. E. WILLIAMS, M. H. L. HECKER et K. D. KRYTER. Articulation-testing method: Consonantal differentiation with a closed response set. *J. of the Acoustical Society of America*, 37:158–166, 1965.

HOWELL (D. C.). *Méthodes statistiques en sciences humaines.* De Boeck Université, 1998.

HUANG (X.), S. ZHANG, Y. WANG, D. METAXAS et D. SAMARAS. A hierarchical framework for high resolution facial expression tracking. *In Proc. of the IEEE Internat. Workshop on Articulated and Nonrigid Motion*, Washington D.C., USA, juin 2004.

ITTI (L.), N. DHAVALE et F. PIGHIN. Realistic avatar eye and head animation using a neurobiological model of visual attention. *In Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, pages 64–78, août 2003.

JANSSON (G.), S. S. BERGSTRÖM et W. EPSTEIN, éditeurs. *Perceiving events and objects*. Lawrence Erlbaum Associates, 1994.

JEBARA (T.), A. AZARBAYEJANI et A. PENTLAND. 3D structure from 2D motion. *IEEE Signal Processing Magazine*, 16(3):66–84, 1999.

JOHANSSON (G.). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973.

JULIER (S.) et G. BISHOP. Guest editors' introduction: Tracking: How hard can it be? *IEEE Computer Graphics & Applications*, 22:22–23, novembre-décembre 2002.

JURIE (F.) et M. DHOME. Real time tracking of 3D objects: an efficient and robust approach. *Pattern Recognition*, 35:317–328, 2002.

KALBERER (G. A.), P. MUELLER et L. V. GOOL. Visual speech, a trajectory in viseme space. *Internat. J. of Imaging Systems and Technology*, 13(1):74–84, juin 2003.

KING (S. A.) et R. E. PARENT. Creating speech-synchronized animation. *IEEE Trans. on Visualization and Computer Graphics*, 11(3):341–352, 2005.

KOENDERINK (J. J.) et A. J. van DOORN. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

KOZLOWSKI (L. T.) et J. E. CUTTING. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

KROONENBERG (P. M.). *Three-mode principal component analysis. Theory and applications*. DSWO press, Leiden University, Netherlands, 1983.

KROOS (C.), S. MASUDA, T. KURATATE et E. VATIKIOTIS-BATESON. Towards the facecoder: Dynamic face synthesis based on image motion estimation in speech. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 24–29, Aalborg, Denmark, septembre 2001.

KROOS (C.), T. KURATATE et E. VATIKIOTIS-BATESON. Video-based face motion measurement. *J. of Phonetics*, 30(3):569–590, juillet 2002.

KURATATE (T.). *Talking head animation system driven by facial motion mapping and a 3D face database.* PhD thesis, Department of Information Processing, Nara Institute of Science and Technology, Nara, Japan, juin 2004.

LA CASCIA (M.), S. SCLAROFF et V. ATHITSOS. Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(4): 322–336, avril 2000.

LACHS (L.) et D. B. PISONI. Specification of cross-modal source information in isolated kinematic displays of speech. *J. of the Acoustical Society of America*, 116(1):507–518, 2004.

LAGARIAS (J. C.), J. A. REEDS, M. H. WRIGHT et P. E. WRIGHT. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

LE GOFF (B.), T. GUIARD-MARIGNY et C. BENOÎT. Read my lips... and my jaw! How intelligible are the components of a speaker's face? *In Proc. of the European Conf. on Speech Communication and Technology*, pages 291–294, Madrid, Spain, 1995.

LEBART (L.), A. MORINEAU et M. PIRON. *Statistique exploratoire multidimensionnelle.* Dunod, 1995.

LI (H.), P. ROIVAINEN et R. FORCHHEIMER. 3-D motion estimation in model-based facial image coding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):545–555, juin 1993.

LI (H.) et R. FORCHHEIMER. Model-based coding: The complete system. *In* PANDZIC (I. S.) et R. FORCHHEIMER, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 11, pages 187–218. Wiley, 2002.

LIN (I.-C.), J.-S. YEH et M. OUHYOUNG. Extracting 3D facial animation parameters from multiview video clips. *IEEE Computer Graphics & Applications*, 22(6): 72–80, novembre-décembre 2002.

LINDBLOM (B.). Role of articulation in speech perception: Clues from production. *J. of the Acoustical Society of America*, 99(3):1683–1692, mars 1996.

LINDEBERG (T.). Feature detection with automatic scale selection. *Internat. J. of Computer Vision*, 30(2):77–116, 1998.

LINDEBERG (T.). Principles for automatic scale selection. *In Handbook on Computer Vision and Applications*, volume 2, pages 239–274. Academic Press, Boston, USA, 1999.

LUCERO (J. C.), S. T. R. MACIEL, D. A. JOHNS et K. G. MUNHALL. Empirical modeling of human face kinematics during speech using motion clustering. *J. of the Acoustical Society of America*, 118(1):405–409, 2005.

LUCERO (J. C.) et K. G. MUNHALL. A model of facial biomechanics for speech production. *J. of the Acoustical Society of America*, 106(5):2834–2848, novembre 1999.

LUETTIN (J.) et N. A. THACKER. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

MAEDA (S.). On articulatory and acoustic variabilities. *J. of Phonetics*, 19:321–331, 1991.

MAEDA (S.), M. TODA, A. J. CARLEN et L. MAFTAHI. Functional modeling of the face during speech production. *In Actes des Journées d'études sur la parole*, pages 341–344, Nancy, France, juin 2002.

MANTEL (C.). *Modélisation 3D non linéaire des lèvres et du visage à partir de données IRM par ACP à noyau*. Master SIPT, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.

MARQUARDT (D.). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441, 1963.

MARR (D.) et E. HILDRETH. Theory of edge detection. *Proc. of the Royal Society of London: Biological Sciences*, 207:187–217, 1980.

MATTHEWS (I.), T. F. COOTES et J. A. BANGHAM. Extraction of visual features for lipreading. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2): 198–213, février 2002.

MATTHEWS (I.) et S. BAKER. Active appearance models revisited. *Internat. J. of Computer Vision*, 60(2):135–164, 2004.

McGurk (H.) et J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, décembre 1976.

McKinnon (K. I. M.). Convergence of the Nelder-Mead simplex method to a non-stationary point. *SIAM Journal of Optimization*, 9(1):148–158, 1998.

Messer (K.), J. Matas, J. Kittler, J. Luettin et G. Maître. XM2VTSDB: The extended M2VTS database. *In Proceedings of the Conference on Audio- and Video-based Biometric Personal Authentification*, pages 72–77, Washington, DC, USA, mars 22–23 1999.

Miller (G. A.) et P. E. Nicely. An analysis of perceptual confusions among some English consonants. *J. of the Acoustical Society of America*, 27(2):338–352, mars 1955.

Moghaddam (B.) et A. Pentland. Probabilistic visual learning for object representation. *In* Nayar (S. K.) et T. Poggio, éditeurs. *Early visual learning*, pages 99–130. Oxford University, 1996.

Morency (L.-P.), P. Sundberg et T. Darell. Pose estimation using 3D view-based eigenspaces. *In Proc. of the IEEE ICCV Internat. Workshop on Analysis and Modeling of Faces and Gestures*, pages 45–52, Nice, France, octobre 2003.

Nelder (J. A.) et R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Nölker (C.) et H. Ritter. Visual recognition of continuous hand postures. *IEEE Trans. on Neural Networks*, 13(4):983–994, juillet 2002.

Ojanen (V.). *Neurocognitive Mechanisms of Audiovisual Speech Perception.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2005.

Olivès (J.-L.), J. Kulju, R. Möttönen et M. Sams. Audio-visual speech synthesis for Finnish. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 157–162, Santa Cruz, USA, 1999.

Pandzic (I. S.), J. Ostermann et D. Millen. User evaluation: synthetic talking faces for interactive services. *The Visual Computer*, 15:330–340, 1999.

Parke (F. I.) et K. Waters. *Computer facial animation.* A K Peters Ltd, Wellesley, USA, 1996.

PARKE (F. I.). Parameterized models for facial animation. *IEEE Computer Graphics & Applications*, 2:61–68, novembre 1982.

PECKELS (J. P.) et M. ROSSI. Le test de diagnostic par paires minimales, adapatation au français du *Diagnostic Rhyme Test* de Voiers. *Revue d'Acoustique*, 27:245–262, 1973.

PELACHAUD (C.). Visual text-to-speech. *In* PANDZIC (I. S.) et R. FORCHHEIMER, éditeurs. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*, chapitre 8, pages 125–140. Wiley, 2002.

PERRET (Y.). *Suivi de paramètres de modèle géométriques à partir de séquences vidéo multi-vues.* Thèse de doctorat, Université Claude Bernard, Lyon, France, décembre 2001.

PIGHIN (F.), R. SZELISKI et D. H. SALESIN. Modeling and animating realistic faces from images. *Internat. J. of Computer Vision*, 50(2):143–169, novembre 2002.

PRESS (W. H.), S. A. TEUKOLSKY, W. T. VETTERLING et B. P. FLANNERY. Downhill simplex method in multidimensions. *In Numerical Recipes in C*, chapitre 10-4, pages 408–412. Cambridge University Press, 1992.

PRÊTEUX (F.) et M. MALCIU. Model-based head tracking and 3D pose estimation. *In Proceedings of SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision*, pages 94–110, San Diego, USA, juillet 1998.

R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL http://www.R-project.org. ISBN 3-900051-07-0.

REVÉRET (L.). From raw images of the lips to articulatory parameters: a viseme-based prediction. *In Proc. of the European Conf. on Speech Communication and Technology*, volume 4, pages 2011–2014, Rhodes, Greece, septembre 1997.

REVÉRET (L.), G. BAILLY et P. BADIN. MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *In Proc. of the Internat. Conf. on Spoken Language Processing*, volume 2, pages 755–758, Beijing, China, octobre 2000.

Revéret (L.) et I. Essa. Visual coding and tracking of speech related facial motion. *In IEEE International Workshop on Cues in Communication*, Hawaii, USA, décembre 2001.

Revéret (L.). *Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, mai 1999.

Revéret (L.) et C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 207–212, Terrigal, Australia, décembre 1998.

Robert-Ribes (J.), J.-L. Schwartz, T. Lallouache et P. Escudier. Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise. *J. of the Acoustical Society of America*, 103(6):3677–3689, juin 1998.

Robert-Ribes (J.). *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, février 1995.

Romdhani (S.), V. Blanz et T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. *In Proc. of the European Conf. on Computer Vision*, volume 2353 de *Lecture Notes in Computer Science*, pages 3–19, Copenhagen, Denmark, mai 2002. Springer-Verlag.

Rosenblum (L. D.), J. A. Johnson et H. M. Saldaña. Visual kinematic information for embellishing speech in noise. *J. of Speech and Hearing Research*, 39(6): 1159–1170, 1996.

Rosenblum (L. D.) et H. M. Saldaña. An audiovisual test of kinematic primitives for visual speech perception. *J. of Experimental Psychology: Human Perception and Performance*, 22(2):318–331, 1996.

Rossi (R.). Scrutation de scènes pour l'animation du regard d'un agent conversationnel. Rapport interne, Institut de la Communication Parlée, Grenoble, France, 2005.

Runeson (S.) et G. Frykholm. Visual perception of lifted weight. *J. of Experimental Psychology: Human Perception and Performance*, 7:733–740, 1981.

Rydfalk (M.). CANDIDE, a parameterized face. Rapport technique LiTH-ISY-I-866, Dept. of Electrical Engineering, Linköping University, 1987.

Santi (A.), P. Servos, E. Vatikiotis-Bateson, T. Kuratate et K. G. Munhall. Perceiving biological motion: Dissociating talking from walking. *J. of Cognitive Neuroscience*, 15(6):800–809, 2003.

Schmid (C.), R. Mohr et C. Bauckhage. Evaluation of interest point detectors. *Internat. J. of Computer Vision*, 37(2):151–172, 2000.

Schwartz (J.-L.), F. Berthommier et C. Savariaux. Auditory syllabic identification enhanced by non-informative visible speech. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 19–24, St Jorioz, France, septembre 2003.

Schwartz (J.-L.), F. Berthommier et C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93: B69–B78, 2004.

Sekiyama (K.) et Y. Tohkura. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. of the Acoustical Society of America*, 90(4):1797–1805, octobre 1991.

Sherrah (J.) et S. Gong. Fusion of perceptual cues for robust tracking of head pose and position. *Pattern Recognition*, 34:1565–1572, 2001.

Shi (J.) et C. Tomasi. Good features to track. *In Proc. of the Internat. Conf. on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

Siciliano (C.), G. Williams, J. Beskow et A. Faulkner. Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. *In Proc. of the Internat. Congress of Phonetic Sciences*, pages 131–134, Barcelona, Spain, août 2003.

Sim (T.), S. Baker et M. Bsat. The CMU pose, illumination, and expression (PIE) database. *In Proc. of the IEEE Internat. Conf. on Automatic Face and Gesture Recognition*, pages 53–58, 2002.

SMINCHISESCU (C.). *Estimation algorithms for ambiguous visual models — Three dimensional human modeling and motion reconstruction in monocular video sequences*. Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, France, juillet 2002.

STRÖM (J.). *Model-based head tracking and coding*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2002.

STRÖM (J.), T. JEBARA, S. BASU et A. PENTLAND. Real time tracking and modeling of faces: an EKF-based analysis by synthesis approach. *In Proc. of the Internat. Conf. on Computer Vision*, Corfu, Greece, septembre 1999.

SUMBY (W. H.) et I. POLLACK. Visual contribution to speech intelligibility in noise. *J. of the Acoustical Society of America*, 26(2):212–215, mars 1954.

SUMMERFIELD (A. Q.). Use of visual information in phonetic perception. *Phonetica*, 36:314–331, 1979.

TAO (H.) et T. S. HUANG. Visual estimation and compression of facial motion parameters — Elements of a 3D model-based video coding system. *Internat. J. of Computer Vision*, 50(2):111–125, novembre 2002.

TERZOPOULOS (D.) et K. WATERS. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):569–579, juin 1993.

THEOBALD (B.-J.), J. A. BANGHAM, I. MATTHEWS et G. C. CAWLEY. Visual speech synthesis using statistical models of shape and appearance. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 78–83, 2001.

THEOBALD (B.-J.), J. A. BANGHAM, I. MATTHEWS et G. C. CAWLEY. Evaluation of a talking head based on appearance models. *In Proc. of the Auditory-Visual Speech Processing Workshop*, pages 187–192, St Jorioz, France, septembre 2003.

TURK (M.) et A. PENTLAND. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–86, 1991.

ULLMAN (S.). The interpretation of structure from motion. AI Memo 476, Massachusetts Institute of Technology, Cambridge, USA, octobre 1976.

van Son (R. J. J. H.). A method to quantify the error distribution in confusion matrices. *In Proc. of the European Conf. on Speech Communication and Technology*, pages 2277–2280, Madrid, Spain, 1995.

Vogelhuber (V.) et C. Schmid. Face detection based on generic local descriptors and spatial constraints. *In Proc. of the Internat. Conf. on Pattern Recognition*, volume 1, pages 1084–1087, Barcelona, Spain, septembre 2000.

Voiers (W. D.). Performance evaluation of speech processing devices: Diagnostic evaluation of speech intelligibility. AF Cambridge Research Laboratories Final Report AFCLR-67-0101, Contract AF19(628)-4987, 1967.

Walden (B. E.), R. A. Prosek, A. A. Mongomery, C. K. Scherr et C. J. Jones. Effects of training on the visual recognition of consonants. *J. of Speech and Hearing Research*, 20:130–145, 1977.

Walker (K. N.), T. F. Cootes et C. J. Taylor. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(5–6):435–440, avril 2002.

Williams (J. J.) et A. K. Katsaggelos. An HMM-based speech-to-video synthesizer. *IEEE Trans. on Neural Networks*, 13(4):900–915, juillet 2002.

Woo (M.), J. Neider et T. Davis. *OpenGL programming guide: The official guide to learning OpenGL, version 1.1.* Addison-Wesley, second édition, 1997.

Yamamoto (E.), S. Nakamura et K. Shikano. Lip movement synthesis from speech based on Hidden Markov models. *Speech Communication*, 26(1-2):105–115, octobre 1998.

Yan (S.), C. Liu, S. Li, H. Zhang, H. Shum et Q. Cheng. Texture-constrained active shape models. *In Proc. of The First Internat. Workshop on Generative-Model-Based Vision*, Copenhagen, Denmark, mai 2002.

Zhang (Y.) et C. Kambhamettu. 3D head tracking under partial occlusion. *Pattern Recognition*, 35:1545–1557, 2002.

# Index des auteurs

Les numéros de page sont composés en gras lorsque la personne est premier auteur. Les numéros de page en italique correspondent à la bibliographie.

## A



## B

## C



## D



## E



## F

## G



## H



## I



## J



## K

## L



## M



## N

## O



## P



## R



## S

## T



## U



## V



## W



## Y

## Z

# Table des figures

# Liste des tableaux

# Table des matières

✠ Cette thèse ✠
est composée en LaTeX ;
plusieurs participants au
groupe de discussion de
Usenet fr.comp.text.tex
ont donné les coups de
main obligés. Les fontes
utilisées sont Utopia,
✠ Fourier et Xipa. ✠

## Résumé

Cette thèse présente un système pour l'estimation en 3D des mouvements du visage d'un locuteur pour chaque image d'une séquence audiovisuelle. Afin de prendre en compte les spécificités de l'articulation du locuteur ainsi que la complexité des déformations faciales lors de la parole, des modèles articulatoires, propres au locuteur, de la géométrie et de l'apparence du visage ont été construits à partir de données soigneusement collectées. Des analyses statistiques supervisées ont fait émerger de ces données un modèle précis en 3D de la géométrie et plusieurs modèles de l'apparence du visage. L'apparence a été vue ici comme la texture de tout le visage ou comme l'apparence locale de points de chair sélectionnés automatiquement sur le visage. L'estimation proprement dite des gestes de la parole a été faite à partir de ces modèles *via* une boucle d'analyse par la synthèse. Les résultats du suivi ont été comparés aux données de référence ; les évaluations basées sur l'erreur de recouvrement de la géométrie 3D et sur les gains d'intelligibilité procurés par les mouvements ont illustré le très bon fonctionnement des systèmes basés sur des descripteurs de l'apparence dépendant de l'articulation.

## Mots clés

Vision par ordinateur ; parole audiovisuelle (production et perception) ;
modélisation articulatoire 3D ; modélisation de l'apparence ; suivi ; évaluation

## Title

Estimation of a speaker's facial movements in an audiovisual sequence

## Abstract

This thesis presents a system that can recover and track the 3D speech movements of a speaker's face for each image of a video sequence. To handle both the specificity of the speaker's articulation and the complexity of facial deformations during speech, speaker-specific articulated models of the face geometry and appearance were built from carefully collected real data. Statistical analyses then led to a precise 3D model of the facial geometry and to several models of the facial appearance. Appearance was considered to be the texture of the entire face or the local appearance of fleshpoints automatically selected over the face. Given these models, the speech gesture estimation was done using an analysis-by-synthesis paradigm. Tracking results were compared with ground truth data not only in terms of recovery errors of the 3D geometry but also in terms of intelligibility enhancement provided by the movements. Results of these evaluations showed a very good performance for systems that used appearance features depending on articulatory movements.

## Keywords

Computer vision; audiovisual speech (production and perception);
3D articulatory modelling; appearance modelling; tracking; evaluation

*Thèse préparée à l'Institut de la Communication Parlée*
*46, avenue Félix Viallet, 38031 Grenoble CEDEX 1, France*