Extraction semi-automatique des mouvements du tractus vocal à partir de données cinéradiographiques

Julie Fontecave

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

THESE

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble
Spécialité : « Signal Image Parole Télécom »

Préparée au laboratoire
Institut de la Communication Parlée, UMR CNRS 5009

dans le cadre de l'Ecole Doctorale
« Electronique Electrotechnique Automatique et Traitement du Signal »

présentée et soutenue publiquement par

Julie FONTECAVE JALLON

le 8 Décembre 2006

Extraction semi-automatique des mouvements du tractus vocal à partir de données cinéradiographiques

---

Directeurs de thèse : Frédéric BERTHOMMIER & Gang FENG

JURY

M.	Gérard Bailly	Président
M.	Yves Laprie	Rapporteur
M.	Rudolph Sock	Rapporteur
M.	Shinji Maeda	Examinateur
M.	Frédéric Berthommier	Co-directeur de thèse
M.	Gang Feng	Co-directeur de thèse

REMERCIEMENTS

Rien de tout ceci n'aurait été possible sans les points marqués manuellement sur les images cinéradiographiques. J'ai évalué à quelques 20000 clics de souris pour marquer les images de langue et de vélum, sur les séquences Wioland, Flament et Laval43. C'est toi, Fred, qui a cliqué ! 20000 mercis ! Mais je te remercie surtout de m'avoir si bien encadrée, de m'avoir fait partager ta culture scientifique et générale. Je crois que tes remarques, parfois critiques, m'ont permis d'avancer, d'évoluer et de me forger peu à peu une identité scientifique. J'ai apprécié de travailler avec toi. Malgré les moments difficiles que tu traverses personnellement, tu as toujours su être disponible et présent, je t'en remercie.

Feng, tu es celui qui m'a fait découvrir et apprécier le traitement du signal d'abord en école d'ingénieurs par tes cours, puis en stage de fin d'études et en DEA. Tu m'as transmis en partie ton amour de l'enseignement et tu m'as fait comprendre les joies et les peines de la recherche. C'est grâce à toi que j'ai découvert l'ICP. Même si nous n'avons pas travaillé ensemble, autant que nous le souhaitions, j'ai toujours su que je pouvais compter sur toi. Merci de cette présence sécurisante dont tu m'as assurée toutes ces années.

Merci à Jean-Luc Schwartz et Pierre Badin de m'avoir accueillie à l'ICP. Je les remercie aussi des conseils qu'ils ont su me prodiguer, scientifiques ou autres, à des moments opportuns de ma thèse.

Je remercie bien évidemment mes deux rapporteurs, d'avoir accepté de lire et juger mon travail.
Yves Laprie a été un des précurseurs en matière d'extraction automatique de contours du conduit vocal. Ces travaux sont en partie à l'origine de cette thèse. C'est un honneur qu'il en soit un rapporteur. Je le remercie également pour les quelques contacts que nous avons pu avoir à propos de son logiciel Snorri. Il a toujours pris le temps de répondre à mes questions, et, même si je n'ai pas poussé plus que ça l'utilisation de Snorri, cela m'a rendu d'immenses services.
Rudolph Sock est, parmi d'autres, à l'origine de cette base de données cinéradiographiques du français, dont sont issues les séquences Wioland et Flament. Je le remercie d'abord grandement pour nous avoir permis de travailler sur ces données. Je le remercie ensuite pour l'intérêt qu'il porte à mon travail.
Merci à Shinji Maeda d'avoir examiné ce travail. Référence dans ce domaine, c'est un immense plaisir et honneur de l'avoir eu dans ce jury. Merci du temps consacré à la lecture de ce manuscrit, au milieu de tant d'autres, la même semaine !
Je remercie Gérard Bailly d'avoir accepté de présider mon jury de thèse. Tout au long de mon travail, il a suivi l'avancement de mes résultats et par ces questions, m'a ouvert certaines pistes de réflexion.

Merci à Denis Beautemps de certaines discussions qui nous ont donné des idées pour avancer et aller plus loin, en particulier pour l'utilisation du modèle acoustique.
A propos de ce modèle, je remercie aussi bien évidemment Pierre Badin et Antoine Serrurier.

Au même titre que Rudolph Sock, Pascal Perrier est à l'origine de la base de données cinéradiographiques du français. C'est lui qui nous a fourni les films numérisés des diverses séquences, et même à deux reprises, suite au crash de mon disque dur ! Je le remercie, ainsi que pour

tous les encouragements et toutes les informations qu'il a pu nous donner, par l'intermédiaire de Frédéric, au cours de cette thèse.
Je remercie aussi Kevin Munhall et Bryan Burt, qui, sur simple demande de notre part, nous ont envoyé sur un DVD l'ensemble de la base de données cinéradiographiques d'ATR. C'est de cette base de données qu'est extraite la séquence Laval43, à l'origine de la plus grande partie de nos résultats.

Je remercie chaleureusement tous les membres de l'ICP. J'ai, grâce à eux, passé de très bons moments au cours de ces dernières années, que ce soit sur le campus ou à la gare. Tous contribuent à créer une atmosphère chaleureuse et conviviale, permettant de travailler dans de bonnes conditions. Il serait trop long ici de citer tout le monde, mais je remercie en particulier les équipes techniques et administratives, aussi bien de la gare et du campus, qui ont toujours été prêtes à m'aider et à répondre à mes questions. J'ai également une pensée particulière pour les adeptes de la pause café, que j'ai toujours retrouvés avec plaisir le matin vers 10h30 après l'habituel et souvent attendu « Café ! », prononcé à l'interphone.
Je tiens à remercier Solange pour son enthousiasme aux dernières JEP à Dinard, qui m'a permis, je crois, d'aborder avec plus de sérénité ma période estivale de rédaction.
J'ai apprécié au cours de ces années les liens entre les jeunes chercheurs, je les remercie tous, actuels, nouveaux, anciens, ceux qui sont partis ou revenus. Je garde de bons souvenirs des réunions et des déjeuners « jeunes-chercheurs », et aussi des conférences, en particulier celle d'Eurospeech à Lisbonne en Septembre 2005, avec Amélie, Marion, Pauline, Mohammad et Antoine.
Je remercie tous les thésards qui ont transité par mon bureau à la gare tout au bout du couloir et avec qui, j'ai partagé des journées de travail : Stephan, Noureddine, Lucie, Julien, Xavier. Merci aussi aux thésards et stagiaires qui ont accepté de passer mes tests de perception.
Parmi ceux qui ne sont plus à l'ICP, je remercie Isabelle pour ma première année de thèse à l'ICP. Tu es la première, dont j'ai suivi la rédaction (de près) et que j'ai vue soutenir. Je n'oublie pas les (longues) discussions avec Claire, Cécile et toi. Mille mercis aussi à Virginie pour ses mails de soutien depuis les Canaries, au cours des dernières semaines, et surtout pour les « ragots » ou autres discussions dans son bureau, avec les bonbons, Antoine, Bertrand, Guillaume et Annemie.
Merci à Guillaume et Virginie d'avoir fait le déplacement pour être là le jour de nos thèses, j'ai été très touchée.
Annemie, je sais que tu n'es plus jeune chercheur, tu es permanente, voire la plus permanente de tous. Toujours là, toujours prête à discuter, à demander comment ça va, merci pour ta bonne humeur et ta présence sans faille (mis à part les obligations sociales et les « mois » de vacances).

Parmi les jeunes chercheurs de l'ICP, il y a mes frères et sœurs de thèse, Claire, Bertrand et Antoine. Même année, moniteur ou non, campus ou gare, nous avons d'une certaine façon, mené nos thèses en parallèle.
Claire, même séparées par une dizaine d'arrêts de tram, nous avons pu partager, par mail ou devant une salade en ville. Probablement qu'Autrans en Janvier 2004 nous a rapprochées. Depuis, tu as su rester disponible en toutes occasions. J'ai particulièrement apprécié de pouvoir vivre avec toi l'atelier « éveil aux sciences à l'école primaire ». Merci pour ton aide et ton amitié. Je te souhaite plein de courage pour finir et pour la suite.
Bertrand, merci de m'avoir permis de vivre une Cherch'ac au labo (j'en rêvais). Même si nos horaires au labo n'ont pas toujours collé (!), tu m'as permis de ne jamais oublier qu'il y a une règle essentielle : Cachan d'abord, le reste après… histoire de se sentir au labo comme à la maison !
Antoine, qui aurait dit il y a quelques années que nous passerions 3 ans dans le même labo ? J'imagine que rares sont les cas où 2 cousins soutiennent le même jour au même endroit. J'ai vécu

ces années avec toi avec un grand plaisir, elles ont permis de mieux se connaître, de parler famille (toujours en bien, n'est-ce pas ?), pour se changer les idées au milieu d'une réflexion articulatoire. Tu as été un soutien et un ami. Merci.

Je n'oublie évidemment pas Caro. Sans être à l'ICP, tu es aussi une « sœur » de thèse, encore plus maintenant que le LIS et l'ICP sont regroupés au sein de GIPSA. Nous avons vécu ensemble cette étape de notre parcours, un point de plus commun à notre CV. Merci d'être cette amie si proche et si à même de comprendre les doutes, les joies et les peines liées à la thèse. Tu me permets ici une transition vers des remerciements plus personnels, puisque avec Jean, vous êtes des amis sur qui l'on sait que l'on peut compter et on n'a pas toujours l'occasion de dire merci.

Je remercie tous mes amis. Sans les citer tous, ils se reconnaîtront.
Anne-Claire, merci de ton soutien depuis le début. Tu as toujours su être là quand il le fallait. On a du shopping en retard, on va se rattraper.
Julie, ça y est, j'en suis venue à bout. Depuis tant d'années, tu es le témoin, la confidente, l'écoute, la meilleure amie. Il est impossible de résumer tout ce que je voudrais te dire, et de toutes façons, ai-je vraiment besoin de l'écrire ?

Merci à Isabelle, Pascal, Emilie et Anne-Laure, de s'être toujours intéressés à l'avancement de cette thèse. Merci à Marine qui rêvait d'assister à ma soutenance pour ne rien comprendre !

Je terminerai avec ceux qui me sont le plus chers et à qui je dédie ce travail.

Thomas, qui réussit tout, tu es le frère dont tout le monde rêve. Tu sembles parti pour commencer une thèse dans quelques années. Bon courage et surtout fais-toi plaisir !

Claire, ce sera probablement le seul livre que j'écrirai dans ma vie. L'écrivain de la famille, c'est toi. Trouve la voie qui t'intéresse vraiment, tu as tellement de talents et de capacités, j'ai confiance en ce que tu peux faire.

Ce n'est pas toujours facile d'être une Fontecave à Grenoble et « pourquoi ne fait-elle pas de chimie ? ». Quant à Google, il ne connaît que toi sur ces 10 premières pages. Papa, tu es mon moteur pour toujours tenter de me surpasser et faire que tu sois fier de moi. J'espère y avoir réussi.

Maman, tu es le soutien infaillible, la présence sécurisante, l'aide matérielle et l'oreille toujours sympathisante, celle qui peut tout entendre et tout comprendre, et dont j'ai tant besoin.

Merci à tous les deux d'être toujours là, de me soutenir dans mes choix et de m'avoir toujours donné les moyens de faire ce que je voulais. Je suis heureuse de rester à Grenoble, pas trop loin de vous. C'est en partie grâce à Pierre.

Merci Pierroux d'avoir vécu cette thèse avec moi, d'avoir affronté, accepté et survécu aux doutes, aux larmes, aux coups de stress, pas toujours compréhensibles. Merci d'y avoir cru, d'avoir su me redonner courage quand il le fallait, d'avoir tout relu en un week-end, d'avoir partagé mes moments de joie et d'être fière de moi. Merci d'être toi. 2006 aura été une sacrée année pour nous deux. Que les suivantes le soient aussi… je nous fais confiance pour ça

Sommaire

Introduction 1

L'imagerie au service de la modélisation articulatoire 9

1. Techniques d'imagerie et d'extraction de données 9

1.1. Techniques d'acquisition de données articulatoires 9

1.2. Extraction de la géométrie du conduit vocal à partir de la cinéradiographie 14

1.2.1. Extraction manuelle 15

1.2.2. Extraction automatique 16

1.2.3. Cas particulier des lèvres 18

2. Modèles articulatoires 19

2.1. Modèles physiques 20

2.2. Modèles géométriques et statistiques 20

2.3. Modèles biomécaniques 22

2.4. Modèles de fonction d'aire 23

Quelques bases de traitement vidéo 25

1. Coefficients DCT 25

2. Indexation vidéo à partir d'images clefs 28

2.1. MPEG et la notion d'images clefs 28

2.2. Indexation vidéo 30

Capture de mouvements 33

1. Paradigme des points lumineux 33

2. Suivi et reconstruction du mouvement humain 34

3. Analyse de scènes dynamiques 35

4. Parole et mouvement 36

PREMIERE PARTIE : METHODE QUASI-AUTOMATIQUE D'EXTRACTION DE MOUVEMENTS A PARTIR DE DONNEES CINERADIOGRAPHIQUES 39

Chapitre 1 : Principe de la méthode et application aux mouvements de la langue dans Wioland 41

1. Base de données Wioland 41

2. Méthode de rétro-marquage 42

2.1. Etape manuelle : Marquage des images clefs 42

2.2. Etape automatique : Rétro-marquage sur l'ensemble de la base 45

2.2.1. Caractéristiques vidéos 45

2.2.2. Indexation automatique 48

2.2.3. Remarque : Rétro-marquage automatique 49

3. Reconstruction du mouvement 50

3.1. Note sur l'ACP 51

3.1.1. Introduction générale 51

3.1.2. Utilisation de l'ACP 52

3.1.3. Application 54

3.2. Continuité temporelle 55

3.2.1. Analyse fréquentielle sur la séquence vidéo 55

3.2.2. Analyse fréquentielle sur les données géométriques 56

3.2.3. Filtrage temporel passe-bas de l'information géométrique 57

3.3. Régularité des projections 58

3.3.1. Observation de la dispersion entre les 2 représentations 58

3.3.2. Moyenne pondérée des voisinages 60

3.4. Interpolation entre les points 62
Chapitre 2 : Evaluation de la méthode 63
1. Mesures des erreurs 63
2. Evaluation quantitative 65
2.1. Nombre d'images clefs 65
2.2. Choix des images clefs 66
2.3. Type d'indexation 68
2.3.1. Taille du voisinage 68
2.3.2. Distance 69
2.4. Effet cumulatif des améliorations apportées 70
2.5. Evaluation finale de l'erreur dans les conditions d'application choisies 71
2.6. Variabilité inter-experts 73
3. Temps d'exécution 74
4. Evaluation qualitative 74
Chapitre 3 : Extensions de la méthode à d'autres articulateurs et à d'autres séquences cinéradiographiques 77
1. Adaptation de la méthode 77
1.1. Principe 77
1.2. Premiers essais sur Wioland 78
1.3. Première extension vers une autre séquence : la séquence Flament 79
2. Estimation des mouvements du conduit vocal sur Laval43, une séquence de la base de données ATR 81
2.1. Extraction séparée des différents articulateurs 81
2.1.1. Réglage des paramètres 82
2.1.2. Pointe de la langue 85
2.1.3. Voile du palais 85
2.1.4. Lèvres 86
2.1.5. Evaluation 88
2.1.6. Mâchoires 90
2.2. Contour complet du conduit vocal 91
2.3. Une étude comparative sur la langue avec une autre méthode d'extraction 93
2.3.1. Méthode d'extraction mise en place à l'IDIAP 93
2.3.2. Comparaison des méthodes 95
Chapitre 4 : Calcul de la fonction d'aire à partir des contours géométriques extraits de la séquence Laval43 99
1. Des contours aux sections sagittales 99
1.1. Mise en place de la grille 99
1.1.1. Définition d'une grille de référence 100
1.1.2. Correction globale de la grille 101
1.2. Mesure directe à partir de la grille statique 102
1.3. Correction image par image 103
1.3.1. Procédure de Yehia 103
1.3.2. Procédure de Beautemps 104
1.4. Représentation des sections 105
1.5. Comparaison des distances sagittales suivant les procédures de mesure 107
2. Des sections à la fonction d'aire 109
2.1. Difficultés rencontrées 111
2.2. Choix des paramètres 112
2.2.1. Modèle *αβ* 112
2.2.2. Rapport pixels/cms 114

2.2.3. Position de la glotte 115

DEUXIEME PARTIE : EXPLOITATION DES DONNEES ARTICULATOIRES ET MISE EN CORRESPONDANCE AVEC LES DONNEES ACOUSTIQUES 117

Relations Articulatoire-Acoustique 119

1. Inversion et synthèse articulatoire 120

1.1. Inversion 120

1.2. Synthèse articulatoire 122

2. Association linéaire 125

Chapitre 5 : Analyse descriptive des données articulatoires de Laval43 127

1. Données phonétiques 127

1.1. Distinction (audio) parole-silence 127

1.2. Voyelles de Laval43 128

2. Observations directes 130

2.1. Configurations géométriques moyennes 130

2.2. Fonctions d'aire moyennes 132

3. Représentations paramétriques 133

3.1. Espaces de représentation 134

3.1.1. Point le plus élevé de la langue 134

3.1.2. Ouverture aux lèvres 134

3.1.3. Constriction 135

3.2. ACP 137

3.2.1. ACP sur les degrés de liberté géométriques 137

3.2.2. ACP sur les sections médio-sagittales et les fonctions d'aire 140

Chapitre 6 : Association linéaire 143

1. Modèle linéaire 143

1.1. Données articulatoires et acoustiques 143

1.1.1. Données articulatoires 143

1.1.2. Données acoustiques 143

1.2. Estimation acoustique linéaire 145

1.3. Comparaison des formants d'origine et des formants estimés 146

1.3.1. Corrélations 146

1.3.2. Espaces formantiques 147

1.3.3. Mesures de biais 148

1.4. Interaction production-perception 149

2. Transformation affine globale 150

2.1. Transformation sur F_1 et F_2 150

2.2. Transformation sur F_3 152

2.3. Apport de la transformation 153

3. Test de perception 154

3.1. Préparation 154

3.2. Résultats 155

4. Approche linéaire sur éléments sélectionnés 158

Chapitre 7 : Synthèse articulatoire 161

1. Données 161

1.1. Formants et signal audio de référence 161

1.2. Estimation de la fonction de transfert du conduit vocal via la fonction d'aire et estimation de formants 163

2. Analyse de formants 165

2.1. Représentations 165
2.2. Comparaison entre les formants d'origine et les formants estimés 168
2.2.1. Corrélation 168
2.2.2. Biais et écart-type 170
2.2.3. Ellipses de dispersion et tables de confusion 175
2.2.4. Discussion sur F_3 179
3. Signal audio synthétisé 180
3.1. Synthèse du signal 181
3.1.1. Source 181
3.1.2. Amplitude 183
3.2. Analyse spectrale 185
3.2.1. Paramétrisation 186
3.2.2. Distance spectrale 186
3.2.3. Réglages des paramètres 189
3.2.4. Analyse des distances spectrales 190
3.3. Test de perception 193
3.3.1. Stimuli 194
3.3.2. Résultats 195
3.4. Conclusion 199
Chapitre 8 : Etude des consonnes de Laval43 201
1. Consonnes alvéolaires et palatales - Etude des points de contact 203
1.1. Méthode d'analyse 204
1.2. Résultats 206
2. Consonnes bilabiales 209
Chapitre 9 : Une nouvelle source d'informations pour l'étude des mouvements du vélum 213
Discussion et Perspectives 219
1. Bilan du travail effectué 220
2. Conclusions 222
3. Perspectives 227
3.1. Perspectives à court terme 227
3.2. Vers de nouvelles voies de recherche 229
Bibliographie 233
ANNEXES 249
A1. Quelques éléments d'anatomie du conduit vocal 251
A2. Corpus des séquences traitées 255
1. Séquence Wioland 255
2. Séquence Flament 256
3. Séquence Laval43 257
A3. Interface de marquage géométrique 259
1. Description 259
2. Remarques 261
A4. Interface d'étiquetage audio 265
A5. Rapport pixels-cms dans Wioland 267

Table des figures

Figure 0 : Principe général de l'algorithme de rétro-marquage [Ber04]. 5

Figure 1 : Vue latérale du conduit vocal complet réalisée manuellement à partir d'une image radiographique, d'après [BSWZ86]. 15

Figure 2 : Exemples d'images de la séquence cinéradiographique Wioland. Il est quasiment impossible de discerner correctement le contour de la langue sur ces images statiques. 43

Figure 3 : (a) Interface de marquage Matlab (explications en annexe). Les bords de l'image ont été supprimés au maximum pour réduire la taille de l'image à 480*490 pixels, de manière à accélérer l'appel aux images avec le curseur. (b) Lignes verticales et horizontales pour fixer les degrés de liberté des points 3 à 10 de la langue..... 44

Figure 4 : Découpages réalisés sur les images redimensionnées (136*108 pixels) pour atténuer les effets de décalages entre les groupes d'images dans la séquence (groupe 1 à 3 de gauche à droite). Les cadres gris correspondent aux images 136*108, les cadres blancs aux images 99*104 pixels, qui sont dites « réduites » et qui sont utilisées pour le calcul des coefficients DCT. 46

Figure 5 : (a) « Grande » image (720*540 pixels) d'origine utilisée pour le marquage. (b) Image « réduite » (99*104 pixels) avec cache noir, pour le calcul des coefficients DCT. 47

Figure 6 : (a) Coefficient DCT numéro 2 calculé sur les 5673 images de la base. (b) Correction sur la moyenne des coefficients DCT, observée sur le coefficient 2..... 47

Figure 7 : (a) Histogramme de répartition des images de la base en fonction de l'indexation par les images clefs. La répartition est uniforme. (b) Les images d'un groupe de la séquence (début, milieu ou fin) ne sont pas uniquement indexées par des images clefs appartenant à ce groupe. 49

Figure 8 : Extraction automatique de gestes de la main à partir de contours actifs. 50

Figure 9 : Exemples de représentations dans le plan principal. (a) Plan principal vidéo résultant de l'ACP sur les coefficients DCT de la base (les points bleus représentent les images clefs et les jaunes les autres images de la base). (b) Plan principal géométrique résultant de l'ACP sur les 12 degrés de liberté de marquage de la langue des 100 images clefs. 54

Figure 10 : (a) Cadres choisis pour l'analyse fréquentielle du mouvement. (b) Moyenne des DSP sur les pixels de chacun des cadres..... 55

Figure 11 : (a) Moyenne des DSP sur les coefficients DCT de chacun des cadres. (b) DSP sur les composantes 1 et 2 du plan principal vidéo. 56

Figure 12 : (a) Moyenne des DSP sur les coordonnées de marquage (12 ddl). (b) Moyenne des DSP sur les composantes 1 et 2 du plan principal géométrique..... 57

Figure 13 : (a) Réponse fréquentielle du filtre choisi et appliqué aux données géométriques. (b) Moyenne sur les 3 cadres des DSP moyennes sur les pixels pour la séquence de départ et la séquence filtrée. 57

Figure 14 : Mise en évidence des irrégularités de projection entre espaces vidéo et géométrique. Les cercles représentent les voisinages (10 plus proches voisins) des points marqués d'une croix rouge. 59

Figure 15 : (a) Trajectoire projetée dans le plan principal de l'ACP vidéo. (b) Trajectoire générée via l'indexation dans le plan géométrique. 59

Figure 16 : Schématisation de la moyenne des voisinages dans les plans principaux vidéo et géométrique. A partir d'une image I et de ses 3 plus proches voisins, on récupère 3 configurations géométriques qu'on moyenne pour trouver la configuration géométrique associée à l'image I. 60

Figure 17 : (a) Trajectoire projetée dans le plan principal vidéo. (b) Trajectoire générée via l'indexation dans le plan géométrique (en trait fin) et trajectoire obtenue après multi-indexation (en trait gras). 61

Figure 18 : Contour de la langue point à point (en pointillés bleus) et estimation de ce contour par un polynôme d'ordre 5 (en trait plein rouge). 62

Figure 19 : Influence du nombre d'images clefs sur l'erreur $Etot_1$, calculée après indexation par 1 voisin et filtrage temporel – (a) représentation linéaire – (b) représentation Log-Log. 66

Figure 20 : (a) Plan principal et 100 images clefs choisies aléatoirement, en respectant la répartition d'origine. (b) Signal (gris) et signal indexé (noir) selon la $2^{ème}$ composante principale. 67

Figure 21 : (a) Plan principal et 100 images clefs choisies uniformément. (b) Signal (gris) et signal indexé (noir) selon la $2^{ème}$ composante principale. 67

Figure 22 : Influence de la taille du voisinage sur l'erreur $Etot_1$, calculée sans filtrage temporel et à partir de 100 images clefs. 68

Figure 23 : Influence du choix de la mesure de similarité sur l'erreur $Etot_1$, pour une estimation calculée sans filtrage temporel et à partir de 100 images clefs. 69

Figure 24 : Bloc-diagramme des traitements postérieurs. 70

Figure 25 : Contribution cumulée des méthodes de réduction d'erreur observée à partir des mesures de $Etot_1$ et $Etot_2$. 71

Figure 26 : Barres d'erreur $Edof_2$ sur une image clef (490*480 pixels). 72

Figure 27 : Distributions des erreurs $Edof_1$ selon les 12 degrés de liberté entre le marquage manuel (en rouge) et le marquage estimé (en bleu). 72

Figure 28 : Contour de la langue estimé à partir de la méthode de rétro-marquage et superposé à l'image d'origine. 75

Figure 29 : Marquage des lèvres et des dents avant de la séquence Wioland (la lèvre supérieure est en haut à droite de l'image). L'image a été découpée pour se focaliser sur ces articulateurs et 10 degrés de liberté ont été définis. 78

Figure 30 : Marquage complet du conduit vocal dans Wioland. 79

Figure 31 : (a) Degrés de liberté pour le marquage du contour de la langue pour la séquence Flament. (b) Contour de langue estimé par la méthode. 80

Figure 32 : (a) Degrés de liberté pour le marquage du vélum pour la séquence Flament. (b) Contour du vélum estimé par la méthode. 80

Figure 33 : Cadres spécifiques pour chaque articulateur. (a) De gauche à droite, les cadres se focalisent respectivement sur le vélum, la mandibule et les lèvres. (b) L'image complète correspond au cadre pris en compte pour l'estimation de la langue, le rectangle blanc délimite le cadre spécifique à l'estimation indépendante de la pointe. 82

Figure 34 : Sachant l'index calculé avec 24*24 coefficients DCT, pourcentage de trames indexées par ce même index en fonction du nombre de coefficients DCT et du voisinage. (a) Cas des lèvres – (b) Cas du vélum. 84

Figure 35 : Erreur $Etot_1$ en fonction du nombre de coefficients DCT pris en compte dans l'indexation (indexation simple et pas de filtrage temporel) pour les lèvres (175 clefs) et le vélum (75 clefs). 84

Figure 36 : Degrés de liberté pour différents articulateurs. (a) Le marquage manuel du vélum est guidé par une grille polaire. (b) Pour chacune des 2 lèvres, 4 lignes définissent 4 ddl sur 8. (c) Le dos et la base de la langue sont décrits à partir de lignes horizontales et verticales, la pointe est représentée par 5 degrés de liberté (cadre pointillé). 84

Figure 37 : Consonne [p] (a) Erreur de marquage géométrique sur les lèvres, qui devraient se toucher. (b) Correction du marquage géométrique des lèvres grâce au masquage des incisives pour l'indexation. 87

Figure 38 : (a) Cadre initial utilisé pour l'indexation commune des incisives et des lèvres. (b) Cadre utilisé pour l'indexation des lèvres : l'influence des incisives a été supprimée à l'aide d'un cache noir. 87

Figure 39 : Comparaison, suivant l'indexation, de l'évolution de l'erreur $Etot_1$ sur les 16 degrés de liberté des lèvres pour différents traitements postérieurs et 175 clefs. 89

Figure 40 : Evolution Log-Log de l'erreur $Etot_1$ sur 16 degrés de liberté des lèvres (sans filtrage, simple indexation) en fonction du nombre d'images clefs. 89

Figure 41 : Marquage des dents et de la mâchoire inférieure. 90

Figure 42 : Contour complet du conduit vocal pour une image de la séquence Laval43. 92

Figure 43 : Contours extraits par la méthode de Thimm et Luettin [TL99] sur une image de Laval43.. 95

Figure 44 : Mesure de distance *D* entre 2 splines. 96

Figure 45 : (a) Seuls 8 degrés de liberté (définis par les lignes horizontales et verticales) sont pris en compte pour la comparaison. (b) Comparaison sur une image test d'un marquage manuel de la langue avec 2 marquages estimés..... 97

Figure 46 : (a) Répartition de la différence *D* entre les 2 estimations, et définition d'un seuil de décrochage *D*>10. (b) Décrochage observé au milieu de la séquence considérée, avec les deux contours estimés à cet instant. 97

Figure 47 : Evolution de l'erreur $Etot_1$ pour notre méthode de marquage avec différents traitements postérieurs (indexation simple ou multiple, sans ou avec filtrage) sur les 8 degrés de liberté considérés. 98

Figure 48 : Définition d'une grille statique de référence, initialisée à partir des parties fixes du conduit vocal. 100

Figure 49 : (a) Correction d'orientation apportée à une des lignes de la grille (en trait fin). La ligne en pointillés permet de visualiser l'angle moyen de la ligne et de la langue. La ligne en gras est celle obtenue après rotation d'angle α de cette ligne, où $\alpha = \frac{1}{2}\left[angle_{ligne/\ palais} + (angle_{ligne/langue})_{moyen}\right]$. (b) Grille définissant 27 sections du pharynx aux incisives, après correction globale des orientations de chacune des lignes de façon à ce que chacune soit le plus possible orthogonale à la fois à la langue et à la ligne palais-pharynx. 101

Figure 50 : Grille corrigée en moyenne. Les mesures relatives aux dents avant et aux lèvres se font indépendamment de cette grille. Les points noirs correspondent aux intersections des lignes avec le contour de la langue et avec la ligne palais-pharynx. Deux configurations sont représentées : à gauche, la langue est en avant, à droite, la langue est en arrière et la distance au plancher sous-lingual est prise en compte pour la cavité avant. 102

Figure 51 : Procédure utilisée par Yehia [Yeh02] pour déterminer la distance sagittale et la longueur de la section pour une section donnée. Les sections de langue et de palais sont représentées en gras. 104

Figure 52 : Détermination des longueurs de sections et de distances sagittales par algorithme de centre de gravité. Si on considère la section p, G_p est son centre de gravité et la zone grisée son aire A_p. La longueur de la section L_p est représentée par la flèche et la distance médio-sagittale est égale à A_p/L_p. 105

Figure 53 : Exemple de représentation en escalier des distances sagittales en fonction de la section. 105

Figure 54 : Exemple de représentation lissée des distances sagittales en fonction de la section.. 106

Figure 55 : Exemple de représentation sous forme de tubes des distances sagittales en fonction de la section. 106

Figure 56 : Pour une configuration, représentation en escalier des distances sagittales en fonction de la distance par rapport au bas du pharynx. 107

Figure 57 : Pour une configuration, représentation des distances sagittales corrigées par la procédure de Yehia et comparaison suivant la grille de départ utilisée (initiale ou corrigée en moyenne). 107

Figure 58 : ACP sur les sections pour la séquence complète après correction par les procédures de (a) Beautemps et (b) Yehia. 109

Figure 59 : Coupes du conduit vocal réalisées dans un moulage de cadavre, à droite coupe sagittale, à gauche sections transversales (d'après Calliope, 1989). 110

Figure 60 : Représentation des différentes régions du conduit vocal sur une vue IRM de profil médio-sagittal pour une locutrice prononçant un [u], d'après [SLMD02]. 112

Figure 61 : Calque inversé d'une image radiographique de la séquence Laval43 avec les 26 sections du conduit vocal et les frontières (en pointillés) délimitant les régions définies par le modèle [SLMD02]. 113

Figure 62 : Points de référence Procuste pour comparer, en terme de rapport pixels-cms, les séquences (a) Wioland (image tournée de 90°) et (b) Laval43. 114

Figure 63 : Ligne médiane moyenne pour la séquence Laval43, prolongée jusqu'à l'estimation de la position de la glotte (point noir). 115

Figure 64 : Sections médio-sagittales et fonction d'aire calculée pour une image de Laval43. 116

Figure 65 : Banc de filtres 4 sous-bandes. 128

Figure 66 : Contour du conduit vocal obtenu en moyenne sur les exemplaires de chaque classe vocalique. Pour la langue, un écart-type est observé de part et d'autre du contour moyen - (a) [a] – (b) [ɔ] – (c) [o] – (d) [e] – (e) [ø] – (f) [ɛ] – (g) [i] – (h) [y] –(i) [u]. 131

Figure 67 : Fonction d'aire obtenue en moyenne sur les exemplaires de chaque classe vocalique (et écart-type en pointillés) - (a) [a] – (b) [ɔ] – (c) [o] – (d) [e] – (e) [ø] – (f) [ɛ] – (g) [i] – (h) [y] –(i) [u].133

Figure 68 : Représentation graphique de la position (X_h, Y_h) du point le plus élevé du dos de la langue et ellipses de dispersion (à un écart-type) des voyelles. 134

Figure 69 : Représentation de la position du point à 2 ddl à l'avant de la lèvre inférieure et du point à 2 ddl à l'avant de la lèvre supérieure et ellipses de dispersion (à un écart-type) pour quelques voyelles. 135

Figure 70 : Recherche du point constriction à partir de la fonction d'aire. 136

Figure 71 : (a) Espace de représentation X_c (lieu de constriction par rapport aux incisives) - A_c (aire de la constriction) et ellipses de dispersion des voyelles. (b) Espace de représentation X_c (lieu de constriction par rapport aux incisives) - A_l (aire aux lèvres) et ellipses de dispersion des voyelles. .. 136

Figure 72 : Cercles de corrélation mettant en évidence la distribution des degrés de liberté de la langue (en noir) et des lèvres (en gris) selon les 3 premières composantes principales de l'ACP. ... 138

Figure 73 : ACP sur les degrés de liberté des points marqués du conduit vocal pour la séquence complète (silence compris) et observation des voyelles – (a) 15 ddl de la langue : 2 premières composantes – (b) 15 ddl de la langue + 16 ddl des lèvres : composantes 1 et 3 – (c) Ellipses de dispersion pour les 15 ddl de la langue – (d) Positions extrêmes de la langue. 139

Figure 74 : Tables de confusion associées aux plans principaux obtenus à partir des données géométriques – (a) 15 ddl de la langue – (b) 15 ddl de la langue + 16 ddl des lèvres. 140

Figure 75 : (a) et (c) ACP sur les 28 distances sagittales du conduit vocal pour la séquence complète et observation des voyelles et ellipses de dispersion associées – mise en parallèle avec des allures de sections. (b) et (d) ACP sur les fonctions d'aire du conduit vocal pour la séquence complète et observation des voyelles et ellipses de dispersion associées – mise en parallèle avec des allures de fonction d'aire. 141

Figure 76 : Relation articulatoire-acoustique définie à partir (a) de la position de la langue (sans échelle) et (b) des formants. 144

Figure 77 : Séquence de formants F_1, F_2, F_3 estimés à partir du modèle linéaire (en noir) superposés aux formants mesurés depuis le signal d'origine et filtrés (en gris). 147

Figure 78 : Espaces (a) F_1-F_2 et (b) F_2-F_3 obtenus à partir données mesurées (en gris) et données estimées (en noir). 148

Figure 79 : A partir des distributions dans l'espace formantique F_1-F_2, 95% de chacun des deux jeux de données sont conservés pour établir les contours. (a) Données acoustiques normalisées. (b) Histogramme de distribution des données acoustiques estimées. (c) Histogramme de distribution des données acoustiques d'origine. 151

Figure 80 : Le contour (en trait gris plein) pour les paramètres audio estimés est dilaté et orienté (grâce à la transformation *M*) pour s'ajuster au contour des paramètres audio mesurés (en trait noir plein). Le contour obtenu est représenté en pointillés. 152

Figure 81 : Espace F_1-F_2 après transformation affine. La couverture de l'espace formantique est meilleure. 152

Figure 82 : Espace F_2-F_3 après transformation affine. 153

Figure 83 : Signaux synthétisés et spectrogrammes ([a], [i], [ɛ], de gauche à droite), pour les formants mesurés (en haut), les formants estimés (au milieu), et les formants estimés puis « transformés » (en bas). 153

Figure 84 : Matrices de confusion obtenues à partir des réponses des sujets pour les 4 modèles de signaux testés. 155

Figure 85 : (a) Résultats du test en terme de taux de reconnaissance entre voyelles périphériques et voyelles centrales. (b) Tables de confusions entre voyelles centrales et périphériques. 156

Figure 86 : Comparaisons sur des triangles vocaliques entre stimuli et perception pour les différentes catégories de voyelles (10 items pour chaque) : les voyelles périphériques ([a], [i], [u], [y]) sont en trait gras. 157

Figure 87 : Classes vocaliques dans l'espace F_1-F_2 pour les données d'origine, les données estimées par modèle linéaire et les données estimées puis soumises à transformation affine. (a) méthode globale – (b) méthode par classe. 159

Figure 88 : (a) Signal audio Laval43 original - (b) Signal audio Laval43 après débruitage - (c) Spectrogramme du signal original – (d) Spectrogramme du signal après débruitage. 162

Figure 89 : Fonction d'aire et fonction de transfert simulée d'un [a]. 165

Figure 90 : (a) Triangle vocalique de la séquence Laval43 à partir des données d'origine : les points gris foncé sont des instants de silence. (b) Zoom sur les trames de parole et les voyelles sélectionnées. 166

Figure 91 : (a) Triangle vocalique de la séquence Laval43 à partir des formants estimés des fonctions d'aire : les points gris foncé sont des instants de silence. (b) Zoom sur les trames de parole et les voyelles sélectionnées pour ces mêmes données estimées. 167

Figure 92 : F_1 et F_2 pour les trames étiquetées de voyelles. Le tracé noir correspond aux valeurs d'origine et le gris aux fréquences estimées via les fonctions d'aire. 169

Figure 93 : Biais entre formants estimés et formants d'origine pour les 3 premiers formants, pour la séquence hors moments de silence et pour différentes valeurs de paramètres : (a) rapport pixels/cm - (b) zoom sur la moyenne des 3 formants - (c) position de la glotte - (d) paramètres α-β. 171

Figure 94 : Comparaison des valeurs moyennes des 3 premiers formants, pour chaque classe vocalique orale du corpus, entre les données d'origine et celles estimées. Mise en parallèle avec des valeurs de référence de ces mêmes formants [Cal89]. 173

Figure 95 : Biais, pour les 3 premiers formants et pour les 9 classes vocaliques, entre les formants extraits de Praat et une référence [Cal89] et entre les formants estimés et cette même référence... 174

Figure 96 : (a) Plan F_1-F_2 des données extraites du signal audio et ellipses de dispersion (un écart-type) des voyelles sélectionnées. (b) Plan F_1-F_2 des données estimées et ellipses de dispersion (un écart-type) des voyelles sélectionnées. 176

Figure 97 : (a) Plan F_1-F_2 des données extraites du signal audio et diagramme de Voronoï des classes vocaliques. (b) Plan F_1-F_2 des données estimées et diagramme de Voronoï des classes vocaliques. 176

Figure 98 : (a) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 d'origine et sur les distances aux valeurs moyennes des classes vocaliques. (b) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 estimés à partir des fonctions d'aire et sur les distances aux valeurs moyennes des classes vocaliques. 178

Figure 99 : (a) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 d'origine et sur *k*-voisins dans l'espace vocalique. (b) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 estimés à partir des fonctions d'aire et sur *k*-voisins dans l'espace vocalique. 178

Figure 100 : Plan F2-F3 pour les trames de parole de Laval43 à partir des données (a) d'origine et (b) estimées. 180

Figure 101 : Schéma de blanchiment de la source. 182

Figure 102 : Spectrogramme d'un segment de la source obtenue par filtrage de Hilbert et filtrage inverse LPC. La structure harmonique est préservée dans les périodes voisées. 183

Figure 103 : Banc de filtres pour le calcul des amplitudes en 2 sous-bandes..... 184

Figure 104 : Schéma-Bloc de synthèse des signaux..... 186

Figure 105 : Spectres LPC du signal original (trait gras) et du signal synthétisé à partir de la fonction de transfert estimée (trait fin) : les 2 spectres sont assez similaires jusqu'à 3KHz puis plus du tout ensuite. 187

Figure 106 : Comparaison de spectres LPC (on se limite à la bande [0-3.5] Hz). (a) Les 2 signaux ont été synthétisés à partir de la source blanchie par filtrage par la fonction de transfert du conduit vocal, le signal (en gras) a été modulé en amplitude en 2 sous-bandes, pas l'autre. (b) Les 2 signaux ont été synthétisés à partir de la source blanchie modulée en amplitude par filtrage par la fonction de transfert du conduit vocal (en trait fin) ou par filtrage par le spectre LPC du signal d'origine (en trait gras). ... 188

Figure 107 : Distances spectrales, sur la séquence complète pour différentes valeurs de paramètres, entre le signal d'origine et le signal synthétisé (à partir d'une source blanchie et modulée en amplitude en 2 sous-bandes, et filtrée par la fonction de transfert du conduit vocal) : (a) rapport pixels/cms – (b) position de la glotte – (c) paramètres α-β..... 189

Figure 108 : Signal d'origine, spectrogramme, spectrogramme LPC et formants estimés (en traits noirs épais) à partir des fonctions d'aire extraites..... 192

Figure 109 : Le spectre LPC (trait gras) obtenu à partir du signal original est comparé au spectre LPC du signal synthétisé depuis la fonction d'aire (trait fin). Le signal de parole, les fonctions d'aire et les profils de conduit vocal sont également donnés, à titre d'exemple, pour 4 trames étiquetées de voyelles du corpus, de gauche à droite [a], [i], [u] et [ø]. 192

Figure 110 : Niveaux d'intelligibilité (MOS) moyens pour les 5 sujets pour les différents types de signaux testés et écart-type suivant les sujets. 196

Figure 111 : Sections (en blanc) prises en compte pour le calcul du minimum de constriction entre la langue et le palais..... 204

Figure 112 : Section supplémentaire (en trait gras et blanc) prise en compte pour le calcul du minimum de constriction entre la pointe de la langue et le palais. Cette mesure supplémentaire permet d'affiner la mesure de la taille de constriction entre la pointe et le palais. 205

Figure 113 : (a) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes dorsales du corpus Laval43. (b) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes alvéolaires du corpus Laval43. 206

Figure 114 : (a) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes dorsales du corpus Laval43 après fenêtrage. (b) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes alvéolaires du corpus Laval43 après fenêtrage. 207

Figure 115 : (a) Répartition des consonnes dorsales en fonction de la taille de constriction, avant et après fenêtrage. (b) Répartition des consonnes alvéolaires en fonction de la taille de constriction. . 207

Figure 116 : (a) Distribution des tailles de constriction suivant les modes d'articulation, pour les consonnes alvéolaires. Les courbes ont été normalisées par le nombre de consonnes de chaque catégorie. (b) Tailles de constriction suivant les 8 consonnes alvéolaires..... 209

Figure 117 : (a) Exemple de configuration d'un [t]. (b) Exemple de configuration d'un [k]..... 209

Figure 118 : Mesure de l'écart entre les lèvres supérieure et inférieure pour une image de la séquence Laval43. 210

Figure 119 : (a) Ecart mesuré entre les lèvres inférieure et supérieure, à partir du marquage géométrique extrait pour la séquence Laval43 et mise en évidence des instants de production des consonnes bilabiales étiquetées du corpus. (b) Répartition des trames en fonction de l'écart aux lèvres et mise en évidence d'un minimum d'ouverture pour les bilabiales. 211

Figure 120 : Aérogramme pour la phrase du corpus « Le léopard réintègre sa cage ». Les constrictions sont représentées par les tâches blanches. 212

Figure 121 : Image de la séquence Wioland sur laquelle le contour du voile du palais, en position haute, n'est pas net. 213

Figure 122 : Marquage manuel de points d'appui anatomiques sur le voile du palais, d'après [SFK80]. 214

Figure 123 : Représentation graphique de la position moyenne de la partie supérieure du vélum (moyenne des 3 ddl indiqués par les étoiles) et distinction entre voyelles orales (en gris) et voyelles nasales (en noir). 216

Figure 124 : Mesure de l'écart entre la paroi pharyngale et le voile du palais pour une image de la séquence Laval43. 216

Figure 125 : (a) Ecart minimal entre le pharynx et le voile du palais pour la séquence complète Laval43. Les consonnes nasales sont représentées par les points noirs, les voyelles nasales par les croix rouges. (b) Ecart minimal entre le pharynx et le voile du palais pour la séquence Laval43 privée des moments de silence. 217

Figure 126 : (a) Répartition des trames en fonction de l'écart minimal entre le pharynx et le vélum pour la séquence, avec ou sans silence. (b) Répartition de l'écart minimal entre le pharynx et le vélum pour les voyelles nasales et orales du corpus Laval43. Les ordonnées ont été normalisées par rapport au nombre de voyelles. 218

Figure 127 : Exemple de cliché radiologique de la séquence Wioland, avec annotations des articulateurs. 252

Figure 128 : Interface Matlab de marquage géométrique manuel d'images clefs. Exemple avec le marquage du contour de la langue de Laval43. 260

Figure 129 : (a) Aucun point de cette image n'a été marqué pour l'instant. (b) Le point 7 a été marqué à l'aide du bouton 7. Une ligne bleue apparaît, elle fixe une des coordonnées (ici Y). Le point est à marquer à l'intersection entre la ligne et le contour de la langue. Une fois marqué, le point est sauvegardé. On peut utiliser le slider pour vérifier le marquage des points. (c) Pour cette image, tous les points ont été marqués. Dès que les points sont tous marqués, un contour est tracé et relie les points entre eux. (d) Le bouton spline sert à lisser le contour obtenu. On utilise des interpolations par des polynômes. 261

Figure 130 : Interface Matlab de marquage géométrique manuel d'images clefs. Exemple avec le marquage du contour du vélum de Laval43. 263

Figure 131 : Interface Matlab d'étiquetage audio. 265

Figure 132 : (a) Croquis de la vue de profil de référence utilisée par Wioland pour l'établissement de sa grille de mesure. (b) Circonférence superposée à une image de la séquence pour l'évaluation du rapport pixels/cms. 267

Table des tables

Table 1 : Construction de modèles combinant les différentes améliorations possibles de la méthode de rétro-marquage. 70

Table 2 : Biais et écart-type de l'erreur (sur 100 images clefs) entre marquage manuel et marquage estimé par degré de liberté. 73

Table 3 : Erreur $Etot_2$ du contour de la langue à partir de jeux d'images clefs différents et de marques manuelles obtenues de 2 experts. 74

Table 4 : Paramètres de la méthode de rétro-marquage réglés pour l'extraction géométrique de divers articulateurs dans la séquence Laval43. 82

Table 5 : Résultats de l'évaluation $Etot_1$ pour différents articulateurs (pour rappel, les grandes images sont de taille 720*480 pixels). 89

Table 6 : Nombre et ordre des polynômes d'interpolation pour les différents articulateurs. 91

Table 7 : Paramètres α et β obtenus par Soquet et al. [SLMD02] pour les 8 régions définies du conduit vocal et les 2 locuteurs. 113

Table 8 : Voyelles sélectionnées dans le corpus de Laval43. 129

Table 9 : Voyelles et signes graphiques associés. 133

Table 10 : Corrélations entre formants d'origine et formants prédits pour la séquence (hors moments de silence) en fonction des degrés de liberté pris en compte. 147

Table 11 : Coefficients de corrélation entre formants d'origine et formants estimés 169

Table 12 : Différence moyenne et écart-type des formants estimés par rapport aux formants d'origine. 172

Table 13 : Comparaison des valeurs moyennes des 3 premiers formants, pour chaque classe vocalique orale du corpus, entre les données d'origine et celles estimées. Mise en parallèle avec des valeurs de référence de ces mêmes formants [Cal89]. 173

Table 14 : Seuils de perception pour différentes valeurs de fréquences de formants F_1 et F_2 d'après Flanagan (1955). 175

Table 15 : Coefficient de corrélation entre l'amplitude du signal original et celle du signal synthétisé, sans et avec modulation d'amplitude en 2 sous-bandes. 185

Table 16 : Distances spectrales *D* entre spectres LPC sur la séquence complète. 191

Table 17 : Coefficients de corrélation entre spectres LPC synthétisés et d'origine. 191

Table 18 : Echelle utilisée pour évaluer le niveau d'intelligibilité des stimuli. 194

Table 19 : Corpus de phrases utilisées pour le test d'intelligibilité. 194

Table 20 : Signaux considérés pour le test d'intelligibilité. 195

Table 21 : ANOVA réalisée sur les réponses au test d'intelligibilité. Les facteurs pris en compte sont les phrases, les sujets, le type de source (bruit blanc ou source blanchie) et la présence ou non de la modulation d'amplitude. 197

Table 22 : ANOVA réalisée sur les réponses au test d'intelligibilité. Les facteurs pris en compte sont les phrases, les sujets et le filtrage. 198

Table 23 : Consonnes sélectionnées dans le corpus de Laval43 et classées par lieux et modes d'articulation. 202

INTRODUCTION

Production de la parole et modélisation

La parole se distingue des autres sons par des caractéristiques acoustiques qui ont leur origine dans les mécanismes de production. Les sons de parole sont produits à partir d'une source, soit par des vibrations des cordes vocales, soit par une turbulence créée par l'air qui s'écoule rapidement dans une constriction du conduit vocal ou lors du relâchement d'une occlusion de ce conduit.

Pour chaque position articulatoire, les spectres des sons produits par la source sont modifiés par les propriétés résonantes particulières du conduit vocal ; le mouvement des articulateurs est soumis à des contraintes physiques dues à l'anatomie.

Le processus de production de parole est complexe (c'est l'un des actes volontaires les plus complexes de l'activité humaine) et nécessite pour être décrit d'avoir recours à la modélisation, c'est-à-dire à une substitution par un processus physique plus facile à maîtriser et à faire évoluer.

Les modélisations ont des finalités différentes, qu'on peut classer en deux classes de recherches :

1/ celles qui tentent de reproduire le signal de parole sans chercher à décrire ou imiter les mécanismes naturels de production de la parole (modèle fonctionnel basé sur des techniques de traitement du signal), comme par exemple la synthèse à formants, la prédiction linéaire, …

2/ celles qui, au contraire, tout en visant une reproduction aussi fidèle que possible du signal de parole, cherchent à connaître les mécanismes naturels qui permettent cette production (modèle physique) : forme 3D, aéro-acoustique, commandes musculaires…

La compréhension des processus de la communication parlée est un objectif majeur, aussi bien pour approfondir les connaissances dans le domaine de la parole que pour parvenir à des systèmes de dialogue vocal ambitieux et réellement utilisables.

Dans les travaux présentés ici, nous nous plaçons en amont de la seconde finalité, dans un plan intermédiaire, en nous intéressant à la modélisation du conduit vocal à partir de données issues d'une observation directe de séquences cinéradiographiques existantes. Un des résultats attendu de la thèse est de savoir si un modèle 2D+t est suffisant dans un contexte de parole continue. Nous négligeons les détails des mécanismes naturels et sortons des conditions idéales habituellement considérées (contexte de voyelles tenues ou de logatomes).

L'étude de la production de la parole et des gestes articulatoires a conduit à l'élaboration de modèles numériques des organes impliqués dans la production de la parole (système respiratoire, cordes vocales, larynx, pharynx, vélum, langue, mâchoire et lèvres) et des déformations du conduit vocal sous l'effet des articulateurs.
Pour ce dernier point, les modèles articulatoires utilisés sont souvent le résultat d'analyses statistiques de films cinéradiographiques qui décrivent la forme du conduit vocal en fonction d'un nombre réduit de paramètres ([Fan60], [Mer73], [Mae79], …). Ces modèles complétés par une simulation acoustique permettent de passer de l'espace des commandes articulatoires à l'espace acoustique.

Pourquoi des données articulatoires ?

La parole est un moyen de communication qui est basé sur la perception et la compréhension par un auditeur du signal sonore produit par un locuteur. Ce signal doit donc être adapté à la perception et contenir des traits invariants par rapport aux caractéristiques phonétiques. Mais, si psychologiquement la parole apparaît comme une succession d'unités discrètes (Théorie Quantique de Stevens [Ste72, Ste89]), physiquement, elle est produite par un flux d'air continu expulsé par les conduits nasal et vocal. Il semble donc judicieux de représenter la parole par un flot continu de mouvements articulatoires. La recherche d'invariants acoustiques permettant de classifier les unités discrètes laisse alors place à une recherche des causes de la variabilité (comme les travaux de Maeda sur le problème de la variabilité articulatoire et acoustique [Mae91]).
Le signal acoustique est le résultat du travail multidimensionnel des articulateurs du conduit vocal. Les observations articulatoires constituent un apport considérable d'informations qui ne peuvent être facilement dérivés des signaux acoustiques.
En partant des données articulatoires que sont la forme et les mouvements du conduit vocal, les chercheurs espèrent arriver à une meilleure compréhension de ce qu'est la parole.
Des efforts intenses ont été fournis pour mettre en œuvre des techniques d'enregistrement de données articulatoires, souvent difficiles à obtenir, en particulier celles sur le contour des organes qui ne sont pas directement accessibles.

Pourquoi la cinéradiographie ?

La radiographie a été pendant longtemps l'une des principales techniques d'acquisition de données articulatoires en offrant la possibilité d'obtenir une vue sagittale complète des articulateurs du conduit vocal, de la glotte jusqu'aux lèvres. Devenue dynamique à la fin des années 1950, sous le terme de cinéradiographie, elle permet l'observation des mouvements des articulateurs de la parole avec une résolution temporelle importante, de l'ordre de 50-60 images par seconde. Depuis quelques années, pour des questions de déontologie, il est de

plus en plus difficile d'enregistrer de nouveaux films radiologiques du conduit vocal à cause de la trop forte radiation à laquelle sont exposés les sujets sains. Les films existants ont besoin d'être préservés, conservés et utilisés avec précaution. La cinéradiographie ayant fait la preuve de son utilité pour la recherche scientifique, il est nécessaire de continuer à utiliser les données existantes.

C'est dans ce contexte que Munhall et al. [MVT95] ont réalisé la base ATR « X-ray film database for Speech Research », à partir de films enregistrés par Rochette (Université Laval), et Stevens et Perkell (M.I.T.). Cette base de données, dont nous reparlerons, comprend 25 films radiographiques totalisant 55 minutes et près de 100000 images. Soutenus par le programme « Ingénierie des Langues » du CNRS, l'Institut de Phonétique de Strasbourg et l'Institut de la Communication Parlée de Grenoble ont aussi élaboré une base de données cinéradiographiques du français [ABB+00] incluant, parmi 4 films d'une durée de quelques minutes, les séquences Wioland, Flament et Zerling.

Les objectifs de ces projets sont multiples :

(1) assurer la sauvegarde de ces données par numérisation et par stockage sur des supports vidéo de haute qualité,

(2) faciliter l'accès et le traitement de ces données par leur intégration dans des bases de données,

(3) apporter une valeur supplémentaire à ces données avec des tracés sagittaux réalisés par des experts phonéticiens et montrant les limites du conduit vocal,

(4) développer de nouvelles techniques de traitement automatique de ces images,

(5) assurer une accessibilité et une diffusion de ces données sur le web.

Dans le cadre de nos travaux, nous avons pu bénéficier de longues séquences cinéradiographiques extraites de ces bases de données. Ces séquences enregistrées dans les années 1970 et numérisées récemment sont une mine d'informations à exploiter grâce aux capacités de traitement apparues depuis.

Nous utilisons ces données dynamiques dans le but d'obtenir des données de type 2D en fonction du temps pour l'ensemble du conduit vocal et sur de longues séquences de parole naturelle.

Extraction des données : vers une méthode quasi-automatique

L'extraction de données géométriques à partir de films radiologiques est généralement réalisée manuellement ([BSWZ86], [Mae79], [Bad91]), les configurations géométriques du conduit vocal sont extraites image par image par des tracés manuels. Mais on doit ici faire

face à de grandes quantités de données pour traiter la moindre séquence et le développement de traitements automatiques semble nécessaire.
En 1994, Tiede et Bateson [TV94] ont présenté des pistes pour automatiser le traitement des images radiologiques pour exploiter au mieux ces grandes bases cinéradiographiques. Une méthode d'extraction automatique des contours de la langue a été proposée par Laprie et Berger [LB96]. Cette méthode concerne le traitement d'images statiques. Plus tard, Thimm et Luettin [TL99] de l'IDIAP se sont intéressés au traitement de séquences radiologiques et ont abouti au traitement complet d'un film issu de la base ATR, la séquence Laval43. Cependant, la qualité des contours extraits avec cette méthode est nettement moins précise que celle obtenue manuellement. Cette qualité est pourtant nécessaire pour utiliser ces données géométriques dans des applications telles que la synthèse articulatoire.
En vue d'améliorer cette situation, nous proposons de ré-introduire une part d'expertise humaine avec une étape limitée de marquage manuel par un expert humain. Nous avons mis en place une méthode semi-automatique applicable film par film. Elle combine un marquage manuel et une reconstruction automatique du mouvement, basée sur la forte redondance temporelle des mouvements de parole, du fait de la pseudo-périodicité des gestes articulatoires du conduit vocal.

Ce contexte d'informations dynamiques et redondantes est adapté au développement de nouveaux algorithmes de capture de mouvements biologiques. Ceci est aussi fortement favorisé par les progrès de ces dernières années en traitement vidéo. Nous disposons d'un contexte riche en applications, de très nombreux algorithmes ont été développés ces dernières années pour analyser et tirer profit de l'immense quantité de données véhiculées par les images et la vidéo. Dans la plupart de ces travaux, la capture du mouvement est réalisée avec des marqueurs (pastilles, billes, maquillage bleu…). Pour éviter l'usage de ces marqueurs, un algorithme appelé rétro-marquage a été proposé [Ber04], dont le principe est d'effectuer l'équivalent de ces marquages sur un nombre limité d'images clefs et d'associer ensuite ces marquages aux images de la séquence d'origine. Cette association est réalisée par l'intermédiaire de paramètres vidéos qui sont supposés être en correspondance avec les marques géométriques. Ces marques peuvent être des contours ou des points caractéristiques des contours. La correspondance entre les paramètres vidéo et les marques dépend des contrastes et des structures basses fréquences présentes dans l'image. Initialement, l'algorithme a été appliqué sur les lèvres à partir d'une base de données vidéo bien cadrée [HBSK00] et sans marqueurs.
Les images clefs étaient en RGB et elles étaient calculées avec l'algorithme SOM (Self Organizing Map). Les paramètres vidéo étaient les 24*12 premiers coefficients de la

Transformée en Cosinus Discrète (DCT en anglais) et les marques géométriques 8 points permettant de définir les paramètres d'ouverture ABS des lèvres.

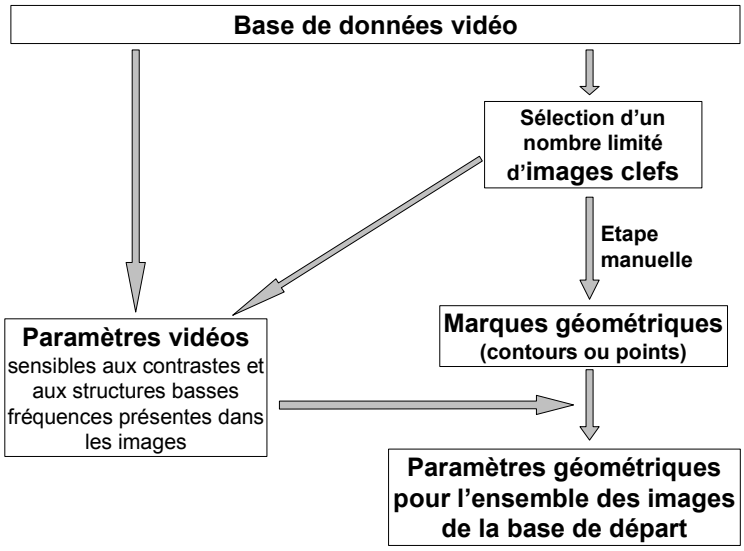


Figure 0 : Principe général de l'algorithme de rétro-marquage [Ber04].

La précision du cadrage est d'une grande importance pour l'utilisation du rétro-marquage et cette condition est généralement bien remplie par les enregistrements cinéradiographiques. C'est de là que vient l'idée d'adapter cet algorithme à des séquences d'images rayons X.

La méthode que nous proposons dans cette thèse est une adaptation de l'algorithme de rétro-marquage à des séquences cinéradiographiques. En effet, il s'agit d'un cadre idéal pour l'application de cet algorithme puisque la plupart des séquences dont nous disposons ont été réalisées sans marqueur et il y a très peu de points d'appui anatomiques utilisables sur les images. Nous cherchons donc à associer des paramètres implicites et extraits du signal vidéo à des paramètres géométriques contrôlés et définis a posteriori, plutôt que d'extraire directement des données géométriques.

La méthode proposée a été mise en place à partir de la séquence « Wioland », extraite de la base de données française [ABB+00] et obtenue grâce à Pascal Perrier sous forme de séquences QuickTime. Elle a été au départ développée pour extraire les mouvements de la langue.

Elle se décompose en 3 étapes :

(1) le traitement manuel d'un nombre restreint d'images clefs qui permet de définir des paramètres géométriques (initialement le contour de la langue),

(2) une étape automatique d'indexation de la base à partir de ces mêmes images clefs réduites et cadrées, qui a pour but d'associer à chacune des images de la base le marquage géométrique,

(3) des traitements postérieurs de régularisation pour améliorer la méthode et la rendre équivalente à un tracé manuel des images.

Le principe de la méthode sera décrit en détail au chapitre 1 pour le marquage des configurations géométriques de la langue et l'évaluation de la méthode sera présentée au chapitre 2.
On montrera ensuite (chapitre 3) que la méthode peut être étendue à d'autres parties du conduit vocal (lèvres, mâchoire, larynx, vélum…) et à d'autres séquences cinéradiographiques, et notamment Laval43. Cette séquence est extraite de la base de données cinéradiographique d'ATR « X-ray film database for Speech Research » [MVT95], dont nous avons pu bénéficier grâce à Kevin Munhall, sous forme d'un DVD.
Pour une séquence donnée, les articulateurs sont marqués un à un de façon indépendante en utilisant la méthode avec des paramètres adaptés et spécifiques à chacun. Le contour complet du conduit vocal est ensuite reconstruit en combinant ces marquages indépendants.

De l'articulatoire à l'acoustique

La description ainsi réalisée de la morphologie du conduit vocal nécessite d'être validée avant de pouvoir être exploitée dans l'étude approfondie des relations articulatori-acoustiques. Cette validation sera l'objet des chapitres suivants.
Aussi, dans l'optique de mettre en correspondance nos données articulatoires extraites avec les données acoustiques et d'effectuer, en particulier, de la synthèse articulatoire, nous terminerons la première partie de ce manuscrit (chapitre 4) avec la mesure de sections dans le plan sagittal et le calcul des fonctions d'aire [HS65], à partir des contours géométriques extraits semi-automatiquement de la séquence Laval43 et d'une grille élaborée à cet effet.

Disposant ainsi d'une description géométrique « classique » du conduit vocal, extraite depuis la cinéradiographie, la seconde partie de cette thèse s'attache à analyser ces données articulatoires estimées et à les mettre en correspondance avec les données acoustiques.
Dans un premier temps, au chapitre 5, les données articulatoires sont analysées en regard avec des données phonétiques extraites du corpus, dans le but de retrouver des résultats classiques en phonétique. L'analyse, principalement orientée vers les voyelles, se situe à 2 niveaux : géométrique (observations directes des configurations estimées) puis paramétrique. Cette seconde analyse permet l'observation des contours extraits dans des espaces de dimensions réduites. On s'intéressera à des représentations classiquement utilisées (lieu et taille de constriction par exemple) ainsi qu'à des observations après analyse en composantes principales.

Tout en étant conscients de la non-linéarité de la relation entre l'articulatoire et l'acoustique de la parole, un certain nombre d'auteurs dont Yehia et al. [YRV98] ont montré que des modèles linéaires d'association entre des données audio et vidéo caractéristiques de la parole permettent de capturer une grande part de l'information, à condition d'adopter les bonnes représentations. Nous nous sommes intéressés à la possibilité, grâce à une transformation linéaire, de prédire de l'audio à partir des contours géométriques extraits de la séquence cinéradiographique Wioland. Ceci sera l'objet du chapitre 6.

Elément indispensable à un modèle de production de parole, la fonction d'aire permet de faire le lien entre les données articulatoires et acoustiques, via l'utilisation d'un modèle acoustique. Si l'on considère le conduit vocal comme un tube uniforme (ou une concaténation de tubes uniformes), l'analogie électrique-acoustique permet d'effectuer la simulation d'un signal acoustique. Le chapitre 7, intitulé synthèse articulatoire, présente les résultats de ces simulations d'abord en terme d'analyse formantique. Dans un deuxième temps, des signaux de parole ont été synthétisés grâce à une estimation complémentaire et nécessaire de la source et de l'amplitude en deux sous-bandes.

Ces chapitres concentrent leur analyse sur la séquence complète et plus spécifiquement sur les voyelles orales des corpus. Les deux derniers chapitres de cette thèse étudient deux autres aspects de ces corpus : les consonnes de Laval43 au chapitre 8 et les voyelles et consonnes nasales de cette même séquence au chapitre 9. Ces analyses ne sont qu'un bref aperçu des possibilités que peut apporter une méthode quasi-automatique d'extraction de données cinéradiographiques appliquée à de longues séquences de parole naturelle.

L'IMAGERIE AU SERVICE DE LA MODELISATION ARTICULATOIRE

La production de parole fait intervenir des mouvements de la mâchoire, des lèvres qui sont immédiatement visibles, mais également des mouvements d'organes sous-jacents tels que le larynx, le vélum ou la langue. La capture d'informations provenant des articulateurs visibles est de fait plus aisée que celles des autres.
Cependant, face à la nécessité de disposer de données articulatoires du conduit vocal pour comprendre le phénomène de production de la parole, l'extraction des formes et mouvements des articulateurs non visibles a toujours été l'objet de nombreuses études.
D'abord, et de front avec la recherche médicale, les techniques d'imagerie sont utilisées sous leurs diverses formes, pour obtenir des images ou films du conduit vocal. Ensuite ces supports visuels enregistrés nécessitent une exploitation faisant appel à des techniques ou algorithmes adaptés. Enfin et grâce à ces données articulatoires, la complexité du processus de production de parole peut être décrite grâce à la modélisation. Au cours des dernières décennies, nombre de modèles ont vu le jour.

1. Techniques d'imagerie et d'extraction de données

1.1. Techniques d'acquisition de données articulatoires

Les techniques d'acquisition de données articulatoires sont nombreuses et ont fait l'objet d'efforts de développement et de diversification très intenses dans la communauté internationale au cours de ces 15 dernières années.

Jusqu'au début des années 70, la palatographie et la radiographie ont été les principales techniques d'acquisition de données articulatoires. La palatographie, simple et peu onéreuse, a l'inconvénient de limiter la zone d'observation à la zone de contact entre la langue et le palais. Au contraire, la radiographie offre la possibilité d'obtenir une vue sagittale complète des articulateurs du conduit vocal, de la glotte jusqu'aux lèvres.
Cette technique utilise un faisceau de rayons X qui émerge uniformément d'un générateur dans un large angle solide. Les rayons X sont une forme de rayonnement électromagnétique à haute fréquence dont la longueur d'onde est comprise approximativement entre 5 picomètres et 10 nanomètres. C'est un rayonnement ionisant. En imagerie médicale, les rayons X sont émis par une source fixe, ils sont regroupés en un faisceau de photons (grain de lumière) qui traversent le corps humain.

La radiographie est réalisée sur film, le film étant disposé dans une cassette protectrice derrière ou sous le corps exposé. L'image est créée par la différence d'opacité des tissus aux rayons X. En effet, le corps est composé de tissus dits "mous", peu opaques aux rayons X (comme la peau, la graisse, les muscles), et de tissus plus opaques (les os, essentiellement). Les différentes parties du corps réagissent donc différemment aux rayons X, qui sont plus ou moins absorbés. Après avoir traversé les tissus composant le corps, les rayons X arrivent sur le film sensible (plus précisément photosensible) situé de l'autre côté du patient. À cet instant précis, les rayons X laissent une trace plus ou moins grise selon la densité des différents organes traversés. Les os absorbent plus les rayons X que les parties molles, c'est la raison pour laquelle ils apparaissent plus opaques et d'une tonalité plus blanche que les autres tissus de l'organisme.
En outre, dès la fin des années 1950, la radiographie devient dynamique, sous le terme de cinéradiographie, ce qui permet l'observation des mouvements des articulateurs de la parole.

Cette technique a permis pendant des années un grand nombre des recherches de référence en phonétique, depuis les premiers travaux de Chiba et Kajiyama [CK41]. A titre d'exemple, la théorie acoustique de Fant en 1960 [Fan60] repose essentiellement sur des analyses de données cinéradiographiques. De la même façon, on peut aussi citer Perkell [Per69] et Wood [Woo79] et les travaux de modélisation articulatoire de Mermelstein [Mer73] et Maeda [Mae79]. Cette technique par rayons-X a été utilisée avec succès dans de très nombreuses études de parole, parmi lesquelles, par exemple, certains travaux de l'Institut de Phonétique de Strasbourg [BSWZ86]. Une bibliographie [Dar87] regroupant plus de 280 sources cinéradiographiques a été réalisée en 1987.
Encore aujourd'hui, les travaux basés sur les rayons X et la cinéradiographie sont nombreux. En enregistrant de nouvelles données dans des conditions médicales surveillées, l'Institut de Phonétique de Strasbourg utilise largement cette technique, afin d'étudier les phénomènes d'anticipation ou de coarticulation ([VSR+03], [CBRH03], [AVFG03]). L'enregistrement de films de grande qualité a donné lieu à des modélisations articulatoires, comme les modèles Bergame [BGB+95] ou Gentiane [VAB98], à partir duquel il est désormais possible d'acquérir des données, sans exposer des sujets aux rayons X.

Le problème de la cinéradiographie est en effet d'exposer des locuteurs sains à une dose de radiation importante pendant une longue durée. Aussi même les mouvements correspondants aux voyelles stationnaires sont difficiles à enregistrer. En raison de l'effet dangereux du rayonnement, la quantité de données enregistrées est réduite et il est de plus en plus difficile d'enregistrer de nouveaux films radiologiques du conduit vocal. Devant la nécessité de préserver et conserver les films existants, au même titre que celle de continuer

à utiliser ces données précieuses, de grandes bases de données cinéradiographiques ont été réalisées ([MVT95], [ABB+00]).

Les raisons d'éthique mais aussi des questions plus techniques comme la résolution temporelle ont poussé à s'intéresser à de nouvelles voies d'investigation expérimentale en matière d'acquisition de données articulatoires. En effet, la résolution temporelle de la cinéradiographie ne dépasse pas 50-60 images par seconde. Cette fréquence d'échantillonnage de 50-60 Hz n'est pas toujours suffisante pour une étude fine de l'organisation temporelle des consonnes, en particulier pour les consonnes plosives alvéolaires dont la durée est de l'ordre de 15-20 ms.
Nous dressons ici un état de l'art de techniques d'imagerie disponibles.

Le **système Xray microbeam** ou micro-faisceau de rayons X permet de suivre les trajectoires de quelques pastilles d'or fixées et minimise l'exposition des sujets aux rayons ionisants en concentrant des faisceaux sur des points très localisés du conduit vocal. Le faisceau de rayons X est extrêmement fin. Un calculateur contrôle d'une manière adaptative la direction du rayon X sur la position cible du capteur collé sur la surface des organes articulatoires, par exemple, la langue. La position du capteur est identifiée et calculée en temps réel. L'analyse trame par trame des données utilisée dans la méthode traditionnelle des rayons X est limitée à celle de la position du capteur. Pendant l'acquisition des données, l'émission automatique des rayons X n'est effectuée que lorsque l'identification de la position du capteur est nécessaire. La dose de rayonnement est extrêmement faible en comparaison à celle produites par les autres systèmes à rayons X. Ce système à été implémenté au début des années 1970 et utilisé dans différentes études ([KIF75], [Sto90], [Wes91]). Cette technique utilise des marqueurs géométriques ponctuels et a une très bonne résolution temporelle : par exemple, les positions de 10 pastilles peuvent être enregistrées 100 fois par seconde. Mais certains articulateurs ou certaines zones du conduit vocal ne sont pas complètement accessibles.

L'**électropalatographie** (EPG) permet d'enregistrer dynamiquement les points de contact entre la langue et le palais pendant la phonation, cette technique utilise un palais artificiel avec des contacts électriques. Elle permet des mesures de contacts palataux entre 100 et 200 Hz. Les électrodes sont activées dès que la langue touche le palais. Les modèles de contact linguopalatal sont inférés à partir des signaux enregistrés par quelques dizaines d'électrodes. L'avantage principal de la palatographie est de donner facilement en temps réel des informations sur le lieu et le mode d'articulation. Son inconvénient majeur est que l'information est binaire d'une électrode à une autre, c'est-à-dire que cela indique simplement

si oui ou non la langue touche l'électrode. L'électropalatographie donne de bonnes informations temporelles mais avec une résolution spatiale assez grossière. Elle donne cependant accès à la 3ème dimension spatiale transversale.
Les mesures sont généralement limitées à 2 dimensions. Le lieu et le mode d'articulation étant d'une grande importance en production de parole, un grand nombre d'études ont été réalisées à partir de l'électropalatographie ([Har72, Eng00a, Dix99]).

Les **techniques à ultrasons** permettent la visualisation du corps humain en coupe et fournissent des observations bidimensionnelles de bonne qualité des organes inaccessibles de parole, tels que la langue et le larynx, sans risque pour le locuteur. Elles mettent en jeu une vibration acoustique de fréquence ultrasonore. Les ultrasons traversent toutes les matières à l'exception de l'os et de l'air contenu dans le corps. Le principe consiste à balayer un organe à l'aide de ces ondes, à partir d'une sonde posée sur la peau. Les sondes émettent et reçoivent des ultrasons à l'aide d'un effet piézo-électrique sur des cristaux. L'informatique recueille les signaux renvoyés à la sonde par les organes et reconstruit une image. Les mesures ultrasons ont été recommandées pour détecter aussi bien les contractions musculaires que la forme de la surface de la langue [SSH+81].
Pour les études en parole, les techniques à ultrasons permettent de recueillir des données 2D ou 3D à des fréquences d'échantillonnage de l'ordre de 100Hz, dans une zone limitée du conduit vocal. Parmi ces études, on trouve celles de Maureen Stone ([SSTR88], [Sto90], [AKS99]) ou d'Ostry ([KO83],[PO93]).

L'articulographie électromagnétique (EMMA : electromagnetic misdagittal articulography ou EMA : electromagnetic articulography) utilise l'électromagnétomètre. Cette technique est utilisée pour suivre l'évolution des mouvements articulatoires lors de la production de la parole, par l'intermédiaire de marqueurs géométriques ponctuels. Des petites bobines sont attachées sur le sujet humain, à des endroits tels que les lèvres, la pointe de la langue ou le dos de la langue. Ces bobines jouent le rôle de capteur ou récepteur. Le sujet porte en plus un casque spécial qui produit un champ magnétique alternatif. Deux ou plusieurs émetteurs magnétiques, qui correspondent à de grandes bobines (fréquence de fonctionnement à 60kHz) sont placés au-dessus et derrière la tête du locuteur, de façon à ce que leurs axes principaux soient parallèles entre eux et perpendiculaires au plan médio-sagittal du conduit vocal. Ainsi la position des articulateurs peut être déterminée et enregistrée.
La difficulté majeure de cette technique est de placer les capteurs sur des axes parallèles aux émetteurs de manière à obtenir des mesures précises.
La technique EMMA permet de mesurer les déplacements de 5 à 10 points du conduit vocal à une fréquence d'échantillonnage pouvant dépasser 1 kHz. Les études en parole utilisent

l'EMA pour enregistrer des données de divers articulateurs [PCS+92], notamment la langue [Eng00a] ou le vélum [RBF00].

Toutes ces techniques offrent des résolutions spatiales bonnes et des résolutions temporelles supérieures à la cinéradiographie. Elles fournissent des données de type 1D+t ou 2D+t, c'est-à-dire que l'on extrait des informations sur une ou 2 dimensions au cours du temps. Cependant elles ont toutes pour défaut de ne pouvoir observer qu'une partie, et non la totalité, du conduit vocal.

L'**Imagerie par Résonance Magnétique** (IRM) est une technique qui permet de pallier ce défaut [BGGN91]. L'IRM permet de recueillir des données 2D et 3D du conduit vocal entier.
On fonde de grands espoirs sur cette technique qui permet d'obtenir une section du conduit vocal selon n'importe quel plan. Même si elle nécessite encore une longue exposition pour collecter une image, ces durées tendent à diminuer et la qualité de l'image ainsi obtenue est bien meilleure à celle des clichés radiographiques.
Le principe technique de l'IRM (milieu des années 1980) est complexe et s'appelle résonance magnétique nucléaire ou RMN. L'IRM utilise les propriétés qu'ont les noyaux d'hydrogène (ou protons) de l'organisme de générer un champ électromagnétique lorsqu'ils ont été excités (par une onde radio) et qu'ils retournent à leur état d'équilibre (relaxation). L'onde émise est ensuite mesurée pour former l'image.
L'IRM est une technique en pleine évolution. L'IRM dynamique commence à se développer avec la ciné-IRM ou l'IRM temps-réel. Le principe de la ciné-IRM est de découper un mouvement répétitif en N images et d'acquérir une nouvelle image à chaque répétition de ce mouvement. Celui de l'IRM temps-réel consiste simplement à acquérir une succession d'images le plus rapidement possible. L'avantage de ce mode d'acquisition est qu'il ne nécessite pas de répétitions du mouvement observé (qui pose des problèmes dans le cas de la ciné-IRM), par contre, la résolution temporelle n'est pas aussi élevée.
Actuellement, la résolution temporelle de l'IRM dynamique est de l'ordre de 10 à 15 images par seconde.

Les études en parole utilisant l'IRM sont de plus en plus nombreuses : parmi elles, par exemple, les études en IRM statique de Badin [BBRS98], Engwall [Eng00b] ou Yehia [YT97], ainsi que celles en ciné-IRM ([SDD+01], [KHM05] ou [MTH+99]) ou en IRM temps réel ([DMS00], [NNL+04]).
Les données 3D obtenues par IRM sont utilisées pour évaluer les modèles de passage des distances sagittales aux fonctions d'aire ([SLMD02] par exemple). Ce qui offre ensuite la

possibilité, dans différentes applications, d'effectuer un passage du 2D vers les fonctions d'aire, voies d'entrée vers l'acoustique.

Les données IRM sont les données de référence en modélisation 3D (images de langue les plus détaillées sans effet néfaste pour le sujet), elles ont besoin d'être complétées avec des données temps réel, pour générer un modèle représentatif de la parole courante. Les enregistrements IRM ont l'inconvénient de produire du bruit et une hyper-articulation, mais ces inconvénients sont considérés comme acceptables. En terme de modélisation 3D, les ultrasons ont un avantage sur l'IRM en ce qui concerne le temps d'acquisition mais il est compensé par le fait que les ultrasons ne permettent pas d'enregistrer la pointe de la langue et donnent moins d'informations sur la surface de la langue.

Pour un certain nombre de ces techniques d'imagerie présentées, la meilleure utilisation est d'en combiner plusieurs entre elles. En particulier, l'électropalatographie est combinée avec les ultrasons ou l'IRM mais le plus souvent avec l'EMMA, du fait de la possibilité d'enregistrer simultanément avec l'EMMA et l'EPG (voir par exemple [Eng00a] ou [Eng03]).

Malgré l'apparition de ces techniques d'imagerie (IRM, EMA), produisant des images de meilleures qualités que les rayons X, les films cinéradiographiques sont encore d'une grande utilité et restent appropriés à l'étude des articulateurs ; ils offrent la meilleure vue dynamique du conduit vocal dans son ensemble dans le plan sagittal et en particulier, ils offrent beaucoup d'information sur les mouvements complexes des articulateurs et sur leur coordination. La cinéradiographie est un bon compromis pour avoir des résolutions spatiale et temporelle correctes pour les données articulatoires. Cette technique permet un débit plutôt rapide avec une résolution suffisante pour récupérer la forme des articulateurs.

1.2. Extraction de la géométrie du conduit vocal à partir de la cinéradiographie

Devant cette grande variété de données disponibles pour étudier la parole, et en particulier ses modes de production, il faut trouver des techniques pour récupérer sur les images l'information géométrique pertinente. L'extraction de données caractéristiques du conduit vocal et de ses articulateurs cherche à s'adapter au type d'imagerie et à la quantité d'images. On s'intéresse ici essentiellement aux données cinéradiographiques.

Le marquage manuel est la seule technique utilisée, en dessinant directement le conduit vocal sur les images ou en y marquant des points à l'aide d'une grille. Mais face à de grandes bases de données vidéos, ce travail, très long, n'est pas envisageable pour extraire

des images toute l'information géométrique relative au conduit vocal. Des techniques, en partie ou totalement, automatiques voient le jour, avec plus ou moins de succès.

1.2.1. Extraction manuelle

Comme nous l'avons déjà dit, les clichés radiographiques sont depuis longtemps une source d'information essentielle en production de parole. Dans de très nombreux travaux, les coupes sagittales, c'est-à-dire les représentations du contour du conduit vocal dans le plan sagittal, ont été tracées manuellement. Le modèle de Maeda [Mae78] a été élaboré à partir de plusieurs dizaines de croquis de la coupe sagittale tirés de films radiocinématographiques. Les contours ont été dessinés sur une table traçante interactive reliée à un ordinateur, afin de transférer les coordonnées des points décrivant ces contours. L'ouvrage de Bothorel, Simon, Wioland et Zerling [BSWZ86] contient plus de 1000 croquis tirés de films radiologiques et labiographiques synchrones tournés à 50 images/seconde. Il s'agit de vues de profils réalisées manuellement par les auteurs, comme on en voit un exemple sur la figure suivante.

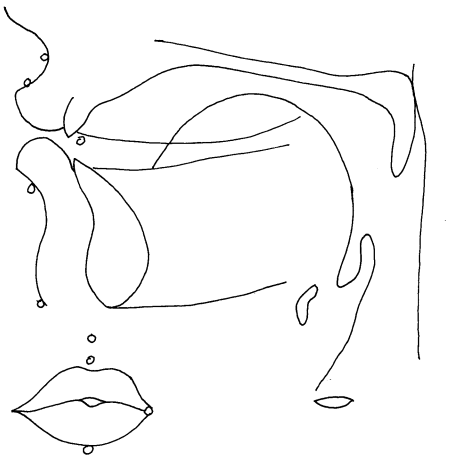


Figure 1 : Vue latérale du conduit vocal complet réalisée manuellement à partir d'une image radiographique, d'après [BSWZ86].

Les travaux de modélisation articulatoire de Pierre Badin ([Bad91], [BBL95]) résultent également de tracés manuels sur des images rayons X. Les profils sont tracés sur des feuilles transparentes puis numérisés en une liste continue de pixels codés par leurs coordonnées x-y.

Plus récemment, Anne Vilain pour sa thèse [Vil00] a établi un modèle anthropomorphique du conduit vocal d'un locuteur français à partir d'un enregistrement cinéradiographique de trente secondes de parole. Ce dernier a demandé un traitement long destiné à en tirer une banque de contours sagittaux, d'abord sur papier puis numérisés. Les images cinéradiographiques ont été projetées sur papier en chambre noire à l'aide d'un agrandisseur photographique

puis les contours du conduit vocal ont été retracés au crayon à la main en taille réelle. Des pastilles métalliques fixées sur le menton et la langue du sujet apportent des repères précieux et de bonne qualité aux images et des informations précises, en particulier pour le tracé de la pointe de la langue. Il résulte de ces travaux 1282 vues sagittales du conduit vocal.

Nombreux sont les travaux de l'Institut de Phonétique de Strasbourg qui utilisent les rayons X. Récemment, une interface de traitement d'images cinéradiographiques, INTRIC[1], a été élaborée [Roy03]. Celle-ci a été développée dans le but de faciliter et d'accélérer l'analyse de ces images, disponibles sous forme numérique. Cette interface n'a pas pour but de traiter automatiquement les images mais simplement d'en permettre leur exploitation directement à l'écran.

Les autres techniques d'imagerie font aussi l'objet de tracés manuels, comme par exemple l'IRM. Dans les travaux d'Engwall [Eng00b], le contour de la langue est extrait manuellement de chaque image IRM en utilisant une interpolation avec des courbes de Bézier, comme le fait Badin [BBB+00] à partir de trois piles d'images acquises pour chaque articulation (axiales, obliques et coronales). La langue est extraite comme une unité entière sur différents plans.

1.2.2. Extraction automatique

Plusieurs auteurs ont suggéré que pour poursuivre et compléter les études en production de parole (co-articulation, différences inter et intra sujets, dynamique…), une amélioration des méthodes d'extraction de caractéristiques articulatoires était nécessaire. En effet, l'extraction manuelle est forcément limitée par le nombre de données pouvant être analysées. L'extraction manuelle est soit limitée à de courtes séquences, soit réalisée pour des images choisies de la séquence, ceci empêchant l'étude dynamique des formes du conduit vocal, qu'on aurait pourtant pu espérer réaliser grâce à la cinéradiographie.

L'extraction automatique d'information articulatoire a alors été envisagée par des techniques d'images qui réduisent le bruit et rehaussent l'intérêt de certains éléments visuels.

En 1994, Tiede et Bateson [TV94] ont présenté des idées pour traiter automatiquement les images radiographiques afin d'exploiter au mieux les grandes bases cinéradiographiques et en particulier celle d'ATR [MVT95], en développant des algorithmes pour normaliser des histogrammes ou détecter des arêtes sur les images, ainsi que différents types de filtres. Les mesures peuvent être soit manuelles, soit automatiques. L'utilisateur positionne manuellement des objets de mesures (points, lignes, courbes de Bézier) pour les aligner

[1] Cette interface a été développée sous Windows dans un environnement Matlab et utilise un système de référence et la grille semi-polaire de Maeda pour les mesures de langue, de façon à mener à une extraction de la fonction sagittale.

avec les régions d'intérêt des images. La mesure automatique est basée sur la détection de changements significatifs d'intensité de pixels le long d'une ligne superposée à des images consécutives.
Ces auteurs soulignent la difficulté d'extraire le contour d'articulateurs « cachés » par des structures opaques aux rayons X ; c'est le cas, par exemple, de la langue qui est parfois difficilement visible à cause des dents (occlusion par la mâchoire supérieure). Mais la position de ces contours « cachés ou occlus » peut être estimée à partir de certaines contraintes physiques connues (par exemple, le palais dur représente une limite supérieure pour la langue) et de contraintes imposées par les positions du contour observées précédemment. Une approche de ce type, permettant d'inclure des contraintes, a été implémentée par Kass, Witkin et Terzopoulos [KWT87], qui utilise des splines minimisant l'énergie (« snakes ») pour extraire l'ouverture des lèvres. Le « snake » est une courbe géométrique qui approche les contours d'un objet par minimisation de fonctions d'énergie. Un « snake » se comporte comme une corde élastique qui serpente vers les contours de l'image grâce à un ensemble de forces locales gérées par un processus itératif. Les forces internes conservent la forme et assurent une continuité spatio-temporelle. Les forces externes tirent et guident le « snake ». L'initialisation et la normalisation des paramètres du « snake » sont difficiles. Le « snake » doit être placé assez proche du contour pour atteindre le minimum global de la fonction d'énergie. Une fois initialisé au contour de la première image, le « snake » suit ce contour d'image en image.

La méthode d'extraction automatique mise en place par Laprie et Berger ([LB96, BL96]) repose sur cette technique. Cependant cette technique « Snake » à contour actif ne permet pas d'extraire le contour de la langue dans la région de la mâchoire supérieure, lorsqu'elle est occluse. Elle nécessite d'être associée à une méthode de flux optique dans les régions où les contours ne sont pas suffisamment isolés, c'est-à-dire quand d'autres contours sont superposés au contour à extraire. Ainsi, au lieu de détecter le contour d'une image en utilisant le contour détecté sur l'image précédente, les auteurs s'efforcent de suivre le mouvement du contour de la langue à partir des variations de niveaux de gris entre deux trames consécutives.
La langue est correctement suivie lorsqu'elle est visible mais une interaction humaine est nécessaire quand elle est masquée par d'autres articulateurs ou quand elle bouge trop vite.
Cet algorithme de suivi du mouvement de la langue a été testé sur un film court mais nous n'avons pas connaissance de résultats d'extraction de contour publiés par les auteurs.

En 1999, Thimm et Luettin [TL99] de l'IDIAP se sont intéressés à l'extraction de l'ensemble des articulateurs du conduit vocal et au traitement de séquences radiologiques longues, par

une méthode automatique. Ils ont abouti au traitement complet d'un film issu de la base ATR, la séquence Laval43. Nous résumons ici en quelques mots la méthode de Thimm et Luettin, elle sera en effet plus longuement détaillée au chapitre 3, dans le but d'une étude comparative. Leurs résultats ont été obtenus à partir d'une technique de traitement d'images avec normalisation d'histogrammes et d'une méthode d'extraction de contours. Cette méthode fait appel à des images d'état, sur lesquelles l'application d'un détecteur de Canny permet de récupérer des contours. La procédure de suivi utilise l'appariement avec ces images et l'information temporelle.
La qualité des contours extraits avec cette méthode est faible en comparaison de celle obtenue manuellement. La tâche d'extraction automatique reste donc difficile et nécessite le plus souvent l'assistance ou la correction d'un expert. C'est dans ce contexte que se place notre méthode qui se veut quasi-automatique avec l'intervention d'un expert humain, précédant un traitement automatique.

Pour finir, nous pouvons mentionner d'autres études concernant l'extraction de contours pour d'autres techniques d'imagerie (IRM, ultrasons), permettant l'analyse des articulateurs de la parole.
Davis et al. [DDS96] utilisent l'IRM et de petites billes attachées à la langue. Les billes, placées sur l'articulateur, permettent d'extraire l'information. Des techniques de traitement d'images sont utilisées sur chaque image pour identifier les pixels qui représentent la surface de la langue. Les mouvements de 4 voyelles dans des fenêtres de 500ms ont été extraits.
Pour extraire le contour de la langue, Engwall [Eng04] utilise une détection automatique de frontières, basée sur un seuillage des images IRM. Cette détection est suivie d'une correction manuelle des points de contrôle des splines ainsi obtenues et permet d'obtenir 910 configurations de formes de langue.
D'autres approches utilisent les ultrasons. Akgul et al. [AKS99] proposent une technique à base de contours déformables, imposant des contraintes sur ces déformations et utilisant des techniques de flux optique et un filtrage spatio-temporel. Denby et al ([DS04], [DODS06]) utilisent un critère de maximum de gradient sur l'intensité pour extraire 14 points du contour de la langue depuis des images acquises par ultrasons. Ce critère est complété par un algorithme local de lissage pour relier les points du contour et par un filtrage temporel rétablissant une continuité entre les trames consécutives.

1.2.3. Cas particulier des lèvres

En tant qu'articulateur directement visible du conduit vocal, les lèvres peuvent être soumises à des traitements différents, en terme d'acquisition puis d'extraction de données. En effet, les

lèvres sont enregistrables à l'aide de simples caméras qui filment le visage de face ou de profil.

L'importance de la vision du visage du locuteur en terme d'intelligibilité dans le bruit ou dans le cas de surdité a largement été démontrée. Aussi, parmi le nombre important d'études portant sur le traitement automatique de la parole visuelle, beaucoup s'intéressent à la lecture labiale automatique. Il existe plusieurs approches de "labiométrie" automatique, issues de recherches sur la reconnaissance de formes, en vision par ordinateur.

Parmi les approches orientées « images » (on traite l'image comme un ensemble de pixels), on trouve notamment les méthodes basées sur des traitements par chroma-key sur des séquences vidéo de locuteur dont les lèvres sont peintes en bleu [Lal91]. Les études utilisant les « lèvres bleues » sont nombreuses à l'ICP, comme par exemple celle de Guiard-Marigny et al. [GAB96]. Si cette méthode est précise, elle est néanmoins tributaire du maquillage et de l'éclairage. Les méthodes à base de flux optique sont aussi à classer dans les approches « image » : l'analyse associe à chaque pixel un vecteur vitesse correspondant au flot d'intensité observé entre deux images consécutives.

Des systèmes de labiométrie [BBB+00] permettent de reconstruire un maillage 3D complet des lèvres et du visage à partir de petites billes collées sur divers points du visage, dont les coordonnées 3D sont ensuite reconstruites à partir des images de face et de profil.

D'autres méthodes ([YHC92], [RGBV97]), motivées par la modélisation articulatoire, utilisent des modèles géométriques des contours labiaux, contrôlés par peu de paramètres. Les paramètres sont déduits par optimisation de telle sorte que le modèle s'adapte au mieux avec les contours réels. Splines et équations polynomiales sont le plus souvent utilisées pour suivre les lèvres dans ce type de modèle.

Pour l'approche « pixel », certaines études (elles seront rappelées un peu plus loin dans ce manuscrit) mettent en œuvre la DCT. Le lien entre l'approche « pixel » et l'approche géométrique est discutée dans les travaux de Potamianos et al. [PGC98] ou Heckmann et al. [HKSB02]. C'est ce lien qui donne lieu à la technique de rétro-marquage [Ber04] mise en jeu dans notre méthode semi-automatique.

2. Modèles articulatoires

La modélisation articulatoire a pour objectif de décrire avec un petit nombre de paramètres les formes possibles du conduit vocal tout en préservant les déformations observées sur un conduit réel. Les différents articulateurs de la parole ont été étudiés dans des travaux spécifiques (lèvres, mâchoire, conduit nasal). La langue reste la structure déterminante et centrale des modélisations globales du conduit vocal.

Parmi le grand nombre d'études dédiées à la modélisation articulatoire, on distingue 3 approches principales [BMP94] : géométrique, physique et biomécanique.
On peut différentier ensuite les modèles 2D et 3D et ceux fonctionnant en temps réel.

2.1. Modèles physiques

Certains auteurs représentent la forme des articulateurs par une équation mathématique. Cette approche est basée sur le calcul de paramètres à partir de données d'observation, notamment radiographiques. Par exemple, Liljencrants [Lil71] représente la langue par les premiers coefficients de la décomposition en série de Fourier de la forme de la langue dans un repère semi-polaire. Ces modèles mettent en évidence le fait que la géométrie de la langue dans le plan sagittal peut être contrôlée avec peu de paramètres. Ainsi les deux ou trois premiers coefficients de la série de Fourier suffisent à Liljencrants pour définir le contour du corps de la langue, dans le cas des voyelles. En utilisant aussi la seconde harmonique, la pointe de la langue est modélisée. Le problème de ce type de modèle est de trouver une interprétation des composantes extraites en termes articulatoires.
En 1982, Hashimoto et Sasaki [HS82] ont décrit les profils de la langue à l'aide d'une décomposition polynomiale quadratique. Cinq paramètres caractérisent ces profils. Des relations polynomiales entre les formants et la forme de la langue ont été proposées et ont abouti à des relations de correspondance (mapping) entre les espaces articulatoire et acoustique.
Bien que ces modèles donnent une bonne représentation de la forme de la langue, ils ont toutefois quelques difficultés à rendre compte de certains aspects anatomiques du conduit, comme par exemple le fait que la langue soit fixée à la mâchoire.

2.2. Modèles géométriques et statistiques

Les modèles géométriques sont construits à partir de coupes sagittales de conduits vocaux et ont tous en commun de simplifier le conduit vocal par une lecture géométrique de ces coupes. Les différents articulateurs sont représentés par des formes géométriques simples (par exemple un cercle pour la langue...) qui sont déformées en utilisant des rotations autour d'axes fixes ou des translations selon des directions prédéfinies. Parmi ces modèles on trouve ceux de Coker [CF66] et de Mermelstein [Mer73]. Ce dernier a utilisé ce type de modèle pour définir des règles de synthèse dynamiques pour produire des séquences de type VCV. En 1976, Coker [Cok76] propose une autre version du modèle 2D étudié avec Fujimura en 1966 et l'intègre dans un système de synthèse articulatoire.
L'idée essentielle de ce type de modélisation est de produire un modèle avec peu de paramètres, capables cependant de produire toutes les articulations observables possibles.

L'inconvénient majeur de cette représentation géométrique est qu'elle ne prend pas correctement en compte la souplesse de la langue (capacité d'aplatissement et abaissement, mouvements complexes de la pointe dans les mouvements consonantiques).

Pour pallier cette lacune, des modèles statistiques ont vu le jour, comme ceux de Lindblom et Sundberg [LS71] et de Maeda [Mae78, Mae79]. Ceux-ci sont aussi bâtis à l'aide de radiographies de conduits vocaux. Les auteurs ont cherché à caractériser les formes de la langue par des valeurs directement mesurées sur les images des films radiologiques. Les paramètres articulatoires du modèle sont issus d'une analyse statistique des contours extraits de ces clichés.

En 1971, Lindblom et Sundberg [LS71] ont utilisé une décomposition purement statistique des variations du contour de la langue en paramètres articulatoires à partir de données radiographiques. Ils ont adopté une représentation où ils montrent trois positions principales pour lesquelles sont reproduites les voyelles [i, u, a] et qui définissent une caractéristique relative au lieu de constriction.

Le moyen le plus efficace de réaliser une analyse statistique est d'échantillonner le contour de la langue le long d'une grille, de représenter ainsi la forme de la langue par un vecteur et ensuite de trouver les composantes qui résument les vecteurs observés dans un grand ensemble de données.

L'analyse des contours permet de construire directement des modèles (paramétriques) contrôlés à l'aide d'un nombre restreint de paramètres (les composantes principales de la coupe sagittale).

Harshman, Ladefoged et Goldstein ont proposé en 1977 une méthode d'analyse factorielle, PARAFAC [HLG77]. En quantifiant les positions de langue de 10 voyelles anglaises prononcées par 5 locuteurs à l'aide de 13 lignes, ils ont montré que les formes de la langue pouvaient être décrites par deux paramètres. L'un décrit un mouvement d'avancement de la base de la langue accompagné du mouvement d'élévation de la partie antérieure du corps de la langue. L'autre décrit un mouvement d'élévation et de recul de la langue. Ces 2 facteurs donnent de fortes corrélations (supérieures à 0,96) entre les données observées et les prédictions du modèle. La PARAFAC est surtout utile pour représenter la variabilité inter-locuteurs ([HGW01], [ZHP03]).

L'analyse en composantes principales (ACP ou PCA) permet aussi d'extraire les facteurs optimaux décrivant le contour de la langue. Dans le modèle de Maeda [Mae90], le contour de la langue est décomposé à l'aide d'une ACP, après avoir supprimé par régression linéaire l'effet de la position de la mandibule. La forme de la langue est alors décrite de manière adéquate par 4 paramètres articulatoires (position de la mandibule, position avant-arrière du corps de la langue, dos de la langue, pointe de la langue), correspondant aux 4 premières

composantes de l'ACP. A ces 4 paramètres viennent s'ajouter 3 autres : deux paramètres pour déterminer l'ouverture et la protrusion des lèvres et un paramètre pour fixer la hauteur du larynx. Au final sept variables suffisent à obtenir une bonne adéquation entre les conduits vocaux humains et synthétiques.

Une difficulté rencontrée lors de la construction de tels modèles est bien sûr la lecture précise sur les radiographies des contours servant à construire le modèle. Néanmoins, ils représentent bien la géométrie effective du conduit vocal.

Il existe plusieurs modèles dérivés de celui de Maeda. Certains modifient la modélisation de l'apex de la langue [Gal97] pour modéliser les fricatives ou bien apportent quelques modifications au modèle afin de l'adapter à d'autres locuteurs [Mat99]. D'autres s'inspirent du travail de Maeda pour créer un nouveau modèle articulatoire en apportant certaines améliorations, pour mieux modéliser la langue par exemple [BBB+96].

La méthode statistique dans la tradition de Maeda a également été utilisée par Badin [BBB+00] pour construire un modèle 3D de la langue basé sur des contours mesurés dans le plan à partir de données IRM. La position de chaque point de contour est contrôlé par 6 paramètres articulatoires, définis par une analyse en composantes principales guidée.

Engwall [Eng03] propose un autre modèle 3D de la langue basé sur une approche similaire. L'objectif de ce modèle est d'être utilisé dans un système de synthèse audio-visuelle à partir de texte. En vue de corriger l'hyper-articulation observée durant l'acquisition d'IRM, les valeurs des paramètres ont été ajustées en comparant les contacts linguo-palataux virtuels et ceux mesurés par électropalatographie. De plus, des données de mouvements ont été mesurées à l'aide d'un articulographe électromagnétique, afin de déterminer le contrôle cinématique du modèle.

2.3. Modèles biomécaniques

Depuis le début des années 1970, des modèles biomécaniques tridimensionnels ont vu le jour. En effet, il est naturel de penser que les caractéristiques physiologiques de la langue, cette masse musculaire non rigide et très déformable (17 muscles extrinsèques et intrinsèques), imposent des contraintes aux déformations pour produire les sons du langage. Les données utilisées dans ces modèles proviennent en général de l'imagerie ultrasonique ou par résonance magnétique. La méthode des éléments finis permet de contrôler les différents muscles. Ces modèles permettent de prendre en compte des contraintes telles que la conservation de la masse de la langue et par conséquent de son volume (en effet on considère qu'elle est incompressible étant donnée l'abondance d'eau qu'elle contient). Les modèles bidimensionnels sont incapables d'intégrer ces contraintes puisque la langue ne se déforme pas uniquement dans le plan sagittal. En particulier, la profondeur du sillon varie fortement, comme on peut l'observer sur des images IRM mais difficilement sur des

radiographies. Les modèles biomécaniques n'ont pas cet inconvénient mais en revanche, ils nécessitent plus de paramètres pour contrôler finement la forme du conduit. La première tentative de modèle biomécanique a été menée par Perkell [Per74]. Son modèle est construit à partir d'éléments musculaires viscoélastiques dont l'organisation reflète la structure musculaire de la langue décrite dans le plan médio-sagittal. Les structures musculaires sont simplifiées et modélisées par un système de ressorts et de masses. La masse de la langue est concentrée en 16 points mobiles, interconnectés entre eux et connectés aux surfaces fixes. D'autres modèles ont suivi. Ces modèles utilisent des modélisations à éléments finis. Par exemple, Payan et al. [PP93] ont modélisé la langue dans le plan médio-sagittal et simulé les transitions vocaliques, contrôlées par une équation du mouvement simulant la dynamique de la langue. Wilhelms-Tricarico [Wil95] a défini en 1995 une méthode mathématique pour des simulations des mouvements et des déformations de la langue. Il s'agit des solutions d'équations différentielles non-linéaires du second-ordre approximant l'énergie des structures.

Bien qu'ils permettent de modéliser très finement les mouvements musculaires, les modèles biomécaniques sont difficiles à utiliser car il faut un grand nombre de paramètres nécessaires à la définition de la dynamique du modèle.

2.4. Modèles de fonction d'aire

Il est également possible de modéliser le conduit vocal par un ensemble de tubes acoustiques [Fan60]. Il est généralement modélisé par un tube droit de section circulaire $A(x,t)$. On considère souvent le tube statique à parois rigides : A ne dépend que de x qui varie de la glotte aux lèvres. Pour modéliser le conduit vocal, il suffit de considérer 3 tubes, le premier correspondant au pharynx, les deux autres respectivement à la bouche et au conduit nasal.

Toutes les fonctions d'aire produites par ce type de modèle ne peuvent certainement pas l'être par un humain. L'espace acoustique résultant de la simulation recouvre l'espace acoustique d'un locuteur quelconque. L'avantage en codage de parole est évident : à chaque son produit par le locuteur humain, il correspond au moins une fonction d'aire et en conséquence, tous les sons de la parole peuvent ainsi être codés. Cependant, l'utilisation de modèles articulatoires a pour but de retrouver les trajectoires des articulateurs du locuteur. Comme Atal l'a remarqué [ACM+78], il existe une infinité de fonctions d'aire qui produisent le même ensemble de formants. Le problème est de choisir parmi elles celle que le locuteur a utilisée pour parler. C'est le problème (classique) du « many-to-one ». L'utilisation d'un modèle articulatoire construit à partir d'images de conduits vocaux a l'avantage de fixer des contraintes réalistes. Néanmoins ces modèles doivent être spécifiquement adaptés au

locuteur qu'ils modélisent afin de faire coïncider les espaces vocaliques du locuteur et du modèle.
Les modèles biomécaniques tridimensionnels permettent de fournir directement la fonction d'aire le long du conduit, contrairement aux modèles de coupes qui s'appuient sur une application pour calculer l'aire à partir du diamètre mesuré dans le plan sagittal afin de résoudre les équations acoustiques. Cependant, comme on l'a dit, les techniques d'acquisition ne permettent pas encore d'étudier la dynamique du conduit dans les trois dimensions.

Pour la validation acoustique, la fonction d'aire est un passage obligé en permettant le passage de la coupe sagittale à une pseudo-estimation 3D. Les modèles articulatoires ont directement stimulé toutes les études concernant la fonction d'aire, à commencer par Heinz et Stevens en 1964 [HS64]. Deux types de stratégies émergent de la littérature sur la détermination de la fonction d'aire du conduit vocal : les méthodes directes impliquant des mesures géométriques du conduit vocal et les méthodes indirectes basées sur l'inversion acoustique.
L'approche indirecte consiste à déterminer la fonction d'aire à partir de données acoustiques, soit à partir du signal de parole lui-même, soit à partir de la réponse acoustique du conduit vocal à une excitation externe (par exemple [Sch67], [Son79]).
L'approche directe utilise une variété de techniques complémentaires dont aucune (excepté l'IRM) ne donne de résultats complets en trois dimensions. A partir de ces données, un certain nombre de modèles pour le passage de la coupe sagittale à la fonction d'aire ont été proposés, tous basés sur la relation $A = \alpha.d^{\beta}$ initiée par Heinz et Stevens en 1965 [HS65], dont on reparlera au chapitre 4. Parmi ces travaux, ceux de Sundberg [Sun69], Baer [BGGN91], Perrier [PBS92] ou Soquet [SLMD02] se limitent aux voyelles. Beautemps et al. [BBL95] proposent un modèle de passage de la coupe sagittale à la fonction d'aire validé sur des consonnes et optimisé pour relier de manière cohérente fonctions sagittales, fonctions d'aire et formants.

Parmi les motivations citées précédemment de l'élaboration de grandes bases de données cinéradiographiques, plusieurs sont mises en jeu dans nos travaux. Pour permettre l'exploitation de ces données et leur apporter une valeur supplémentaire, nous pensons qu'il est nécessaire de développer de nouvelles techniques de traitement automatique de ces images. En facilitant l'extraction de données articulatoires sur de longues séquences, celles-ci pourront ensuite être mises à profit pour développer de nouvelles applications, dans la lignée de certains modèles articulatoires existants.

QUELQUES BASES DE TRAITEMENT VIDEO

D'un point de vue technique, travailler avec des images, ou des séquences d'images dans le cas de la vidéo, demande de manipuler de grandes quantités d'information, généralement plus que lorsque l'on manipule du son. Ce volume de données peut cependant être maîtrisé par le développement de traitements nouveaux et appropriés. Certains de ces traitements peuvent être exploités dans les algorithmes de capture de mouvements articulatoires que nous cherchons à établir à partir des données d'imagerie présentées précédemment. Nous présentons ici des traitements généraux qui serviront de base à la méthode que nous proposons. Il ne s'agit évidemment que d'un bref état de l'art compte-tenu de la quantité de traitements vidéos développés ces dernières années.

1. Coefficients DCT

La Transformée en cosinus discret ou TCD (de l'anglais : DCT ou Discrete Cosine Transform) est très utilisée en traitement du signal et de l'image [Mak80], et spécialement en compression. La DCT possède en effet une excellente propriété de "regroupement" de l'énergie : l'information est essentiellement portée par les premiers coefficients qui correspondent aux basses fréquences.
C'est une transformation similaire à la transformée de Fourier discrète (DFT). Le noyau de projection est un cosinus et génère donc des coefficients réels, contrairement à la DFT, dont le noyau est une exponentielle complexe et qui génère des coefficients complexes.
Le passage par la DCT a été l'idée majeure pour la compression JPEG (*Joint Photographic Experts Groups*, le groupe qui a créé ce format en 1987). JPEG est un format d'image très utilisé notamment dans le codage de vidéos MPEG (*Moving Photographic Experts Groups*). Ce format [Wal91] a l'avantage de fournir des images de bonne qualité avec un fort taux de codage, il est donc particulièrement répandu sur Internet. En outre il permet de choisir quelle qualité d'image on veut obtenir, sachant que plus la qualité est faible, plus la taille de stockage de l'image est faible. Cette compression est donc une compression avec pertes appliquée par petits blocs de 8*8.

Le processus de la DCT est un opérateur mathématique, tout comme la Transformée de Fourier. Cet opérateur est bijectif, c'est-à-dire qu'il permet un changement de domaine d'étude, tout en gardant exactement la même fonction étudiée. On étudie une image, c'est à dire une fonction de 2 variables : X et Y indiquant les coordonnées du pixel, et f(X,Y) la

valeur du pixel en ce point. Dans le cas d'une image couleur, il faut donc considérer indépendamment 3 fonctions, pour chacun des canaux RGB.
L'application de la DCT, ou d'une Transformée de Fourier, fait passer l'information de l'image du domaine spatial en une représentation identique dans le domaine fréquentiel. La DCT décompose l'image en coefficients, chacun contenant une information de fréquence spatiale. La DCT s'applique à une matrice, le résultat fourni est représenté dans une matrice de même dimension. Plus on s'éloigne de l'origine et plus les fréquences correspondantes sont élevées. Les basses fréquences se trouvent en haut à gauche de la matrice, et les hautes fréquences en bas à droite.

Formule pour calculer la DCT sur une matrice NxN

$$DCT(i,j) = \frac{1}{\sqrt{2N}} C(i)C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} pixel(x,y) \cos\left(\frac{(2x+1)i\pi}{2N}\right) \cos\left(\frac{(2y+1)j\pi}{2N}\right)$$

avec $C(x) = \frac{1}{\sqrt{2}}$ si $x = 0$

et $C(x) = 1$ si $x > 0$

Ce changement de domaine est intéressant. En effet, une image classique admet une grande continuité entre les valeurs des pixels. Les hautes fréquences étant réservées à des changements rapides d'intensité du pixel, ceux-ci sont en général minimes dans une image.
La plupart des images sont donc composées d'informations de basses fréquences. Quand on s'éloigne de l'origine, on trouve des coefficients faibles et qui sont peu importants pour la vision. Ainsi la force de la DCT est de "choisir" les éléments importants de l'image (ce qui semble difficile à réaliser sur une image non transformée) ; on parvient à représenter l'intégralité de l'information de l'image sur très peu de coefficients, correspondant à des fréquences plutôt basses.

La transformation DCT étant bijective, elle s'accompagne d'une méthode d'inversion (IDCT) pour revenir dans le domaine spatial. Ainsi après avoir fait des modifications dans le domaine fréquentiel, éliminer des variations de l'image quasiment invisibles par l'œil humain, on retourne à une représentation sous forme de pixels.

On a dit précédemment que la DCT était dans la même classe d'outils mathématiques que la Transformée de Fourier. Alors pourquoi les membres du groupe JPEG ont-ils fait le choix de la DCT ? Bien qu'étant plus longue à calculer que la transformée de Fourier (on peut calculer la DCT à partir de la FFT), la DCT comporte certains avantages. La DCT n'utilise pas de coefficients complexes, ceci rend la programmation légèrement plus simple. La DCT répartit

l'énergie du signal sur une bande de fréquences plus basses que pour la transformée de Fourier et on a donc besoin de moins de coefficients. La DCT permet ainsi d'extraire les coefficients basses fréquences en image pleine.

On voit sur le schéma suivant que la reconstruction à partir des premiers coefficients de la DCT (en tronquant les autres) donne de meilleurs résultats qu'avec la transformée de Fourier. En effet, le nombre de coefficients significatifs de la DCT est inférieur à celui de la FFT. Pour restituer convenablement l'information, on a besoin de beaucoup plus de coefficients pour la FFT que pour une DCT ; la décroissance des coefficients de la FFT n'est pas suffisante pour négliger rapidement les coefficients de grands indices. De plus en tronquant les derniers coefficients, on risque de voir se produire le phénomène de Gibbs, qui se traduit par une oscillation au niveau des discontinuités.

Dans le cadre de la reconnaissance automatique de la parole, l'utilisation de séquences vidéos des lèvres du locuteur a été l'objet de nombreuses études. La plupart s'intéressent aux moyens de combiner l'information vidéo avec la partie audio et surtout s'interrogent sur l'extraction de caractéristiques visuelles appropriées, contenant l'information utile à propos des mots prononcés.

L'étude de Potamianos et al. [PGC98] compare différentes caractéristiques visuelles sur la base de performance en lecture labiale. Ils séparent les caractéristiques basées sur une approche contour (une paramétrisation des lèvres est réalisée à partir des contours extraits de la séquence) et celles basées sur une approche pixels (l'image entière est considérée et des transformations adaptées sont utilisées à partir des pixels). L'approche DCT en images pleines et bien cadrées permet de bonnes performances et est recommandée. Elle a été utilisée dans des études comme celle de Heckman et al. [HKSB02] qui évalue plusieurs stratégies pour choisir les coefficients DCT les mieux adaptés à un système de reconnaissance et qui compare les caractéristiques DCT aux caractéristiques géométriques des lèvres (extraites à partir de lèvres colorées en bleu). Les paramètres vidéos sont extraits en utilisant une DCT sur les images des lèvres converties en échelle de gris. Un petit nombre de coefficients (une trentaine sur environ 5000) est suffisant pour permettre un score optimal de reconnaissance : il s'agit des coefficients basses fréquences. L'usage d'une approche basée sur les pixels a l'avantage d'éviter l'utilisation de marqueurs mais elle a

l'inconvénient d'être sensible aux mouvements d'ensemble à l'intérieur du cadre, qui se composent avec le mouvement des lèvres.

Les coefficients DCT globaux (en image pleine) sont sensibles aux mouvements et c'est cette propriété que nous allons exploiter dans la méthode d'extraction proposée dans cette thèse.

2. Indexation vidéo à partir d'images clefs

L'indexation vidéo fait actuellement l'objet de recherches très abondantes dans le domaine du traitement d'images et de la vision par ordinateur. En effet, la généralisation des supports numériques, l'apparition de formats vidéos compacts, la chute du coût des média de stockage a engendré une augmentation vertigineuse de la quantité des données multimédia. Pour que ces données soient exploitables, il faut qu'elles puissent être consultées efficacement comme par le biais d'un catalogue.

(1) Le passage de la forme d'enregistrement analogique à la forme numérique permet un accès non linéaire au contenu. En effet, il devient possible en posant des index sur le flux audiovisuel numérique d'accéder directement au contenu sans avoir à parcourir séquentiellement celui-ci. Ces index permettent de pointer ou de montrer où se trouve le contenu et servent à la recherche d'information. Un index est donc vu comme pointeur de contenu.

(2) Indexer le matériel audiovisuel requiert la sélection d'images clefs suffisamment représentatives. Cette notion d'images clefs est une notion habituelle en traitement vidéo, elle est utilisée dans des modes de compression et notamment dans la compression MPEG [LeG91], où certaines images clefs servent de référence, et où une compensation de mouvements sert à coder les images intermédiaires. Il est en effet inutile d'entrer dans des calculs compliqués et longs pour toutes les images d'une séquence afin de correctement en extraire les caractéristiques visuelles, puisqu'il sera par la suite impossible de conserver et d'utiliser cette information, par ailleurs redondante. Le processus de simplification de la vidéo passe par la sélection d'une ou plusieurs images, dites images clefs, représentatives des plans de la séquence. Ici, la notion d'index est associée à celle de représentant qui permet de réduire la redondance.

2.1. MPEG et la notion d'images clefs

MPEG constitue l'algorithme de compression le plus utilisé pour la vidéo. Il combine plusieurs algorithmes différents : une compression temporelle et une compression spatiale.

En effet, les séquences vidéo contiennent une très grande redondance statistique et subjective entre images successives, dans le domaine temporel comme dans le domaine spatial. La propriété statistique de base des techniques de compactage est la corrélation entre pixels. On suppose que l'importance d'un pixel particulier de l'image peut être prévue des pixels voisins de la même trame ou des pixels d'une trame voisine.
Intuitivement il est clair que lors d'un changement de scène au cours d'une séquence vidéo, la corrélation temporelle entre pixels des trames voisines est petite et peut même disparaître. Dans ce cas les techniques de codage « intra-frame » sont appropriées pour explorer la corrélation spatiale afin de réaliser une compression efficace des données. Les algorithmes de compression vidéo de type MPEG utilisent la compression JPEG avec application de la DCT, sur des blocs de 8x8 pixels, pour analyser efficacement les corrélations spatiales entre pixels voisins de la même image.
Si la corrélation entre pixels dans des trames voisines est grande, c-à-d. lorsque deux trames consécutives ont un contenu semblable ou identique, il est souhaitable d'utiliser la technique de codage « inter-frame » DPCM qui utilise la prévision temporelle (prévision compensée du mouvement entre trames). L'objet de cette compression temporelle est de ne stocker que ce qui est modifié lors du passage d'une image à une autre dans une séquence vidéo. Dans le schéma classique du codage vidéo, une combinaison adaptative entre les deux mouvements (temporel et spatial) de l'information est utilisée pour réaliser une grande compression de données (codage vidéo hybride DPCM/DCT).

La technique de compression de base de MPEG-1 (et de MPEG-2) est basée sur une structure de macro-blocs. La première trame d'une séquence vidéo (trame I) est encodée avec le mode « intra-frame » sans aucune référence à des trames passées ou futures. Chaque trame suivante est codée en utilisant la prédiction « intra-frame » (trames P) ; seules les données de la trame codée juste précédemment (trames I ou P) seront utilisées pour la prédiction. Pour accéder à un support média, l'algorithme MPEG-1 a été pensé pour supporter différentes fonctionnalités comme l'accès aléatoire, la recherche en vitesse avant et arrière dans le flux vidéo, etc. Pour incorporer ces fonctionnalités et pour tirer plus d'avantages de la compensation et de l'interpolation de mouvement, l'algorithme MPEG-1 introduit le concept d'images prédictées et interpolées bidirectionnellement (trames B), i.e. en avant et en arrière dans le temps.
Pour résumer, trois types de trames sont considérés :

- Trames I : Ces trames sont codées indépendamment de leur contexte, c'est à dire sans aucune référence à autre image de la séquence vidéo, comme expliqué juste avant. Les trames I ne permettent qu'un très bas taux de compression.

- Trames P : Ces trames sont codées avec une référence à l'image précédente (trame I ou trame P). Ces trames sont utilisées pour la prédiction de trames futures ou passées. Elles utilisent la compensation de mouvement pour un meilleur taux de compression. Par ailleurs, elles ont l'inconvénient de propager les erreurs, du fait qu'elles réutilisent les informations de l'image précédente.
- Trames B : Elles ont besoin des trames futures et passées comme référence pour être codées. Elles sont utilisées pour obtenir un très haut taux de compression. Elles ne sont jamais utilisées comme référence, ce qui permet de ne pas propager les erreurs.

Les images I sont des images clefs ou images de référence, qui contiennent en elles-même, une grande partie de l'information. Les autres images ne contiennent que les pixels modifiés vis à vis de l'image précédente, qui est une image clef ou non. La première image est nécessairement une image clef. La compression spatiale (type DCT) s'applique exclusivement à une image donnée, sans tenir compte des images environnantes.

2.2. Indexation vidéo

Indexer le matériel audiovisuel requiert la sélection de séquences comme pointeurs vers des unités de plus haut niveau (comme les scènes) et la sélection d'images clefs.

L'unité la plus basique après l'image est le plan, c'est-à-dire une séquence audiovisuelle enregistrée de manière continue. La détection des transitions entre les séquences est opérée grâce à des méthodes de différence d'images basées sur la comparaison pixel à pixel ou sur la distribution (histogrammes) des valeurs colorimétriques sur les images entières (ou sur un ensemble de sous-régions de l'image).

La segmentation en plans pour l'aide à l'indexation vidéo s'accompagne du problème de l'extraction d'images clefs dans chaque plan, c'est-à-dire les images « les plus représentatives » du plan. Idéalement les images clefs doivent capturer le contenu sémantique du plan. Cependant les techniques de traitement de l'information ne sont pas assez avancées pour déterminer de telles images clefs. Les algorithmes utilisent donc les caractéristiques brutes obtenues sur les images (couleur, texture, mouvement). Lorsqu'un plan est statique, les images le composant sont souvent très similaires. Théoriquement il suffit alors de choisir l'image qui est la plus similaire aux autres, mais malheureusement cette recherche exhaustive est difficilement réalisable en pratique. Les approches empiriques sélectionnent simplement la première, la dernière ou l'image médiane du plan.

Des techniques spécifiques de détection, reconnaissance, identification sont utilisées pour effectuer des tâches particulières d'aide à l'indexation. Ce sont typiquement la détection et le suivi des objets mobiles, la détection d'objets particuliers ou l'identification. Les techniques

d'indexation proposent d'attacher à une image ou à une vidéo un ensemble de descripteurs de leur contenu, dans le but de mesurer la ressemblance avec les descripteurs correspondant à une requête.
D'une manière générale, étant donnée une image, on peut calculer un index de description, puis réaliser une mesure de similarité de l'index inconnu avec les indices de la base. Ceci permet d'obtenir les adresses des meilleures images au sens de la mesure de similarité.
Les questions qui se posent sont quels descripteurs et quelles mesures de similarité choisir, pour prendre en compte les difficultés qui peuvent se présenter (rotation, translations, 2D-3D, visibilité partielle, changement de luminosité…).

Devant le nombre d'études, plusieurs auteurs ont réalisé des aperçus des techniques existantes en indexation vidéo. Parmi eux, Veltkamp et Tanase [VT00] s'intéressent aux systèmes de traitement d'images par le contenu.
De nombreuses techniques d'indexation vidéo se placent dans le domaine des pixels. Zhang et al. [ZLSW95] présentent des techniques d'analyses visuelles et d'extraction d'images clefs. La représentation du contenu d'une image est basée sur différents types de caractéristiques, incluant des histogrammes de couleurs, des caractéristiques de texture, de forme, de contours. Des travaux similaires ont été effectués par Nagasaka et Tanaka [NT91] qui étudient différentes mesures pour détecter des changements de scènes, la meilleure étant un test normalisé du χ^2 pour comparer la distance entre deux histogrammes. De plus, pour réduire certains bruits, les trames sont divisées en sous-trames et les comparaisons se font sur ces sous-trames. Flickner et al. [FSN+95] décrivent le système QBIC (Query by Image Content) qui récupère les données par le contenu (formes, textures, couleurs) dans de grandes bases d'images et de vidéos.
Cependant, compte-tenu du temps nécessaire pour décoder une vidéo compressée, d'autres techniques ont vu le jour, qui analysent les données directement dans le flot vidéo compressé. Chang [Cha95] décrit de manière globale les techniques d'indexation dans le domaine compressé vidéo et soulève les difficultés rencontrées dans les domaines de la DCT, des ondelettes et des transformations en sous-bandes.
L'étude de Arman et al. en 1993 [AHC93] utilise les coefficients DCT des séquences codées pour détecter les changements de scène et pour d'autres traitements bas-niveau. Les coefficients DCT sont analysés pour sélectionner systématiquement les trames représentatives pour chaque plan. En utilisant quelques coefficients, un vecteur est formé puis utilisé pour détecter les variations dans le contenu des trames. Kobla et al. [KDLF97] présentent également une approche dans le domaine compressé, qui mesure le déplacement de la caméra dans le plan et découpe ce dernier en sous plans afin de limiter l'amplitude du mouvement. Ils considèrent le problème d'indexation à partir de trames

codées MPEG. L'information disponible depuis ces trames est le type de chaque image (I, P, B), les coefficients DCT de chaque image et le vecteur de mouvement pour les images prédites (P et B). Le but de l'approche est d'extraire un jeu d'images clefs de façon à capturer l'essentiel du contenu vidéo tout en excluant les trames redondantes. Après un partage de la séquence en scènes, ils choisissent la première image de chaque scène comme image clef. Les caractéristiques sont extraites de ces images. Les coefficients DCT (qui représentent l'information spatiale des trames individuelles dans l'espace compressé) basses fréquences des images clefs sont transformés en vecteurs de faibles dimensions qui sont utilisés pour l'indexation. La similitude entre deux trames est déterminée par la distance euclidienne entre les vecteurs.

Pour finir avec l'indexation vidéo, on note l'émergence de nouveaux standards de codage vidéo tels que MPEG7 qui intègre dans le codage des données explicites relatives aux contenus audiovisuels, dans le but de faciliter à la fois la recherche d'information dans une base de données vidéo et la navigation « intelligente » dans une vidéo.

CAPTURE DE MOUVEMENTS

L'analyse vidéo des mouvements humains est un domaine important de recherche, dédié à la détection de personnes et à la compréhension de comportements dynamiques humains dans des environnements complexes. La modélisation de ces comportements peut être utile pour toutes sortes d'applications, comme les animations graphiques, la compréhension de comportements normaux ou pathologiques ainsi que l'analyse de ces données pour des applications médicales. De plus, le domaine de la vision par ordinateur s'intéresse au problème de la détermination de la forme 3D à partir d'images 2D. Une approche classique est la *forme d'après le mouvement* (une traduction (peu utilisée) de l'anglais *shape-from-motion*), dans laquelle la forme 3D, ou structure spatiale, est déterminée à partir de positions dans le plan image de primitives 2D (par ex. des points ou des contours).

L'estimation et la modélisation des mouvements humains à partir d'images a déjà fait l'objet de nombreuses études. D'un côté, il existe depuis longtemps des systèmes dit de « motion capture », qui sont utilisés dans la réalité virtuelle ou augmentée, dans les applications médicales, et plus récemment pour l'animation des personnages virtuels dans les films et les jeux vidéo. Ces systèmes sont à présent très performants, mais ils ont besoin de plusieurs caméras calibrées et synchronisées, d'une illumination contrôlée, et surtout de vêtements et de maquillages spéciaux, muni de cibles actives ou passives qui facilitent les étapes de suivi et de mise en correspondance. D'un autre côté, il y a les études qui tentent de travailler avec des séquences vidéo plus « naturelles » : non-calibrées, non-synchronisées, prises dans un milieu naturel et non-instrumenté, et sans cibles ou autres aides à l'appariement.

Dans cette revue des études en capture de mouvement, on distingue également les études qui s'intéressent à la dynamique du corps humain complet et celles qui cherchent à identifier les actions de parties spécifiques du corps, comme les mouvements de la main ou les expressions du visage.

1. Paradigme des points lumineux

Afin de comprendre la perception visuelle de ce qu'il a, par la suite, appelé le mouvement biologique, Gunnar Johannsson [Joh73] a étudié l'interprétation d'affichages de points lumineux ("point-light displays") en mouvement. Ces points lumineux étaient situés sur les principales articulations (chevilles, genoux, hanches, poignets, coudes, épaules) du corps de personnes, au cours d'activités diverses. Il a ainsi démontré l'extraordinaire capacité des sujets humains à interpréter le mouvement d'un nombre limité de points éclairés dans une

action telle que marcher, danser, courir ou lancer un javelot. Chaque image prise séparément est difficile, sinon impossible, à interpréter, ce qui montre que le mouvement est bien l'élément primordial de la perception.

Depuis, il a été démontré que cette capacité remarquable permet de faire des discriminations encore plus riches, comme déterminer le sexe d'une personne à partir de sa façon de marcher. Kozlowski et Cutting [KC77, CPK78] ont montré que des stimuli cinématiques de cette nature permettent de distinguer un marcheur d'une marcheuse et d'autre part, que des stimuli «moins biologiques » (en changeant la fréquence de la marche ou l'amplitude de la course des bras) rendent l'interprétation plus difficile. Cette sensibilité aux mouvements humains semble innée ou très précoce [BPC84, BPK87] : des enfants âgés de 3 à 5 mois sont capables de discriminer une disposition cohérente de ces points lumineux d'une disposition incohérente ou « au hasard ». Enfin toute déviation par rapport à des gestes attendus est facilement interprétée comme une intention délibérée. On estime facilement le poids d'un objet à partir de la cinématique d'une personne qui le porte et on infère aussi une différence entre le poids perçu et réel [RF81] : un acteur soulevant une boite ne peut tromper un observateur sur le poids de la boite. En ce qui concerne la perception du visage, Bassili [Bas78] 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. De nombreux systèmes de capture de mouvement très performants utilisent une technologie dite « instrumentée », c'est-à-dire que le sujet doit porter un harnais spécial ou des vêtements spéciaux munis de cibles géométriques réfléchissantes, et il faut travailler sous l'illumination stroboscopique dans un espace muni d'un nombre suffisant de caméras synchronisées.

Cependant, ce genre de système se révèle insuffisant pour un grand nombre d'applications qui nécessitent l'utilisation de vidéos « plus naturelles ». Par exemple, la surveillance vidéo, l'indexation de vidéo par action, l'analyse sportive sont des domaines d'application potentiels de cette capture de mouvement « non-instrumentée ». Mais sans cibles réfléchissantes et sous une illumination quelconque, l'extraction d'indices de l'image devient une tâche plus délicate.

2. Suivi et reconstruction du mouvement humain

Parmi les travaux sur le suivi et la reconstruction 2D et 3D du mouvement humain (détection de personnes, suivi 2D sans reconstruction 3D, reconstruction du mouvement 3D), on voit

apparaître deux approches. Une première, dite « articulée », est basée sur un modèle géométrique explicite, comme le « scaled prismatic model » [CR99]. Une seconde approche est basée essentiellement sur la sélection d'exemples similaires dans une base d'apprentissage et sans modèle géométrique explicite. Celle-ci est dite « par apparence » ou « exemplaire ». A partir d'un ensemble d'exemplaires de positions, on estime une nouvelle pose en cherchant parmi ces exemplaires la ou les images les plus proches et en interpolant les positions. Par exemple, en détection 2D de personnes, Gavrila [Gav00] développe des algorithmes efficaces pour comparer explicitement l'image avec une série d' « exemplaires », afin de décider si oui ou non il y a une personne présente. Le suivi 2D peut se faire par l'enchaînement d'une série d'exemples types d'apparence, sans modèle explicite [TB01].
Les travaux de Triggs se sont intéressés à l'estimation de la position et du mouvement articulaire du corps humain, à partir d'une seule image ou d'une séquence monoculaire d'images. Ils sont reportés dans son Habilitation à Diriger des Recherches [Tri05], ainsi qu'une revue d'autres travaux dans le domaine.

Une autre revue [WS03] présente diverses recherches dans le domaine de la capture de mouvements, en fonction des différentes parties du corps humain. Les techniques diverses s'adaptent à la partie du corps concernée. La détection de têtes et de visages est un problème complexe dans la mesure où il faut trouver le même visage dans les images de la séquence. Les études utilisent l'information de couleur ou des caractéristiques du visage, mais aussi des méthodes de contours actifs et de flux optiques, ou une combinaison de plusieurs méthodes. Le suivi des mouvements de la main et des doigts utilise souvent des gants et des marqueurs. La capture des mouvements du corps humain complet a été analysée en 2D et en 3D, ainsi que celle de plusieurs corps simultanément.

3. Analyse de scènes dynamiques

On peut concevoir deux façons distinctes pour analyser une scène dynamique. La première consiste à étiqueter les objets intéressants dans chaque image l'une après l'autre. Ainsi, supposons que l'on puisse localiser un objet précis (la tête d'une personne, par exemple) dans la première image d'une séquence aux coordonnées XY et que, dans l'image suivante, une tête aux coordonnées X'Y' soit identifiée. On suppose alors que la tête a bougé entre les deux images. Une telle approche dépend de façon critique de la précision du processus d'étiquetage. De plus, il ne faut pas que l'image contienne trop d'objets similaires, sinon le problème classique de mise en correspondance apparaît.

La deuxième méthode passe par une étape de traitement visant à analyser les mouvements au niveau local dans l'image et plus particulièrement en calculant le flux optique à chaque point de l'image. Une fois que ces mouvements sont analysés, nous pouvons chercher des séquences de mouvements spécifiques qui caractérisent une action donnée, sans passer par l'étiquetage des objets concernés. C'est ce qui se passe dans la perception des point-light displays, car aucune forme précise n'est identifiable ; ce n'est que le pattern de mouvement qui est utilisé.

L'identification des formes et l'analyse du mouvement sont deux processus séparés. Récemment, Giese et Poggio (2003) ont mis au point un modèle simplifié du système visuel qui intègre ces deux types de mécanismes [GP03]. Ils utilisent un réseau de neurones multicouche et deux voies de traitement indépendantes, l'une spécialisée dans l'analyse des formes et capable d'analyser chaque pose isolément et une autre spécialisée dans le traitement des mouvements seuls. Ce découpage trouve sa justification dans l'organisation des voies visuelles chez le primate où la distinction entre traitement des formes et du mouvement se manifeste dans l'organisation des aires visuelles extra-striées. Le fait d'utiliser deux voies indépendantes donne une robustesse supplémentaire au système lui permettant d'interpréter à la fois des poses isolées et des patterns de mouvements primitifs.

Notre modèle est basé sur un principe équivalent. Les mouvements sont capturés dans l'espace DCT et associés à des formes dans l'espace géométrique pour reconstruire des données géométriques dynamiques.

4. Parole et mouvement

L'analyse du mouvement trouve aussi sa place dans les études en parole et notamment dans la perception auditive et visuelle de la parole.

Des expériences utilisant le paradigme des « points lumineux » [Joh73] ont été effectuées sur de la parole audiovisuelle depuis les travaux de Summerfield [Sum79], où quatre points situés sur le contour des lèvres n'avaient pourtant pas permis de procurer un gain d'intelligibilité significatif.

Les travaux de Rosemblum et Saldaña [RJS96, RS98] ont placé des points lumineux sur le visage du locuteur et ont montré que le mouvement pouvait permettre de spécifier le geste et l'intention. Ainsi, des observateurs non entraînés sont capables de faire de la lecture labiale de voyelles, de syllabes et de mots courts sur un visage en points lumineux [RS98] avec des bénéfices en terme de rapport SNR (Signal to Noise Ratio) presque identiques à la même tâche sur un visage en vidéo complète [RJS96]. Même si cette présentation des stimuli ne contient aucune information sur les critères faciaux tels que la peau, les dents ou les ombres

produites par les mouvements de la bouche, elle produit des stimuli de « mouvement pur » dont l'information visuelle s'intègre parfaitement avec le signal acoustique correspondant.
Une étude plus récente de Bergeson et al. [BPR03] confirme que ces stimuli visuels peuvent être intégrés avec le signal audio pour rehausser l'intelligibilité. De plus, dans une autre étude récente, Santi et al. [SSV+03] 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.
Largement utilisé pour les lèvres et le visage (comme dans les travaux de Bailly et al. [BGO02]), le paradigme des points lumineux est plus difficilement applicable aux articulateurs de la parole, qui ne sont pas directement visibles, comme la langue ou le vélum. Comme on en a déjà parlé avec l'acquisition de données articulatoires, la capture des mouvements biologiques du conduit vocal est un domaine de recherche à part entière, qui fait appel, pour automatiser le processus d'extraction de données, à des techniques vidéo (flux optique, contours actifs, images clefs ou exemplaires…).

Pour beaucoup, les représentations dynamiques sont essentielles en parole et la théorie « shape-from-motion » initiée par Ullman [Ull76] est au cœur des débats. En 1976, Ullman prouve que trois vues orthographiques de quatre points non-coplanaires permettent de déterminer les distances entre ces points à un facteur d'échelle près. Cette théorie consiste donc à dire que lorsqu'un objet se déplace relativement à l'observateur, ou lorsque celui-ci bouge par rapport à l'objet observé, il est possible de recueillir les informations induites par ce mouvement relatif afin de reconstruire les surfaces ou les arêtes de cet objet. Mais le débat en parole entre forme et mouvement lancé par Strange et al. dans les années 1970 [SVSE76] autour de spécifications dynamiques des voyelles a évolué grâce aux études sur la perception visuelle en général. Les travaux de Cathiard [Cat94] montrent que même si le mouvement s'avère utile pour percevoir la forme (vision de face), il n'est pas toujours nécessaire (vision de profil). Ceci a mené à proposer une théorie combinée « shape-from-shading and shape-from-motion » qui s'inscrit bien dans les travaux en modélisation neurophysiologique de Giese et Poggio [GP03].

De plus, la production de la parole se place dans le cadre du contrôle du mouvement. Replacer le signal de parole dans ce cadre se situe dans l'héritage des idées de Stetson, qui écrivait dans *Motor Phonetics* en 1928 [Ste28] : *« Speech is rather a set of movements made audible than a set of sounds produced by movements. ».* Cela rejoint la théorie motrice de Liberman et Mattingly [LM85].

Parler n'est pas simplement émettre des phonèmes ou des syllabes les uns après les autres ; la production de la parole consiste à modeler et moduler rapidement le conduit vocal, par rotation et translation de la mandibule, portant les articulateurs essentiels que sont la langue et les lèvres. Les articulateurs se positionnent de façon à effectuer le geste acoustique attendu. La théorie motrice postule que la perception de la parole repose sur la perception des gestes articulatoires à partir des sons de la parole. Cette hypothèse est à la base des nombreux travaux sur l'inversion. Mais remonter des sons aux gestes est une entreprise beaucoup plus complexe et moins intuitive que dans le domaine visuel. Cette approche sera facilitée si l'on se replace dans un cadre équivalent à celui de Johansson [Joh73]. Pour cela, il faut disposer d'un procédé de marquage des données audio et des données gestuelles. A priori, les marqueurs audio, équivalent aux points Light Display, sont les pics formantiques.

Grâce à la cinéradiographie, nous avons à disposition des données dynamiques qui permettent d'observer les articulateurs du conduit vocal en mouvement. Dans la méthode que nous présentons, ce mouvement a été très largement exploité pour récupérer la forme de ces articulateurs sur une séquence complète. On différencie deux aspects :

(1) La première étape de la méthode correspondant à la phase manuelle de marquage utilise le mouvement pour faciliter le marquage d'images statiques. Par exemple, la forme de la langue qui est très souvent difficile à voir sur des images radiographiques statiques, a pu être marquée uniquement grâce à la possibilité de l'observer en mouvement, dans son contexte.

(2) La méthode elle-même reconstruit des données géométriques (de forme) dynamiques à partir de données vidéo dynamiques et de données géométriques statiques.

Nous rentrons maintenant dans le vif du sujet de cette thèse et présentons la méthode d'extraction semi-automatique qui a été développée.

Première partie : Méthode quasi-automatique d'extraction de mouvements à partir de données cinéradiographiques

CHAPITRE 1 : PRINCIPE DE LA MÉTHODE ET APPLICATION AUX MOUVEMENTS DE LA LANGUE DANS WIOLAND

1. Base de données Wioland

La base de données Wioland est une base de données cinéradiographiques du conduit vocal de l'Institut de Phonétique de Strasbourg. Cette base correspond à l'enregistrement d'un corpus élaboré dans le cadre du doctorat d'état de François Wioland, soutenu en Juin 1985 et portant sur les faits de jointure en français [Wio85]. Le corpus des phrases prononcées est en annexe A2.
L'enregistrement radiocinématographique a été réalisé le 28 février 1977 au Centre Médico-Chirurgical de Strasbourg-Schiltigheim, dans le service d'hémodynamique et d'exploration cardiovasculaire. 65 phrases ont été prononcées par un sujet de sexe féminin, âgé de 23 ans, né dans les Hauts de Seine et ayant grandi dans les Vosges. Le sujet avait pris connaissance du corpus et lors d'essais, les conditions réelles de l'enregistrement ont été simulées afin de faciliter au plus l'élocution et d'éviter le style récitatif.

Dans le cadre du programme « Ingénierie des Langues » du CNRS, cette base de données a été numérisée. Ceci permet de disposer d'une durée totale de 3 minutes 51 secondes, ce qui correspond à un ensemble de 5779 images radiologiques à la cadence de 25 images par seconde. Les images ont été tournées de 90° vers la droite. Ceci revient alors à considérer que la locutrice était en position allongée, le haut de la tête vers la droite et les lèvres vers le haut, lors de l'enregistrement.
Le film a été découpé de manière à ce que chaque phrase prononcée corresponde à une séquence (une seule répétition de chaque phrase a été conservée). On dispose ainsi de 65 séquences vidéo avec le signal audio correspondant. Les séquences ne sont pas de durée fixe, elles durent entre 2 et 8 secondes.
Au cours de l'enregistrement, la pellicule cinématographique s'est déchirée à 2 moments, ce qui a entraîné 2 sauts (un premier entre les images 2229 et 2230 et un second entre les images 4204 et 4205). Ces sauts ont provoqué des décalages au niveau du cadrage des images. Dans notre traitement de cette base, il a fallu prendre garde à ces décalages, comme nous le signalerons plus loin. On notera groupe 1 le groupe des images numérotées 1 à 2229, groupe 2 celui des images 2230 à 4204 et groupe 3 les images 4205 jusqu'à la dernière.

La séquence 63 est entièrement noire et empêche l'utilisation de 102 images. De la même façon, les images 3966 à 3969 sont à éliminer. Nous avons pour finir 5673 images utilisables. On note aussi qu'après l'image 4205, il y a une surexposition de la pellicule, les images sont « beaucoup plus blanches », mais nous intégrons ces images dans notre étude, comme nous l'expliquerons plus loin.
Les images au format bitmap (.bmp) sont de taille 720*540 pixels. Pour la suite, on nommera ces images « grandes images».

2. Méthode de rétro-marquage

Le principe de cette méthode est, dans un premier temps, de marquer manuellement le contour de la langue sur un nombre limité d'images de la base, avec quelques points bien choisis. Ces images sont appelées « images clefs ». Ensuite, de façon automatique, on marque l'ensemble des images de la base par association entre les marques géométriques des images clefs et des paramètres extraits du signal vidéo. C'est cette étape qu'on appelle rétro-marquage et qui consiste en une indexation automatique de la base entière via les images clefs marquées manuellement.

On utilisera les notations suivantes :
Soit N le nombre d'images de la base, N=5673
Soit $S = (S_t)_{t=1:N}$ la base entière, chaque image est notée S_t
Soit n le nombre d'images clefs
Soit $K = (K_i)_{i=1:n}$, $(K_i) \in \{1..N\}, K_i \in S$, K est l'ensemble des n images clefs

2.1. Etape manuelle : Marquage des images clefs

La première étape est une étape de marquage manuel, visant à décrire le contour de la langue. Nous choisissons un nombre limité de points pour définir ce contour. Le marquage d'une image consiste à placer correctement ces points sur l'image affichée à l'écran ; il s'agit, pour chacun d'eux, de fixer à la main une marque le long du contour de la langue, par simple clic de la souris.
Le choix de ces points repose principalement sur la détermination de degrés de liberté. Considérant la séquence complète et un point à déterminer, on fixe une coordonnée (abscisse ou ordonnée) pour ce point. Le choix de cette coordonnée est fait de telle manière que pour chaque image de la séquence, le point puisse toujours être marqué sur le contour de la langue. En effet, pour chaque image à marquer, la tâche consistera à placer une marque sur le contour à la coordonnée fixée, et ainsi à déterminer l'autre coordonnée. Ce

procédé est appliqué successivement pour le choix de plusieurs points. On choisit un espacement régulier entre eux pour permettre le tracé d'un contour de la langue réaliste lorsqu'on relie ces points. Chacun des points ainsi définis est alors spécifié par un degré de liberté (i.e. la coordonnée non fixée initialement).
Il est ensuite généralement souhaitable de laisser des points libres, c'est-à-dire sans aucune coordonnée fixée, il s'agit alors de points à 2 degrés de liberté. C'est le cas de la pointe, dans le cas présenté ici de la langue. On détaillera cela juste après. Laisser des points libres permet de prendre en compte plus de variabilité dans le mouvement, mais complique la tâche de marquage. En effet, il est alors nécessaire de prendre en compte un critère de marquage à appliquer sur toutes les images à traiter.

L'étape manuelle concerne, comme on l'a dit, un nombre restreint d'images, que l'on appelle images clefs. Nous choisissons aléatoirement n=100 images clefs (K_i) parmi les 5673 images S_t de la base, ce qui correspond à 1.75% de la base. Le choix aléatoire des images et le choix du nombre d'images seront justifiés plus loin lors de l'évaluation de la méthode.
Les points qui ont été préalablement choisis pour définir le contour sont tracés manuellement sur chacune des images clefs avec grande attention par un expert.

Une interface Matlab, décrite en annexe (A3), a été réalisée pour permettre et faciliter le marquage des images. En effet, grâce à cette interface et surtout au curseur (ou slider) que l'on actionne manuellement (Fig. 3a), il est possible de voir l'image à marquer dans son contexte, c'est-à-dire qu'on peut voir défiler les images adjacentes à l'image considérée. On visualise ainsi la langue en mouvement et dans de nombreux cas, ceci permet d'associer à l'image clef un contour qui était quasiment indiscernable sur l'image statique (Fig. 2). La figure suivante montre quelques exemples d'images de la base sur lesquelles il est difficile de distinguer le contour de la langue.

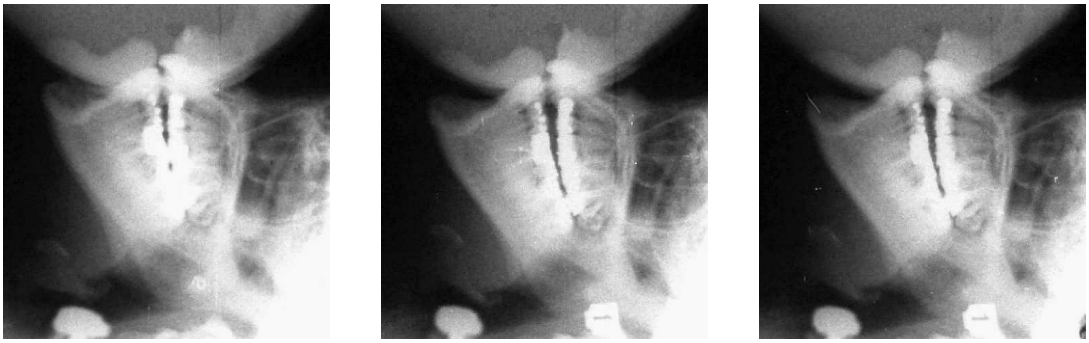


Figure 2 : Exemples d'images de la séquence cinéradiographique Wioland. Il est quasiment impossible de discerner correctement le contour de la langue sur ces images statiques.

Dans le cas de la langue pour Wioland, l'interface permet de marquer 10 points qui représentent le contour de l'articulateur.

Elle permet l'affichage de lignes verticales et horizontales (Fig. 3b), correspondant aux coordonnées fixées pour réduire à 1 le nombre de degrés de liberté (ddl). Ici, le choix de coordonnées fixes a été fait pour 8 des 10 points : les points 3 à 10 sont donc marqués à l'intersection entre les lignes et le contour de la langue. Pour chacun de ces 8 points, la ligne (horizontale ou verticale) fixe une des coordonnées, le degré de liberté porte alors sur l'autre coordonnée. Par exemple, pour le point 6, la coordonnée Y étant fixée, le degré de liberté porte sur la coordonnée X.

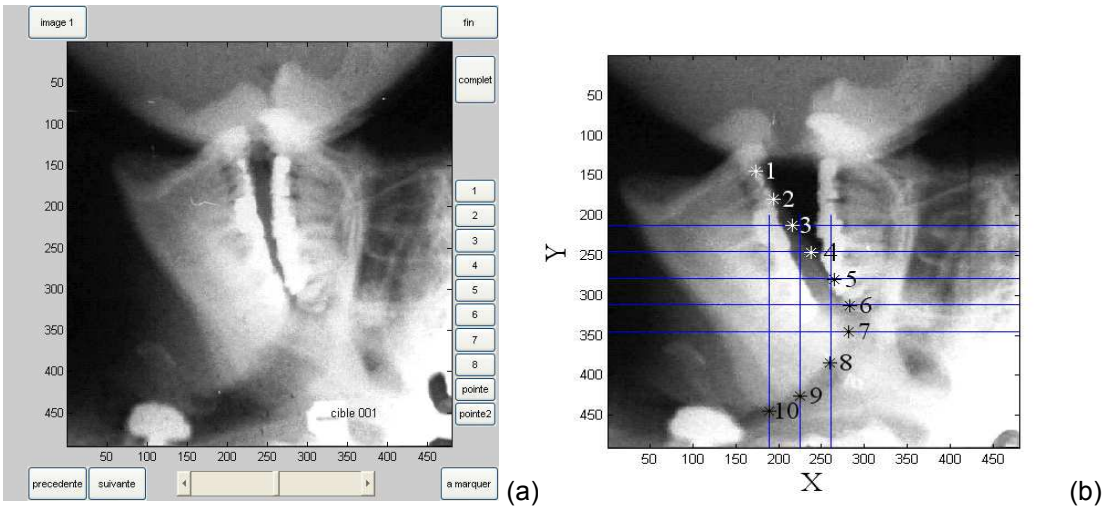


Figure 3 : (a) Interface de marquage Matlab (explications en annexe). Les bords de l'image ont été supprimés au maximum pour réduire la taille de l'image à 480*490 pixels, de manière à accélérer l'appel aux images avec le curseur.
(b) Lignes verticales et horizontales pour fixer les degrés de liberté des points 3 à 10 de la langue.

Les points 1 et 2 marquent la pointe de la langue. Le point 1 est libre, le point 2 est placé en fonction des points 1 et 3. Lorsque le premier point est marqué, l'ordonnée du point 2 est fixée, comme la moyenne des ordonnées des points 1 et 3. Le point 1 a deux degrés de liberté, le point 2 n'en a qu'un, son abscisse. Cependant nous considérerons par la suite l'ordonnée de ce point comme un degré de liberté supplémentaire, du fait de sa variabilité suivant les images.

Pour résumer, la position du contour de la langue est définie par 10 points, mais on parlera plutôt en termes de degrés de liberté. 8 ddl déterminent le dos et la base de la langue (les abscisses des points 3 à 7 et les ordonnées des points 8 à 10) et 4 déterminent la pointe. Le contour de la langue est donc spécifié par 12 degrés de liberté.

Le choix des lignes de marquage et des points libres a été fait de telle sorte qu'il n'y ait jamais de données manquantes, pour aucune des images de la base. En particulier, l'intersection entre une ligne fixée et le contour de la langue existe toujours.

Un contour de la langue, parfois irrégulier, est alors obtenu en reliant les 10 points marqués.
A ce stade, pour chaque image clef, nous disposons ainsi d'une configuration géométrique brute pour la langue. Chaque configuration correspond à l'ensemble des coordonnées X et Y des 10 points, mais plutôt, pour la plupart des points, aux coordonnées X **ou** Y. La configuration géométrique associée à l'image K_i est notée G_i.

Remarquons ici que la qualité du résultat final est largement conditionnée par la qualité de ce marquage manuel. Ce dernier est réalisé par l'un d'entre nous de façon à rendre compatible le marquage et le traitement automatique. Dans la suite du manuscrit, nous appellerons expert la personne ayant marqué manuellement les images clefs. La très grande majorité des données marquées ont été obtenues par l'intermédiaire d'un unique expert. Ceci limitera fortement l'évaluation inter-expert que l'on pourrait réaliser.

2.2. Etape automatique : Rétro-marquage sur l'ensemble de la base

L'étape principale du rétro-marquage est l'indexation automatique de la séquence S de départ (les 5673 images) à partir des 100 images clefs K. Ceci permet l'association entre les caractéristiques géométriques et les caractéristiques vidéos.

2.2.1. Caractéristiques vidéos

Les caractéristiques vidéos considérées dans notre étude sont les coefficients DCT basses fréquences des images de la base.
Les coefficients DCT d'une image correspondent aux coefficients de la Transformée en Cosinus Discrète de l'image. La Transformée en cosinus discret ou TCD (de l'anglais : DCT ou Discrete Cosine Transform) a été décrite précédemment et permet de faire passer l'information de l'image du domaine spatial en une représentation identique dans le domaine fréquentiel. Ce changement de domaine permet de représenter l'intégralité de l'information de l'image sur très peu de coefficients, correspondant à des fréquences plutôt basses. Le résultat fourni est représenté dans une matrice, les basses fréquences se trouvant en haut à gauche de la matrice, et les fréquences augmentant vers la droite et vers le bas.

Au préalable, pour notre étude les images sont soumises à quelques traitements simples, pour le calcul de ces coefficients. Tout d'abord les images sont identiques sur 3 plans, aussi un seul de ces plans est suffisant pour le calcul DCT. Ensuite, les images de départ (720*540 pixels) sont redimensionnées (réduction par 4,5 en abscisse et en ordonnée).
Nous nous intéressons au mouvement de la langue et cherchons à minimiser l'interférence avec d'autres articulateurs ou des parasites. Les images sont donc restreintes à un cadre

minimal d'observation de la langue (l'articulateur cible) pour tout le film. Les inscriptions manuelles visibles ou les bords qui ne comportent pas d'information intéressante sur le conduit vocal sont supprimés en découpant les images.

Les imagettes ainsi obtenues (136*108 pixels) sont, de plus, recadrées entre les 3 groupes d'images de la base et donc re-découpées de manière à éliminer « les décalages » dus aux déchirements de la pellicule. On constate un décalage de 1 pixel vers le haut et 3 vers la droite entre les 2 premiers groupes et un décalage de 3 pixels vers le haut et 3 pixels vers la droite entre le groupe 2 et le groupe 3. Ces décalages sont pris en considération dans le découpage, schématisé ci-dessous (Fig. 4).

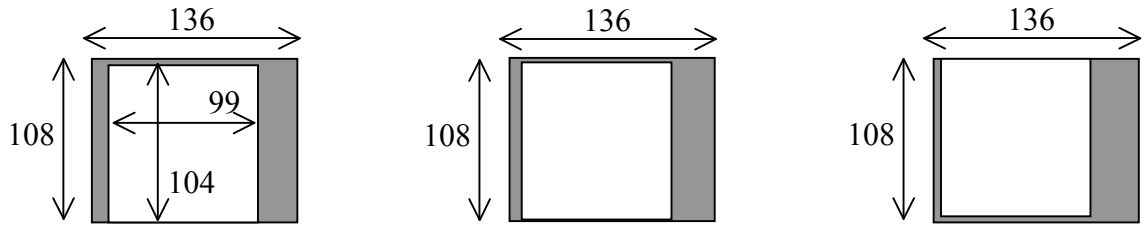


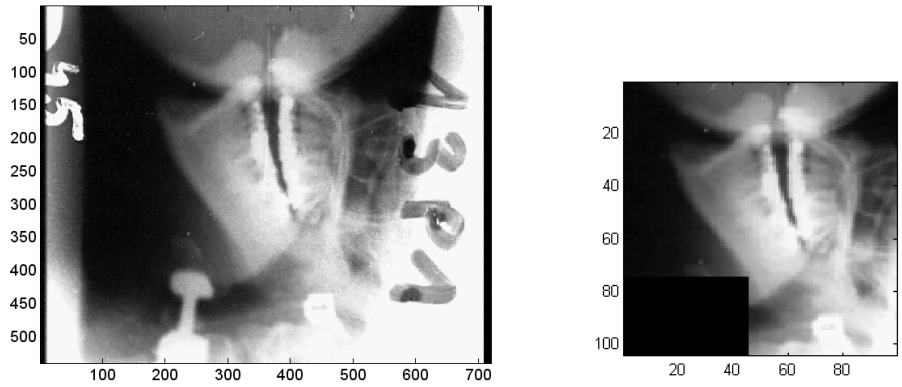
Figure 4 : Découpages réalisés sur les images redimensionnées (136*108 pixels) pour atténuer les effets de décalages entre les groupes d'images dans la séquence (groupe 1 à 3 de gauche à droite). Les cadres gris correspondent aux images 136*108, les cadres blancs aux images 99*104 pixels, qui sont dites « réduites » et qui sont utilisées pour le calcul des coefficients DCT.

Le cadrage des données est indispensable car la paramétrisation DCT y est très sensible [HBSK03]. Mis à part ces « décalages » dus aux déchirements de la pellicule, on constate, sur chacun des trois segments de la séquence (correspondant aux 3 groupes d'images), un des avantages de la cinéradiographie : elle est généralement bien cadrée, ce qui motive ici l'utilisation de la DCT en image pleine.

On dispose désormais d'images de taille 99*104 pixels, que l'on nommera par la suite « images réduites » ou « imagettes » (Fig. 5b).

Sur les images, on observe aussi une sorte de piston « blanc », qui correspond à la tige de plomb du synchronisateur. Cette tige se relève au début de chacune des séquences puis se rabaisse ensuite. Pour éviter de prendre en compte le mouvement de cette tige, nous avons mis en place, pour le calcul des coefficients DCT, un cache noir de taille 45*30 pixels pour masquer totalement l'effet de la tige (il s'agit de mettre à 0 les pixels de la zone considérée). Ce cache est placé dans le coin en bas à gauche de l'imagette.

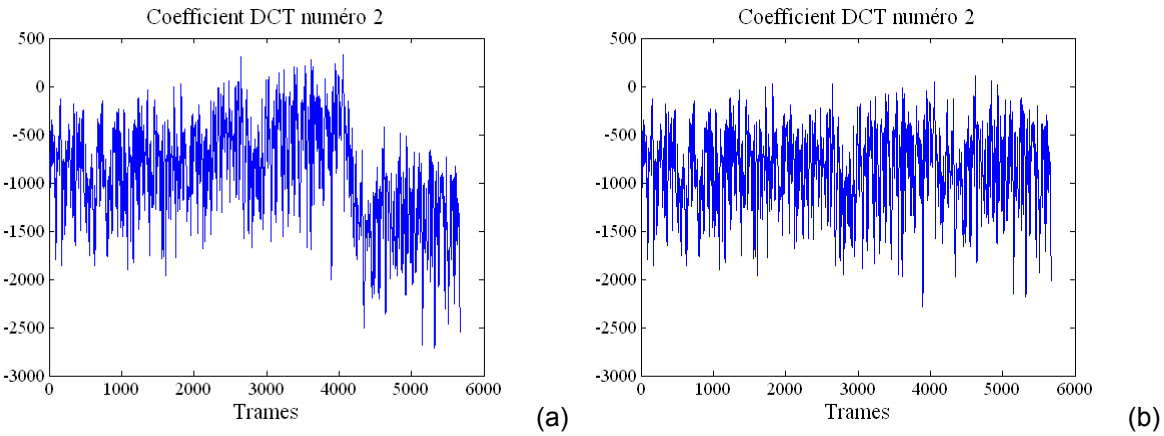
C'est finalement sur ces images réduites avec le cadre noir que nous calculons une DCT sur chacune des dimensions.



(a) (b)

Figure 5 : (a) « Grande » image (720*540 pixels) d'origine utilisée pour le marquage.
(b) Image « réduite » (99*104 pixels) avec cache noir, pour le calcul des coefficients DCT.

Pour obtenir une base homogène de coefficients DCT sur les 5673 images, il reste à moyenner chaque coefficient sur toutes les images, ceci permet en particulier de diminuer l'effet de surexposition de la fin de la base de données (à partir de l'image 4205).
Les 2 graphes suivants (Fig. 6) mettent en évidence l'effet du moyennage sur les coefficients DCT, en visualisant le coefficient numéro 2. Sur le graphe de gauche (Fig. 6a), avant moyennage, on distingue assez nettement les 3 « morceaux » de la base, sous la forme de 3 paliers (le 3ème étant très distinct des 2 autres du fait de la forte différence de contraste). Le graphe de droite (Fig. 6b) après moyennage rétablit une uniformité entre les 3 groupes.
Cette correction de moyenne est réalisée pour l'ensemble des coefficients DCT calculés.



(a) (b)

Figure 6 : (a) Coefficient DCT numéro 2 calculé sur les 5673 images de la base.
(b) Correction sur la moyenne des coefficients DCT, observée sur le coefficient 2.

Seuls les coefficients basses fréquences sont utiles pour suivre le mouvement de la langue. On se limite à ces coefficients-là et plus exactement au cadre supérieur gauche 24*24 des coefficients de la DCT. Le premier coefficient, qui correspond à la moyenne de l'image, n'est

pas informatif et est donc écarté ; nous conservons donc 575 coefficients DCT basses fréquences. Le choix du nombre de coefficients DCT pris en compte sera analysé plus loin.

2.2.2. Indexation automatique

Les caractéristiques vidéos sont donc les coefficients DCT Basses Fréquences de chaque image, comme on vient de le voir. Les caractéristiques géométriques sont les coordonnées des points marqués sur la langue pour les *K* images clefs.

L'association entre les caractéristiques géométriques et les caractéristiques vidéos se fait par indexation automatique, de façon à ce que pour chaque image de la séquence, l'index de l'image clef la plus proche lui soit assigné.

L'indexation automatique consiste à quantifier la séquence vidéo S à partir des images clefs K. On recherche pour chaque image S_t de la base l'image clef K_j qui est la plus proche. Ceci fait appel à une notion de distance et on définit alors une mesure de similarité, qui correspond à la distance euclidienne entre les coefficients DCT basses fréquences des deux images. Le premier coefficient est omis dans le calcul de cette distance.

Pour chaque image S_t de la base, on mesure la similarité entre l'image et chacune des *n* images clefs. L'image clef correspondant à la distance la plus faible permet de définir pour l'image S_t un index j correspondant au numéro de cette image clef.

$$j = index_K(S_t) = \arg\min_i \sqrt{\sum_{p=2}^{24*24} (DCT_p(S_t) - DCT_p(K_i))^2}$$

Ainsi, à chaque image de la base est associé l'index de l'image clef la plus proche.

La répartition des images de la base sur les 100 index des images clefs est relativement uniforme, comme le montre l'histogramme de répartition (Fig. 7a) qui présente une allure assez plate. Chaque index est représenté et pour chacun des 3 groupes de la base (relatifs aux déchirements de la pellicule), les images sont indexées par des images clefs appartenant aux 3 groupes. Ceci est mis en évidence par l'histogramme de la figure 7b. Les images des 3 groupes sont représentées en fonction de la position dans la séquence des images clefs qui les indexent. Si on considère par exemple, les images du second groupe (milieu de la séquence), on observe qu'elles sont indexées par des images clefs présentes n'importe où dans la séquence, avec néanmoins une petite tendance à être indexées par des images clefs du groupe 2.

Nous avons donc corrigé l'hétérogénéité apparente de la base d'origine.

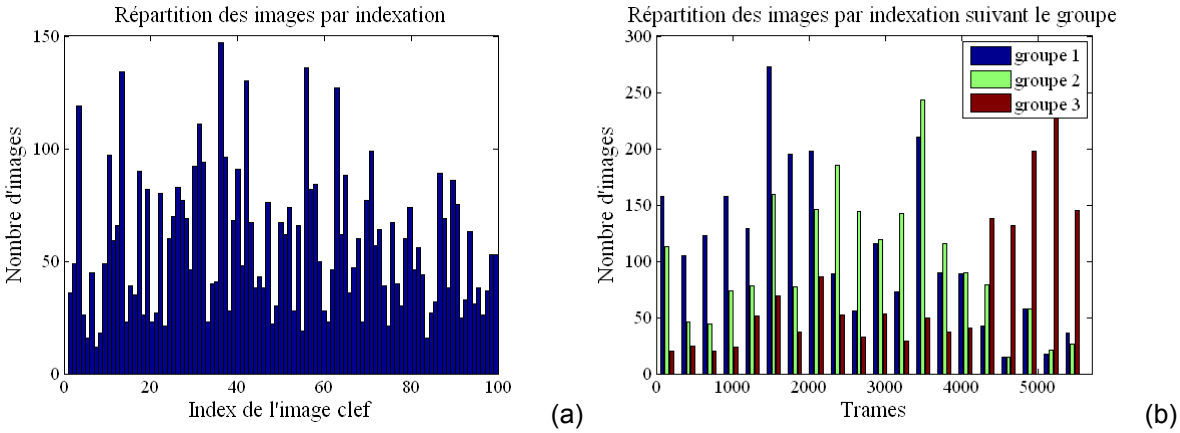


Figure 7 : (a) Histogramme de répartition des images de la base en fonction de l'indexation par les images clefs. La répartition est uniforme.
(b) Les images d'un groupe de la séquence (début, milieu ou fin) ne sont pas uniquement indexées par des images clefs appartenant à ce groupe.

La seconde étape du rétro-marquage est le marquage géométrique de la base d'origine.

On associe sur les images de départ (S_t), via l'indexation, l'information géométrique récupérée sur les images clefs. A chaque index i correspond une image clef K_i et la configuration géométrique associée G_i (les 10 points marqués pour définir la position de la langue sur cette image). Cette configuration géométrique G_i est associée à toutes les images de la base qui portent cet index i.

Cette association par indexation permet ainsi de restaurer de l'information géométrique à partir de l'information vidéo et de reconstituer le mouvement à partir d'images clefs.

2.2.3. Remarque : Rétro-marquage automatique

On note que le rétro-marquage peut être rendu entièrement automatique lorsque les informations géométriques sont extractibles dans les images clefs par des méthodes automatiques. En effet, pour le cas de la main par exemple, il est tout à fait possible de marquer automatiquement les degrés de liberté. On définit des points à des endroits stratégiques de la main (articulations, bouts des doigts…) qu'il est relativement aisé de suivre automatiquement. Des méthodes, difficilement applicables sur des séquences vidéo entières (par ex. contours actifs) à cause de leur complexité algorithmique, deviennent accessibles.

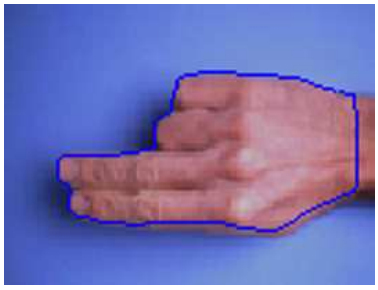


Figure 8 : Extraction automatique de gestes de la main à partir de contours actifs.

Mais dans le cas de la langue, la tâche d'extraction géométrique est très difficile même pour l'expert humain. Elle est donc dévolue au marquage manuel, qui est réalisé dans les conditions de facilitation que nous avons décrites.
Ensuite, plutôt que de placer l'information géométrique et de la suivre, nous allons reconstruire le mouvement par interpolation temporelle. A partir des données géométriques statiques extraites et de l'information vidéo dynamique disponible, nous reconstruisons des données géométriques (de contours) dynamiques.

3. Reconstruction du mouvement

A ce stade du développement, les mouvements de la langue ont été en partie reconstruits. En superposant les configurations géométriques obtenues sur les images de la séquence de départ, on réalise des vidéos et on observe que ces dernières sont saccadées. En effet, nous ne disposons que de 100 configurations des points du contour de la langue ; la quantification est grossière, puisque l'on est passé de 5673 à 100 états. On cherche à atténuer cette erreur de quantification pour améliorer la reconstruction du mouvement de la langue. Pour cela, il faut rétablir la continuité, au niveau de l'information géométrique.
Dans ce paragraphe, à titre d'observations, on utilisera des représentations intermédiaires. En effet, nos données vidéos (coefficients DCT des images) et nos données géométriques (degrés de liberté de marquage des images clefs) sont de dimensions respectives 575 et 12. Pour représenter les données des espaces vidéo et géométrique, qui sont de grande dimension, on calcule des analyses en composantes principales (ACP). En ne conservant que les 2 premières composantes principales, on représente alors nos données dans un plan, appelé plan principal. On dispose ainsi d'un plan principal vidéo (résultant de l'ACP sur les coefficients DCT des images) et un plan principal géométrique (résultant de l'ACP sur les 12 degrés de liberté des configurations géométriques de la langue).

L'amélioration de la reconstruction du mouvement se fait à 2 niveaux.

Tout d'abord, on s'intéresse au rétablissement de la continuité temporelle et on cherche à réduire les effets de quantification en filtrant temporellement les caractéristiques géométriques. Ce filtrage est choisi après une analyse fréquentielle détaillée, comme on le verra juste après.
Ensuite, nous essayons de compenser les irrégularités de projection entre les représentations vidéo et géométrique. Comme on le verra un peu plus loin, en observant les ACP sur les données vidéo et géométriques, on met en évidence des irrégularités entre les projections. Ce sont ces irrégularités que l'on tente de diminuer pour réduire les erreurs de reconstruction de mouvement.
Ces deux niveaux d'amélioration sont indépendants et vont permettre de jouer sur deux aspects pour reconstruire le mouvement des points de marquage.

3.1. Note sur l'ACP

L'analyse en composantes principales, communément appelée ACP, est une méthode statistique multidimensionnelle qui permet de représenter un ensemble de données en identifiant la redondance dans celles-ci. Il s'agit, comme la DCT ou la FFT, d'une transformation bijective, qui permet d'analyser les données observées dans une base où elles sont décorrélées. Elle fournit notamment une synthèse graphique des résultats. L'analyse en composantes principales fait partie d'une famille de techniques statistiques - les méthodes multidimensionnelles - utilisées pour traiter des données provenant de situations où plusieurs variables sont mesurées. Lorsque plusieurs mesures continues sont observées sur un ensemble d'individus ou d'objets, il est rare que toutes les mesures prises soient indépendantes. L'ACP permet de mesurer la redondance dans ces mesures et le nombre de paramètres nécessaires pour les caractériser. L'observateur a intérêt à multiplier les capteurs afin de rendre ses observations plus robustes et à ne pas manquer des informations importantes.

3.1.1. Introduction générale

L'analyse en composantes principales est une méthode mathématique qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre n variables aléatoires. L'ACP est une méthode statistique qui a pour but de comprendre et de visualiser comment les effets de phénomènes plus ou moins indépendants se combinent. Lorsque l'on ne considère que deux effets, il est usuel de caractériser leurs effets conjoints via le coefficient de corrélation (son seul défaut est de ne prendre en compte que des effets conjoints linéaires).

Lorsque l'on se place en dimension deux, les points disponibles sont représentés sur un plan. Le résultat d'une ACP sur ce plan est de déterminer les deux axes qui expliquent le mieux la dispersion des points disponibles.

Lorsqu'il y a plus de deux effets, par exemple trois effets X_1, X_2, X_3, il y a trois coefficients de corrélations à prendre en compte: $C(X_1,X_2)$, $C(X_1,X_3)$, $C(X_2,X_3)$. La question qui a donné naissance à l'ACP est : "comment avoir une intuition rapide des effets conjoints ?".

En dimension plus grande que deux, une ACP va toujours déterminer les axes (si on est en dimension N, il y aura N axes à déterminer), qui expliquent le mieux la dispersion du nuage des points disponibles. Elle va aussi les ordonner par « inertie expliquée ».

Si on ne décide de ne retenir que les deux premiers axes de l'ACP, on pourra alors projeter notre nuage de dimension N sur un plan, et le visualiser.

L'ACP est une méthode statistique descriptive, mais même si elle est majoritairement utilisée pour visualiser des données, il ne faut pas oublier que c'est aussi un moyen :

- de décorréler ces données (dans la nouvelle base, constituée des nouveaux axes, les points ont une corrélation nulle),
- de débruiter ces données (en considérant que les axes que l'on décide d'oublier sont des axes bruités),
- et même de classifier ces données (en clusters corrélés).

Les autres méthodes de compressions statistiques habituelles sont:

- l'Analyse en Composantes Indépendantes,
- les cartes auto-adaptatives (SOM : self organizing maps), appelées aussi cartes de Kohönen,
- l'Analyse en Composantes Curvilignes,
- la compression par ondelettes.

3.1.2. Utilisation de l'ACP

Soit un ensemble de données, soumis à une analyse en composantes principales. Il est constitué de p variables aléatoires $X_1, \ldots, X_p$ connues à partir d'un échantillon de N observations de ces variables[2]. Cet échantillon de p variables aléatoires peut être structuré dans une matrice à N lignes et p colonnes.

[2] Dans notre étude, à titre d'illustration (§3.1.2.), on dispose de N=5673 images et on cherche la corrélation entre les 575 coefficients DCT (p=575) ou entre les 12 degrés de liberté (p=12).

$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,p} \end{bmatrix}$ où $x_{j,i}$ représente l'observation j de la variable i.

La matrice X est généralement centrée et réduite. En la multipliant par sa transposée, on obtient la matrice de corrélation des $X_1, \ldots, X_p$, estimée sur la base des N observations. Cette matrice est carrée (de taille p), symétrique et réelle. Elle est donc diagonalisable dans une base orthonormée.

Le principe de l'ACP est de trouver un axe u, issu d'une combinaison linéaire des X_p, tel que la variance du nuage selon cet axe soit maximale. C'est-à-dire que nous cherchons le vecteur u tel que la projection du nuage sur u ait une variance maximale.

On calcule les valeurs propres et les vecteurs propres de la matrice de corrélation. Les valeurs propres calculées, et les vecteurs propres qui leur correspondent, sont ordonnés par valeurs décroissantes : le vecteur qui explique le plus d'inertie du nuage est le premier vecteur propre. De même le deuxième vecteur qui explique la plus grande part de l'inertie restante est le deuxième vecteur propre, et ainsi de suite.

On construit alors une matrice Q, telle que la $k^{\text{ième}}$ colonne de Q corresponde au $k^{\text{ième}}$ vecteur propre. On obtient ainsi une transformation linéaire du vecteur observé :

$Y = Q^T (X - \overline{X})$, où $\overline{X}$ représente le vecteur moyen de X.

$Y = (y_1, y_2 ....., y_p)^T$, où les y_k sont appelés composantes principales.

Comme la matrice Q est orthogonale, la transformation inverse est donnée par : $X = QY + \overline{X}$.

Le fait remarquable des composantes principales est le suivant : alors que la série complète des p composantes principales représente exactement le vecteur original X, il est possible de ne garder pour une approximation donnée que les n premières composantes ($n<p$), en sachant que ces n facteurs contribuent plus à la variance de X que n'importe quelle autre série de n facteurs orthogonaux. L'approximation selon cette méthode et avec cette contrainte sera la meilleure.

A des fins de représentation, il est habituel de ne conserver que les 2 ou 3 premières composantes principales, on parle alors de plan principal avec une première composante en abscisse et une seconde en ordonnée.

Soit y_1 et y_2 les deux premières composantes principales calculées, en notant Y_{12} le vecteur constitué de ces 2 composantes ($Y_{12} = (y_1, y_2)^T$), on obtient la projection des données de départ dans le plan principal par la formule $Z = X * Y_{12}$.

$Z = \begin{bmatrix} z_{1,1} & z_{1,2} \\ \vdots & \vdots \\ z_{N,1} & z_{N,2} \end{bmatrix}$ où $z_{j,1}$ (resp. $z_{j,2}$) représente l'abscisse (resp. l'ordonnée) de l'observation j dans le plan principal.

3.1.3. Application

Ce type d'analyse est couramment utilisé dans les études en parole qui traitent de modélisation articulatoire statistique. Entre autres, les travaux de Maeda [Mae90] ou Badin [BBB+00] utilisent des analyses en composantes principales pour extraire de leurs données des paramètres articulatoires.

Nous appliquons l'ACP sur nos données. L'objectif n'est pas d'identifier des commandes articulatoires, mais plutôt d'accéder à une représentation de nos données (de grandes dimensions) dans un plan. Pour les données vidéo (5673 observations de 575 variables, correspondant aux coefficients DCT basses fréquences), les 2 premières composantes de l'ACP expliquent 77% de la variance (66 % pour la 1ère composante et 11% pour la seconde). On projette (Fig. 9a) les données vidéo dans le plan principal, constitué par ces 2 premières composantes. De même pour les données géométriques (5673 observations de 12 variables, correspondant aux degrés de liberté sur la langue), les 2 premières composantes de l'ACP expliquent 90% de la variance (78%+12%). On projette aussi (Fig. 9b) ces données géométriques dans le plan principal, constitué par ces 2 premières composantes.

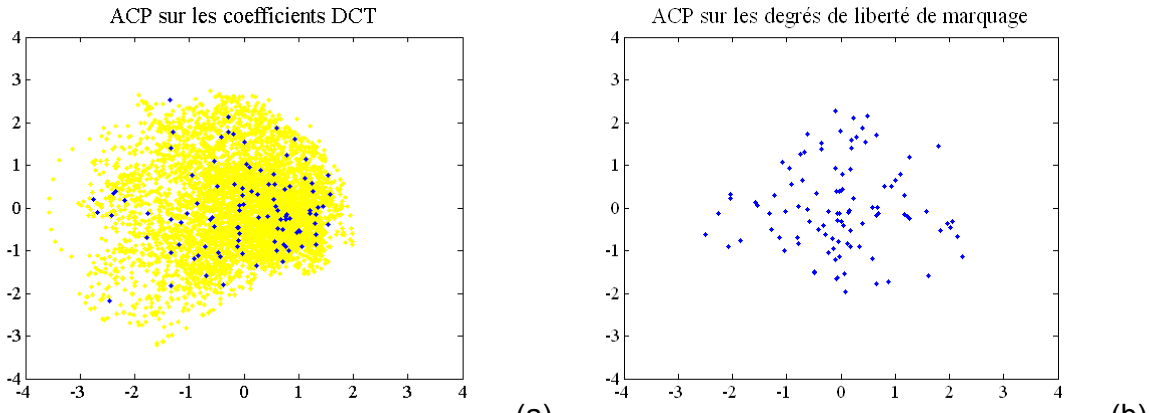


Figure 9 : Exemples de représentations dans le plan principal.
(a) Plan principal vidéo résultant de l'ACP sur les coefficients DCT de la base (les points bleus représentent les images clefs et les jaunes les autres images de la base).
(b) Plan principal géométrique résultant de l'ACP sur les 12 degrés de liberté de marquage de la langue des 100 images clefs.

Les plans principaux présentés ci-dessus et dans la suite de ce manuscrit sont obtenus à partir de données centrées et réduites. Les axes de tous ces graphes seront donc toujours sans dimension, normalisés et centrés autour de 0.

3.2. Continuité temporelle

Pour réduire le bruit de quantification dû à l'indexation et les sauts dans la séquence reconstruite à partir des images clefs, on cherche à filtrer le signal indexé par un filtre passe-bas. La valeur de la fréquence de coupure de ce filtre est déterminée à partir d'une analyse fréquentielle du mouvement. On calcule de façon systématique les densités spectrales de puissance (DSP) des données dont on dispose, d'abord sur les données vidéos puis sur les données géométriques.

3.2.1. Analyse fréquentielle sur la séquence vidéo

L'analyse fréquentielle sur les données vidéo est réalisée sur les images réduites (99*104 pixels), celles sur lesquelles on a calculé les coefficients DCT qui ont ensuite été utilisés dans l'étape de rétro-marquage.

Des régions spécifiques de ces images (ou cadres) [PYS+06] ont été choisies pour représenter le mouvement (Fig. 10a). Le calcul des densités spectrales de puissance se fait pour chacun des pixels des 3 cadres. On moyenne ensuite pour chaque cadre les DSP des pixels concernés (Fig. 10b).

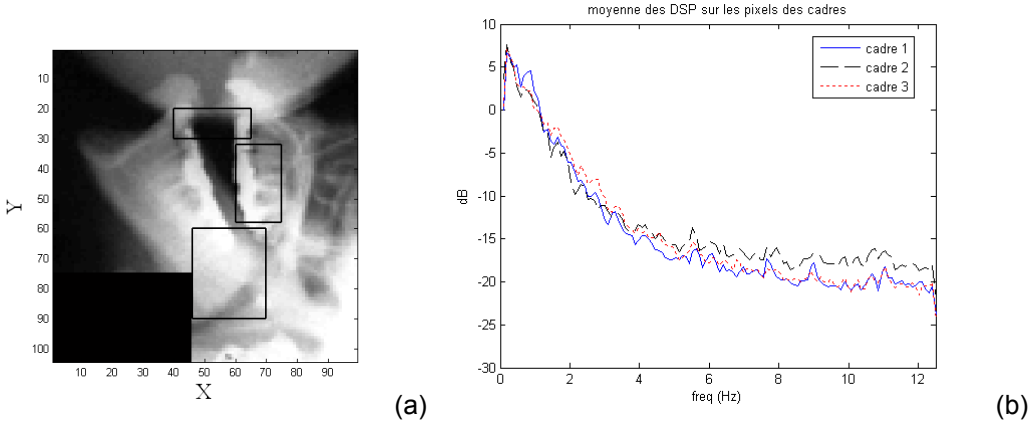


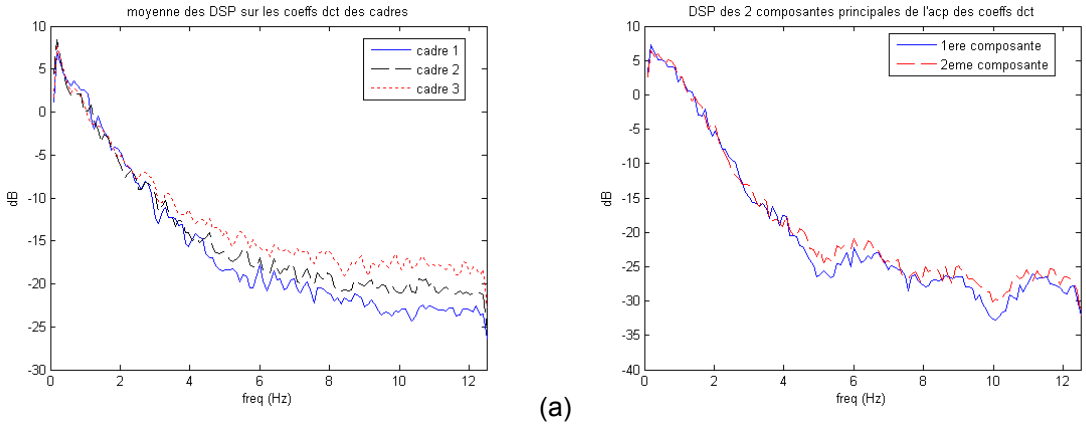
Figure 10 : (a) Cadres choisis pour l'analyse fréquentielle du mouvement.
(b) Moyenne des DSP sur les pixels de chacun des cadres.

Le profil spectral observé permet de dire que jusque vers 4 Hz le mouvement est identique pour les 3 cadres observés de l'image : les 3 courbes se superposent jusqu'à cette valeur de fréquence. Au-delà, les courbes sont moins bien superposées, on estime qu'il s'agit de bruit qu'il n'est pas utile de prendre en compte. On estime à –15 dB le point de séparation entre le

mouvement et le bruit, ce qui correspond à une fréquence de 3,75 Hz. Cette analyse permet ainsi de choisir une fréquence de coupure à 3,75 Hz.

La même allure spectrale est obtenue en calculant de nouvelles densités spectrales de puissance, mais cette fois, sur des coefficients DCT de ces mêmes régions choisies. On retrouve cette valeur de fréquence de coupure en moyennant sur chaque cadre les DSP des coefficients DCT de ces cadres (Fig. 11a).

Comme une grande partie de la variance des données vidéo est expliquée par les 2 premières composantes de l'ACP sur ces données, le calcul des densités spectrales de puissance est également réalisé sur les composantes du plan principal des coefficients DCT et on observe à nouveau la même allure spectrale (Fig. 11b).



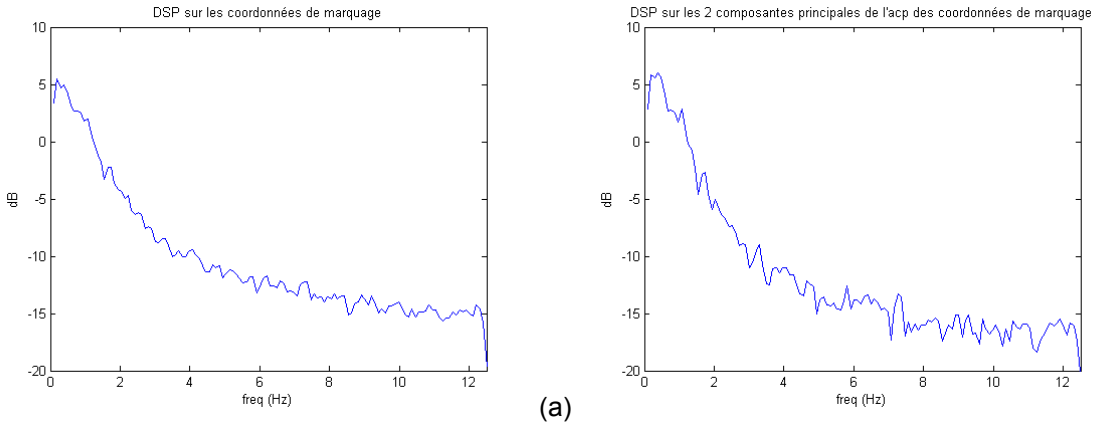
(a) (b)

Figure 11 : (a) Moyenne des DSP sur les coefficients DCT de chacun des cadres.
(b) DSP sur les composantes 1 et 2 du plan principal vidéo.

Cette fréquence de coupure à 3,75Hz est obtenue ici pour la séquence vidéo numérisée à 25 images par seconde Si on la rapporte à la fréquence de l'enregistrement d'origine (66 im/s), la fréquence de coupure est alors de 10Hz. Ce résultat est comparé à celui trouvé par [PYS+06]. L'étude réalisée sur les mouvements de la face de locuteurs anglais ou japonais met en évidence une fréquence de coupure de l'ordre de 6Hz au delà de laquelle le filtrage n'a pas d'influence sur l'intelligibilité. Les mouvements de la langue sont un peu plus rapides.

3.2.2. Analyse fréquentielle sur les données géométriques

Une analyse fréquentielle analogue est réalisée sur les données géométriques, d'abord en moyennant les densités spectrales de puissance des 12 degrés de liberté des configurations géométriques (Fig. 12a) puis en moyennant les DSP des 2 composantes principales de l'ACP sur les données géométriques issues de l'indexation (Fig. 12b).



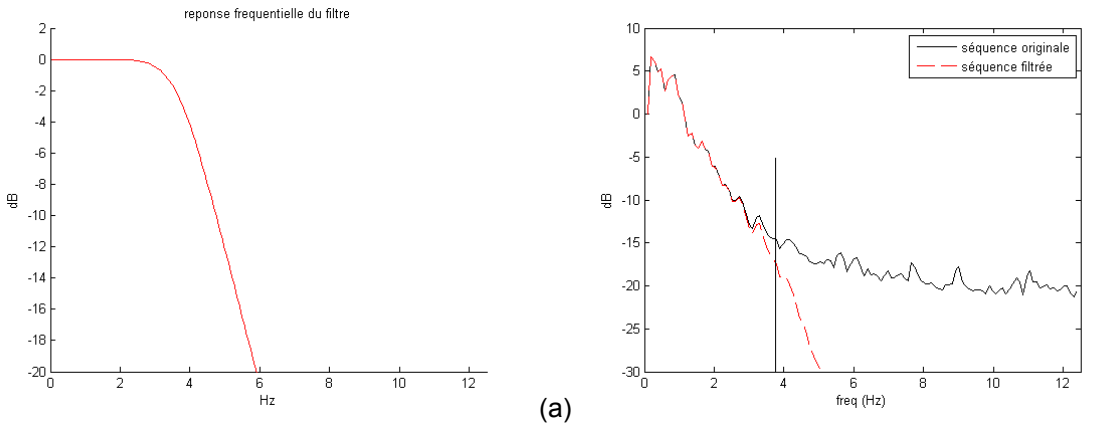
(a) (b)

Figure 12 : (a) Moyenne des DSP sur les coordonnées de marquage (12 ddl).
(b) Moyenne des DSP sur les composantes 1 et 2 du plan principal géométrique.

L'allure spectrale obtenue est équivalente. La bande passante (Fc=3,75 Hz) observée précédemment pour les données vidéo est approximativement la même pour l'information géométrique.

3.2.3. Filtrage temporel passe-bas de l'information géométrique

Cette étude permet ainsi de choisir correctement le filtre temporel à appliquer à la séquence reconstruite à partir des images clefs pour réduire les sauts de quantification. On choisit un filtre passe-bas de type Butterworth d'ordre[3] 4 de fréquence de coupure 3,75 Hz, permettant de conserver intact les fréquences inférieures à 3,75Hz et de supprimer celles supérieures (Fig. 13a).



(a) (b)

Figure 13 : (a) Réponse fréquentielle du filtre choisi et appliqué aux données géométriques.
(b) Moyenne sur les 3 cadres des DSP moyennes sur les pixels pour la séquence de départ et la séquence filtrée.

[3] Le filtrage a été effectué par la fonction Matlab « *filtfilt* ». Il s'agit d'un filtrage non causal qui annule le déphasage entre l'entrée et la sortie et qui double l'ordre du filtre ; tout se passe comme si on filtrait le signal deux fois par le même filtre, une fois dans une direction et une fois dans l'autre.

L'information géométrique est filtrée temporellement (Fig. 13b). Le mouvement est moins saccadé, la reconstruction est meilleure. L'apport de ce filtrage sera évalué plus loin, dans la partie évaluation.

3.3. Régularité des projections

Nous nous intéressons maintenant à l'autre aspect d'amélioration de reconstruction du mouvement et cherchons à compenser les irrégularités du « mapping » entre les représentations vidéo et géométrique. Pour cette étude, nous travaillons avec les données de départ, non filtrées et nous oublions pour un temps le filtrage temporel qui vient d'être expliqué.

3.3.1. Observation de la dispersion entre les 2 représentations

Les représentations utilisées sont les projections de nos données vidéos et géométriques dans les plans principaux associés, obtenus par analyse en composantes principales sur les coefficients DCT des 100 images clefs et sur les degrés de liberté de marquage de ces 100 images clefs. Ces plans sont centrés et normalisés. Ces représentations ont pour intérêt de permettre une visualisation 2D de nos données multi-dimensionnelles.

Pour un point P donné de l'espace vidéo, correspondant à une image I parmi les images clefs, nous considérons, autour de ce point, son voisinage, c'est-à-dire ses k voisins les plus proches, du point de vue des données vidéos. La notion de distance considérée est la distance euclidienne entre les points de cet espace.

En prenant k=10 par exemple, nous traçons le cercle centré sur le point P en question et dont le rayon est égal à la distance moyenne de ces 10 plus proches voisins de l'espace vidéo. De la même façon, nous construisons un cercle équivalent dans l'espace géométrique.

Soit le point Q de l'espace géométrique correspondant au point P de l'espace vidéo, c'est-à-dire le point de l'espace géométrique correspondant à la configuration géométrique associée à l'image clef I. Le voisinage géométrique du point Q est représenté par le cercle centré sur le point Q et dont le rayon est égal à la distance moyenne de ces 10 plus proches voisins de l'espace géométrique. La distance considérée est toujours la distance euclidienne dans l'espace en question.

Nous observons, à titre d'exemples, les cercles de voisinage de certains points sur la figure 14. Cette représentation permet d'observer une augmentation de la longueur des rayons des cercles dans l'espace géométrique par rapport à l'espace vidéo et un recouvrement plus important des cercles dans l'espace géométrique que dans l'espace vidéo. Ceci est une première mise en évidence des irrégularités entre les 2 projections.

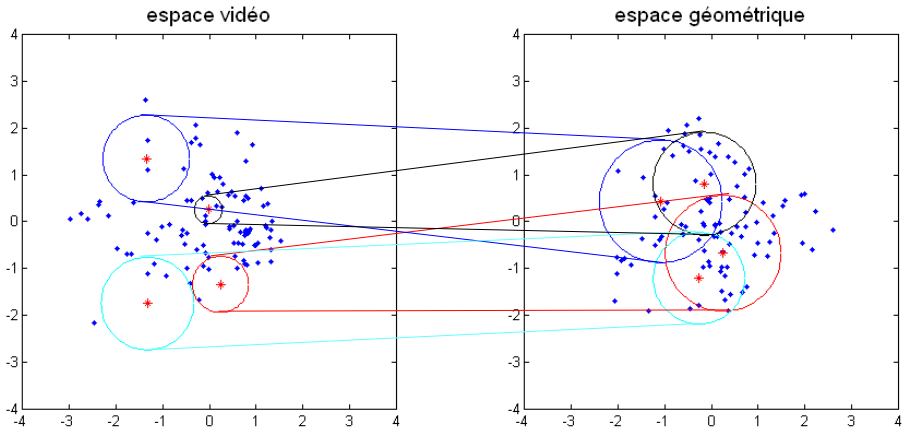


Figure 14 : Mise en évidence des irrégularités de projection entre espaces vidéo et géométrique. Les cercles représentent les voisinages (10 plus proches voisins) des points marqués d'une croix rouge.

Cette irrégularité est révélée d'un point de vue dynamique, en observant la projection des trajectoires vidéo et géométrique.

En effet, on peut mettre en parallèle d'un côté le plan principal des coefficients DCT de toutes les images de la base et d'un autre celui sur les degrés de liberté de marquage de ces 100 images clefs (ces plans sont toujours centrés et normalisés). Dans chacun de ces plans, des trajectoires sont projetées figure 15. Dans le plan vidéo (Fig. 15a), une trajectoire est générée à partir de quelques images de la séquence vidéo de départ en reliant les points associés à ces images. Dans le plan géométrique (Fig. 15b), la trajectoire équivalente est générée via l'indexation en reliant les points associés aux index des clefs pour les images du bout de séquence considéré.

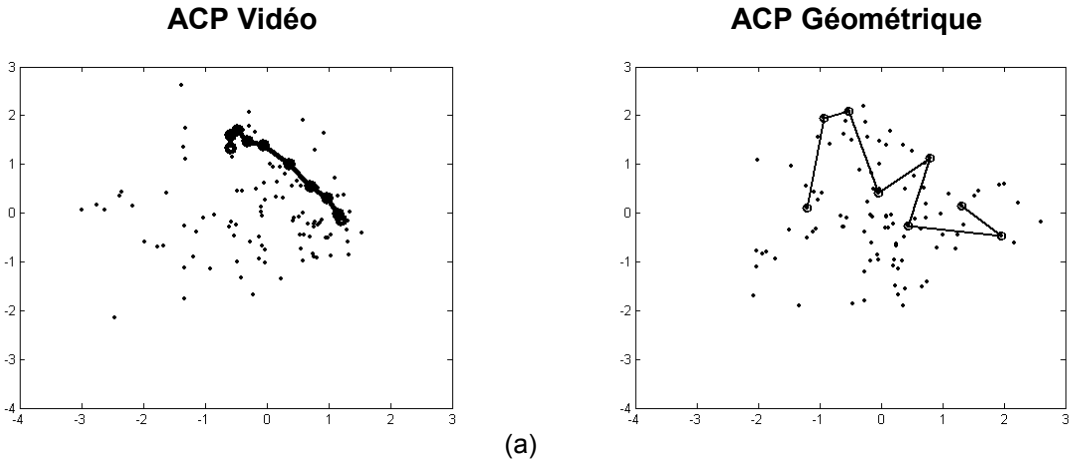


Figure 15 : (a) Trajectoire projetée dans le plan principal de l'ACP vidéo. (b) Trajectoire générée via l'indexation dans le plan géométrique.

Deux points proches dans l'espace vidéo ne le sont pas forcément dans l'espace géométrique. La trajectoire dans l'espace géométrique est nettement plus discontinue que

dans l'espace vidéo. Ce sont ces discontinuités que l'on veut atténuer pour réduire les erreurs de reconstruction de mouvement.

3.3.2. Moyenne pondérée des voisinages

Cette atténuation des discontinuités des trajectoires géométriques est obtenue en moyennant les configurations géométriques des 3 plus proches voisins pris dans l'espace vidéo.

Pour chaque image S_t, on récupère ses 3 plus proches voisins K_{i1}, K_{i2} et K_{i3} parmi les images clefs. La notion de distance utilisée est la mesure de similarité définie précédemment dans l'étape d'indexation, c'est-à-dire la distance euclidienne sur les 575 coefficients DCT. La figure 16 permet d'illustrer le procédé. L'image S_t et ses 3 voisins K_{i1}, K_{i2} et K_{i3} sont projetés dans l'espace vidéo (points I, P_1, P_2 et P_3). A K_{i1} (respectivement K_{i2} et K_{i3}) est associée une configuration géométrique GK_{i1} (respectivement GK_{i2} et GK_{i3}). Ces configurations sont observées par projection dans l'espace géométrique (points Q_1, Q_2 et Q_3).

Les 3 vecteurs de configurations géométriques GK_{i1}, GK_{i2} et GK_{i3} sont moyennés pour générer une nouvelle configuration, qu'on note $\tilde{GK}_t$. La projection de ce nouveau point dans le plan principal géométrique est le point J (Fig.16).

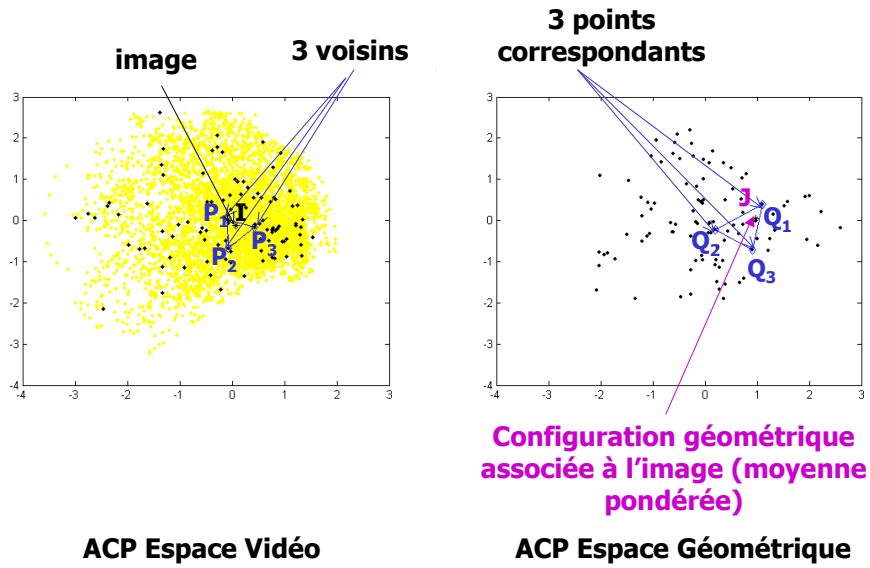


Figure 16 : Schématisation de la moyenne des voisinages dans les plans principaux vidéo et géométrique. A partir d'une image I et de ses 3 plus proches voisins, on récupère 3 configurations géométriques qu'on moyenne pour trouver la configuration géométrique associée à l'image I.

La moyenne est pondérée de manière à prendre en compte la distance entre S_t et chacun de ses voisins K_{i1}, K_{i2} et K_{i3}, toujours en terme de distance euclidienne sur les coefficients DCT. Plus l'image voisine est proche, plus la configuration géométrique associée à ce voisin aura de poids dans la configuration géométrique de l'image considérée.

$$\tilde{GK}_t = \frac{\frac{GK_{i1}}{d(\mathrm{S}_t, K_{i1})} + \frac{GK_{i2}}{d(S_t, K_{i2})} + \frac{GK_{i3}}{d(S_t, K_{i3})}}{\frac{1}{d(S_t, K_{i1})} + \frac{1}{d(S_t, K_{i2})} + \frac{1}{d(S_t, K_{i3})}}$$

La formule ci-dessus, dans le cas de 3 voisins, est généralisée et utilisée pour k voisins, comme on le verra par la suite.

$$\tilde{GK}_t = \frac{\sum_{j=1}^{k} \frac{GK_{ij}}{d(\mathrm{S}_t, K_{ij})}}{\sum_{j=1}^{k} \frac{1}{d(S_t, K_{ij})}}$$

Suite à ce moyennage géométrique, réalisé sur les données de marquage (les 10 points de contour de langue de chaque image), on projette à nouveau dans le plan principal géométrique et on observe les nouvelles trajectoires (Fig. 17b). On constate que les irrégularités ont été compensées. La nouvelle trajectoire obtenue dans le plan géométrique est bien plus lisse. En pratique, les voisins sont pré-calculés dans une phase de multi-indexation.

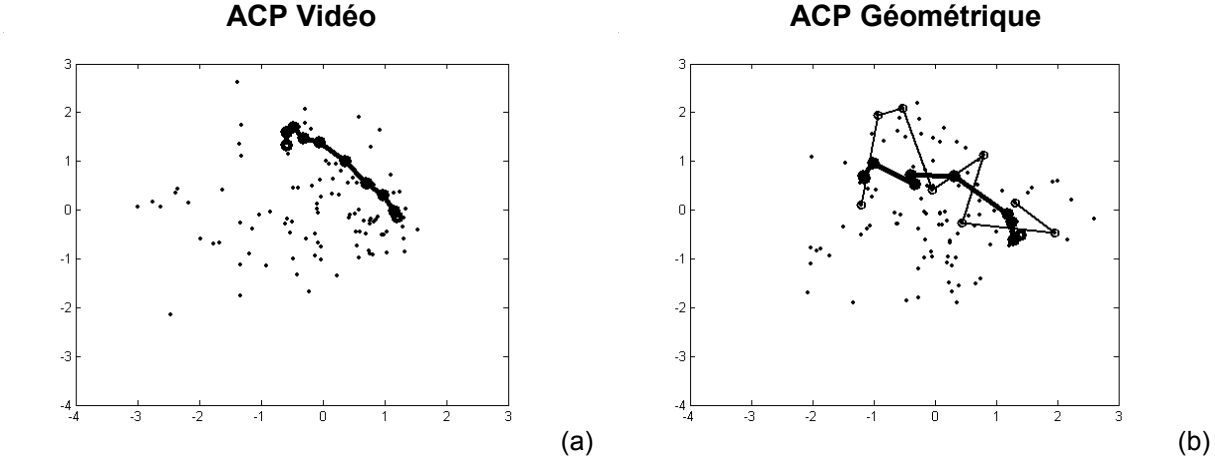


Figure 17 : (a) Trajectoire projetée dans le plan principal vidéo.
(b) Trajectoire générée via l'indexation dans le plan géométrique (en trait fin) et trajectoire obtenue après multi-indexation (en trait gras).

Cette régularisation des données entraîne des améliorations sur la reconstruction du mouvement que l'on peut observer visuellement et que l'on quantifiera plus loin.

Nous avons mis en place 2 méthodes pour réduire l'erreur d'estimation et ainsi améliorer la reconstruction du mouvement de la langue. Ces 2 méthodes, indépendantes, peuvent être combinées. La multi-indexation peut être suivie par le filtrage temporel des configurations géométriques résultantes au même titre que la simple indexation initiale.

3.4. Interpolation entre les points

Pour une image S_t, en reliant les points d'une configuration géométrique GK_t , résultant d'une simple indexation ou $\tilde{GK}_t$ résultant d'une multi-indexation, avec ou sans filtrage temporel, on obtient un contour de langue parfois irrégulier. Pour améliorer l'estimation géométrique du contour de la langue pour chaque image (Fig. 18), nous réalisons une interpolation de type spline par un polynôme $\tilde{SK}_t$ qui s'ajuste aux points de la configuration GK_t ou $\tilde{GK}_t$ au sens des moindres carrés. L'ordre du polynôme est à choisir pour suivre au mieux le contour. Dans le cas de la langue sur Wioland, un polynôme d'ordre 5 donne de bons résultats.

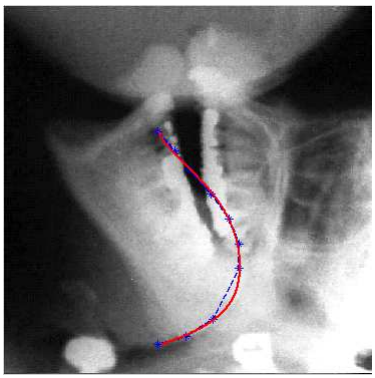


Figure 18 : Contour de la langue point à point (en pointillés bleus) et estimation de ce contour par un polynôme d'ordre 5 (en trait plein rouge).

A ce stade de l'étude, nous observons qualitativement le résultat obtenu, à l'aide de vidéos reconstruites. On superpose les configurations géométriques estimées aux images cinéradiographiques de départ et on rétablit le signal audio correspondant. Le mouvement est correctement reconstruit. On observe cependant des erreurs, que l'on cherche à évaluer quantitativement.

CHAPITRE 2 : EVALUATION DE LA METHODE

L'étape d'évaluation est un point essentiel. De par la nature même de cette méthode, on s'attend à avoir des erreurs. La partie qui suit s'attache à évaluer d'un point de vue purement géométrique l'erreur d'estimation, due au traitement automatique d'indexation. Cette méthode quasi-automatique ne peut pas prétendre à des résultats comparables aux méthodes manuelles. Les erreurs sont forcément plus importantes avec cette méthode en comparaison à un marquage totalement manuel, où l'expert humain intervient pour chaque image. Tout au long de l'évaluation qui suit, il faut garder en tête que le gain en temps est considérable par rapport à une méthode manuelle.
L'évaluation est effectuée par l'intermédiaire de deux mesures d'erreur. La première s'attache à garder la référence manuelle et compare le marquage estimé avec un marquage manuel. Nous construisons ensuite une seconde mesure avec une référence semi-automatique.

1. Mesures des erreurs

Une méthode pour mesurer l'erreur de reconstruction consisterait à marquer manuellement avec précision de longues séquences d'images de la base et comparer ensuite sur chacune des images ce marquage manuel avec le marquage estimé par la méthode mise au point. Ceci est laborieux !
Pour permettre l'évaluation quantitative, deux mesures d'erreur ont été mises en place et nécessitent l'utilisation d'un deuxième jeu de données (100 images et les marques manuelles associées). En pratique, le même expert a marqué manuellement et dans les mêmes conditions 200 images différentes, de façon à éviter la variance inter-expert. Cette question sera traitée plus tard. On considère ici dans cette évaluation que le marquage manuel des images clefs a été réalisé avec soin et sans erreur majeure.
Pour les évaluations qui suivent, des simulations ont été réalisées. Pour chacune d'elles, parmi les 200 images marquées, 100 images sont considérées comme images clefs $(K_i) \in \{1..N\}$, et sont associées aux jeux de marques (GK_i) et les 100 autres images sont considérées comme images tests $(T_j) \in \{1..N\}$ et sont associées aux jeux de marques (GT_j).

1) La première mesure est basée sur l'idée évoquée précédemment et consiste à comparer le marquage manuel et le marquage estimé sur 100 images tests, c'est-à-dire 100 images de la base, différentes des 100 images clefs et marquées manuellement par le

même expert et dans les mêmes conditions que pour les images clefs. La référence est donc un marquage manuel, mais limité à 100 images. Cette mesure d'erreur sera appelée par la suite « erreur sur le groupe d'images tests ».

Les erreurs $(Edof_1)$ sont les erreurs de reconstruction RMS (root mean square) degré de liberté par degré de liberté sur les n=100 images tests (T_j) uniquement, entre les marques $(\tilde{GK}_j)$ estimées à partir des images clefs (K_i) (comme une référence externe) et les marques « manuelles » (GT_j) des images tests correspondantes (T_j).

$$Edof_1(x) = \sqrt{\frac{1}{n}\sum_j (\tilde{GK}_j(x) - GT_j(x))^2}$$

L'erreur $(Etot_1)$ est la moyenne des erreurs RMS $(Edof_1)$ sur les 12 degrés de liberté et sur les 100 images tests (T_j).

$$Etot_1 = \frac{1}{12}\sum_{x=1}^{12} Edof_1(x)$$

2) La deuxième méthode de mesure d'erreur proposée est une comparaison entre 2 séquences géométriques estimées sur toute la base et générées chacune à partir d'un jeu différent d'images clefs. La référence est donc cette fois un marquage semi-automatique étendu à toute la séquence. Cette mesure d'erreur sera appelée par la suite « erreur entre paires de trajectoires».

On évalue l'erreur de reconstruction RMS entre 2 jeux de configurations géométriques estimées ; d'un côté les données géométriques estimées à partir des images clefs et de l'autre les données géométriques estimées à partir des images tests.

$(Edof_2)$ et $(Etot_2)$ sont calculées en comparant sur toutes les images de la base les marques estimées via 2 ensembles différents d'images clefs.

Les erreurs $(Edof_2)$ sont les erreurs de reconstruction RMS (root mean square) degré de liberté par degré de liberté sur les N=5673 images de la séquence (S_t) entre les marques $(\tilde{GK}_t)$ estimées à partir des images clefs (K_i) et les marques $(\tilde{GT}_t)$ estimées à partir des images tests (T_j).

$$Edof_2(x) = \sqrt{\frac{1}{N}\sum_t (\tilde{GK}_t(x) - \tilde{GT}_t(x))^2}$$

L'erreur $(Etot_2)$ est la moyenne des erreurs RMS $(Edof_2)$ sur les 12 degrés de liberté et sur la séquence entière.

$$Etot_2 = \frac{1}{12}\sum_{x=1}^{12} Edof_2(x)$$

Les erreurs peuvent aussi être calculées après l'interpolation spline par le polynôme de degré de 5.
Dans ce cas, on utilise les mêmes mesures d'erreur en remplaçant les 12 degrés de liberté par 12 points répartis sur le contour interpolé.
Pour les mesures qui vont suivre, l'erreur est exprimée en pixel par degré de liberté pour des images de taille 490*480 pixels. Il s'agit des « grandes images » 720*540 dont les bords ne sont pas pris en compte, correspondant aux images sur lesquelles le marquage manuel a été réalisé.

2. Evaluation quantitative

L'évaluation quantitative permet de choisir les paramètres de la méthode, de façon à la rendre optimale en minimisant l'erreur de reconstruction RMS.
Cette évaluation, comme on l'a déjà dit, nécessite l'utilisation d'un deuxième jeu de données (des images clefs et les marques manuelles associées). Ce deuxième jeu de données a été réalisé dans les mêmes conditions que le premier. Le même expert a marqué manuellement 100 nouvelles images.
Pour l'évaluation de chaque paramètre, des simulations ont été réalisées à l'aide du technique de Jackknife. Plusieurs mesures sont réalisées et moyennées ; à chaque fois, 2 jeux d'images sont constitués aléatoirement à partir des images marquées (les 100 images marquées au départ et les 100 nouvelles images marquées pour l'évaluation). Un ensemble de n_1 images est considéré comme le jeu d'images clefs et est utilisé pour l'application de la méthode de rétro-marquage. Les images restantes sont utilisées comme le jeu d'images tests pour l'évaluation de l'erreur de reconstruction.

2.1. Nombre d'images clefs

Il est logique de penser que plus le nombre d'images clefs marquées est important, plus la reconstruction du mouvement sera précise. En effet, augmenter le nombre de clefs revient à diminuer le pas de quantification lors de l'étape d'indexation. En contrepartie, marquer un nombre plus grand d'images rallonge l'étape manuelle. Le choix du nombre d'images clefs doit être un compromis entre le coût du marquage manuel (en temps) et le taux d'erreur de reconstruction RMS. La figure 19 montre l'évolution de l'erreur $Etot_1$ en fonction du nombre de clefs ; la méthode appliquée correspond au modèle avec simple indexation (prise en compte du premier voisin) et filtrage temporel. L'erreur obtenue pour 25 clefs est de 20,5 pixels/ddl, de 17,3 pixels /ddl pour 100 clefs et de 16,3 pixels/ddl pour 200 clefs (on rappelle

que la taille des images est de 720*540 pixels). L'erreur diminue donc de plus de 3 pixels/ddl entre 25 et 100 images clefs puis seulement de 1 pixel entre 100 et 200 clefs. Avec une échelle Log-Log (Fig. 19b), on note que la relation entre l'erreur et le nombre de clefs est linéaire avec une pente p = -0,1. La relation entre l'erreur et le nombre d'images clefs est exponentielle, nous voyons qu'une augmentation très importante du nombre d'images clefs n'entraîne pas une diminution significative de l'erreur. Compte-tenu de ces remarques, nous avons estimé que 100 images clefs était un bon compromis. A partir de là, la méthode est appliquée avec 100 clefs.

Nous sommes bien sûr conscients que le résultat obtenu sera toujours meilleur avec plus d'images clefs, mais l'intérêt de notre méthode repose sur cette relation coût-performance qui permet de trouver facilement un compromis.

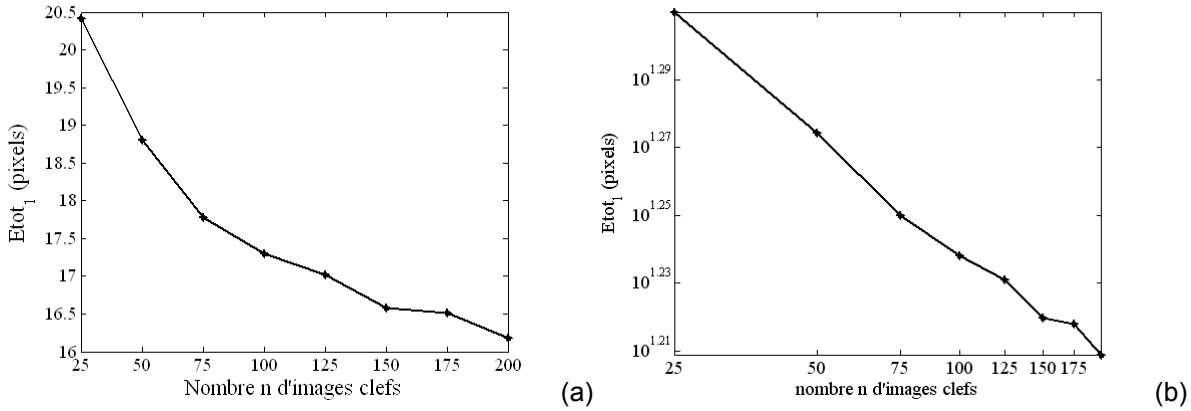


Figure 19 : Influence du nombre d'images clefs sur l'erreur $Etot_1$, calculée après indexation par 1 voisin et filtrage temporel – (a) représentation linéaire – (b) représentation Log-Log.

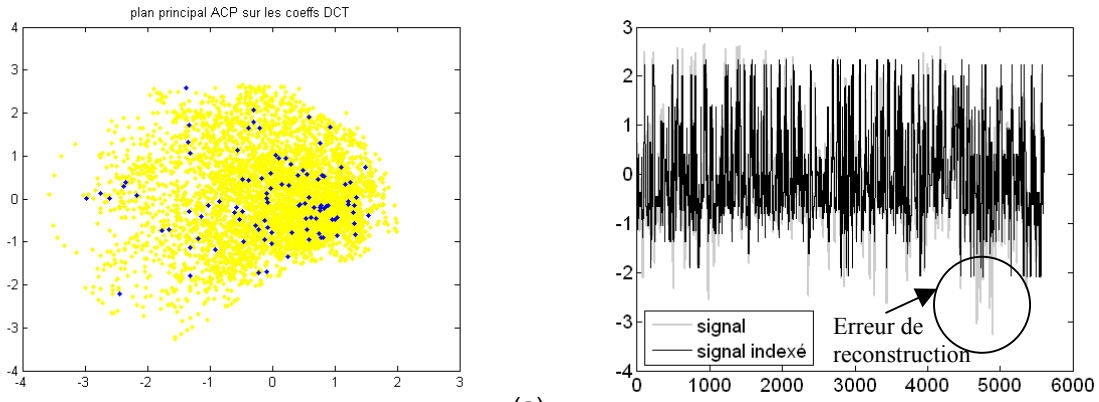
2.2. Choix des images clefs

Grâce au calcul des erreurs de reconstruction, différentes stratégies de choix des images clefs sont comparées.

La stratégie utilisée pour la méthode consiste en un choix aléatoire de n images clefs. Il s'agit de réaliser une permutation des 5673 images de la base et de choisir un segment de n images dans cette permutation.

Cette stratégie permet de respecter la répartition des données de départ, mais de ce fait, elle néglige les extrêmes, ce qui entraîne pour ces points de grands écarts entre la séquence initiale et celle indexée. La couverture de l'espace des données par les images clefs dépend de la fréquence ou densité des configurations. C'est ce qu'on observe avec le plan principal des données vidéos (Fig. 20a), c'est-à-dire la projection des 5673 images de la base sur les 2 premières composantes de l'analyse en composantes principales des coefficients DCT des images. La densité d'images clefs est plus grande là où la densité de données est importante. A partir de ce plan principal, nous suivons, par exemple, l'évolution de la 2ème

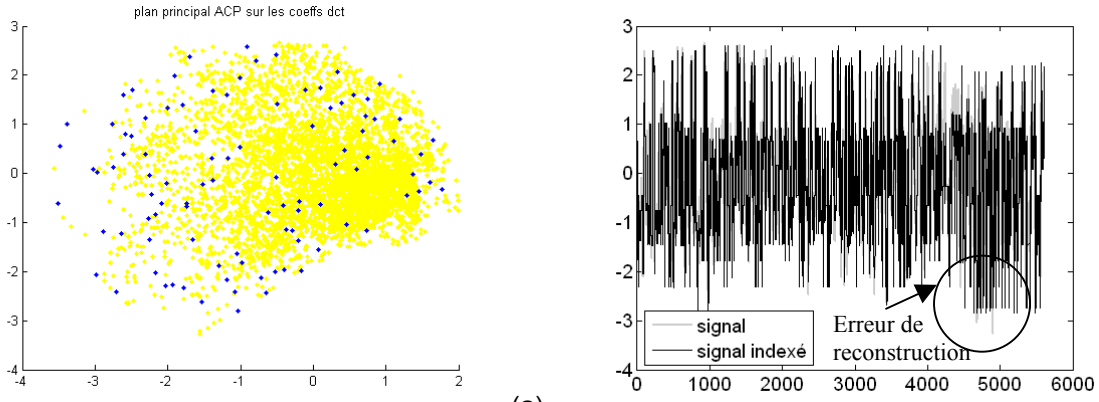
composante sur toute la base pour la séquence de départ (via les points jaunes) et pour la séquence indexée (via les points bleus), et nous observons, sur la figure 20b, que les points extrêmes ne sont généralement pas atteints par l'indexation.



(a) (b)

Figure 20 : (a) Plan principal et 100 images clefs choisies aléatoirement, en respectant la répartition d'origine.
(b) Signal (gris) et signal indexé (noir) selon la 2ème composante principale.

Pour prendre en compte les points extrêmes, on peut envisager pour le choix des images clefs un tirage uniforme sur l'enveloppe convexe des données dans le plan principal (Fig. 20a), de façon à rendre la couverture indépendante de la fréquence. On réduit ainsi l'erreur de reconstruction RMS pour les points extrêmes, comme on l'observe sur la figure 21b avec l'évolution de la 2ème composante de l'ACP vidéo. Mais on augmente alors toutes les petites erreurs de reconstruction des points dans la partie dense de la répartition.



(a) (b)

Figure 21 : (a) Plan principal et 100 images clefs choisies uniformément.
(b) Signal (gris) et signal indexé (noir) selon la 2ème composante principale.

Au final, ces 2 stratégies de choix d'images clefs présentent des valeurs moyennes d'erreur de reconstruction RMS assez proches. Cependant l'erreur est légèrement plus faible pour un choix aléatoire d'images clefs, qui est la stratégie utilisée dans l'étude.

On observe une différence d'erreur $Etot_1$ (erreur sur un groupe d'images tests) de 1 pixel entre le choix aléatoire et le choix uniforme. Cette mesure a été réalisée pour mettre en

évidence quantitativement l'apport de la stratégie de choix aléatoire. Pour cette mesure, 100 images clefs ont été choisies uniformément, c'est-à-dire de façon à mieux prendre en compte les positions extrêmes. Elles ont été marquées par le même expert que celui qui a marqué les 200 premières images et toujours dans les mêmes conditions.
La mesure d'erreur a été faite sur un jeu de 100 images tests. On a comparé l'erreur de reconstruction RMS sur ces images tests entre le marquage manuel et l'estimation à partir des images clefs choisies aléatoirement d'une part et entre le marquage manuel et l'estimation à partir des images clefs choisies uniformément d'autre part. Les estimations considérées ont été obtenues après simple indexation (1 seul voisin) et filtrage temporel.

2.3. Type d'indexation

2.3.1. Taille du voisinage

La méthode initialement mise au point avec une simple indexation a été enrichie d'une « multi-indexation » ; on tient compte, non plus de l'image clef la plus proche uniquement, mais des trois images clefs les plus proches. On prend de plus en considération les distances de l'image considérée avec ses clefs voisines, comme on a pu l'expliquer précédemment (voir ch. 1 §3.3.2. Moyenne pondérée des voisinages).
Le gain de cette multi-indexation est mis en évidence avec une évaluation de $Etot_1$. On observe sur la figure 22, l'évolution de l'erreur en fonction de la taille du voisinage, la méthode étant appliquée sans filtrage temporel. Nous avons fait varier le nombre de voisins de 1 à 10. Augmenter ce nombre de 1 à 3, et même à 4, implique une forte décroissance de l'erreur. Au delà de 4, il n'y a plus réellement de gain supplémentaire.

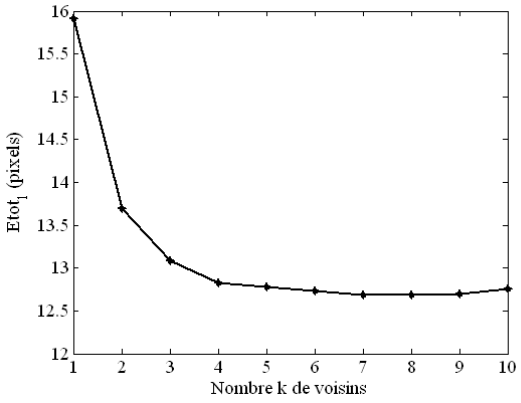


Figure 22 : Influence de la taille du voisinage sur l'erreur $Etot_1$, calculée sans filtrage temporel et à partir de 100 images clefs.

L'idée de multi-indexation, d'abord suggérée avec 3 voisins, est, compte-tenu de cette évaluation, appliquée avec 4 voisins. Le passage de 1 à 4 voisins permet un gain de plus de 3 pixels/ddl. Il s'agit encore d'un compromis entre performance (erreur de reconstruction RMS) et coût (ici, coût en temps de calcul).

2.3.2. Distance

L'indexation automatique fait appel, comme on a pu le voir dans la description de la méthode, à une notion de distance. La mesure de similarité utilisée pour l'assignation des index est la distance euclidienne entre les coefficients DCT des images. C'est aussi cette mesure qui est utilisée pour pondérer les configurations géométriques des images clefs voisines, dans le cas de la multi-indexation.

Néanmoins, d'autres distances pourraient être adoptées ; une mesure comparative a été réalisée de manière à justifier l'utilisation de la distance euclidienne entre les coefficients DCT basses fréquences.

On a parlé d'analyse en composantes principales : à la place de la distance euclidienne entre les 575 coefficients DCT, il est envisageable de considérer cette distance dans le plan principal vidéo. Utiliser l'une ou l'autre de ces 2 distances dans la pondération des configurations géométriques n'a que peu d'influence (erreur $Etot_1$ supérieure de quelques dixièmes de pixels/ddl avec la distance dans le plan principal). Par contre, l'indexation automatique avec la distance euclidienne dans le plan principal vidéo comme mesure de similarité implique une reconstruction de l'information géométrique nettement moins bonne qu'avec la distance euclidienne sur les coefficients DCT. L'erreur $Etot_1$ calculée pour une méthode sans filtrage à partir de 100 images clefs (Fig. 23), pour 2 à 4 voisins, présente un gain de près de 7 pixels/ddl pour la mesure de similarité utilisant les coefficients DCT.

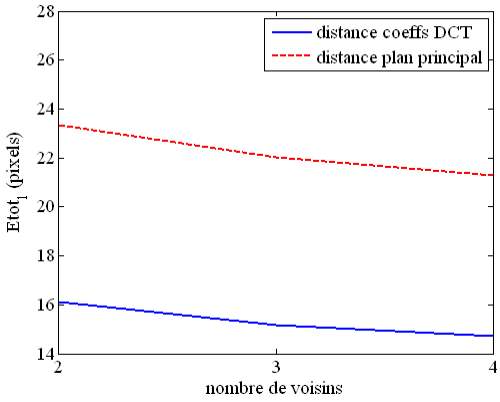


Figure 23 : Influence du choix de la mesure de similarité sur l'erreur $Etot_1$, pour une estimation calculée sans filtrage temporel et à partir de 100 images clefs.

D'autres mesures ont été réalisées avec d'autres paramètres afin d'évaluer les différents aspects de la méthode. Notamment, on a étudié l'influence du nombre de coefficients DCT pris en compte dans l'indexation. Ceci ne sera pas détaillé pour l'instant, mais sera présenté plus tard avec d'autres articulateurs.

2.4. Effet cumulatif des améliorations apportées

On construit plusieurs modèles en combinant les méthodes de réduction d'erreur de reconstruction RMS. On note ces modèles M1 à M5, ils sont définis dans le tableau et par le schéma suivants.

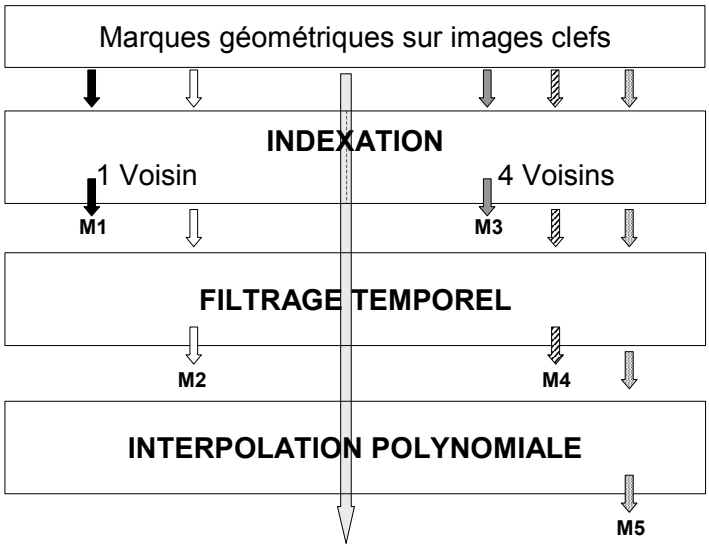


Figure 24 : Bloc-diagramme des traitements postérieurs.

Modèles / Méthodes	M1	M2	M3	M4	M5
Taille du voisinage	1	1	4	4	4
Filtrage temporel	-	+	-	+	+
Interpolation spline	-	-	-	-	+

Table 1 : Construction de modèles combinant les différentes améliorations possibles de la méthode de rétro-marquage.

En appliquant les méthodes de réduction d'erreur, on montre que l'on est capable (figure 25) de réduire graduellement l'erreur de reconstruction RMS. L'erreur $Etot_1$ diminue de 5 pixels/ddl. L'erreur $Etot_2$ est plus optimiste, la réduction après application des 3 méthodes est de 9 pixels/ddl, mais ceci s'explique par le fait que les 2 séquences sont traitées de façon similaire, les erreurs qu'elles produisent sont corrélées.

Cette diminution graduelle de l'erreur montre que l'effet des méthodes est cumulatif et qu'elles sont donc complémentaires.

Au final, l'erreur standard RMS est évaluée à 11 pixels/ddl pour $Etot_1$ et à 8 pixels/ddl pour $Etot_2$, en sachant que la longueur moyenne de la langue est estimée à 350 pixels.

Il est habituel d'estimer les erreurs de marquage en millimètres. Nous n'avons pu réaliser qu'une calibration approximative pour passer de pixels aux centimètres. En effet, cette

information n'est pas disponible directement sur le film. A partir de la thèse de Doctorat d'Etat de François Wioland et de certains des schémas relatifs à cette base de données, nous avons pu estimer 1 cm pour 38 pixels (voir annexe A5). Ainsi nous évaluons l'erreur standard de reconstruction des mouvements de la langue sur cette base de données à 3 mm par degré de liberté.

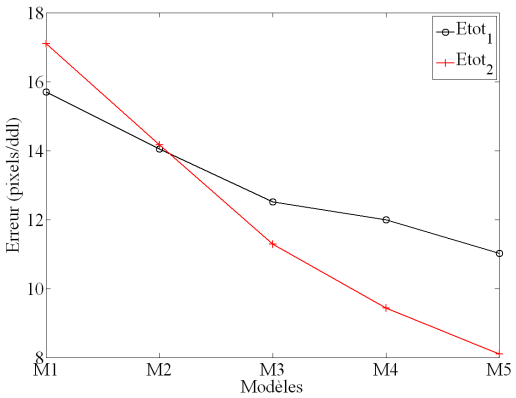


Figure 25 : Contribution cumulée des méthodes de réduction d'erreur observée à partir des mesures de $Etot_1$ et $Etot_2$.

2.5. Evaluation finale de l'erreur dans les conditions d'application choisies

La méthode est finalement appliquée avec des conditions d'application choisies pour permettre un compromis entre performance et coût, de façon à minimiser conjointement l'erreur de reconstruction RMS et le temps de traitement (essentiellement manuel). Nous utiliserons pour la suite, avec 100 images clefs, le modèle M4, celui avec multi-indexation (4 voisins) et filtrage temporel. L'effet du lissage géométrique (spline) sera aussi ajouté, mais pour les mesures qui vont suivre, nous laisserons un temps ce lissage de côté, dans la mesure où nous allons nous intéresser aux erreurs en fonction du degré de liberté.

La figure qui suit permet de mettre en évidence l'erreur de reconstruction RMS degré de liberté par degré de liberté. Sur une image clef, le contour marqué de la langue est affiché ainsi que les barres d'erreur correspondant à l'erreur $Edof_2$ pour le modèle M4. Ces barres correspondent à l'écart type de l'écart entre 2 estimations obtenues à partir de 2 jeux d'images clefs différents (ici on a tracé un demi écart-type de chaque coté du contour pour chaque ddl). On observe ainsi que l'erreur est à peu près uniformément répartie le long du contour de la langue.

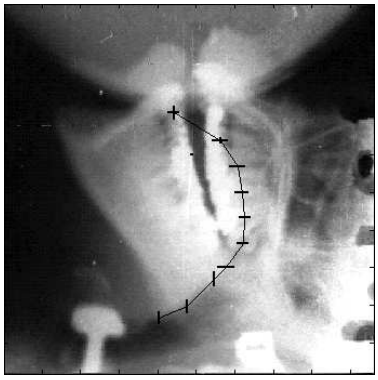


Figure 26 : Barres d'erreur $Edof_2$ sur une image clef (490*480 pixels).

L'observation de la répartition des erreurs pour le modèle M4 montre que les différents degrés de liberté présentent des profils assez similaires. L'erreur a une allure gaussienne (histogrammes sur la figure 27). Cette erreur n'est pas toujours centrée.

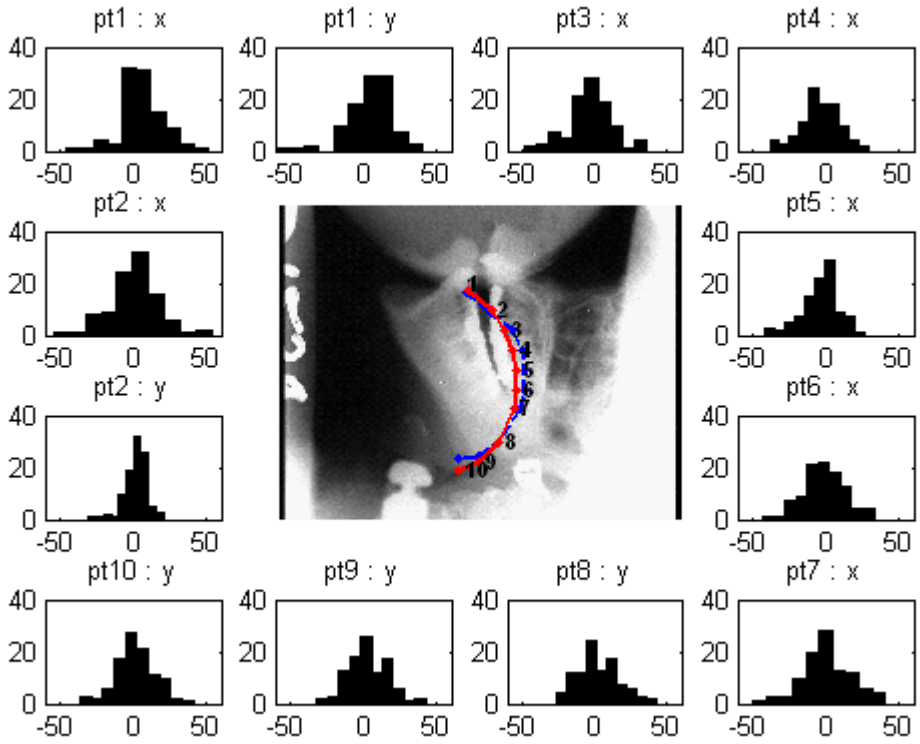


Figure 27 : Distributions des erreurs $Edof_1$ selon les 12 degrés de liberté entre le marquage manuel (en rouge) et le marquage estimé (en bleu).

Il y a donc deux types d'erreurs : le biais (moyenne de l'erreur) et la dispersion (écart-type). L'erreur de reconstruction que nous considérons est l'erreur de dispersion, l'erreur RMS. Il existe un biais différent pour chaque degré de liberté, il est exprimé dans le tableau qui suit. Nous ne prenons pas en compte le biais dans nos mesures d'erreurs.

L'erreur maximale atteinte est de 50 pixels (env. 1,3 cm) ; dans ces situations, on parlera de « décrochages ».

	Pt1 : x	Pt1 : y	Pt2 : x	Pt2 : y	Pt3 : x	Pt4 : x	Pt5 : x	Pt6 : x	Pt7 : x	Pt8 : y	Pt9 : y	Pt10 : y
Biais	6	4,1	1,4	1,9	-3,3	-3,9	-3,6	-0,5	-0,7	2,9	3	1,7
Ecart-Type	13,2	15,8	14,4	7,7	14	12	11,5	13	15,6	14	12,8	12,2

Table 2 : Biais et écart-type de l'erreur (sur 100 images clefs) entre marquage manuel et marquage estimé par degré de liberté.

2.6. Variabilité inter-experts

Le problème de la variabilité entre experts pour le marquage manuel reste encore en suspens. Nous n'avons pas eu la possibilité de multiplier le nombre d'experts pour pouvoir effectuer des comparaisons sur leurs tracés manuels.

Dans une étude récente de Soquet et al. [SLMD02] en IRM contrastée, il a été montré que cette variabilité est faible. En effet, un test sur la reproductibilité de mesures de tracés a été effectué sur des contours de sections de régions connues. Ces tracés sont ensuite numérisés et traités par ordinateur pour obtenir l'aire de la section. Les résultats sur 10 répétitions de chaque tracé montrent que l'écart-type est inférieur à 0.005 cm². Mais il s'agit d'IRM, les images sont beaucoup plus contrastées que les images radiographiques.

En ce qui concerne nos données sur la séquence Wioland, nous avons pu faire marquer par un autre expert les 100 images clefs initialement marquées.

A partir de ces nouvelles marques, à titre indicatif, quelques mesures ont été effectuées en appliquant la méthode semi-automatique sur les différents jeux de marques géométriques. Les résultats présentés ici sont obtenus avec une simple indexation et sans filtrage temporel.

La mesure considérée est l'erreur $Etot_2$ (erreur entre paires de trajectoires) qui compare 2 jeux de marques estimées sur la séquence complète à partir de 2 ensembles différents d'images clefs.

Notons F le premier expert qui a marqué 2 jeux de 100 images et J le second expert qui a donc marqué 100 images correspondant au 1er jeu de l'expert F. Nous disposons pour F des marques $(\tilde{GK}_t)_F$ estimées à partir des images clefs (K_i) et des marques $(\tilde{GT}_t)_F$ estimées à partir des images tests (T_j), et pour J des marques $(\tilde{GK}_t)_J$ estimées à partir des images clefs (K_i).

L'erreur $Etot_2$ est calculée entre $(\tilde{GK}_t)_F$ et $(\tilde{GK}_t)_J$, pour la variabilité inter-experts, ainsi qu'entre $(\tilde{GK}_t)_F$ et $(\tilde{GT}_t)_F$ et $(\tilde{GK}_t)_J$ et $(\tilde{GT}_t)_F$ pour l'erreur liée à la méthode. Les

résultats sont résumés dans le tableau qui suit et permettent de montrer que l'erreur liée à la variabilité entre experts est plus faible que celle liée à la méthode. L'erreur obtenue avec 2 jeux d'images clefs varie peu si l'on considère les marques de l'un ou l'autre des experts (erreur évaluée respectivement à 17,5 et 17,9 pixels par ddl).

Ces résultats semblent être en faveur d'une variabilité faible entre experts, à condition bien sûr que chaque expert réalise de façon précise et appliquée le marquage manuel des images clefs.

$Etot_2$ (pixel/ddl)	$(\tilde{GK}_t)_F$	$(\tilde{GK}_t)_J$	$(\tilde{GT}_t)_F$
$(\tilde{GK}_t)_F$	X	9,7	17,5
$(\tilde{GK}_t)_J$	9,7	X	17,9
$(\tilde{GT}_t)_F$	17,5	17,9	X

Table 3 : Erreur $Etot_2$ du contour de la langue à partir de jeux d'images clefs différents et de marques manuelles obtenues de 2 experts.

3. Temps d'exécution

En terme de temps de traitement, le marquage manuel de 10 points sur 100 images clefs est estimé à 4 heures environ. En effet, la langue est parfois difficilement visible et il est nécessaire d'activer le curseur pour faire défiler les images adjacentes. Généralement, l'expert réalise un premier marquage qu'il corrige point par point grâce à la visualisation de la langue en mouvement.

Une fois le traitement manuel réalisé, le temps d'exécution de la méthode pour la base complète (5673 images) est de quelques minutes pour le calcul des coefficients DCT et l'application de l'algorithme de rétro-marquage. Notons bien évidemment que plus le nombre de coefficients DCT utilisés est important, plus le temps d'exécution est allongé.

Rappelons que nous évaluons une méthode semi-automatique. Les erreurs existent, elles ont été évaluées à quelques pixels/ddl. Mais notons simplement que s'il avait fallu marquer manuellement 5673 images, près de 10 jours 24h/24h auraient été nécessaires !

4. Evaluation qualitative

Une autre façon d'évaluer la reconstruction du mouvement de la langue repose sur un aspect qualitatif. En superposant sur les images de la séquence d'origine le marquage géométrique estimé, on est capable de construire des séquences vidéos et suivre le mouvement de la langue.

Des films ont été réalisés à différents stades de l'élaboration de la méthode pour mettre en évidence l'apport des techniques d'amélioration proposées. Les films sont élaborés à partir des images d'origine et du marquage estimé pour chacune de ces images. On reconstruit la séquence vidéo en concaténant les images d'origine sur lesquelles on a tracé le contour géométrique estimé. Le signal audio peut également être intégré au film ainsi reconstitué. Les séquences sont enregistrées à une vitesse de 25 images par seconde. Les films sont conçus au format .avi mais ils peuvent aussi être encodés en .wmv ou en .swf pour permettre un partage sur le web.

L'observation de ces films permet d'un point de vue tout à fait qualitatif, d'apprécier le suivi des mouvements de la langue. On observe que le marquage estimé à l'aide de la méthode quasi-automatique mise en place suit globalement les mouvements de cet articulateur, mais on observe aussi quelques décrochages.

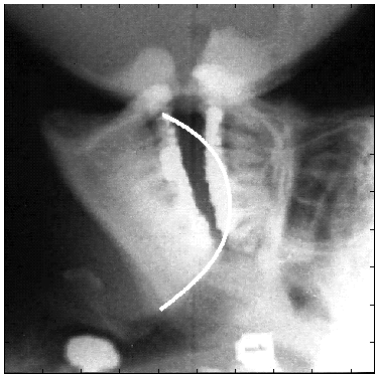


Figure 28 : Contour de la langue estimé à partir de la méthode de rétro-marquage et superposé à l'image d'origine.

CHAPITRE 3 : EXTENSIONS DE LA METHODE A D'AUTRES ARTICULATEURS ET A D'AUTRES SEQUENCES CINERADIOGRAPHIQUES

1. Adaptation de la méthode

1.1. Principe

Comme on l'a vu, cette méthode quasi-automatique a été mise au point à partir de la base de données Wioland pour en extraire les mouvements de la langue ; elle peut être adaptée à d'autres articulateurs du conduit vocal, comme les lèvres ou le vélum, ainsi qu'à d'autres bases de données cinéradiographiques.

Méthodologiquement, nous considérons que chaque articulateur est indépendant et partant de là, un traitement spécifique est appliqué à chacun d'eux. Le procédé est le même mais utilise des paramètres différents et adaptés pour chaque articulateur :

(1) Les images d'origine sont découpées, par cadrage ou par la pose d'un masque, de façon à inclure, pour tout le film, l'élément à marquer, à exclure les interférences avec d'autres articulateurs et à éliminer les éléments parasites. C'est à cette étape qu'il faut prendre en compte d'éventuels décalages des images au cours du temps, comme nous l'avons mis en évidence précédemment.

Ensuite, les paramètres de la méthode (nombre d'images clefs, degrés de liberté, nombre de coefficients DCT nécessaires pour l'indexation) sont définis de façon indépendante pour chaque articulateur.

(2) Chaque traitement est suivi d'une évaluation (comme cela a été défini à la partie précédente, nous mesurons l'erreur RMS par degré de liberté), de manière à quantifier sur des images tests l'écart entre le marquage manuel et le marquage estimé avec la méthode.

Ces évaluations font appel à une technique de Jackknife pour éviter de marquer manuellement un nouveau jeu d'images. Plusieurs mesures sont réalisées et moyennées ; à chaque fois, 2 jeux d'images sont constitués à partir des images clefs. Un jeu de n_1 images, laissant de côté un petit nombre n_2 d'images, est utilisé comme images clefs pour l'application de la méthode. Les n_2 images restantes sont alors considérées comme images test sur lesquelles l'évaluation $Etot_1$ est réalisée.

(3) De façon qualitative, pour apprécier la qualité de la reconstruction du mouvement des différents articulateurs, des vidéos sont aussi réalisées en superposant les

marques estimées sur les images d'origine. Cette reconstruction de séquences vidéos est effectuée de façon tout à fait similaire à celle décrite au chapitre 2 (§4). Les images d'origine sur lesquelles on affiche le contour estimé de l'articulateur sont concaténées à la cadence de la séquence d'origine. Le signal audio peut ou non être intégré aux vidéos.

(4) Enfin, et nous le verrons en détail plus loin, chaque articulateur ayant été marqué indépendamment, le tracé des contours du conduit vocal complet est obtenu par reconstruction, pour une séquence complète, en combinant l'ensemble de ces marquages indépendants.

1.2. Premiers essais sur Wioland

Avant d'étendre la méthode à d'autres séquences cinéradiographiques, nous avons d'abord appliqué la méthode à d'autres parties du conduit vocal de la séquence Wioland.

Les lèvres, le vélum, la mâchoire et le larynx ont été traités à l'aide de notre méthode quasi-automatique d'extraction de mouvements. Nous ne détaillerons pas ici les paramètres utilisés pour chaque articulateur ainsi que l'évaluation propre à chacun d'eux. Nous préciserons simplement le marquage réalisé sur les lèvres. Le traitement de rétro-marquage a été appliqué aux lèvres. Un cadre a été défini autour des lèvres et des incisives avant comme on l'observe sur l'image suivante. Il s'agit d'un cadre de taille 140*200 pixels. Ce cadre a servi pour le calcul des coefficients DCT dont le nombre a été limité à 71 (6*12 coefficients, mis à part le premier) pour cet articulateur. Le marquage des points a été réalisé sur ces images réduites. 8 points, avec 10 degrés de liberté, définissent les lèvres et les dents. Chaque dent est marquée avec 1 point à 2 ddl, chaque lèvre est marquée avec 3 points à 1ddl chacun. Cette définition des degrés de liberté est visualisée sur l'image ci-dessous.

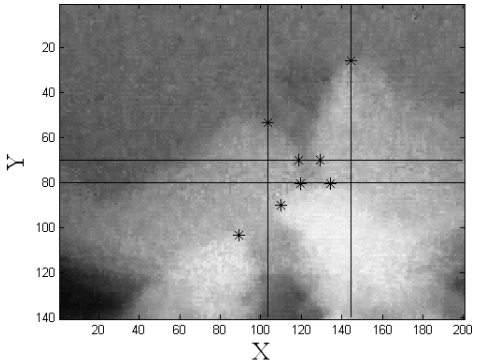


Figure 29 : Marquage des lèvres et des dents avant de la séquence Wioland (la lèvre supérieure est en haut à droite de l'image). L'image a été découpée pour se focaliser sur ces articulateurs et 10 degrés de liberté ont été définis.

Les méthodes d'amélioration pour la reconstruction du mouvement ont été appliquées à cet articulateur pour permettre la meilleure extraction possible des mouvements des lèvres.

Des traitements semblables ont été réalisés sur le vélum, sur la mâchoire ou sur le larynx. Nous n'en dirons pas plus ici car un exemple plus complet de marquage du conduit vocal entier sera réalisé avec la séquence Laval43 de la base de données cinéradiographiques d'ATR.

Remarquons simplement qu'en marquant les parties fixes (le palais, le pharynx), on reconstruit ici, pour la séquence Wioland, la forme du conduit vocal entier, comme on le voit sur la figure 30.

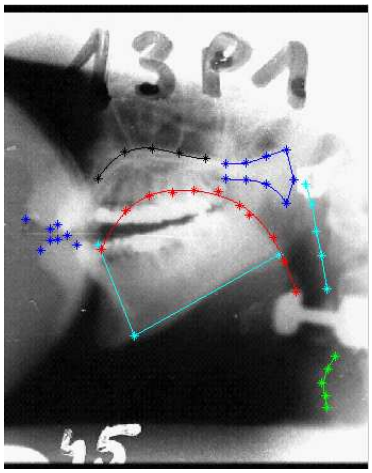


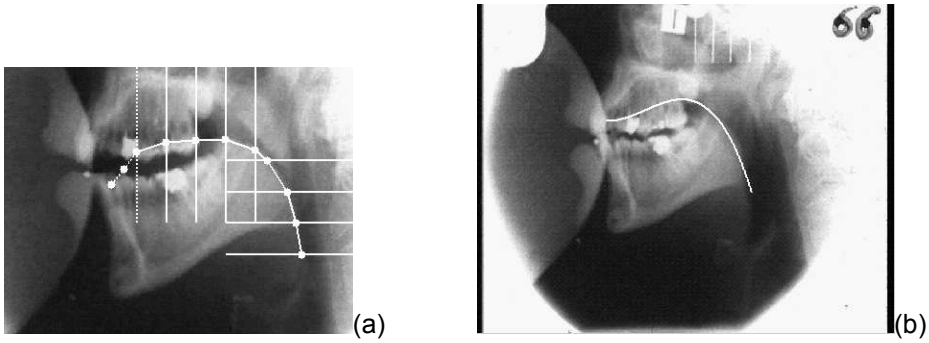
Figure 30 : Marquage complet du conduit vocal dans Wioland.

1.3. Première extension vers une autre séquence : la séquence Flament

La séquence Flament a été enregistrée dans des conditions proches de Wioland (66 images par seconde) et a été numérisée dans le même contexte. Elle est composée de près de 5000 images (720*540 pixels) pour une durée de 3 minutes et 18 secondes (25 im/s). Nous n'avons pas observé de décalages au cours de cette séquence, comme cela était le cas pour la séquence Wioland, il n'a donc pas été nécessaire de réaliser plusieurs cadrages au sein de la séquence. La séquence Flament nous a permis de nous intéresser principalement à la langue et au vélum.

La méthode semi-automatique d'extraction géométrique a été appliquée au contour de la langue, de façon assez similaire à Wioland. Pour cela, 13 degrés de liberté ont été définis (Fig. 31) : 9 points à 1 ddl pour la base et le dos et 2 points à 2 ddl pour la pointe. La pointe de la langue est nettement plus visible dans ce film et une adaptation a été réalisée afin de mieux capturer ses mouvements rapides et parfois relativement indépendants. Cette adaptation ne sera pas détaillée ici mais un peu plus loin avec le marquage du contour de la langue de la séquence Laval43, qui a été réalisé de façon analogue. Pour améliorer encore les performances de la méthode, 200 images clefs ont été marquées manuellement. Ainsi,

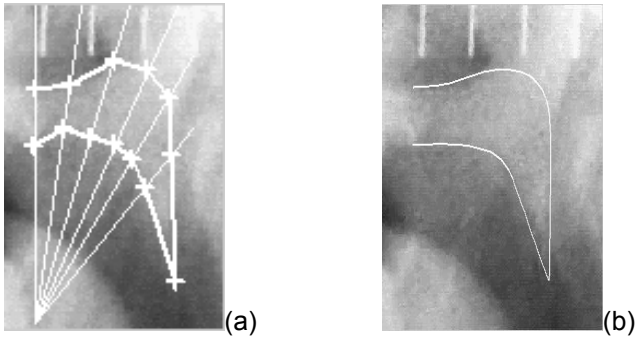
pour la séquence Flament, l'erreur moyenne de reconstruction du contour de la langue a été, dans les meilleures conditions, évaluée à 10 pixels/ddl.



(a) (b)

Figure 31 : (a) Degrés de liberté pour le marquage du contour de la langue pour la séquence Flament.
(b) Contour de langue estimé par la méthode.

D'autre part, le corpus de la base Flament est dédié à la question des nasales du Français [Fla84] – le corpus est disponible en annexe (A2) – et le vélum est bien visible sur le film. Un traitement spécifique à cet articulateur a aussi été réalisé avec succès, à partir de 100 images clefs marquées manuellement. Comme on l'observe sur la figure 32, 13 points ont été marqués sur le vélum avec 14 degrés de liberté, à l'aide d'une grille de marquage polaire. L'allure de cette grille sera ré-utilisée pour le marquage du voile du palais de la séquence Laval43.



(a) (b)

Figure 32 : (a) Degrés de liberté pour le marquage du vélum pour la séquence Flament.
(b) Contour du vélum estimé par la méthode.

Il serait évidemment possible de marquer successivement tous les articulateurs du conduit vocal de cette séquence Flament. Nous n'avons pas ici traité les lèvres et les parties fixes. Nous nous sommes plutôt concentrés sur une autre séquence cinéradiographique, Laval43, tirée d'une grande base de données et déjà traitée par une autre méthode d'extraction automatique.

2. Estimation des mouvements du conduit vocal sur Laval43, une séquence de la base de données ATR

La base de données cinéradiographiques d'ATR, "X-ray film database for Speech Research", est la plus grande disponible pour les recherches en parole, avec 25 films différents totalisant une durée de 55 minutes et près de 100000 images[4]. Elle a été constituée par Munhall, Vatikiotis-Bateson et Tohkura [MVT94, MVT95] à partir de films enregistrés par Rochette (Université Laval) et Perkell et Stevens (M.I.T.). Parmi les 25 films, 24 ont été enregistrés par Claude Rochette au Département de Radiologie de l'Hôtel Dieu de Québec, à Québec au Canada en 1974. Ils ont été enregistrés à 50 images par seconde. Le dernier film est désigné comme film M.I.T. et a été enregistré en 1962 au laboratoire de recherche Wenner-Gren à l'hôpital de Nortull à Stockholm (Suède) sous l'investigation de Stevens et Öhman et a été analysé en partie par Perkell en 1969. Il a été enregistré à 45 images par seconde.

La base de données peut être récupérée sur demande auprès de Kevin Munhall et a déjà été le support de travaux récents ([TL99], [Isk05]). Nous nous sommes procurés la base complète et avons extrait les images à partir du DVD reçu. Les séquences disponibles sont enregistrées au format NTSC au rythme de 29,97 images par seconde. Le signal audio associé est disponible, il est échantillonné à 44,1 KHz. La synchronisation est réalisée à l'aide d'un bip sonore et d'un marqueur visuel.

Il n'est pas possible pour l'instant de réaliser le traitement complet de cette base à cause de l'étape de marquage manuel qui est propre à chaque film, nous en rediscuterons à la fin de ce manuscrit. La méthode s'applique film par film et notre étude ici se limite à la séquence Laval43. Cette séquence a été enregistrée en 1974 et elle est composée de 3973 images du conduit vocal, provenant d'une séquence vidéo de 2 minutes 14 secondes. Cette séquence correspond à des phrases en français lues par un locuteur mâle québécois de 19 ans. Les images d'origine sont de taille 720*480 pixels.

2.1. Extraction séparée des différents articulateurs

A partir de cette séquence cinéradiographique et en utilisant la méthode de rétro-marquage, nous avons analysé de façon indépendante les mouvements de différents articulateurs du conduit vocal.

Une estimation indépendante a été réalisée pour les articulateurs suivants :

- la langue, de manière assez similaire à Wioland
- la pointe de la langue, par une estimation spécifique que nous décrirons en détail
- les lèvres

[4] En réalité, une fraction de ces images (de l'ordre de 30%) n'est pas utilisable.

- le vélum, de manière assez similaire à Flament
- la mandibule.

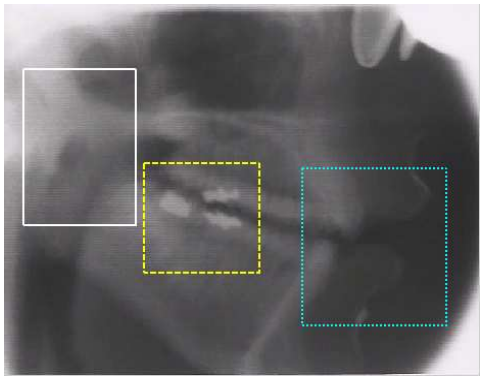
Les paragraphes qui suivent détaillent les réglages des paramètres pour chaque articulateur, ainsi que quelques spécificités observées et propres à chacun d'eux.

2.1.1. Réglage des paramètres

Nous précisons ici les paramètres utilisés pour l'extraction des contours des articulateurs. Le tableau qui suit présente le choix des paramètres : le nombre de points marqués et le nombre de degrés de liberté, le nombre d'images clefs et la taille du cadre défini pour l'indexation, et enfin le nombre de coefficients DCT considérés dans les calculs de distance euclidienne. Une présentation plus précise des degrés de liberté sera présentée un peu plus loin. Les figures 33a et 33b montrent le cadrage des images, défini pour chaque articulateur.

Articulateur	**Paramètres**				
	Points	**Degrés de liberté**	**Images clefs**		**Coefficients DCT**
			Nombre	**Taille du cadre**	
Langue	13	15	200	105*95	24*24
Pointe de la langue	3	5	200	48*75	24*24
Vélum	13	14	100	142*186	24*24
Lèvre supérieure	6	8	200	182*186	24*24
Lèvre inférieure	6	8			
Mandibule	4	4	60	131*131	12*12

Table 4 : Paramètres de la méthode de rétro-marquage réglés pour l'extraction géométrique de divers articulateurs dans la séquence Laval43.


(a)

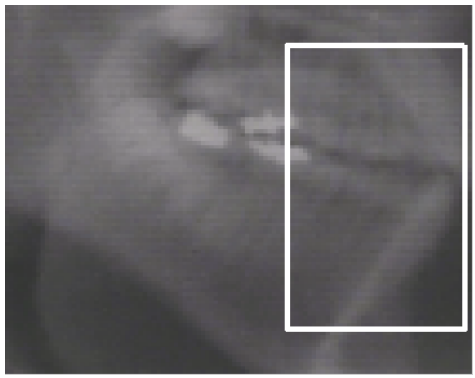

(b)

Figure 33 : Cadres spécifiques pour chaque articulateur.
(a) De gauche à droite, les cadres se focalisent respectivement sur le vélum, la mandibule et les lèvres.
(b) L'image complète correspond au cadre pris en compte pour l'estimation de la langue, le rectangle blanc délimite le cadre spécifique à l'estimation indépendante de la pointe.

Le calcul des coefficients DCT relatifs à la langue ou à la pointe est réalisé à partir d'un cadre défini sur des images qui ont été décimées au préalable (Fig. 33b), ce qui n'est pas le cas pour les autres articulateurs. Dans ces autres cas, les cadres ont été définis directement sur les grandes images d'origine, comme sur la figure 33a.

Pour chaque articulateur (mis à part la mandibule), la mesure de similarité utilisée pour l'indexation tient compte de 575 coefficients DCT, c'est-à-dire du bloc de 24*24 coefficients dans le coin en haut à gauche de la matrice DCT calculée pour chaque image (cadre spécifique à l'articulateur) et le premier coefficient (moyenne) n'est pas pris en considération. Cependant, ce nombre de coefficients aurait pu être réduit pour certains articulateurs, compte-tenu de la taille des images et de l'apparence propre de chaque articulateur. Le résultat n'aurait été que peu différent. Les graphes qui suivent (Fig. 34) montrent, pour les lèvres et le vélum, les similitudes d'indexation en fonction du nombre de coefficients DCT pris en compte. Considérant comme référence l'indexation faite avec 24*24 coefficients, nous comparons le pourcentage d'index communs sur la séquence avec d'autres blocs de coefficients (à chaque fois, le premier coefficient, correspondant à la moyenne, n'est pas pris en considération). Avec un bloc de 12*12, l'indexation est similaire à 90% pour les lèvres. En considérant les 2 premiers voisins, l'indexation est similaire à plus de 90% dès 6*6 coefficients DCT, aussi bien pour les lèvres que pour le vélum. Et avec les 4 premiers voisins, on atteint 100% dès 6*6 coefficients pour le vélum et 12*12 coefficients pour les lèvres (on a déjà 97% d'index communs avec un bloc de 6*6). En prenant en compte 4 voisins, comme c'est le cas dans nos conditions d'application de la méthode, le résultat sera quasiment identique avec 24*24 coefficients ou moins : la multi-indexation atténue fortement l'influence du nombre de coefficients DCT. Les index voisins sont identiques, la différence qui subsiste est liée à la pondération de la moyenne en fonction de l'ordre de ces index.
Des mesures de type $Etot_1$, que nous avons décrites au chapitre 2 et dont nous reparlerons à la fin de ce paragraphe, sont réalisées sur les lèvres et le vélum en faisant varier le nombre de coefficients DCT pris en compte dans l'indexation. Le graphe 35 montre qu'il n'y a quasiment pas d'influence du nombre de coefficients pour le vélum, un peu plus pour les lèvres. Aussi, on aurait pu, sans changer beaucoup les résultats, prendre en compte moins de coefficients DCT, ce qui diminuerait légèrement le temps de calcul. En pratique, ce temps de calcul n'étant pas rédhibitoire, nous avons conservé 24*24 coefficients.

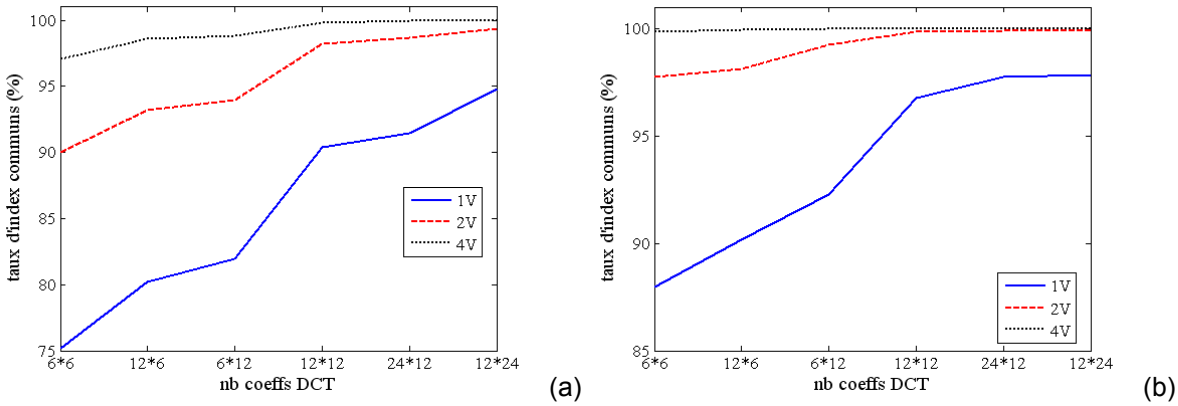


Figure 34 : Sachant l'index calculé avec 24*24 coefficients DCT, pourcentage de trames indexées par ce même index en fonction du nombre de coefficients DCT et du voisinage. (a) Cas des lèvres – (b) Cas du vélum.

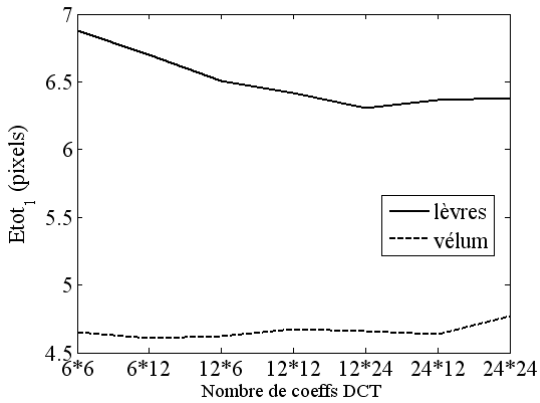


Figure 35 : Erreur $Etot_1$ en fonction du nombre de coefficients DCT pris en compte dans l'indexation (indexation simple et pas de filtrage temporel) pour les lèvres (175 clefs) et le vélum (75 clefs).

L'organisation des degrés de liberté pour chaque articulateur est illustrée ci-dessous. Pour chacun d'eux, le choix est réalisé de telle sorte que chaque point puisse être marqué pour toutes les images de la séquence et ainsi qu'il n'y ait jamais de données manquantes.

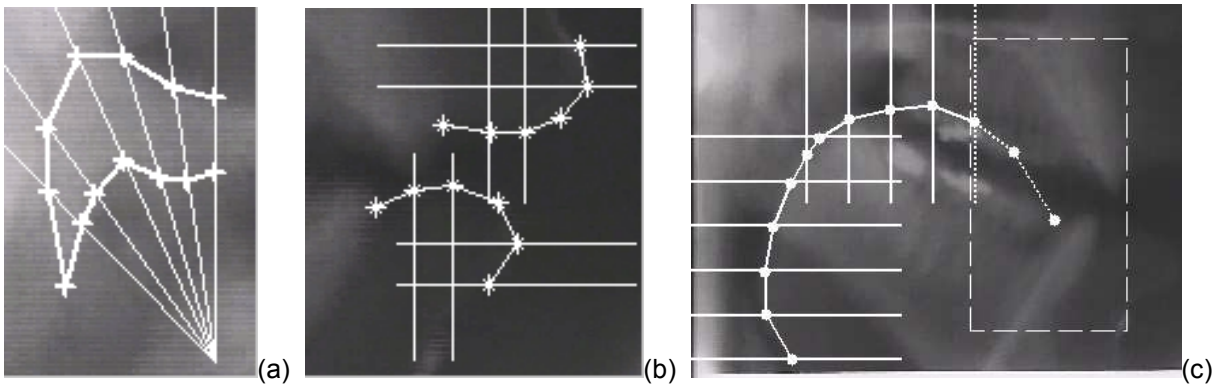


Figure 36 : Degrés de liberté pour différents articulateurs.
(a) Le marquage manuel du vélum est guidé par une grille polaire.
(b) Pour chacune des 2 lèvres, 4 lignes définissent 4 ddl sur 8.
(c) Le dos et la base de la langue sont décrits à partir de lignes horizontales et verticales, la pointe est représentée par 5 degrés de liberté (cadre pointillé).

2.1.2. Pointe de la langue

Les contours de la langue sont extraits de manière analogue à la base Wioland mais un traitement spécifique est appliqué à la pointe. En effet, la pointe de la langue est nettement plus visible dans ce film et une adaptation de la méthode a été réalisée afin de mieux capturer ses mouvements rapides et parfois relativement indépendants, même si la pointe est portée par le dos et la base de la langue. Cette adaptation consiste en un double marquage associant une estimation globale des 15 degrés de liberté définis dans le tableau 4 et sur la figure 36c, et une seconde estimation, spécifique de la pointe.

L'estimation globale correspond au traitement réalisé similairement sur Wioland ou Flament, à partir des coefficients DCT calculés sur le cadre présenté figure 33b.

La seconde estimation est spécifique à la pointe et inclut seulement 3 points ou 5 degrés de liberté, ceux représentés en pointillés sur la figure 36c. Cette estimation est calculée à partir d'un cadre focalisé sur la pointe. Ce cadre spécifique est défini en délimitant une zone réduite à l'avant du conduit vocal (on l'observe sur la figure 33b ou 36c).

La fusion de ces deux estimations est réalisée par substitution des 5 ddl de la pointe dans l'estimation globale : on combine ainsi les 10 points (ou 10 ddl) les plus en arrière de la langue estimés globalement, et les 3 points (ou 5 ddl) les plus en avant estimés spécifiquement.

Pour reconstruire les mouvements de la langue et de la pointe, la méthode semi-automatique a été appliquée avec 200 images clefs marquées.

2.1.3. Voile du palais

Le vélum est un articulateur qui est généralement difficile à enregistrer. Dans le film Laval43, il est bien visible et l'extraction des mouvements de cet articulateur est ainsi une source nouvelle de données, comme nous en reparlerons en fin de manuscrit. Il est marqué, de façon tout à fait similaire au vélum dans la séquence Flament. Une grille polaire est utilisée pour permettre le marquage de 13 points avec 14 degrés de liberté (Fig. 36a). Cet articulateur est relativement facile à marquer. A ce propos, une étude cinéradiographique du vélum [SFK80] montre que des mesures directes, sans utilisation de marqueurs, sont fiables pour le vélum.

La variabilité des mouvements étant plus réduite que celle de la langue, 100 images clefs sont tout à fait suffisantes pour permettre une reconstruction correcte du mouvement du vélum.

2.1.4. Lèvres

Les points de marquage des lèvres sont définis de façon analogue pour la lèvre supérieure et la lèvre inférieure. Pour chaque lèvre, 4 points sont définis avec 1 degré de liberté chacun et 2 points sont libres. Un de ces points libres correspond au point d'intersection sur l'image entre le contour de la lèvre et celui de l'incisive. Les lèvres sont plus difficiles à marquer à cause du faible contraste de la zone de l'image à marquer.
La qualité de la reconstruction du mouvement n'est pas tout à fait équivalente pour les 2 lèvres. En effet, alors que les mouvements de la lèvre supérieure sont relativement bien suivis, on remarque visuellement que ceux de la lèvre inférieure sont reconstruits avec moins de précision. Comment expliquer ces différences et les erreurs de reconstruction que l'on observe ?

Avant tout, une observation minutieuse des configurations géométriques estimées des lèvres, superposées aux images d'origine, nous a permis de mettre en évidence des cas particuliers d'erreurs. L'image qui suit (Fig. 37a) montre la configuration du conduit vocal pour une bilabiale plosive [p]. Malgré la difficulté d'observation des lèvres à cause du contraste, l'image radiographique montre bien qu'à cet instant les lèvres supérieure et inférieure se touchent, alors que le marquage géométrique estimé montre tout à fait le contraire ! On voit cependant que les incisives sont, elles, correctement marquées. Plusieurs images similaires ont montré le même résultat : pour des configurations du conduit vocal de type lèvres fermées avec les incisives écartées, les lèvres ne sont jamais marquées fermées. Ces erreurs de marquage sont dues à l'indexation qui semble s'appuyer sur la position des dents, du fait de leur fort contraste comparé à celui des lèvres. Pour pallier ce défaut de marquage, nous avons réalisé une nouvelle indexation automatique des lèvres à partir d'un cadre modifié. Les coefficients DCT, précédemment calculés sur le cadre présenté figure 38a, sont désormais calculés sur le cadre, de même dimension mais sur lequel un cache noir oblique a été inséré de façon à supprimer l'influence des dents dans l'indexation. Ce nouveau cadre est présenté figure 38b. L'application sur les lèvres de l'algorithme de rétro-marquage et des traitements postérieurs de la méthode, à partir de ces nouveaux coefficients DCT, améliore le marquage géométrique des lèvres. La figure 37b montre la même image que la figure 37a avec le nouveau marquage des lèvres, qui sont cette fois représentées quasiment fermées. C'est ce nouveau marquage géométrique qui est désormais considéré pour les lèvres. On montrera au paragraphe suivant l'apport de cette indexation sur l'erreur de reconstruction.

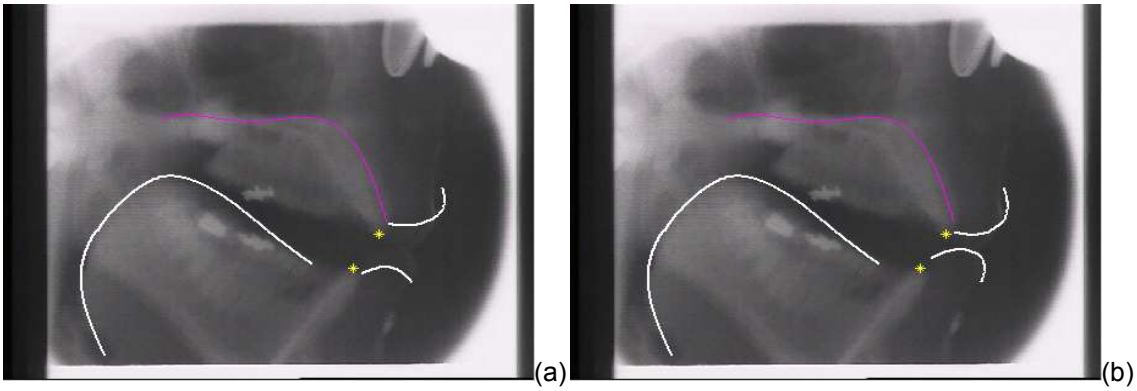


Figure 37 : Consonne [p]
(a) Erreur de marquage géométrique sur les lèvres, qui devraient se toucher.
(b) Correction du marquage géométrique des lèvres grâce au masquage des incisives pour l'indexation.

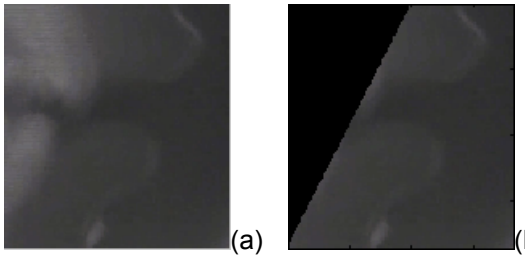


Figure 38 : (a) Cadre initial utilisé pour l'indexation commune des incisives et des lèvres.
(b) Cadre utilisé pour l'indexation des lèvres : l'influence des incisives a été supprimée à l'aide d'un cache noir.

Une erreur de reconstruction persiste malgré tout et peut s'expliquer par plusieurs raisons.
Tout d'abord, la technique de rétro-marquage est appliquée sur les 2 lèvres simultanément ; or, la lèvre supérieure est moins mobile que la lèvre inférieure, cette différence de mobilité provoque des erreurs de reconstruction d'amplitude différente (plus importantes pour la lèvre inférieure). L'erreur de reconstruction est en effet fonction de la mobilité. Pour les lèvres, comme pour les autres articulateurs, nous vérifions que la corrélation est forte (> 0,8) entre la variance de l'erreur et celle du mouvement, pour les différents degrés de liberté.
Ensuite, le faible contraste du cadre observé pour les lèvres est aussi une cause possible aux erreurs d'indexation, dans la mesure où les distances calculées pour l'indexation (à partir des coefficients DCT) dépendent du contraste. La variance moyenne des coefficients DCT du cadre des lèvres est 2 fois (resp. 6 fois) moins grande que celle du cadre de la langue (resp. du vélum).
Enfin, il faut aussi remarquer que les erreurs de marquage de la lèvre inférieure observées sur les images cadrées, plus petites, semblent évidemment moins importantes quand le marquage des lèvres est superposé aux grandes images d'origine, le suivi des mouvements est alors globalement correct. L'erreur absolue par degré de liberté est comparable à celle obtenue pour les autres articulateurs.

Cette différence de reconstruction du mouvement entre les 2 lèvres est mise en évidence avec une mesure d'erreur $Etot_1$, comme on le voit au tableau 5.

2.1.5. Evaluation

Pour chaque articulateur, nous évaluons l'erreur d'estimation RMS, en faisant varier un certain nombre de paramètres, comme le nombre de clefs ou les traitements postérieurs. Cette évaluation a été réalisée en détail pour la langue dans Wioland (chapitre 2). Nous avons également avec l'analyse du nombre de coefficients DCT effectué une mesure de cette erreur pour le vélum et les lèvres au paragraphe 2.1.1..

On observe ici, à titre d'exemple (Fig. 39), l'évolution de l'erreur pour les lèvres (supérieure et inférieure, soit les 16 degrés de liberté) dans le cas d'une simple indexation ou d'une multi-indexation d'ordre 4, avec ou sans filtrage, pour 175 clefs. On constate que la multi-indexation a ici aussi pour effet de réduire l'erreur. Cependant, on note que, contrairement aux résultats sur la langue dans Wioland, le filtrage temporel a assez peu d'influence sur l'erreur de reconstruction. La figure permet de plus de comparer les indexations suivant le cadre utilisé. La « nouvelle » indexation ou indexation « corrigée », obtenue à partir du cadre défini sur les lèvres avec masquage des incisives permet de diminuer l'erreur de reconstruction de 0,8 pixels/ddl par rapport à l'indexation réalisée initialement (on passe de 5 à 4,2 pixels/ddl en moyenne pour les 2 lèvres). En séparant l'influence des 2 lèvres, on constate que $Etot_1$ passe de 6 à 5 pixels/ddl pour la lèvre inférieure et de 4 à 3,4 pixels/ddl pour la lèvre supérieure.

Les erreurs moyennes de reconstruction dans les meilleures conditions d'application de la méthode sont résumées dans le tableau 5 pour différents articulateurs. L'algorithme de rétro-marquage est appliqué avec 4 voisins et filtrage temporel. Le nombre d'images considérées pour le Jackknife varie suivant l'articulateur : n_1 est le nombre d'images clefs utilisées par la méthode, n_2 est le nombre d'images tests sur lesquelles on évalue l'erreur.

La mesure donnée pour la langue combine l'estimation globale de la langue et la mesure spécifique de la pointe : l'erreur moyenne est calculée en prenant en considération les 5 ddl de la pointe par l'estimation spécifique et les 10 ddl restants par l'estimation globale. La mesure donnée pour la pointe ne concerne bien sûr que les 5 ddl de pointe (à noter que l'erreur pour ces 5 degrés de liberté obtenue en prenant en compte l'estimation globale, et non spécifique, est de 12 pixels, soit un gain de 2 pixels grâce à la mesure indépendante de la pointe).

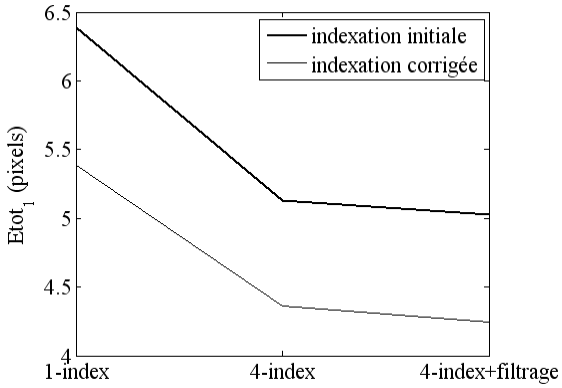


Figure 39 : Comparaison, suivant l'indexation, de l'évolution de l'erreur $Etot_1$ sur les 16 degrés de liberté des lèvres pour différents traitements postérieurs et 175 clefs.

	Degrés de liberté	$Etot_1$ (pixels/ddl)	Jackknife n_2 / n_1
Pointe	5	10	25 / 175
Langue	15	8,8	25 / 175
Vélum	14	3,4	25 / 75
Lèvre supérieure	8	3,4	25 / 175
Lèvre inférieure	8	5	25 / 175

Table 5 : Résultats de l'évaluation $Etot_1$ pour différents articulateurs (pour rappel, les grandes images sont de taille 720*480 pixels).

Une analyse du nombre de clefs, similaire à celle menée sur Wioland, est aussi présentée ici. Elle concerne les lèvres, sans filtrage temporel et avec simple indexation. Nous avons tracé la valeur de l'erreur de reconstruction RMS pour 100, 150 et 175 clefs, en représentation Log-Log.

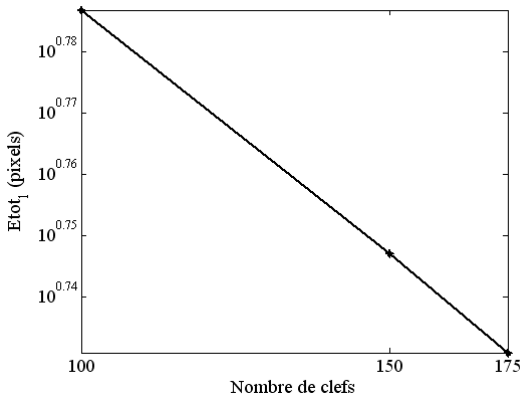


Figure 40 : Evolution Log-Log de l'erreur $Etot_1$ sur 16 degrés de liberté des lèvres (sans filtrage, simple indexation) en fonction du nombre d'images clefs.

A nouveau, nous obtenons une droite (Fig. 40). De là, nous tirons une estimation de l'évolution de l'erreur : étant donnée l'équation de la droite, nous évaluons à 428 le nombre d'images clefs qu'il serait nécessaire de marquer pour réduire de 1 pixel l'erreur par rapport

à la valeur obtenue pour 175 clefs. Cette même interpolation a été effectuée pour les autres articulateurs, et en particulier pour la pointe, où pour réduire de 1 pixel la valeur de l'erreur obtenue avec 175 clefs, on estime à 250 le nombre d'images clefs à marquer.

2.1.6. Mâchoires

Le marquage de la mandibule est aussi réalisé par rétro-marquage. Une première étape consiste à marquer les dents supérieures et inférieures. Pour cela, quelques points sont marqués sur les molaires supérieures (4 points) et inférieures (4 points) de 50 images clefs. Ce faible nombre est suffisant compte-tenu des mouvements simples des mâchoires. Les points (Fig. 41) sont marqués sur les dents à fort contraste. L'indexation automatique de la base complète est réalisée à partir de coefficients DCT d'un cadre (Fig. 33a) basé sur le fort contraste de ces molaires.

Ceci nous permet de tracer deux lignes, une représentant la limite supérieure des dents inférieures et une autre représentant la limite inférieure des dents supérieures.

Les incisives avant supérieure et inférieure sont marquées manuellement sur les mêmes 200 images clefs que pour les lèvres, puis elles sont marquées automatiquement sur l'ensemble de la base à partir du cadre initial utilisé pour le marquage des lèvres (Fig. 38a), c'est-à-dire la cadre englobant les incisives avant et les lèvres.

La mâchoire inférieure étant rigide, on considère que les distances entre les dents de la mâchoire inférieure et le bas de cette mâchoire sont constantes au cours de la séquence. Ceci nous permet, à partir du marquage des dents inférieures, de définir quelques points sur le bas et l'avant de la mâchoire inférieure et de marquer ainsi sommairement la mandibule avec 2 droites, comme on l'observe sur la figure 41.

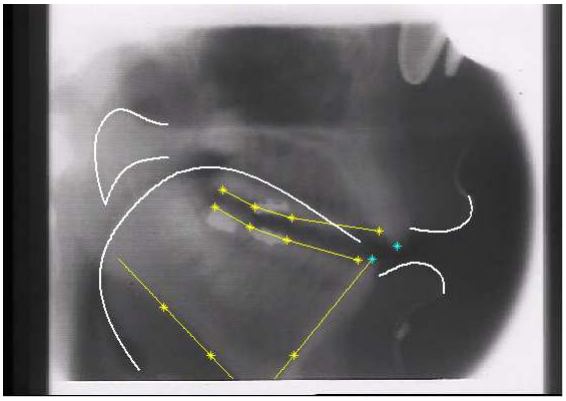


Figure 41 : Marquage des dents et de la mâchoire inférieure.

2.2. Contour complet du conduit vocal

De façon à reconstruire les mouvements du conduit vocal complet, les parties fixes du conduit vocal sont également marquées. Pour le palais, on définit une fois pour toutes le marquage qui correspond au mieux à la majorité des formes de cette partie dans la base.

Le pharynx n'est pas tout à fait fixe, on observe que sa partie haute bouge un peu. Pour prendre en compte ce mouvement, le pharynx a été marqué, à partir de la méthode semi-automatique appliquée avec 50 images clefs, un cadre spécifique et 5 points à 1 ddl (ordonnées fixées).

A ce stade, chaque articulateur a été marqué indépendamment et est décrit par quelques points ou degrés de liberté, pour chaque image de la base. En reliant ces points, on obtient une première représentation du contour de chaque articulateur.

Une étape de mise en forme de la représentation des contours a ensuite été élaborée. Il s'agit d'améliorer l'apparence des configurations géométriques obtenues en les lissant. Des interpolations de type spline ont été effectuées, indépendamment pour chaque articulateur. Cette notion de lissage spline a déjà été mentionnée. Il est spécifique à chaque articulateur et consiste à interpoler les points estimés par un polynôme dont l'ordre varie suivant l'articulateur. Les ordres choisis pour ces polynômes sont résumés dans le tableau 6.

Chaque lèvre est représentée par un polynôme d'ordre 3, le pharynx par un polynôme d'ordre 2 et le palais par un polynôme d'ordre 4.

Compte-tenu de la forme du vélum, un seul polynôme n'est pas envisageable pour marquer le contour de cet articulateur. Le marquage a été réalisé en 3 parties avec 3 polynômes d'ordre 2 : un pour la partie supérieure, un pour la partie inférieure et un pour représenter la « pointe » du vélum (en prenant en compte le point à 2 ddl et les 2 points adjacents).

Les points estimés pour la langue forment un contour qui n'est pas nécessairement une fonction de l'abscisse ou de l'ordonnée. L'utilisation d'un unique polynôme pour interpoler les points n'est pas adaptée. Nous avons choisi de représenter la langue avec deux polynômes d'ordre 3, que nous avons ensuite combinés. La première interpolation prend en compte 9 points en partant de la pointe, la seconde prend en compte 8 points en partant du point le plus bas dans le pharynx.

	Langue	Lèvre inférieure	Lèvre supérieure	Palais	Pharynx	Vélum
Nombre de polynômes	2	1	1	1	1	3
Ordre des polynômes	3	3	3	4	2	2

Table 6 : Nombre et ordre des polynômes d'interpolation pour les différents articulateurs.

En combinant tous ces contours et en prêtant attention aux jonctions entre les différents articulateurs, on parvient ainsi à réaliser un contour complet du tractus vocal et cela pour chacune des images de la base.

Les jonctions entre les articulateurs sont pour la plupart de simples interpolations linéaires entre les points estimés.

Un effort particulier a été mené pour la langue et principalement pour la pointe, ainsi que pour définir la cavité sublinguale. Le plancher sous-lingual a été marqué sur quelques images où il est visible. Ensuite en considérant que ce plancher bouge de façon solidaire avec les mouvements de fermeture et d'ouverture de la mâchoire, sa position a pu être estimée pour les images de la séquence complète.

Il est nécessaire de définir ce plancher pour la complétion du conduit vocal, car il permet d'évaluer ce qu'on appelle « la cavité avant », qui correspond à la cavité en aval de la pointe de la langue. En effet, lors de la production de certains sons de parole, le recul de la langue et du point de constriction entraîne un changement de volume de la cavité en avant de la constriction. Une cavité sublinguale est alors créée si la langue s'éloigne des incisives. La cavité antérieure est importante dans la mesure où il a été montré qu'un changement mineur dans sa morphologie pouvait induire des différences en terme de production, de consonnes notamment ([Tod06, Sha91]).

La pointe de la langue est finalement représentée sous forme d'un demi-cercle et la jonction entre cette pointe et le plancher sous-lingual estimé est réalisée par un polynôme d'ordre 2.

Tous ces raccords sont observables sur la figure 42 et permettent ainsi de disposer d'une représentation complète du conduit vocal, de la glotte jusqu'aux lèvres, pour la séquence Laval43.

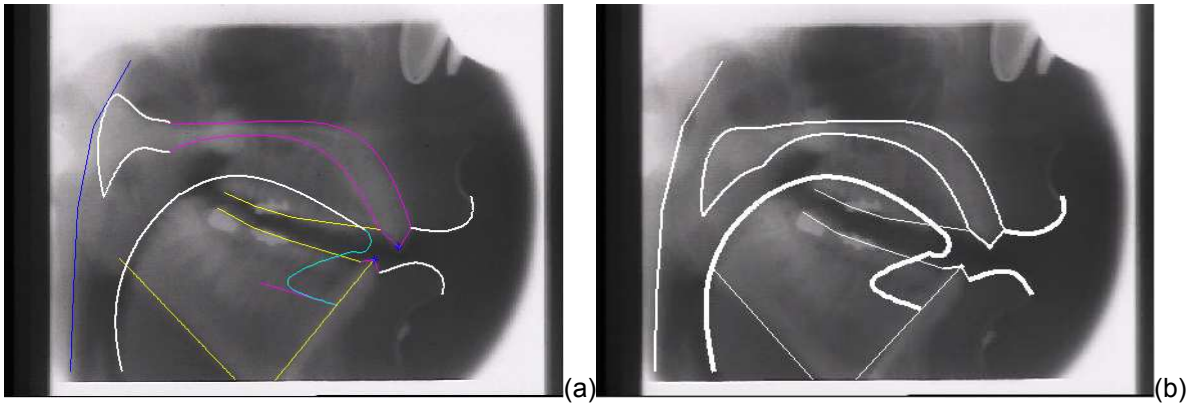

(a) (b)

Figure 42 : Contour complet du conduit vocal pour une image de la séquence Laval43.

2.3. Une étude comparative sur la langue avec une autre méthode d'extraction

Le choix de la séquence Laval43 parmi toutes celles de la base de données ATR est lié au fait que cette séquence a été entièrement traitée par une autre méthode semi-automatique d'extraction de contours, celle de Thimm et Luettin en 1999 à l'IDIAP [TL99]. A titre d'évaluation, nous comparons leur estimation de contour de langue avec notre estimation obtenue par rétro-marquage.

Cette étude comparative permet de mettre en parallèle deux approches :

- Thimm et Luettin proposent une extraction directe de l'information géométrique des contours puis un suivi temporel pour reconstruire le mouvement.
- Nous proposons une extraction indirecte des contours et reconstruisons ensuite le mouvement par interpolation temporelle.

2.3.1. Méthode d'extraction mise en place à l'IDIAP

Le film Laval43 a été marqué en totalité par Thimm et Luettin et les données géométriques sont disponibles sur le web. Ils proposent un algorithme de suivi de contours qui peut être appliqué à des objets dont la position générale est connue (ou au moins limitée à un nombre restreint de positions) mais qui sont sujets à des déformations non linéaires rapides. Cette approche associe les contours à des états et les déformations à des transitions entre les états.

Partant du constat que les dents sont plus faciles à distinguer du fait d'un plus fort contraste (il s'agit en fait de marqueurs anatomiques), la première étape consiste à les localiser.

Un traitement préliminaire est réalisé sur les images pour réduire la variabilité de l'éclairage au cours de la séquence. Les images sont filtrées par un filtre gaussien (les pixels des images ayant une valeur inférieure à un seuil g sont assignés à cette valeur g) puis normalisées de telle sorte que leurs histogrammes couvrent toute la palette possible de niveaux de gris. Cette normalisation des histogrammes n'est pas suffisante à cause de distorsions non-linéaires auxquelles ils sont sujets. Ces distorsions sont compensées en modifiant l'histogramme des dents au cours de la procédure de suivi des dents.

Une fois les dents localisées, l'extraction des autres articulateurs utilise les bénéfices de la normalisation d'histogrammes réalisée et les contraintes imposées par la position des dents.

L'algorithme de tracking est basé sur une procédure d'appariement et fait appel à des images dites d'état. Pour cela, les images normalisées sont soumises à un détecteur de Canny qui en extrait les contours. Des contours représentatifs de l'articulateur à suivre sont extraits de ces images et constituent les images d'état. Ces images sont inversées et convoluées par un filtre gaussien, dont la variance est directement liée à la variabilité de

l'articulateur en question. Ces images sont alors utilisées dans la procédure d'appariement qui recherche un score optimal entre ces images et les images d'origine.
Pour assurer de bons résultats, les contours utilisés pour les images d'état doivent être sélectionnés avec soin. Le nombre de ces images est variable suivant l'articulateur (226 pour la langue, 75 pour la lèvre supérieure, 105 pour la lèvre inférieure…) et chaque contour est défini spécifiquement (par exemple, le contour de la langue est considéré du point le plus bas du pharynx jusqu'à la pointe, à condition qu'elle soit visible, et il peut parfois être invisible par endroits à cause des dents). La sélection d'un ensemble représentatif de ces images d'état est réalisée de manière itérative : dans un premier temps, les images d'état sont choisies aléatoirement, puis le tracking de l'articulateur est réalisé à partir de cet ensemble. Si ce suivi n'est pas jugé correct, i.e. si l'articulateur est mal localisé dans certaines images, les contours associés à ces images sont alors ajoutés à l'ensemble d'images d'état et la procédure est ré-itérée. Une part d'intervention manuelle est donc mise en jeu.
Pour compléter la procédure simple de suivi, l'information temporelle est utilisée de façon à réduire le nombre d'erreurs. En considérant que la déformation d'un articulateur entre 2 trames consécutives est faible, les états atteignables depuis un état donné est un petit sous-ensemble de l'espace complet d'apprentissage. Il est alors nécessaire de connaître les transitions possibles entre les états et d'estimer pour cela des distances entre ces états. Pendant la procédure de suivi, les transitions seront limitées à celles qui correspondent aux petits mouvements. Ces derniers sont déterminés par le calcul de distance entre les splines qui caractérisent les contours.

Les résultats obtenus par Thimm et Luettin sur la séquence Laval43, par cette méthode que nous notons TL par la suite, concernent plusieurs articulateurs du conduit vocal, mais ne permettent pas de reconstruire sa forme complète, en particulier car la pointe de la langue est souvent manquante à cause de l'occlusion par les mâchoires.
Nous nous intéresserons ici aux résultats concernant la langue, dans le but de comparer directement cette méthode à la nôtre, que nous notons FB.
Pour le suivi de la langue, Thimm et Luettin complètent leur méthode par une soustraction de l'arrière-plan (mâchoires supérieure et inférieure), pour compenser le fait que la langue est souvent cachée par ces mâchoires et que de fait, son contour est difficilement visible.

Les résultats de Thimm et Luettin pour la séquence Laval43 ont été récupérés sur internet sur le site *http://www.idiap.ch/machine_learning.php?project=64*. A chaque image de la séquence est associé un fichier texte donnant, par articulateur, les coordonnées de points. Ces points correspondent à des points de contrôle de splines qui représentent les contours

de différents articulateurs (langue, lèvres supérieure et inférieure, pharynx, dents …), comme on en voit un exemple pour une image sur la figure 43. La fonction qui permet l'obtention des splines à partir des points de contrôle est fournie dans les codes sources qui sont documentés grâce aux rapports techniques également disponibles sur le site.

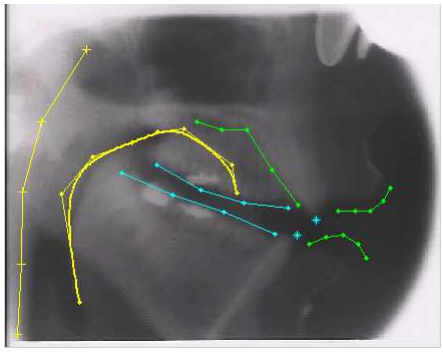


Figure 43 : Contours extraits par la méthode de Thimm et Luettin [TL99] sur une image de Laval43.

Ces documents nous permettent de conditionner les résultats ainsi que le film Laval43, de façon à travailler dans des conditions analogues pour les 2 jeux de données (TL et FB) et à aboutir à une comparaison objective des données TL avec les nôtres.

En particulier, les résultats de Thimm et Luettin ont été obtenus pour un format d'images légèrement différent de celui que nous utilisons directement à partir du DVD. A partir des données du rapport technique [Thi99], nous avons redimensionné[5] nos images et par suite notre marquage pour la langue (puisque c'est l'articulateur que nous considérons dans la comparaison), de façon à comparer des résultats obtenus dans les mêmes conditions (images de taille 564*460).

Concernant le contour de la langue, pour la méthode TL, nous disposons pour chaque image d'un jeu de splines. Ces splines sont définies par des points de contrôle mentionnés précédemment, mais une fonction permet d'obtenir les coordonnées de plusieurs centaines de points directement sur ces splines. Nous noterons par la suite, $S_{\mathrm{TL}i}$ le jeu de splines, pour la séquence, définissant le contour de la langue estimé par Thimm et Luettin.

2.3.2. Comparaison des méthodes

Notre méthode de rétro-marquage, notée FB, appliquée à la langue avec 200 images clefs avec le modèle M5 (défini au chapitre 2 §2.4., et correspondant aux conditions d'application :

[5] La transformation réalisée sur les images est la suivante : les images disponibles sur le DVD au format 720*480 sont redimensionnées au format 640*480 (réduction en largeur) puis découpées en laissant 50 pixels à gauche, 26 à droite, 10 en haut et 10 en bas. Les images finalement considérées sont de taille 564*460, ce qui correspond exactement aux dimensions mentionnées par Thimm et Luettin. Le marquage estimé sur les images 720*480 est appliqué sur ces images 564*460 en redimensionnant les marques dans la largeur (rapport 720/640) et en tenant compte du décalage de 50 pixels à gauche et de 10 pixels à droite.

multi-indexation d'ordre 4, filtrage temporel et interpolation polynomiale) a permis d'obtenir un jeu de splines, notés S_{FBi}.

Pour comparer les 2 estimations à partir des 2 jeux de splines S_{FBi} et S_{TLi}, 2 types de mesures sont considérés :

- une mesure relative *D* calcule pour chaque image la distance entre les 2 splines.

La méthode de mesure que nous utilisons a été proposée par Thimm [Thi99] et fait appel à une notion de surface entre les splines. Pour comparer 2 splines, il faut en général considérer que les points de départ et d'arrivée de la première spline ne correspondent pas forcément à ceux de la seconde. Les parties extrêmes sont négligées dans le calcul de la distance entre les splines. On ne conserve que la partie commune, représentée en gris sur la figure suivante. La distance calculée pour mettre en évidence l'écart entre 2 splines est l'aire comprise entre les 2 courbes splines, normalisée par la somme de leurs longueurs.

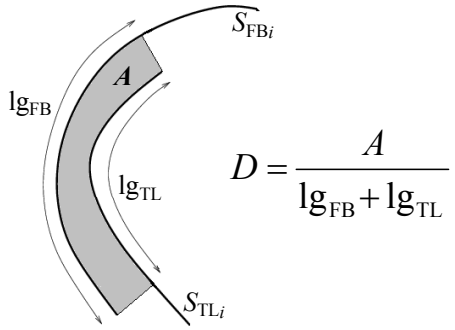


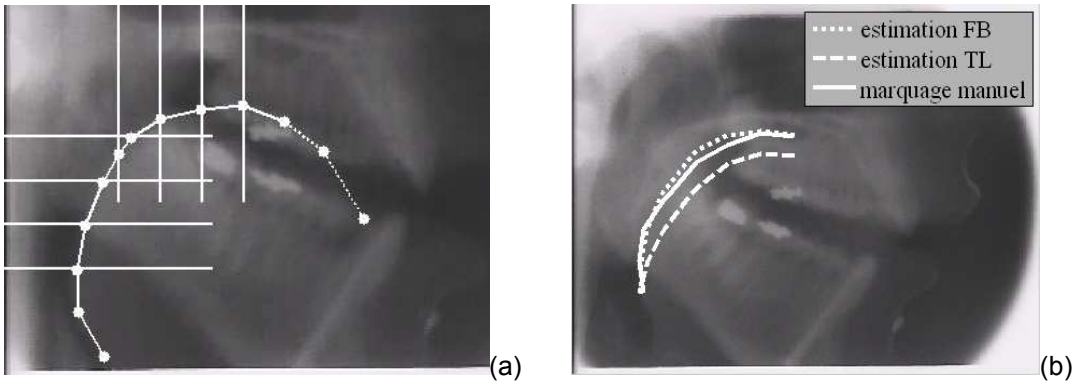
Figure 44 : Mesure de distance *D* entre 2 splines.

- une mesure d'erreur ($Etot_1$), basée sur des images tests, mesure l'écart entre le marquage manuel et chacune des 2 estimations (Fig. 45b).

Pour cette mesure, nous avons limité le nombre de degrés de liberté considérés. En effet, la pointe de la langue n'est pas toujours représentée dans l'estimation de Thimm et Luettin, il y a des données manquantes qui sont dues à la difficulté d'estimation par une approche contour.

Dans notre estimation FB, nous avons donc écarté, pour la comparaison, les 5 degrés de liberté relatifs à la pointe, c'est-à-dire les 2 points libres et le premier point avec un ddl fixé. De plus, nous omettons les derniers points marqués dans le bas du pharynx. Au final, nous limitons la comparaison à 8 degrés de liberté: il s'agit des 8 points d'intersection estimés du contour de la langue avec les lignes de marquage horizontales et verticales, représentées sur la figure 45a. Ces 8 points de marquage (à 1 degré de liberté) définissent le contour du dos et de la base de la langue.

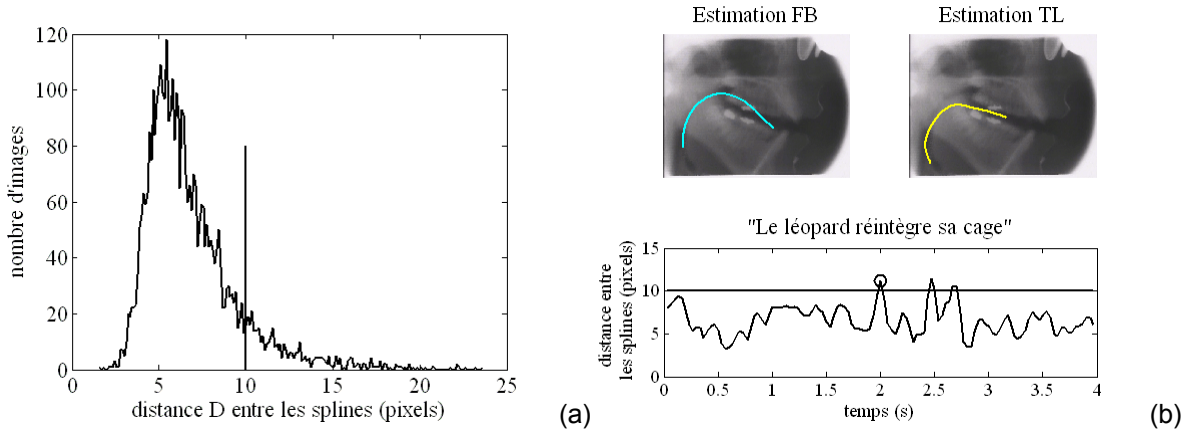
Dans l'estimation TL, ces degrés de liberté, non directement disponibles, ont été mesurés à partir des splines estimées. L'extraction des 8 degrés de liberté a été réalisée, image par image, grâce aux mêmes droites verticales et horizontales que pour FB (Fig. 45a), en déterminant, droite par droite, le point d'intersection avec la spline.



(a) (b)

Figure 45 : (a) Seuls 8 degrés de liberté (définis par les lignes horizontales et verticales) sont pris en compte pour la comparaison.
(b) Comparaison sur une image test d'un marquage manuel de la langue avec 2 marquages estimés.

La différence moyenne *D*, entre splines, entre notre estimation et celle de Thimm et Luettin est évaluée à 6,8 pixels. La distribution de *D* (Fig. 46 a) montre que pour environ 10% de la base, il existe un décrochage entre les 2 méthodes, une incohérence entre les 2 estimations. Nous caractérisons ce décrochage par un seuil fixé à 10 pixels. Au dessus de ce seuil, l'écart moyen entre les 2 estimations est alors de 12,7 pixels. Nous observons visuellement un exemple de décrochage sur la figure 46b. On peut noter que dans ce cas, le contour associé par rétro-marquage est correct, et qu'il inclut la pointe.



(a) (b)

Figure 46 : (a) Répartition de la différence *D* entre les 2 estimations, et définition d'un seuil de décrochage *D*>10.
(b) Décrochage observé au milieu de la séquence considérée, avec les deux contours estimés à cet instant.

Avec notre méthode de rétro-marquage, l'erreur $Etot_1$ calculée sur 8 degrés de liberté (pour le dos et la base de la langue) varie en fonction du nombre d'images clefs et des traitements

postérieurs (Fig. 47). Dans les meilleures conditions d'application de la méthode (c'est-à-dire avec 175 clefs, une indexation avec 4 voisins et un filtrage temporel), nous obtenons une erreur inférieure à 8 pixels/ddl. L'erreur d'estimation pour la méthode de Thimm et Luettin (1999) est, elle, évaluée autour de 20 pixels/ddl, sachant que cette section de la langue a une longueur estimée de 250 pixels.

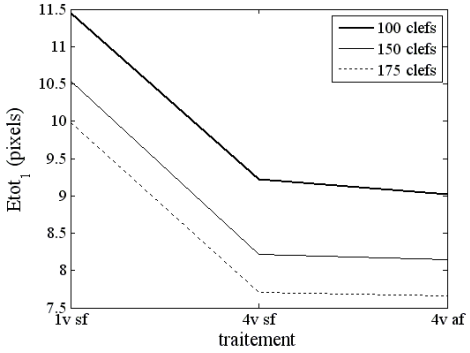


Figure 47 : Evolution de l'erreur $Etot_1$ pour notre méthode de marquage avec différents traitements postérieurs (indexation simple ou multiple, sans ou avec filtrage) sur les 8 degrés de liberté considérés.

On peut remarquer que notre méthode préserve le contour de la langue sur toutes les images, même dans les cas où elle n'est pas entièrement visible. Ceci n'est pas toujours possible avec une approche de type contours, comme celle de Thimm et Luettin, et la perte d'information est très nuisible pour la pointe de la langue, puisqu'elle ne permet pas d'accéder à l'un des 3 paramètres articulatoires de la langue.

CHAPITRE 4 : CALCUL DE LA FONCTION D'AIRE À PARTIR DES CONTOURS GÉOMÉTRIQUES EXTRAITS DE LA SÉQUENCE LAVAL43

En production de parole, nous l'avons dit, les représentations de données articulatoires sont nombreuses, il existe diverses possibilités pour décrire la forme du conduit vocal. La modélisation la plus directe est définie par un certain nombre de sections réparties de la glotte aux lèvres et l'aire pour chacune de ces sections. Cette représentation sous forme de sections médio-sagittales est une des mieux adaptées et des plus utilisées ([HS64], [Ohm66], [Wak73]), dans la mesure où la connaissance de la fonction d'aire du conduit vocal est indispensable pour relier disposition articulatoire et signal acoustique de parole. Le calcul de cette fonction d'aire se décompose en 2 parties : une mesure des distances médio-sagittales suivie d'une estimation pour passer à l'aire.

Comme on va le voir dans ce chapitre, les contours géométriques extraits par la méthode semi-automatique permettent de mesurer pour chaque image les distances sagittales le long du conduit vocal. Ces résultats nous permettent d'éviter l'élaboration d'un modèle articulatoire de commande et nous ne cherchons donc pas dans notre étude à obtenir une représentation paramétrique des mouvements du conduit vocal.

1. Des contours aux sections sagittales

Le calcul des fonctions d'aire passe nécessairement par l'évaluation des sections et distances médio-sagittales. Ceci consiste à mesurer pour différents points le long d'une grille, la hauteur du conduit vocal, c'est-à-dire la distance entre la langue d'une part et le palais ou l'arrière du pharynx d'autre part. La mesure des distances sagittales s'appuie sur l'utilisation d'une grille de référence. Aussi, avant d'expliciter la mesure effective des distances sagittales, nous présentons l'élaboration de la grille.

1.1. Mise en place de la grille

L'objectif est, classiquement, d'obtenir des sections perpendiculaires à la ligne médio-sagittale. C'est le cas des grilles semi-polaires ([HS65], [Mae79]) que la plupart des auteurs utilisent et adaptent.

Il n'est pas toujours facile d'appliquer directement une grille existante. Aussi, nous proposons une méthode plus souple et plus adaptée au découpage en sections.
La construction de la grille s'effectue en 2 étapes. D'abord nous posons une grille que nous initialisons à partir des parties fixes du conduit vocal. Dans un deuxième temps, une correction globale est appliquée à chaque ligne de la grille statique, consistant en une rotation autour de la ligne médiane pour rétablir l'orthogonalité.

1.1.1. Définition d'une grille de référence

L'initialisation de la grille est réalisée à partir des parties fixes du conduit. La grille est élaborée ligne par ligne. Les lignes sont définies pour être perpendiculaires au palais ou au pharynx. Pour cela, nous déterminons comme étant fixe une ligne décrivant le palais et le pharynx. La position de cette ligne est obtenue en faisant la moyenne de toutes les positions de ces contours au cours de la séquence.
Les positions des lignes sagittales de la grille de référence sont établies à partir d'un cercle positionné pour être tangent au palais et au pharynx. Le centre de ce cercle est estimé comme le point d'intersection de quelques droites choisies normales au palais ou au pharynx.
A partir de ce cercle, un choix de rayons est effectué pour permettre une répartition équidistante des lignes le long de la ligne palais-pharynx. Sur la figure 48, on observe le cercle ainsi que la répartition des lignes. Au moment de la reconstruction du conduit vocal complet, nous avons pris soin de représenter convenablement la cavité avant. Pour cette zone délimitée par les incisives, le plancher sous-lingual, le palais et la pointe de la langue, les lignes de la grille ont été définies, à la main, à peu près parallèles entre elles. Le choix des lignes est complété dans le pharynx avec 4 lignes parallèles entre elles.

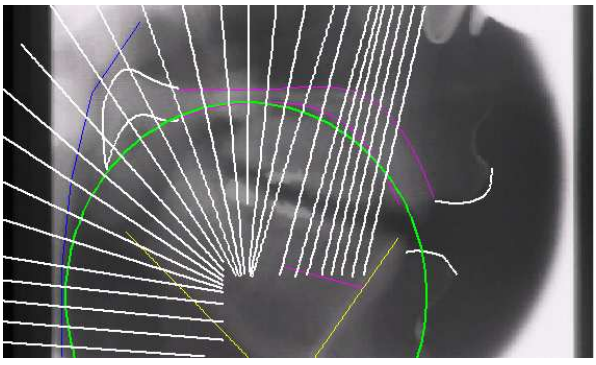


Figure 48 : Définition d'une grille statique de référence, initialisée à partir des parties fixes du conduit vocal.

La grille statique se compose au final de 27 lignes, du pharynx jusqu'aux incisives. Elle est ensuite complétée par une ligne au niveau des incisives avant et une autre pour les lèvres.

Nous disposons ainsi d'une grille permettant la mesure de 29 distances médio-sagittales des lèvres jusqu'au pharynx.

1.1.2. Correction globale de la grille

L'objectif recherché pour la mise en place d'une grille est d'obtenir des sections perpendiculaires à la ligne médio-sagittale. La grille statique définie juste avant est basée sur l'orthogonalité aux parties fixes. Nous cherchons maintenant à ajuster l'orthogonalité à la ligne médiane. La ligne médiane ou médio-sagittale représente la position « centrale » entre la langue et la ligne palais-pharynx. Elle est estimée en moyenne sur la séquence à partir des positions extraites de contours de la langue et de la position de la ligne palais-pharynx.

L'ajustement de l'orthogonalité des lignes se base sur des corrections d'angle autour de cette ligne médiane. Nous évaluons, à partir de la grille statique initiale posée, les angles entre les lignes de la grille et le contour de la langue, pour la séquence complète. Dans la mesure où la langue bouge au cours de la séquence, nous considérons les angles moyens entre la langue et les lignes. Les angles entre les lignes et la ligne palais-pharynx, fixes pour toute la séquence et proches de 90° par définition, sont également calculés.
La correction que nous effectuons sur les lignes de la grille consiste, section par section, à modifier l'angle de chaque ligne, de façon à ce que chacune soit, en moyenne, la plus orthogonale possible à la fois à la langue et à la ligne palais-pharynx, supposée fixe. Ceci implique une légère rotation des lignes autour de la ligne médiane dans le plan sagittal.
La figure suivante présente un schéma de la correction d'orientation pour une ligne ainsi que la nouvelle grille obtenue.

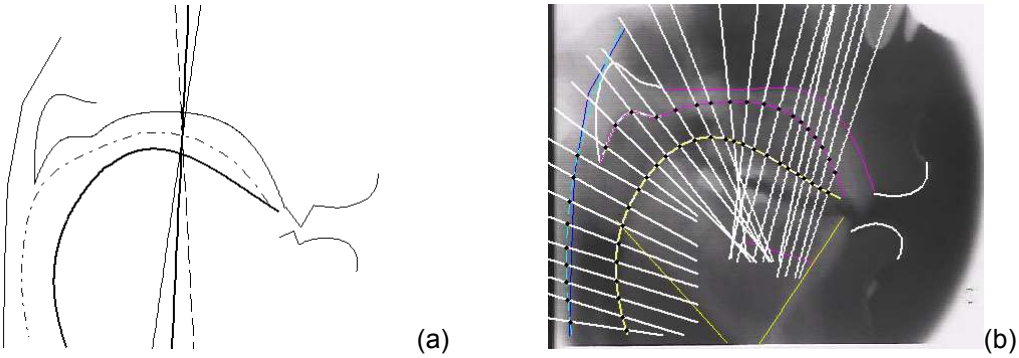


Figure 49 : (a) Correction d'orientation apportée à une des lignes de la grille (en trait fin). La ligne en pointillés permet de visualiser l'angle moyen de la ligne et de la langue. La ligne en gras est celle obtenue après rotation d'angle α de cette ligne,

$$\text{où } \alpha = \frac{1}{2}\left[angle_{ligne/\,palais} + (angle_{ligne/\,langue})_{moyen}\right].$$

(b) Grille définissant 27 sections du pharynx aux incisives, après correction globale des orientations de chacune des lignes de façon à ce que chacune soit le plus possible orthogonale à la fois à la langue et à la ligne palais-pharynx.

L'angle considéré pour la rotation d'une ligne est la moyenne de l'angle entre la ligne et le palais et de l'angle moyen entre la ligne et la langue.

1.2. Mesure directe à partir de la grille statique

La mesure des distances sagittales est effectuée image par image pour la séquence complète Laval43, à partir de la grille établie. Les distances sagittales sont les diamètres apparents le long du conduit, mesurés au niveau des lignes de la grille.

(1) Pour les lèvres, la distance sagittale correspond à l'écart entre le point le plus bas de la lèvre supérieure et le point le plus haut de la lèvre inférieure.

(2) Pour les incisives avant, la distance correspond à l'écart entre les 2 incisives.

(3) Toutes les autres mesures utilisent directement les lignes de la grille. A chaque section, il s'agit de calculer, le long de la ligne, la distance entre la ligne palais-pharynx et le contour spline de la langue. Pour cela, on cherche d'une part le point d'intersection de la ligne correspondante avec la ligne palais-pharynx et d'autre part le point d'intersection de la ligne avec la spline figurant la langue. La distance sagittale associée à la ligne est alors la distance entre ces 2 points.

   a. Remarquons ici que la ligne palais-pharynx que nous avons considérée fixe pour l'élaboration de la grille ne l'est plus pour le calcul des distances sagittales. En effet, nous prenons en considération dans cette ligne les mouvements du vélum. Ainsi pour la zone vélaire et les sections correspondantes, il faut mesurer la distance entre la spline représentant la partie inférieure du vélum et celle représentant le contour de la langue.

   b. Notons aussi que dans certains cas, pour la cavité avant, la distance sagittale à mesurer correspond à la distance entre le palais et le plancher sous-lingual et non à celle entre le palais et la langue.

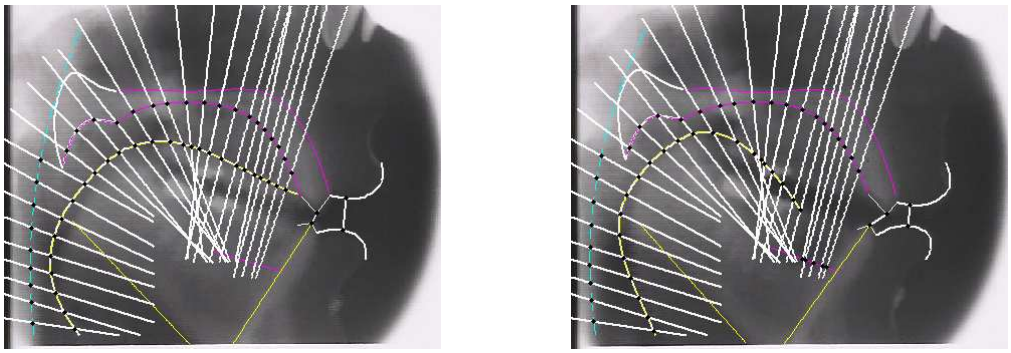


Figure 50 : Grille corrigée en moyenne. Les mesures relatives aux dents avant et aux lèvres se font indépendamment de cette grille. Les points noirs correspondent aux intersections des lignes avec le contour de la langue et avec la ligne palais-pharynx. Deux configurations sont représentées : à gauche, la langue est en avant, à droite, la langue est en arrière et la distance au plancher sous-lingual est prise en compte pour la cavité avant.

1.3. Correction image par image

A ce stade, nous disposons, pour chaque image de la séquence, des sections et des distances sagittales associées. Ces distances ont été obtenues à partir d'une grille de marquage élaborée de façon globale pour assurer l'orthogonalité en moyenne des lignes par rapport au palais et à la langue.

Nous appliquons ensuite la correction classique image par image ; le but est de rétablir l'orthogonalité par rapport à la ligne médiane, pour chacune des sections et pour chaque image. Dans les approches classiques qui calculent des fonctions d'aire, il est en effet habituel de vérifier le calcul de sections et de distances sagittales image par image.

Nous utilisons les procédures de corrections image par image proposées par Yehia [Yeh02] ou Beautemps [BBL95] et nous appliquons une correction à chaque image de la séquence.

1.3.1. Procédure de Yehia

Dans sa thèse [Yeh02], Yehia utilise une grille semi-polaire avec le palais dur comme référence. Les lignes de cette grille sont espacées de 0,5 cm dans les régions linéaires et de 11° dans les régions polaires et permettent pour chaque image la représentation des contours avec le même nombre de points ; à savoir, 29 paires de points, chaque paire constituée d'un point sur la partie antérieure et d'un point sur la partie postérieure du conduit vocal.

Pour chaque image, les distances sagittales mesurées ne sont pas directement les distances entre les 2 points de chaque paire. Une procédure géométrique appropriée est mise en place pour représenter chaque section du conduit vocal par une distance médio-sagittale et une longueur de section. Elle repose sur l'idée que la distance médio-sagittale est la distance entre les points où un front d'onde acoustique longitudinal idéal se propageant dans le conduit vocal touche le palais et la langue.

Les 29 paires de points définissent 28 sections qui peuvent être vues comme un morceau d'une corne conique infinie (ou d'un cylindre si les parois sont parallèles). La direction de propagation est déterminée par la bissectrice de l'angle formé par les bouts de parois antérieure et postérieure (palais et langue) de cette section. L'intersection de cette bissectrice avec les lignes de la grille de marquage détermine 1 segment dont la longueur sera nommée « longueur de section». Ensuite, la ligne orthogonale à ce segment en son milieu coupe les contours du palais et de la langue. Ces 2 points d'intersection définissent un nouveau segment dont la longueur sera la distance médio-sagittale de la section. Le schéma suivant illustre cette procédure.

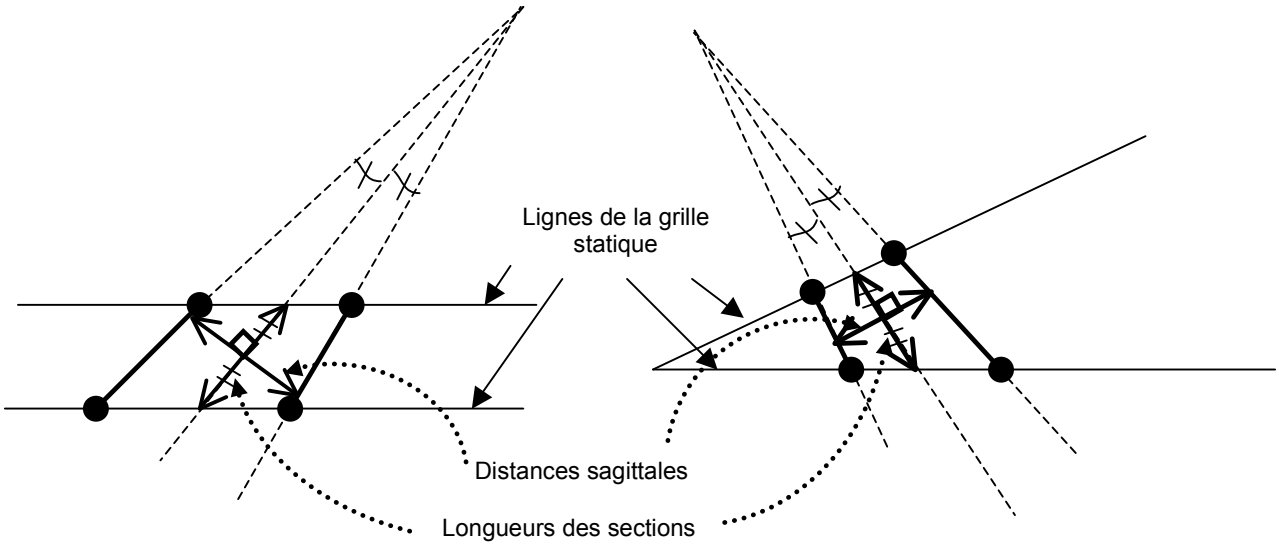


Figure 51 : Procédure utilisée par Yehia [Yeh02] pour déterminer la distance sagittale et la longueur de la section pour une section donnée. Les sections de langue et de palais sont représentées en gras.

Cette procédure est appliquée à partir de notre grille image par image. A partir des 27 lignes de notre grille, nous définissons 26 sections dont nous calculons pour chacune la longueur et distance médio-sagittale. Les mesures de distance pour les lèvres et les incisives sont inchangées.

1.3.2. Procédure de Beautemps

D'autres procédures visant à mesurer les distances médio-sagittales sont proposées, comme celle utilisée par Beautemps et al. [BBL95]. Cette dernière considère que la ligne médiane devrait être idéalement la ligne pour laquelle quelque soit le point le front d'onde est perpendiculaire à la tangente à cette ligne en ce point. La procédure proposée ici consiste à déterminer cette ligne et les distances sagittales pour chaque image, en s'appuyant sur le centre de gravité. Pour chaque section, on évalue en pixels l'aire et la position du centre de gravité de la section considérée. Cette procédure appliquée par Beautemps à partir d'une grille semi-polaire est adaptée à notre grille. Notons ici que comme précédemment avec la procédure de Yehia, nous disposons de 26 sections (les lèvres et les incisives ne sont pas soumises à ce traitement).

La ligne médiane est alors la ligne reliant les centres de gravité. La longueur de chaque section est la somme des longueurs des 2 segments de la ligne médiane contenus dans la section. La distance sagittale est obtenue par le rapport de l'aire de la section sur la longueur de cette dernière. On illustre cette procédure par la figure suivante.

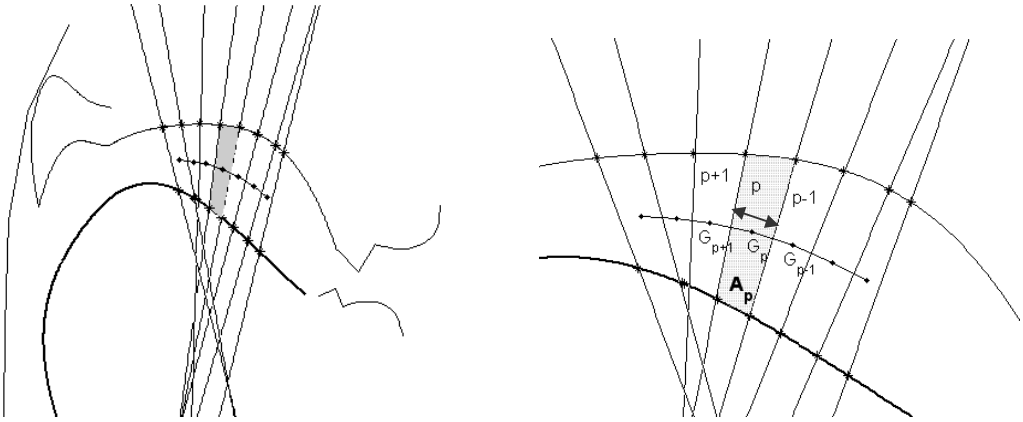


Figure 52 : Détermination des longueurs de sections et de distances sagittales par algorithme de centre de gravité. Si on considère la section p, G_p est son centre de gravité et la zone grisée son aire A_p. La longueur de la section L_p est représentée par la flèche et la distance médio-sagittale est égale à A_p/L_p.

1.4. Représentation des sections

Pour comparer les mesures de distances sagittales, celles obtenues à partir de la grille et celles après correction image par image, plusieurs représentations sont possibles :

- Une représentation en escalier des distances sagittales en fonction du numéro de la section, du pharynx jusqu'aux lèvres.

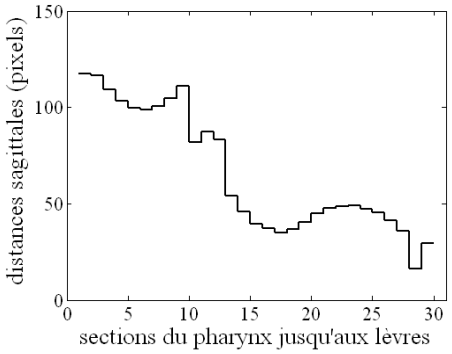


Figure 53 : Exemple de représentation en escalier des distances sagittales en fonction de la section.

- Une représentation lissée, plus lisible. Cette représentation interpolée permet, en outre, de corriger la différence de nombre de sections entre l'estimation à partir de la grille initiale et celle obtenue après correction image par image (réduction de 1 section).

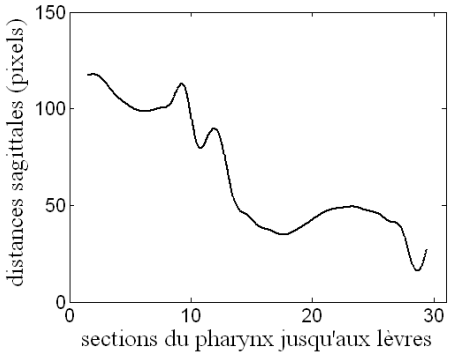


Figure 54 : Exemple de représentation lissée des distances sagittales en fonction de la section.

- Une représentation symétrisée qui permet de visualiser un tube et qui est mieux adaptée pour observer les constrictions.

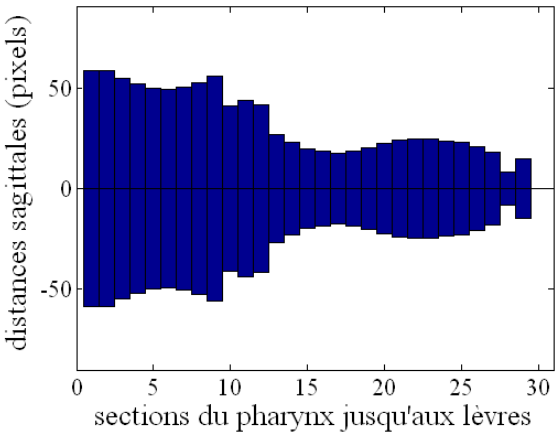


Figure 55 : Exemple de représentation sous forme de tubes des distances sagittales en fonction de la section.

La longueur des sections peut être prise en compte dans la représentation, en remplaçant en abscisse le numéro de la section par la distance par rapport à un point bas du pharynx par exemple. Ceci permet de mieux mettre en évidence la position exacte de la constriction dans le conduit vocal. Il est classique en général de représenter les sections en fonction de leur distance à la glotte. Mais il est nécessaire de signaler ici que la glotte n'est pas visible sur les images de la séquence Laval43. C'est pourquoi, pour l'instant dans notre représentation, la distance des sections est rapportée au point du pharynx le plus bas dont nous disposions et non à la glotte.

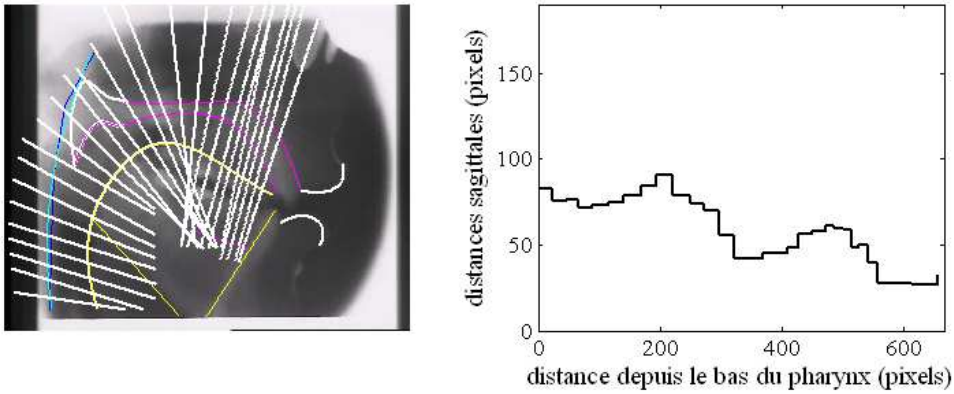


Figure 56 : Pour une configuration, représentation en escalier des distances sagittales en fonction de la distance par rapport au bas du pharynx.

1.5. Comparaison des distances sagittales suivant les procédures de mesure

Nous avons proposé une méthode pour définir une grille de mesure des distances sagittales du conduit vocal. Nous comparons dans cette partie les distances sagittales obtenues à partir de cette grille et celles obtenues après correction image par image, selon deux procédures (Yehia [Yeh02], Beautemps [BBL95]).

Pour toutes ces mesures, les distances estimées aux lèvres et aux incisives sont identiques : elles consistent à mesurer l'écart entre la lèvre supérieure et la lèvre inférieure et l'écart entre les incisives supérieure et inférieure. Nous ne disposons pas de mesures de référence pour savoir quelle procédure de mesure s'approche le plus des mesures exactes. Nous ne pouvons que comparer ces mesures entre elles.

Nous avons effectué les corrections image par image sur les distances sagittales mesurées soit à partir de la grille initiale, soit à partir de la grille corrigée globalement. La figure 57 illustre les résultats obtenus avec la procédure de Yehia en partant de l'une ou l'autre des 2 grilles.

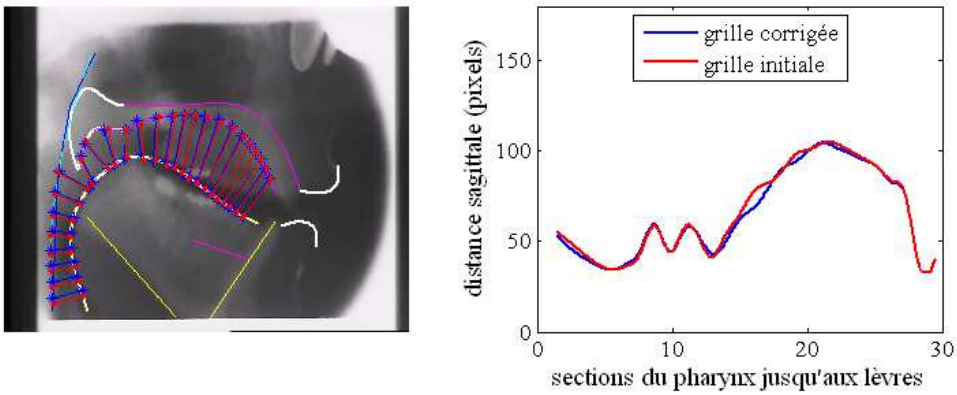
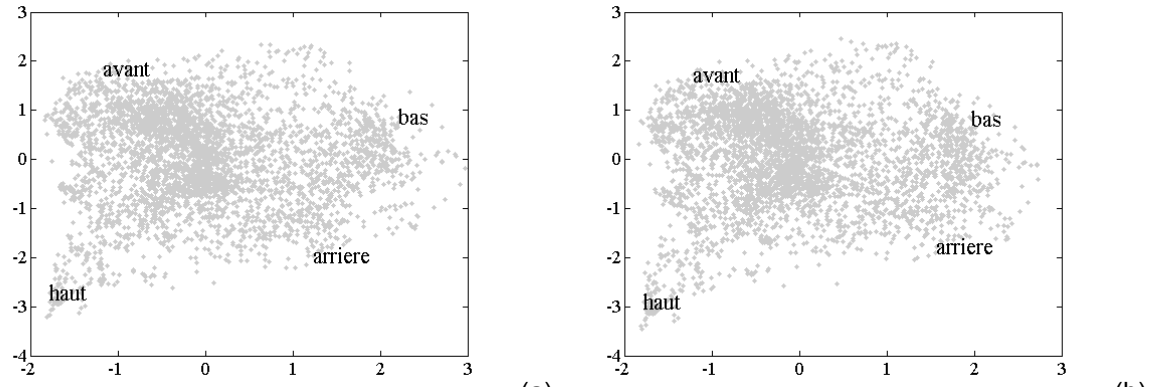


Figure 57 : Pour une configuration, représentation des distances sagittales corrigées par la procédure de Yehia et comparaison suivant la grille de départ utilisée (initiale ou corrigée en moyenne).

L'intérêt de la correction globale que nous proposons permet d'être moins sensible à la correction image par image. En effet, il y a peu de différences entre les distances mesurées à partir de l'une ou l'autre des 2 grilles (Fig. 57 à droite). Si on calcule, en moyenne sur les sections, l'écart moyen RMS sur toute la séquence entre ces 2 distances, on trouve moins de 2 pixels.
Le même calcul est effectué sur les distances obtenues à partir de chacune des 2 grilles et après correction image par image type centre de gravité (Beautemps) : l'écart moyen RMS est lui aussi de l'ordre de 2 pixels (2,3 pour être précis).
Nous avons fait le choix de travailler avec la procédure type Yehia, mais nous aurions aussi bien pu utiliser celle de Beautemps. Les procédures sont proches et les résultats sont très comparables en terme de géométrie, c'est-à-dire de mesures de distances sagittales le long du conduit vocal.

Etant donné le nombre de dimensions (28 sections pour 4043 images), une analyse en composantes principales permet de représenter de manière synthétique ces données. Des études plus détaillées à base d'ACP seront réalisées au chapitre 5. L'ACP nous permet simplement de montrer ici que les deux procédures présentées, celle de Yehia et celle de Beautemps, procurent des distributions de distances sagittales comparables. En effet, l'analyse en composantes principales sur chacun de ces 2 jeux de données pour la séquence complète permet de représenter les données dans deux plans principaux d'allures identiques, comme on peut le voir sur les figures 58a et 58b. Les 2 premières composantes de l'ACP sur les sections corrigées par la procédure de Beautemps expliquent respectivement 68% et 12% de la variance (soit 80% de la variance totale). Celles de l'ACP sur les sections corrigées par la procédure de Yehia expliquent aussi 80% de la variance totale (67%+13%). Nous avons de plus annoté chacun de ces graphiques par avant-arrière et haut-bas de façon à montrer l'organisation générale des données selon 2 axes correspondant aux mouvements de la langue, nous en reparlerons au chapitre 5.



(a) (b)

Figure 58 : ACP sur les sections pour la séquence complète après correction par les procédures de (a) Beautemps et (b) Yehia.

2. Des sections à la fonction d'aire

Les mesures de sections médio-sagittales sont réalisées dans le but de mettre en correspondance nos données géométriques avec les données acoustiques et en particulier les formants. Pour la validation acoustique et l'estimation des formants, la fonction d'aire est le passage obligé pour passer de la coupe sagittale à une estimation 3D.

La génération de ces fonctions à partir des mesures des sections sagittales est donc une étape importante dans l'étude de la relation entre la géométrie du conduit vocal et l'acoustique.

En effet, un modèle physique capable de produire un signal de parole à partir de commandes dites articulatoires est composé de plusieurs parties :

1. un modèle articulatoire capable de produire des coupes sagittales à partir de commandes articulatoires,
2. un modèle de passage de la coupe sagittale à la représentation en un ensemble de tuyaux équivalents au conduit vocal (fonction d'aire),
3. un modèle acoustique permettant de passer de la fonction d'aire aux formants associés.

A ce stade de l'étude, nous n'avons pas déterminé de commandes articulatoires, nous n'avons donc pas défini un modèle, au sens classique du terme ou comme l'entend le point 1 cité juste avant. Mais nous disposons des données nécessaires pour produire des coupes sagittales et n'avons donc pas besoin d'un modèle articulatoire. Les points géométriques des contours du conduit vocal permettent, comme on vient de le voir, de produire des coupes sagittales du conduit vocal pour toutes les configurations de la séquence. La question du modèle acoustique sera traité au chapitre 7.

Nous nous intéressons ici au passage des sections à la fonction d'aire (i.e. au point 2 cité ci-dessus). La fonction d'aire représente l'aire transversale à chaque section de la coupe sagittale, elle est fonction de la distance à la glotte. C'est une représentation bidimensionnelle du conduit vocal. Ce passage de la coupe sagittale du conduit vocal à sa fonction d'aire n'est pas facile à cause de la forme irrégulière du conduit vocal.

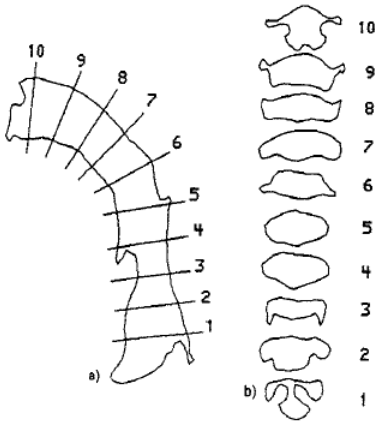


Figure 59 : Coupes du conduit vocal réalisées dans un moulage de cadavre, à droite coupe sagittale, à gauche sections transversales (d'après Calliope, 1989).

A titre d'exemple, sur la figure 59, on observe des coupes du conduit vocal réalisées dans un moulage de cadavre. Ce moulage a été découpé transversalement en dix sections. Il est clair que les formes des dix sections sont irrégulières. Le modèle de passage influence directement la fidélité de la production des sons du synthétiseur par rapport à la production de parole d'un conduit vocal humain. L'information contenue dans la coupe sagittale est insuffisante pour obtenir une bonne estimation de la fonction d'aire.

Aussi, le modèle de passage de la coupe sagittale à la fonction d'aire a été l'objet de nombreuses études, dont la plupart ont été inspirées par les résultats de Heinz et Stevens en 1965 [HS65]. Lors de ce travail réalisé à partir de mesures faites sur des cadavres, la fonction d'aire est calculée d'après l'équation $A(x) = \alpha(x)d(x)^{\beta(x)}$ où x est la distance depuis la glotte, $A(x)$ est l'aire transverse de la section, $d(x)$ est la distance médio-sagittale (distance entre la langue et le palais) et $\alpha(x)$ et $\beta(x)$ sont des coefficients déterminés de manière ad hoc et qui varient selon la région du conduit vocal.
Ce modèle, basé sur une relation exponentielle (ou de puissance), est couramment utilisé sous le terme de modèle « alpha-beta ».

2.1. Difficultés rencontrées

Disposer des sections et d'un modèle de passage de ces sections à la fonction d'aire n'est qu'un début, nous allons être confrontés à un certain nombre de difficultés. Le modèle $\alpha\beta$ est un modèle spécifique à chaque locuteur, comme le montre les travaux menés par de nombreux auteurs ([Mae90], [PBS92], [BBL95], [SLMD02]) qui ont cherché à rendre ce modèle plus élaboré et à définir les paramètres α et β.

Pour utiliser ce modèle, nous serons contraints d'utiliser les paramètres élaborés pour d'autres locuteurs, avec toutes les imperfections que cela implique. En effet, nous ne sommes pas en mesure d'élaborer nos propres coefficients α et β, car pour notre locuteur, nous ne disposons que des images radiographiques, de profil, et aucune indication anatomique supplémentaire.

Perrier, Boë et Sock [PBS92] utilisent l'équation $A(x) = \alpha(x)d(x)^{\beta(x)}$ avec β=1.5 et α dépendant de la région du conduit vocal d'une part et de la valeur de la variable d (distance sagittale) d'autre part. Le conduit vocal est divisé en 7 régions, chacune de ces zones étant associées à 2 valeurs de α, une valeur pour les distances sagittales de grande dimension et une autre pour celles de petite dimension.

Beautemps, Badin et Laboissière [BBL95] utilisent le modèle pour étudier un locuteur produisant des consonnes fricatives et des voyelles. Dans cette étude, les valeurs de α varient de façon continue le long du conduit vocal, ceci permet « de ne pas avoir à déterminer manuellement (et parfois arbitrairement) les limites entre les différentes zones et de limiter les discontinuités entre les régions».

Plus récemment, Soquet et al. [SLMD02] ont mené une étude comparée de différentes transformations utilisées pour calculer l'aire du conduit vocal à partir de la distance sagittale et proposé de nouvelles valeurs de paramètres α et β, obtenues sur 2 locuteurs prononçant des voyelles orales du français. Ce sont ces paramètres que nous utilisons pour obtenir une estimation de la fonction d'aire. Nous détaillerons la méthode au paragraphe suivant.

Un autre problème, déjà cité, et à nouveau en cause ici, est l'information manquante de la glotte. Dans le modèle « alpha-beta », la distance x considérée est définie par rapport à la glotte. Dans la mesure où la glotte n'est pas visible sur les images de la séquence Laval43, nous pouvons uniquement l'interpoler, ce qui va encore un peu ajouter aux imprécisions.

Enfin et ce n'est pas la moindre des difficultés, les fonctions d'aire sont à évaluer en cm² pour permettre ensuite un passage vers l'acoustique. Or le rapport entre les pixels et les cm nous manque, nous pouvons bien sûr l'estimer indirectement, mais sans certitude.

Ces différents réglages et estimations compliquent le passage des sections à la fonction d'aire et encore plus, à terme, aux formants. Néanmoins, nous choisissons au mieux ces divers paramètres (position de la glotte, rapport pixels/cm, coefficients α et β) et obtenons ainsi les fonctions d'aire associées à chaque image de la séquence Laval43.

2.2. Choix des paramètres

Nous faisons ici les approximations nécessaires de façon à estimer pour chaque image de la séquence une fonction d'aire associée à la configuration géométrique extraite par notre méthode semi-automatique. Nous disposons, pour chaque trame, de 28 sections depuis les lèvres jusqu'au pharynx ; ce sont les sections corrigées avec la procédure type Yehia à partir de la grille statique corrigée globalement.

2.2.1. Modèle *αβ*

Nous utilisons un modèle type « alpha-beta », basé sur une transformation exponentielle, avec les coefficients α et β, publiés dans l'article de 2002 de Soquet et al. de l'Université Libre de Bruxelles [SLMD02]. Ces paramètres ont été calculés à partir de coupes obtenues par Résonance Magnétique pour les voyelles orales du français prononcées par 2 locuteurs (un homme et une femme). La valeur de ces paramètres dépend de la région du conduit vocal. Dans leur modèle, les auteurs ont partagé le conduit vocal en 8 régions, visibles sur la figure 60. Les frontières entre les zones sont placées en accord avec les articulateurs qu'elles représentent. La limite entre les régions 3 et 4 est définie par les auteurs à la moitié de la distance entre le sommet de l'épiglotte et le vélum : ces régions sont respectivement nommées pharynx moyen et oropharynx. La valeur des paramètres α et β dépend du locuteur, le tableau 7 présente ces valeurs pour chacun des 2 locuteurs.

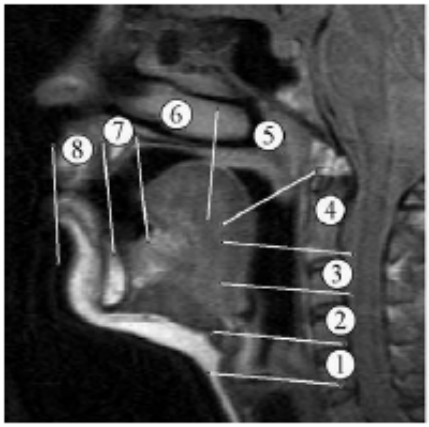


Figure 60 : Représentation des différentes régions du conduit vocal sur une vue IRM de profil médio-sagittal pour une locutrice prononçant un [u], d'après [SLMD02].

Numéro de la zone	Zone	Femme		Homme	
		α	β	α	β
1	Larynx	0,78	0,70	1,11	2,35
2	Pharynx bas	1,86	1,22	1,79	1,38
3	Pharynx moyen	1,74	1,23	1,34	1,62
4	Oropharynx	1,99	0,81	0,73	1,81
5	Vélum	1,84	0,93	1,39	1,08
6	Palais dur	1,82	1,43	1,34	1,51
7	Zone alvéolaire	2,67	1,48	1,92	1,20
8	Zone labiale	2,42	1,67	4,72	2,48

Table 7 : Paramètres α et β obtenus par Soquet et al. [SLMD02] pour les 8 régions définies du conduit vocal et les 2 locuteurs.

Pour utiliser ces paramètres avec nos 28 distances sagittales, il faut avant tout répartir nos sections entre les 8 régions définies par ce modèle. La région 8 est la zone labiale, elle concernera nos 2 sections avant, celle entre les lèvres et celle entre les incisives. La région 1 (larynx) n'est pas visible sur nos images cinéradiographiques, aucune de nos sections ne sera concernée par cette région. Le calque qui suit (inversé par rapport aux images Laval43) représente les 26 sections du conduit vocal (hormis les lèvres et les dents), délimitées par 27 lignes. Les lignes en pointillés correspondent aux frontières entre les régions 2 à 7 du modèle.

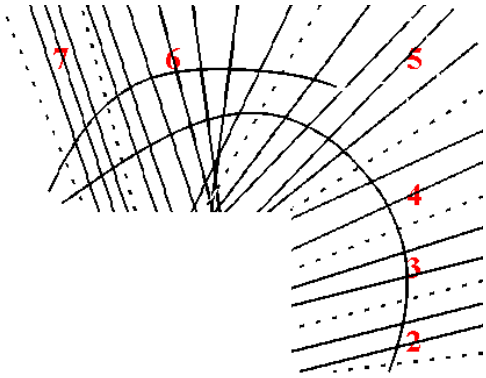


Figure 61 : Calque inversé d'une image radiographique de la séquence Laval43 avec les 26 sections du conduit vocal et les frontières (en pointillés) délimitant les régions définies par le modèle [SLMD02].

Le locuteur de la séquence Laval43 est un homme, nous choisissons de travailler, pour commencer, avec les paramètres α et β du locuteur homme du modèle de Soquet et al.. Nous validerons ce choix plus loin, au chapitre 7.

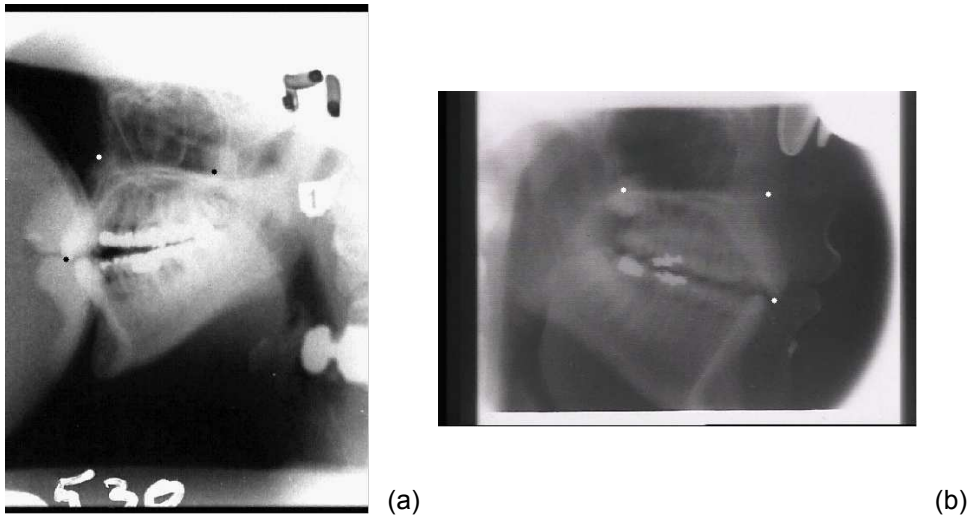
2.2.2. Rapport pixels/cms

La relation exponentielle de transformation des distances sagittales aux fonctions d'aire s'applique avec des distances en cm, or, pour l'instant nos distances sont en pixels (des images 720*480 pixels). Le passage des pixels aux centimètres est une donnée manquante. Il n'y a pas de repère de calibration dans la séquence.

Notre première estimation est de considérer que pour 1 cm, nous avons 38 pixels. Cette grandeur correspond à celle que nous avions estimée pour la base Wioland, à partir des croquis de François Wioland dans son doctorat d'état. Pour cette analogie, nous avons utilisé une référence de Procuste pour montrer que les rapports entre pixels et centimètres des séquences Laval43 et Wioland sont comparables.

Procuste[6] est le nom d'une méthode de superposition utilisée en orthodontie et qui prend en compte la variabilité individuelle. Cette méthode[7] commence par calculer une référence personnalisée en fonction des caractéristiques crânio-faciales (typologie et croissance) du patient à traiter. Dans un second temps, le tracé du patient est superposé avec celui de sa référence personnalisée. Des anomalies peuvent alors être mises en évidence par les écarts entre les tracés.

Dans notre cas, nous nous limitons à 3 points (Fig. 62) : la pointe de l'incisive maxillaire, l'épine nasale antérieure et l'épine nasale postérieure.



(a) (b)

Figure 62 : Points de référence Procuste pour comparer, en terme de rapport pixels-cms, les séquences (a) Wioland (image tournée de 90°) et (b) Laval43.

En marquant ces 3 points sur 2 images de référence de chacune des séquences (bouche fermée en période de silence), on établit la matrice de transformation d'une image à l'autre et

[6] Procuste était un bandit grec qui allongeait ses victimes sur un lit métallique pour les raccourcir si elles dépassaient et les étirer si elles étaient trop courtes. Procuste est devenu le symbole du conformisme et de l'uniformisation.

[7] www.procuste.com

constater que les rapports de distance sont proches de 1, de même que le rapport des aires des 2 triangles. Compte-tenu de ce résultat et en négligeant le fait qu'il s'agit de 2 locuteurs différents, nous admettons qu'une estimation de 38 pixels pour 1 cm est valable pour la séquence Laval43. Cette estimation pourra éventuellement être modifiée par la suite.

2.2.3. Position de la glotte

Si le rapport pixels/cms influe sur le calcul de l'aire, il intervient aussi pour le calcul de la longueur du conduit vocal. Or, il y a un second problème pour cette estimation qui est la non-visibilité de la glotte sur les images de la séquence Laval43. Nous ne disposons pas de la position précise de la glotte. A nouveau, nous faisons une estimation et nous plaçons la glotte sur la ligne médio-sagittale prolongée après le bas du pharynx, comme on le visualise avec le point noir sur la figure 63. La hauteur moyenne de la glotte pourra être ajustée par la suite, au besoin. Mais un problème persiste, nous ne pouvons produire qu'une estimation de la position moyenne alors qu'il faudrait avoir une estimation en fonction du temps.
Nous extrapolons la valeur de la fonction d'aire entre la glotte et la section mesurée la plus proche (la plus basse dans le pharynx) avec la valeur déterminée pour cette section, comme on l'observe sur la figure 64.

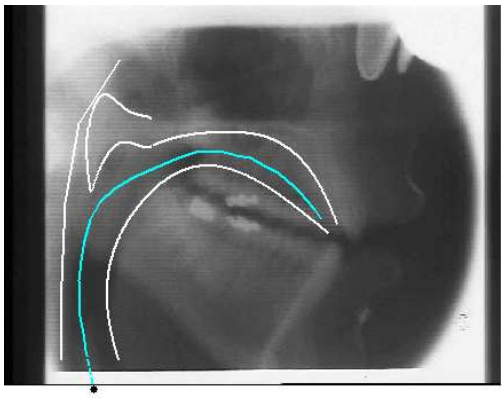


Figure 63 : Ligne médiane moyenne pour la séquence Laval43, prolongée jusqu'à l'estimation de la position de la glotte (point noir).

La position de la glotte et le rapport pixel/cm permettent d'évaluer la longueur du conduit vocal de la glotte jusqu'aux lèvres. Compte-tenu des estimations faites, nous obtenons une longueur moyenne de conduit de 17,5 cm, ce qui est à peu près cohérent avec la longueur moyenne de tractus vocal chez les hommes généralement autour de 17 cm. Nous vérifierons cependant ce résultat en faisant varier le rapport pixels/cms entre 36 et 42, plus loin au chapitre de synthèse articulatoire.
Plus le conduit vocal est long, plus les formants sont bas : on observe cette différence en particulier entre la voix d'un homme et celle d'une femme. Quand je fais un « é » puis « eu » mes lèvres s'allongent, je rallonge la longueur de mon conduit vocal, donc tous les formants

baissent. La longueur du conduit vocal est donc importante pour l'estimation des formants, que nous considérerons plus loin, à partir de la fonction d'aire ici calculée.

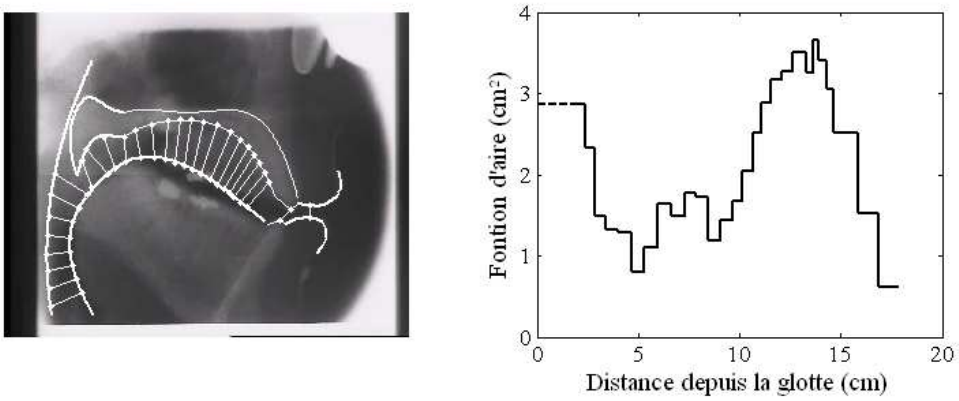


Figure 64 : Sections médio-sagittales et fonction d'aire calculée pour une image de Laval43.

A partir des contours géométriques extraits par la méthode semi-automatique, ce chapitre a permis de calculer pour chaque image de la séquence les distances médio-sagittales ainsi que la fonction d'aire associée. Ceci a été réalisé à partir d'une grille élaborée en plusieurs étapes : d'abord initialisée à partir des contours fixes du palais et du pharynx, elle a ensuite été corrigée globalement en moyenne.

Au terme de cette première partie, nous disposons donc, pour toutes les images de la séquence, à la fois des configurations géométriques du conduit vocal et des fonctions d'aire.

Deuxième partie : Exploitation des données articulatoires et mise en correspondance avec les données acoustiques

RELATIONS ARTICULATOIRE-ACOUSTIQUE

La parole permet de transmettre un message ou code. Décoder le signal de parole ne peut se faire sans aborder les tenants et les aboutissants, la production et la perception de la parole. Comprendre les liens entre espace articulatoire, acoustique et perceptif est donc un enjeu important des études en parole.

La plupart des études menées sur la modélisation de la production de la parole s'appuient sur la théorie source/filtre [Fan60]. Cette théorie décrit la production de la parole sous la forme de deux composantes principales :

- un signal source, encore appelé source d'excitation ;
- un filtre modélisant la perturbation de cette source par les différents articulateurs et cavités du conduit vocal.

Ainsi, la modélisation articulatoire s'applique à simuler la production de parole en utilisant un modèle régi par des paramètres physiologiques ou articulatoires.
A partir de là, la plupart des approches en production de la parole se fondent sur la théorie de l'acoustique linéaire, sur une propagation du son à une dimension et sur des approximations de non-linéarité.
Les modèles de production de la parole peuvent être utilisés pour tester des hypothèses phonologiques selon l'existence d'une explication du phénomène étudié au niveau articulatoire. Si une telle explication existe, des stimuli sont créés par synthèse articulatoire, et des tests de perception sont utilisés pour isoler et contrôler le phénomène. Dans le cas contraire, il peut être intéressant de collecter des données relatives à la production du phénomène étudié, soit par mesure directe (IRM, aérodynamique), soit indirectement par utilisation d'une méthode d'inversion acoustico-articulatoire.

La description précise de la morphologie du conduit vocal permet d'approfondir l'étude des relations articulatori-acoustiques. La méthode mise en place a conduit à l'extraction de données géométriques dans un contexte spécifique de parole continue synchrone et de corpus longs de quelques dizaines de phrases. Cette grande quantité de données récupérées de ces séquences radiographiques, en regard avec les données audio associées, permettrait d'aborder les questions classiques : Peut-on récupérer les paramètres acoustiques à partir des mouvements du conduit vocal ? Est-il possible de déterminer la configuration des articulateurs à partir de l'acoustique de la parole ?

Ces questions ont été au cœur d'un grand nombre d'études parues dans la littérature et qu'on présente sous les termes de synthèse articulatoire et d'inversion acoustico-articulatoire.
En outre, un certain nombre d'auteurs ([YRV98], [BB99b]) se sont intéressés aux relations linéaires qui pouvaient exister entre les configurations du conduit vocal, les attitudes faciales et les paramètres acoustiques, en examinant de possibles associations linéaires entre ces différents jeux de données.

1. Inversion et synthèse articulatoire

L'étude de la parole se place résolument au cœur de la relation entre l'acoustique et l'articulatoire, autrement dit dans le « speech mapping » entre production et perception [ABS+94].

1.1. Inversion

Si les études de simulation permettent de connaître le lien direct entre une configuration articulatoire et le signal de parole, l'inversion de la parole s'intéresse au cheminement inverse : retrouver à partir du signal de parole des informations sur la forme du conduit vocal. Les problèmes du passage de l'articulatoire à l'acoustique ne sont pas tous résolus et cette « inversion du conduit vocal » suscite toujours des questions et se révèle riche en interprétations. L'inversion en parole consiste donc à récupérer des gestes articulatoires à partir de leurs conséquences acoustiques, ou en d'autres termes à retrouver les paramètres articulatoires et les coupes sagittales correspondantes à partir du signal acoustique produit par le sujet. Ce travail se heurte à plusieurs difficultés. La première est liée à l'extraction des indices pertinents dans la parole, et notamment au problème classique d'estimation des fréquences de résonance du conduit vocal, ou formants : cette question est à l'origine de très nombreux travaux de recherche, depuis le début des années 1970 (avec ceux de Flanagan [Fla72] par exemple) et encore aujourd'hui. Mais la difficulté majeure est que plusieurs configurations des articulateurs peuvent produire un signal acoustique équivalent. Il existe de nombreuses configurations articulatoires associées à un même point de l'espace acoustique. La relation entre l'articulatoire et l'acoustique n'est pas biunivoque (elle est dite « many-to-one »).
Il est nécessaire de régulariser le problème de l'inversion par l'adjonction de contraintes supplémentaires. Dès les premiers travaux sur l'inversion en parole, les tentatives de régularisation apparaissent de manière plus ou moins formelle. Schroeder [Sch67] propose par exemple de contraindre son système d'inversion en limitant la fonction d'aire par un développement en série de Fourier dont le nombre de composantes est égal au nombre de

formants fournis en basse fréquence. Cette étude, comme beaucoup d'autres, utilise, pour analyser l'inversion, le lien entre le spectre et la fonction d'aire. La fonction d'aire consiste à représenter le conduit vocal comme une succession de plusieurs cavités de volume variable dont la disposition dépend de l'articulation. Ces cavités produisent des résonances qui vont amplifier ou atténuer le spectre sonore produit par la source acoustique.

L'inversion a été étudiée depuis plusieurs décennies. Outre l'intérêt en reconnaissance, l'extraction de paramètres articulatoires qui évoluent lentement dans le temps permet de coder très efficacement la parole [SS87]. Différentes méthodes sont utilisées pour conduire l'inversion acoustique-articulatoire.

- L'inversion par tabulation consiste à créer une table (ou dictionnaire) reliant les paramètres acoustiques (généralement, fréquences des formants) aux formes du conduit vocal (paramètres articulatoires). Ensuite, pour chaque forme acoustique, on recherche la configuration correspondante des articulateurs. Or comme l'explique Atal [ACM+78], la correspondance n'est pas biunivoque. Il faut donc imposer des contraintes sur la cinématique et/ou la dynamique des articulateurs afin de choisir les paramètres articulatoires qui donnent la meilleure trajectoire.
- L'inversion par optimisation consiste en des méthodes itératives qui optimisent des paramètres d'un modèle articulatoire jusqu'à ce que les caractéristiques acoustiques de la parole prononcée et celles de la parole produite par le modèle soient proches. Les paramètres initiaux peuvent être choisis au hasard, mais il est aussi possible de partir d'une solution approchée obtenue, par exemple, par inversion par tabulation. Ces méthodes font appel à un certain nombre de contraintes permettant de réduire le nombre de formes de conduit vocal possibles.

Comme notre contribution ne porte pas sur l'inversion, nous ne détaillerons que très peu ici les études qui ont porté ou portent sur l'inversion acoustico-articulatoire. On peut consulter une revue de 2003 de l'inversion en parole [TM03]. On y retrouve notamment les travaux de Laprie et al. ([ML97], [OL00], [OL05]) avec l'élaboration conjointe d'une table articulatoire acoustique hypercubique et de la méthode d'inversion associée. Ils ont également travaillé sur l'amélioration de la proximité acoustique entre les données de départ et les formants donnés par les paramètres articulatoires récupérés.

Le projet européen « Speech MAPS » [ABS+94] a eu pour but de répondre, théoriquement et technologiquement, à la question : « est-ce qu'un robot articulatoire peut produire des gestes articulatoires à partir des sons ? ». Il concerne l'inversion acoustique de la parole à travers la connaissance du phénomène de production de parole. L'objectif a consisté à élaborer un agent virtuel, vu comme un système sensori-moteur capable d'articuler et de percevoir des gestes de parole. Le problème est posé à la fois dans les termes de la Robotique et du

Contrôle Moteur. Etant donné un système articulatoire de production de parole réaliste (*plant*), un système de contrôle (dit contrôleur) est capable d'agir sur ce *plant* de façon à produire les résultats désirés en sortie, tout en se servant des retours sensoriels par les voies de rétroaction. Ce contrôleur est la pièce maîtresse de l' « Articulotron », un robot articulatoire capable de synthétiser des séquences motrices à partir de prototypes sonores. Cette approche nécessite d'imposer des contraintes empruntées à l'articulatoire et de prendre en compte la perception auditive et visuelle. Aussi, la stratégie repose sur le couplage de l'Articulotron avec un Perceptron (audiovisuel) pour incorporer la vision, dont l'importance en parole a été facilement montrée (sourds, milieu bruyant…). L'approche robotique permet ainsi de lier Action et Perception et de voir la communication parlée comme un compromis entre le coût de production et la compréhension.

Parmi les travaux sur l'inversion, ceux reposant sur une approche d'analyse par synthèse articulatoire sont les plus représentés dans la littérature car ils se prêtent bien à l'utilisation d'un modèle articulatoire.

1.2. Synthèse articulatoire

La synthèse articulatoire est basée sur des connaissances physiologiques, articulatoires et mécaniques du conduit vocal humain.

A partir de la forme du conduit vocal donnée par un modèle articulatoire (spécifiant par ex. la position et la forme de la langue, la position de la mâchoire, du larynx et des lèvres), la synthèse articulatoire produit le signal acoustique grâce à une simulation numérique utilisant l'analogie acoustique-électrique.

La production de la parole et sa modélisation numérique grâce à un synthétiseur articulatoire permettent ainsi de passer des paramètres articulatoires au signal acoustique. Ce passage n'est pas simple. En effet, il faut développer trois modèles : un modèle articulatoire, un modèle de passage de la coupe sagittale à la fonction d'aire et un modèle acoustique.

Le modèle articulatoire permet d'interpréter en gestes articulatoires les déformations du conduit vocal. Il est donc utile de disposer des paramètres de contrôle du modèle qui correspondent à une réalité physique, c'est-à-dire à des articulateurs existant réellement dans le système humain de production de la parole.

Le modèle de passage de la coupe sagittale à la fonction d'aire permet d'obtenir une représentation du conduit vocal par une série de tubes qui sont pris en compte sous la forme de quadripôles électriques, grâce à l'analogie acoustique électrique, pour calculer le spectre de parole, dont nous reparlerons au chapitre 7.

Comme nous l'avons déjà fait remarquer, nous ne construisons pas un modèle articulatoire. Mais nous avons montré au chapitre précédent que nous pouvions mesurer des distances

sagittales à partir des contours géométriques estimés par notre méthode. C'est dans l'optique de réaliser de la synthèse articulatoire, que cette mesure a été réalisée, ainsi que le passage aux fonctions d'aire.

La démarche la plus courante pour la synthèse articulatoire consiste donc à mesurer la fonction d'aire à partir des données, puis à comparer pour chaque son analysé, le signal synthétisé à partir de cette fonction d'aire et le signal de parole effectivement acquis, en termes de spectre et/ou d'erreur relative entre les formants respectifs. La simulation acoustique, qui utilise un modèle acoustique, peut prendre ou non en compte les pertes (par conduction thermique, viscosité et vibration de parois) dans le conduit vocal. Il convient de noter que la description de la complexité géométrique du conduit vocal est, à ce niveau, différente selon les auteurs, et qu'elle fait déjà partie de la modélisation.

Remarque : si la plupart des travaux consacrés au développement de modèles de passage de la coupe sagittale à la fonction d'aire se basent sur le modèle alpha-beta initié par Heinz et Stevens [HS65], la fonction d'aire obtenue n'est alors qu'une approximation. Ceci explique qu'un certain nombre de spectres de parole produits par le locuteur ne peuvent pas être reproduits par le synthétiseur articulatoire. Une meilleure solution à ce problème est l'élaboration d'un modèle articulatoire tridimensionnel à condition qu'il soit suffisamment précis. De tels modèles exploitent des images IRM ([YT97], [Eng99], [BBRS98]) et envisagent aussi une « optimisation » des fonctions d'aire par inversion. L'algorithme choisi pour la mettre en œuvre vise à minimiser l'erreur entre les formants mesurés sur le signal enregistré et ceux qui sont détectés sur le signal synthétisé, tout en préservant l'allure globale initiale des fonctions d'aire.

Les travaux en synthèse articulatoire sont nombreux. Nous faisons une distinction parmi les études ci-dessous, suivant qu'elles s'intéressent aux consonnes ou aux voyelles.

- Consonnes

Les travaux de Badin et al. [MBVB96] concernent la synthèse articulatoire des consonnes fricatives. Les paramètres de contrôle sont déterminés par inversion à partir d'un signal de parole audiovisuel, et la synthèse articulatoire est évaluée à la fois sur le plan objectif et sur un plan perceptif. Les interactions entre forme du conduit vocal et sources acoustiques ont été étudiées par simulation : ainsi une diminution de l'aire de la constriction orale entraîne une augmentation du bruit de friction généré, associée à une diminution de l'amplitude du mouvement des cordes vocales. Une augmentation de l'aire de la constriction orale entraîne une diminution du bruit de friction et une augmentation de l'amplitude de voisement. Ce phénomène explique la difficulté de produire les consonnes fricatives voisées qui nécessitent

un équilibre très subtil entre aire de constriction orale et aire de glotte pour maintenir simultanément friction et voisement.
Narayanan et al. se sont aussi, et entre autres, penchés sur la question des consonnes fricatives, dans leurs études de synthèse articulatoire à partir de données IRM et audio [NA00].

- Voyelles

La thèse de Yehia [Yeh02] s'intéresse à la mise en correspondance de données acoustiques et articulatoires et a pour objectif d'estimer cette correspondance, dans le cadre très limité des voyelles orales, en utilisant des contraintes imposées par la morphologie humaine et par la dynamique du conduit vocal. Dans ces travaux, le conduit vocal est représenté d'un point de vue articulatoire par le logarithme de la fonction d'aire et d'un point de vue acoustique, par les 3 premiers formants. L'information morphologique est extraite de profils sagittaux obtenus à partir de données cinéradiographiques et est utilisée pour limiter l'espace articulatoire à des fonctions d'aire compatibles avec un conduit vocal humain. Il parvient ainsi à réduire l'ambiguïté de la relation articulatori-acoustique. Le passage de la fonction d'aire aux formants est réalisé en considérant le conduit vocal comme une concaténation de tubes uniformes avec pertes : ceci permet de récupérer une fonction de transfert associée à la configuration géométrique du conduit vocal, dont les maxima sont les fréquences des formants.
Une méthode de simulation du conduit vocal dans le domaine temporel est décrite par Maeda, dans [Mae82]. Le modèle adopté comporte une source de pression d'air constante, une section étroite variable dans le temps représentant la glotte et un tube correspondant au conduit vocal couplé à la cavité nasale. Onze voyelles françaises, synthétisées présentent un haut degré de naturel et d'intelligibilité même si une trace de distorsion fréquentielle est relevée dans la région du 3ème formant.
Dans [Eng01], Engwall présente la première évaluation de son modèle articulatoire (élaboré à partir de données IRM, EPG et EMA) en synthétisant neuf voyelles statiques, à partir de sa fonction d'aire échantillonnée sur 23 plans.

On peut également s'intéresser aux synthétiseurs articulatoires disponibles en ligne sur internet, comme celui des laboratoires Haskins. CASY [RSG$^+$96], pour Configurable Articulatory Synthesizer, permet à l'utilisateur de superposer sur une image sagittale (typiquement IRM) un contour de leur modèle 2D de conduit vocal et d'ajuster graphiquement les paramètres du modèle pour s'adapter aux dimensions de l'image. Les fonctions de transfert et paramètres acoustiques sont ensuite générés à partir de ces paramètres. Ce modèle combine les paramètres du modèle ASY et ceux du modèle original

de Mermelstein. De plus, les surfaces fixes du conduit vocal sont également représentées de façon paramétrique et peuvent ainsi être ajustées à n'importe quel locuteur.
Le logiciel TractSyn [BJ03] est, quant à lui, un modèle 3D du conduit vocal. Les parois du conduit vocal et la langue sont représentées par trois grilles individuelles. La forme des grilles est déterminée par un jeu de paramètres spécifiant la forme et la position de la langue, des lèvres, du vélum, du larynx et de la mandibule. Le logiciel permet de visualiser la production de parole et de synthétiser des sons.

Pour constituer un synthétiseur articulatoire complet, les modèles articulatoires doivent être associés à des modèles d'écoulement d'air, de sources et de propagation acoustique. Ceci permet de contrôler le synthétiseur par deux jeux de paramètres : les paramètres supra-laryngés qui commandent le modèle articulatoire, et un jeu de paramètres qui pilotent les cordes vocales. La synthèse articulatoire à partir des données géométriques extraites de la cinéradiographie sera l'objet du chapitre 7.

2. Association linéaire

Même si un associateur non-linéaire serait meilleur (on trouve une étude comparative des modèles linéaires et non linéaires dans [BB99a]), compte-tenu des non-linéarités entre l'acoustique et la géométrie du conduit vocal, Yehia et al. en 1998 [YRV98] montre qu'un modèle linéaire permet une approximation sur les corrélations entre audio et géométrie.
L'étude de Yehia et al. examine les associations linéaires entre paramètres acoustiques, géométrie du conduit vocal et attitudes du visage. Elle indique que près de 80% de la variance observée dans les configurations du conduit vocal peut être estimée à partir de la position 3D de points fixés sur la surface du visage (lèvres comprises). De plus, l'associateur linéaire permet également d'estimer l'acoustique en terme de représentations LSP (Line Spectrum Pair) : 73% et 69% de la variance observée dans les LSP sont respectivement récupérées à partir des mouvements de la face et des mouvements de la langue.
Dans un modèle de multi-régression, chaque paramètre spectral est estimé comme une combinaison linéaire des paramètres d'entrée. Une base d'apprentissage sert à prédire les paramètres linéaires qui sont ensuite utilisés sur une base de test pour estimer de nouvelles données.
Cette approche a été l'objet d'autres études entre paramètres acoustiques et lèvres ou entre géométrie de la langue et caractéristiques du visage. En 2002, Bailly et Badin [BB02] étudient la corrélation entre mouvements faciaux et mouvements de la langue à l'aide d'un modèle articulatoire basé sur des enregistrements vidéos et cinéradiographiques. Beskow et

al. [BEG03] prédisent des données de langue à partir du visage à l'aide d'estimateurs linéaires, à partir de mesures simultanées des mouvements de la langue et du visage, dans l'optique d'améliorer l'articulation d'une tête parlante. Ces études concluent que le visage procure de l'information sur les articulateurs de la parole, mais Bailly et Badin alertent sur l'insuffisance de cette information pour récupérer le lieu d'articulation.
Barker et Berthommier [BB99b] emploient des estimateurs linéaires pour évaluer dans quelles mesures les caractéristiques acoustiques (représentation LSP aussi) peuvent être récupérées à partir des caractéristiques vidéos (paramètres géométriques de lèvres), et vice et versa, pour 54 non-mots français répétés 10 fois. 75% de la variance totale des configurations labiales est estimée à partir des LSP, alors que 55% seulement de la variance des données acoustiques est expliquée à partir des données géométriques et labiales. Ceci indique que pour les lèvres, l'acoustique n'est pas bien inférée à partir de la géométrie et on peut légitimement espérer que ce résultat soit meilleur à partir de la langue.
L'étude de Jiang et al. [JAB+02] se rapproche de l'étude initiale de Yehia et al. et utilise des données enregistrées simultanément par capture de mouvement optique et articulographie électromagnétique.

Ces études permettent de montrer que l'association linéaire entre paramètres audio et vidéo de la parole capture une grande part de l'information mais on observe également qu'elle n'est pas totalement suffisante, les propriétés non-linéaires articulo-acoustiques et la relation « many-to-one » étant aussi à prendre en compte.
Nous analysons cependant, au chapitre 6, à l'aide d'un estimateur linéaire dans quelle mesure les paramètres audio peuvent être déterminés à partir de nos données géométriques extraites et nous étudierons les corrélations entre géométrie et acoustique.

Mais avant toute chose, nous commençons, au chapitre 5, par observer les configurations articulatoires estimées et nous réalisons une analyse descriptive de ces données, essentiellement orientée vers les voyelles du corpus.

CHAPITRE 5 : ANALYSE DESCRIPTIVE DES DONNÉES ARTICULATOIRES DE LAVAL43

Avant de réaliser le passage des données articulatoires aux données acoustiques par synthèse articulatoire ou association linéaire, une observation directe des données géométriques extraites depuis la vidéo est réalisée ici. Une mise en correspondance des contours géométriques estimés et d'éléments acoustiques extraits du signal est présentée. Il ne s'agit pas encore d'estimer l'acoustique à partir de l'articulatoire mais simplement d'observer certaines configurations estimées du conduit vocal connaissant le contenu phonétique du corpus. Nous nous intéressons dans ce chapitre principalement aux voyelles afin de vérifier que nous retrouvons des observations classiques. Une étude spécifique aux consonnes sera l'objet du chapitre 8.

1. Données phonétiques

Nous disposons, avec la séquence des images Laval43, du signal audio correspondant au corpus des phrases prononcées (en annexe A2). Ce signal et ce corpus nous permettent de connaître le contenu phonétique disponible dans la séquence.

Avant d'en extraire des informations utiles pour la mise en correspondance avec les données articulatoires estimées, nous nous sommes assurés de la bonne synchronisation entre audio et vidéo. Cette vérification a été réalisée à partir de la séquence vidéo : nous avons vérifié qu'aux instants précis de plosives bilabiales [b] ou [p], nous observions une fermeture des lèvres sur les images radiographiques.

1.1. Distinction (audio) parole-silence

A l'écoute et à l'observation du signal Laval43, nous constatons la présence de nombreux instants de silence, notamment entre chacune des phrases du corpus. Pour certains aspects des différentes études qui vont suivre, il est intéressant de se limiter parfois aux trames « de parole », c'est-à-dire à la séquence privée des trames de silence.

La distinction des trames de parole et de silence est réalisée à partir d'une décomposition en 4 sous-bandes du signal. On utilise pour cela un banc de 4 filtres quasi-rectangulaires à gain unitaire (Fig. 65), définis en échelle Bark. Ils sont construits à partir du regroupement et de la sommation de 16 filtres initiaux pondérés par une fenêtre de Hanning répartis aussi selon une échelle Bark.

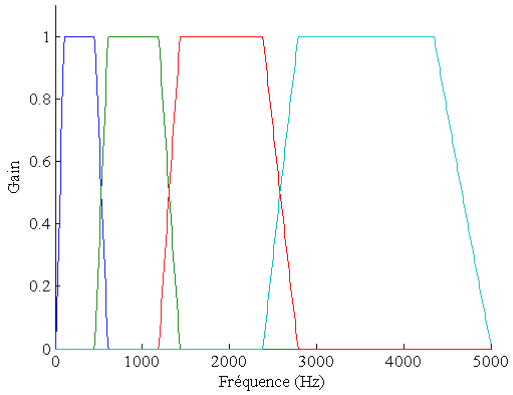


Figure 65 : Banc de filtres 4 sous-bandes.

Grâce à un seuil fixé pour l'amplitude du signal dans la première sous-bande ([0-500 Hz]), nous sommes en mesure de séparer les trames de silence des autres. A un instant donné, une amplitude faible dans la première sous-bande correspond à une faible énergie du signal. Nous considérons que si l'amplitude du signal est inférieure à -30dB alors il s'agit d'un instant de silence. Cette distinction parole/silence sera largement utilisée par la suite.
Compte-tenu de ce seuil, sur les 4043 trames (images) de la séquence, plus de la moitié des trames (2114) sont des trames de silence.

1.2. Voyelles de Laval43

Le but de ce chapitre est de retrouver, à partir de nos données géométriques extraites, des résultats et observations classiques qui ont été établis pour les voyelles, dans la littérature.
Un étiquetage des voyelles du signal audio Laval43 a été réalisé. L'étiquetage consiste à signaler les moments où les voyelles sont produites, en associant à la trame correspondante la classe vocalique. Chaque voyelle étiquetée est donc associée à une image de la séquence. Une des propriétés intéressantes des voyelles est qu'elles sont quasi-stationnaires, la mise en correspondance des données spectrales et géométriques est facilitée. Dans les parties stationnaires des noyaux vocaliques, plusieurs trames consécutives peuvent être sélectionnées pour étiqueter la voyelle : chaque trame a en effet une durée courte de 33ms (nous disposons de 29,97 images par seconde) et un noyau vocalique couvre alors plusieurs trames. Cependant, nous avons choisi d'étiqueter chaque noyau vocalique une seule fois, c'est-à-dire par une seule trame, au centre du noyau.
Nous utilisons pour l'étiquetage une interface de marquage audio, présentée en annexe A4. Ceci nous permet de disposer du corpus décrit dans le tableau suivant.
La plupart des voyelles du corpus ont été étiquetées, mais quelques-unes ont été laissées de côté lorsque nous avons estimé qu'il pouvait y avoir un doute sur leur perception, notamment entre des [e] et des [ɛ] ou des [o] et des [ɔ].

Le système vocalique du français se présente sous différentes formes. Traditionnellement on fait la distinction entre onze voyelles orales et quatre voyelles nasales.

Le français connaît quatre degrés d'ouverture : fermé, mi-fermé, mi-ouvert et ouvert. Les voyelles sont également classées dans la dimension de l'antéro-postériorité (antérieur : [i], [e], [ɛ], [a], [y], [ø] et [œ] ; postérieur : [u], [o], [ɔ] et [ɑ]). La série antérieure est divisée en deux en fonction de l'arrondissement des lèvres ; les voyelles [y], [ø] et [œ] sont arrondies et les [i], [e], [ɛ] et [a] sont non-arrondies.

Les voyelles fermées sont [i], [y] et [u]. Les voyelles mi-fermées sont [e], [ø] et [o], et les voyelles mi-ouvertes sont [œ], [ɛ] et [ɔ]. Il y a en plus la voyelle centrale non-arrondie moyenne [ə], qui est acoustiquement presque identique au [œ]. Dans notre étude, nous considérerons ensemble [ø], [ə] et [œ]. Nous avons de plus choisi [a] pour représenter les deux voyelles ouvertes du français, tout en étant conscients que sur le plan phonétique [a] et [ɑ] représentent deux voyelles différentes.

La notion d'arrondissement est difficile à prendre en compte dans notre travail, compte-tenu du manque d'information à ce propos. Nous n'avons à disposition que la vue sagittale du conduit vocal. Nous considérons ensemble les 2 voyelles nasales [ɛ̃] et [œ̃].

Nous nous trouvons donc finalement avec 3 catégories de voyelles nasales et 9 de voyelles orales, présentées dans le tableau 8 et comptabilisées en nombre d'occurrences vocaliques.

	Voyelles	**Nombre**	**Total**
Orales	[a], [ɑ]	54	227
	[i]	33	
	[y]	16	
	[u]	10	
	[o]	20	
	[ɔ]	15	
	[e]	38	
	[ɛ]	20	
	[ø], [ə], [œ]	21	
Nasales	[ɛ̃], [œ̃]	27	65
	[ɑ̃]	27	
	[ɔ̃]	11	

Table 8 : Voyelles sélectionnées dans le corpus de Laval43.

2. Observations directes

Chaque image de la séquence est associée à un contour géométrique de conduit vocal, obtenu par combinaison des divers articulateurs estimés dans les conditions d'application du modèle M5 (multi-indexation par 4 voisins, filtrage temporel et interpolations splines entre les points). Avant de passer à des représentations paramétriques de ces données géométriques, nous présentons l'allure des données, en moyenne selon chaque classe vocalique étiquetée. Nous montrons d'abord les configurations géométriques directes du conduit vocal puis les fonctions d'aire déduites.
Notre observation se limite aux trames de voyelles sélectionnées.

2.1. Configurations géométriques moyennes

Nous considérons les configurations géométriques associées à chaque voyelle étiquetée du corpus. La figure qui suit montre une représentation du conduit vocal pour chaque classe de voyelles orales (nous nous limitons à celles-ci pour l'instant, les voyelles nasales seront analysées à part, au chapitre 9). Chacun de ces schémas représente le contour estimé moyen du conduit vocal pour les différents exemplaires de la classe ainsi que l'écart type (en pointillés) des contours de langue de part et d'autre du contour moyen (en trait plein).

Nous retrouvons des observations classiques sur les voyelles que d'autres auteurs ont constatés ([HLG77], [Fan73], [BSWZ86]…). Avant d'analyser chaque classe vocalique plus en détail, nous donnons ici quelques observations générales.
Sur chacun des 9 graphiques pour les 9 classes orales des voyelles du corpus, le vélum est en position haute et en contact avec la paroi pharyngale.
Si on considère la propriété d'antériorité/postériorité (position de la langue en avant ou en arrière dans le conduit vocal), on observe que les voyelles [ɔ], [o] et [u] sont postérieures (le dos de la langue se masse en arrière) alors que les voyelles [i], [y], [ɛ] et [e] sont antérieures (la langue est en avant).
Le degré d'ouverture des lèvres est moins facile à observer. Le mouvement des lèvres sur la séquence est faible, l'écart aux lèvres varie peu au fur et à mesure des images. Les lèvres ne sont jamais très ouvertes et ceci se vérifie sur la vidéo d'origine où l'on observe une tendance à l'hypo-articulation, probablement pour accélérer le temps d'enregistrement de la séquence. Il est donc difficile à partir des contours moyens estimés et observables figure 66 de faire une réelle distinction entre voyelles fermées, mi-fermées, mi-ouvertes et ouvertes. Cependant, on peut noter que pour les voyelles [u] et [y], la langue est haute et les lèvres se referment, il s'agit de voyelles fermées. De même, nous observons que pour la voyelle [a], la

langue est abaissée et les lèvres sont plus ouvertes que pour les autres voyelles, nous avons donc bien une voyelle ouverte (ou voyelle basse).

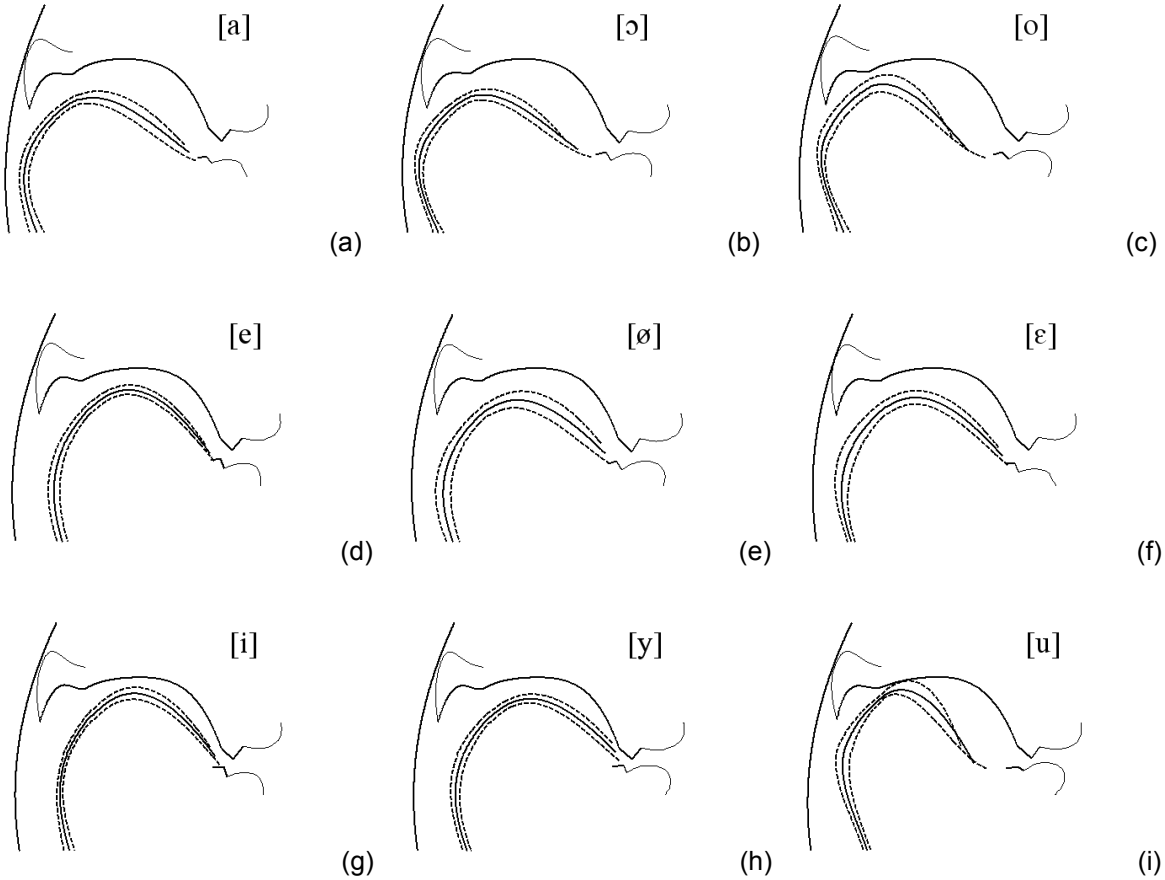


Figure 66 : Contour du conduit vocal obtenu en moyenne sur les exemplaires de chaque classe vocalique. Pour la langue, un écart-type est observé de part et d'autre du contour moyen - (a) [a] – (b) [ɔ] – (c) [o] – (d) [e] – (e) [ø] – (f) [ɛ] – (g) [i] – (h) [y] –(i) [u].

En suivant la première ligne de graphiques et les contours des voyelles [a], [ɔ] et [o], on constate que l'écart aux lèvres se resserre et que la langue se rétracte en arrière. La voyelle [o] est une voyelle postérieure mi-fermée.

La seconde ligne concerne les contours moyens pour les voyelles [e], [ø] et [ɛ]. La classe vocalique [ø] présente une position centrale de la langue et une ouverture de lèvres moyenne, en comparaison aux autres classes. On observe également une variabilité plus grande, l'écart-type sur la langue est plus important, ceci est en accord avec la littérature et l'analyse du schwa ([SBVA97a]). Pour [e], la langue est haute et en avant, les lèvres sont peu ouvertes, il s'agit d'une voyelle antérieure mi-fermée. Quant à [ɛ], la langue est en avant, un peu plus basse et les lèvres sont plus ouvertes. C'est une voyelle antérieure mi-ouverte.

La dernière ligne montre les profils des voyelles [i], [y] et [u]. La langue est en position haute. Pour [i], elle est en avant, un peu moins pour [y] et en arrière pour [u]. L'écart-type observé pour la langue sur cette dernière voyelle met en évidence une position de bascule avec 2 types d'articulation pour le [u].

Ces observations sont en accord avec les données classiques de phonétique, nous distinguons correctement les positions principales avant-arrière et haut-bas habituelles des voyelles du français.

2.2. Fonctions d'aire moyennes

D'une façon similaire, nous observons la fonction d'aire moyenne par classe vocalique, ainsi que l'écart-type (en pointillés sur les figures qui suivent, Fig. 67).

Nous constatons que l'écart-type est particulièrement grand dans la cavité avant, pour les voyelles arrière [u], [o] ou [ɔ]. Ceci s'explique par le mode de mesure des distances sagittales dans cette cavité. En effet, comme on a pu le voir au chapitre 4, les premières sections (hormis les incisives avant et les lèvres) permettent la mesure de la distance entre le palais et la langue ou entre le palais et le plancher sous-lingual, ce qui implique rapidement de fortes variations dans le calcul des distances. De plus, la variance augmente avec le passage des distances sagittales à la fonction d'aire par la transformation exponentielle $\alpha\beta$. L'écart-type est également important pour le [ø] qui est une classe où nous considérons en réalité [ø], [ə] et [œ]. De plus, comme on l'a déjà fait remarquer avec la position du contour moyen de la langue, cette classe vocalique correspond au schwa, qui a une variabilité plus importante. Par contre, nous observons une variabilité plus faible pour le [i] et le [y].

La fonction d'aire permet de mettre en évidence la présence de la constriction, sa position et sa taille. La constriction sépare le conduit vocal en 2 cavités (cavité arrière ou pharyngale et cavité avant ou buccale). Les résonances dans ces cavités sont à l'origine des formants, dont nous reparlerons.

Avec l'observation de ces fonctions d'aire moyennes, nous vérifions géométriquement que le rapprochement du corps de la langue vers la région palatale crée, pour [i], une cavité avant petite et une cavité arrière de volume important. Au contraire la voyelle ouverte [a] présente une cavité antérieure plus grande.

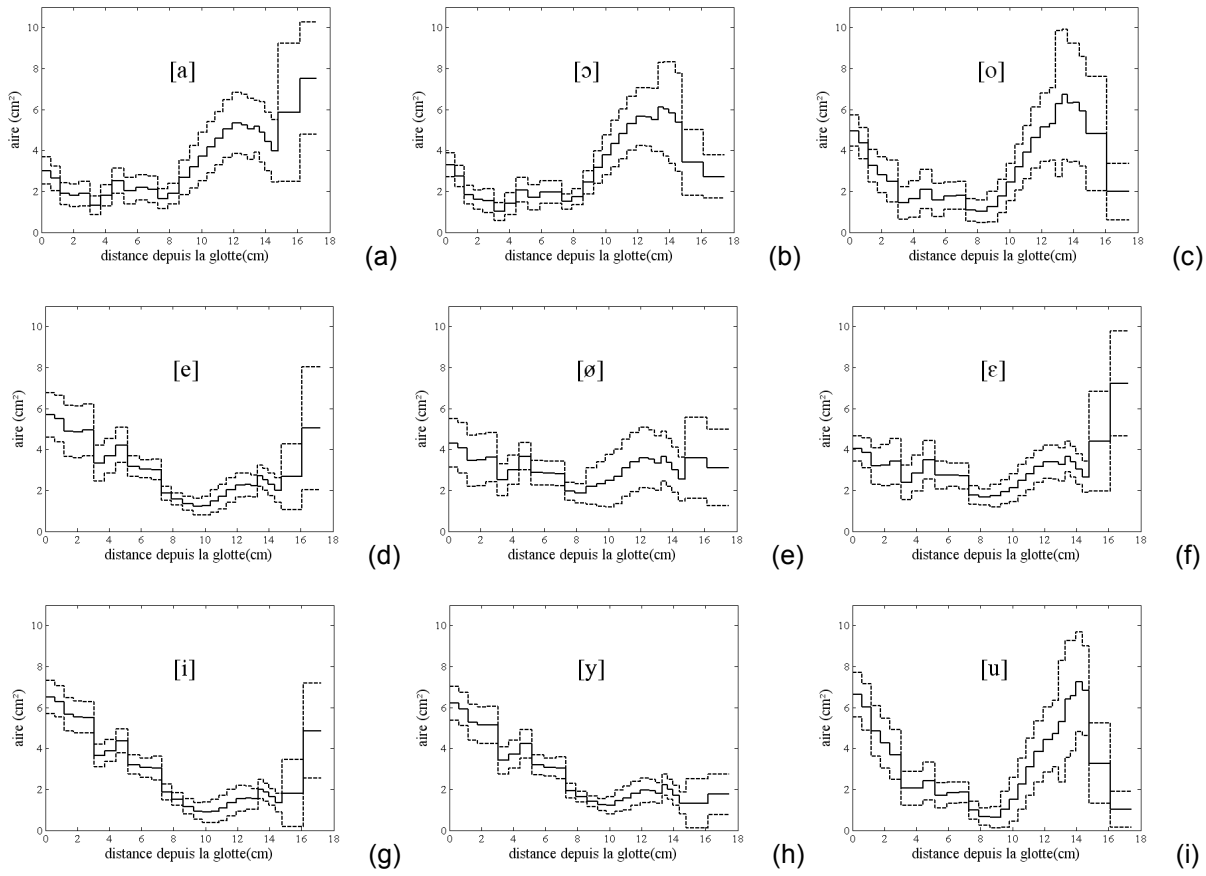


Figure 67 : Fonction d'aire obtenue en moyenne sur les exemplaires de chaque classe vocalique (et écart-type en pointillés) - (a) [a] – (b) [ɔ] – (c) [o] – (d) [e] – (e) [ø] – (f) [ɛ] – (g) [i] – (h) [y] –(i) [u].

3. Représentations paramétriques

Les représentations paramétriques que nous allons présenter cherchent à mettre en évidence une certaine organisation des données articulatoires, sachant le contenu phonétique. Ces représentations pourront s'appliquer aux données de la séquence complète, de la séquence sans silence, elles pourront aussi se limiter aux voyelles sélectionnées. Pour différencier les neuf classes de voyelles orales, nous avons choisi un code graphique couleur défini ci-dessous. Pour toute la suite de ce manuscrit, le choix des signes graphiques est le suivant (Table 9).

[a]	[i]	[y]	[u]	[e]	[ɛ]	[ø],[ə],[œ]	[o]	[ɔ]
▲	▼	●	*	♦	◀	▶	☆	■

Table 9 : Voyelles et signes graphiques associés.

3.1. Espaces de représentation

Les espaces de représentation permettent d'étudier, de comparer, de classer les systèmes vocaliques de différents locuteurs pour différentes langues. Au niveau articulatoire, deux approches sont en concurrence [BGP+95]. Une première privilégie la position du dos de la langue, alors que la seconde est fondée sur le lieu et la dimension de la constriction et l'aire aux lèvres.

3.1.1. Point le plus élevé de la langue

La représentation phonétique traditionnelle qui consiste à observer la position du dos de la langue est pratiquement isomorphe avec l'espace de représentation acoustique (dérivé des 2 premiers formants). La reproduction approximative de la forme de la langue à partir des deux paramètres représentants ce plus haut point sur l'axe horizontal et l'axe vertical a été jugée suffisamment pertinente par un certain nombre d'auteurs ([SH55], [LS71]).

Nous observons (Fig. 68) l'évolution du point le plus élevé du dos de la langue (autrement dit de la zone la plus proche du palais) au cours de la séquence et mettons en évidence une position basse pour le [a], avant pour le [i] et haute et arrière pour le [u]. On observe cependant une certaine dispersion des données. Les ellipses représentant un écart-type pour chaque voyelle ont tendance à se recouvrir, même si l'organisation générale est respectée et si certaines classes (notamment le [u]) sont bien dégagées.

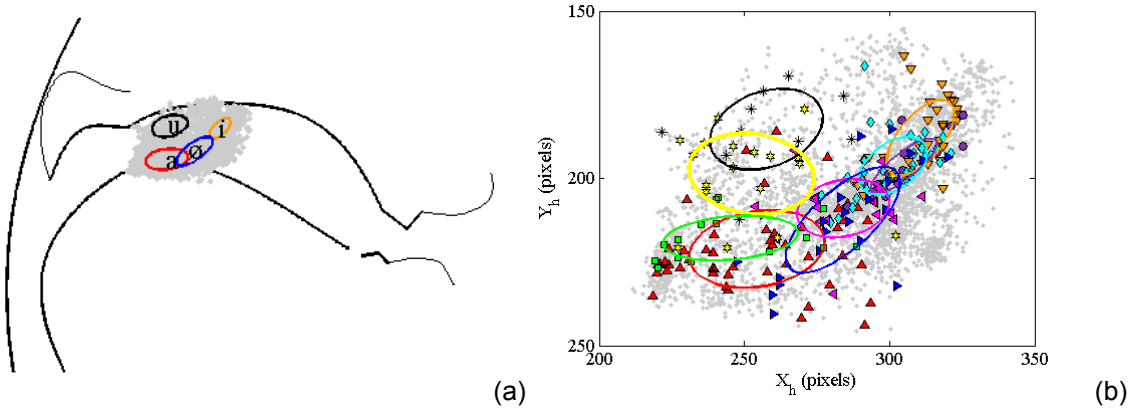


Figure 68 : Représentation graphique de la position (X_h, Y_h) du point le plus élevé du dos de la langue et ellipses de dispersion (à un écart-type) des voyelles.

3.1.2. Ouverture aux lèvres

De façon similaire au point le plus élevé du dos de la langue, nous observons un point sur la lèvre inférieure et sa position au cours de la séquence. Le point en question correspond au point de marquage à 2 degrés de liberté à l'avant de la lèvre (voir figure 36b). Les ellipses de dispersion à un écart-type de quelques voyelles sont représentées sur la figure 69. Nous

avons également représenté les positions du point analogue de la lèvre supérieure (le point de marquage à 2 ddl à l'avant de la lèvre). Pour ce point, les ellipses sont quasiment superposées, mettant en évidence le peu de déplacement de la lèvre supérieure. Compte-tenu de cette position presque fixe de la lèvre supérieure, l'observation du point considéré de la lèvre inférieure permet de visualiser l'ouverture aux lèvres. Le maximum d'ouverture aux lèvres est observé pour la voyelle [a] et le minimum pour la voyelle [u]. La figure permet aussi de remarquer la protrusion avant des lèvres pour les voyelles [ø] et [u], de façon marquée pour la lèvre inférieure, mais aussi (dans une moindre mesure évidemment) pour la lèvre supérieure.

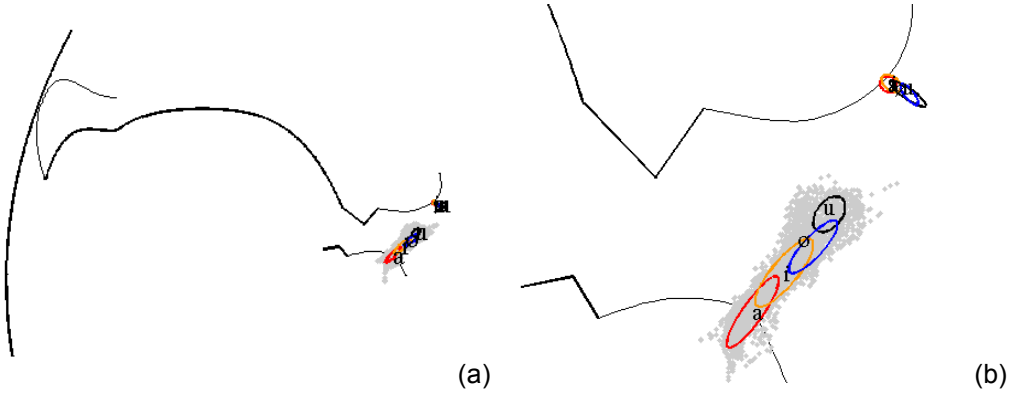


Figure 69 : Représentation de la position du point à 2 ddl à l'avant de la lèvre inférieure et du point à 2 ddl à l'avant de la lèvre supérieure et ellipses de dispersion (à un écart-type) pour quelques voyelles.

3.1.3. Constriction

La théorie acoustique de la production de la parole [Fan60] permet de décrire les dispositions du conduit vocal à l'aide de 3 paramètres de la fonction d'aire, directement interprétables en terme de paramètres articulatoires de contrôle : X_c, A_c et A_l, soit, la position de la constriction, son aire et l'aire aux lèvres.

Ces paramètres ont été très vite adoptés pour la description articulatoire des voyelles. Stevens & House (1955) et Fant (1960) ont conçu des modèles à partir de ces trois paramètres, exploités pour établir des nomogrammes ([SH55], [Fan60]). En variant ces paramètres, on peut simuler l'ensemble des voyelles orales.

Cette description est utilisée pour proposer, avec la théorie quantique (Stevens, 1972, [Ste72]), un cadre très général de l'organisation des systèmes sonores ou pour étudier les problèmes d'inversion ([ACM+78], [BPB92]). Wood conforte cette approche par des mesures radiographiques [Woo79]. La relation entre les paramètres de constriction et la sensibilité au niveau acoustique a été largement étudiée ([GBP+91] entre autres).

Nous observons le lieu et la taille de constriction sur les fonctions d'aire. Pour cela (Fig. 70), nous recherchons pour chaque image le minimum de la fonction d'aire et le numéro de la section pour laquelle ce minimum est atteint, en omettant dans cette recherche les sections relatives à la pointe de la langue et au bas du pharynx. Ceci nous permet de disposer pour chaque trame, de la taille de constriction A_c (il s'agit de l'aire de la constriction) et du point de constriction X_c.

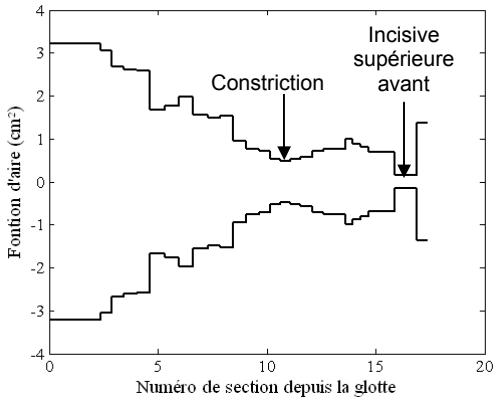


Figure 70 : Recherche du point constriction à partir de la fonction d'aire.

Nous représentons alors (Fig. 71) les données articulatoires en traçant l'aire de la constriction en fonction de la position de la constriction par rapport à l'incisive supérieure avant ou l'aire aux lèvres en fonction de cette position. La position de la constriction est fixée par rapport aux incisives à cause de la variabilité du point de mesure aux lèvres (protrusion) et de la non-visibilité de la glotte.

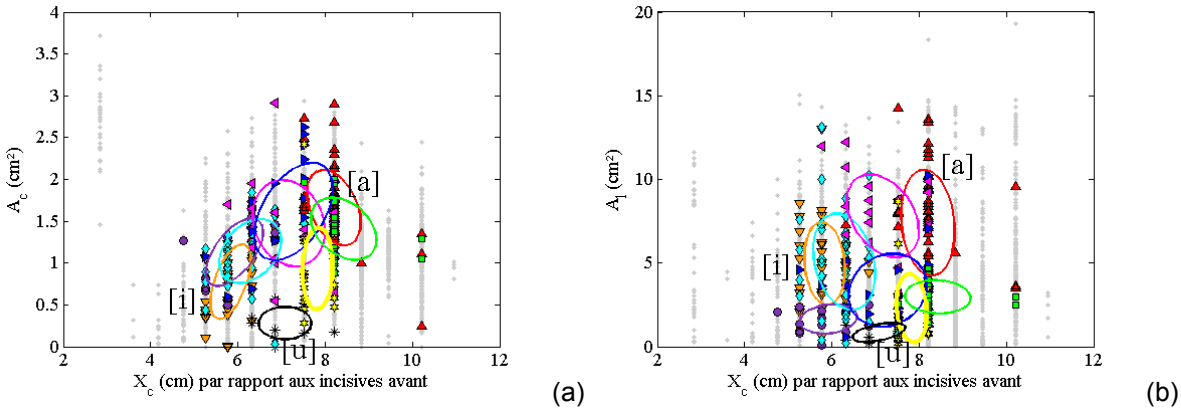


Figure 71 : (a) Espace de représentation X_c (lieu de constriction par rapport aux incisives) - A_c (aire de la constriction) et ellipses de dispersion des voyelles. (b) Espace de représentation X_c (lieu de constriction par rapport aux incisives) - A_l (aire aux lèvres) et ellipses de dispersion des voyelles.

La représentation dans le plan X_c-A_l est plus ordonnée que celle dans le plan X_c-A_c, où les ellipses présentent un plus fort recouvrement. Ces mesures permettent dans une certaine mesure de classifier l'information géométrique : on observe à peu près correctement les

résultats attendus pour les voyelles extrêmes. Les [a] présentent une grande constriction arrière et une grande aire aux lèvres. Les [i] se caractérisent par une constriction de faible dimension à l'avant et une aire aux lèvres plus petite. Quant aux [u], la constriction est petite et plutôt en arrière du conduit vocal et l'aire aux lèvres est faible.
La variabilité est trop grande pour que l'on puisse paramétriser nos données selon des points locaux. Il vaut mieux utiliser des techniques « globales » comme l'ACP.

3.2. ACP

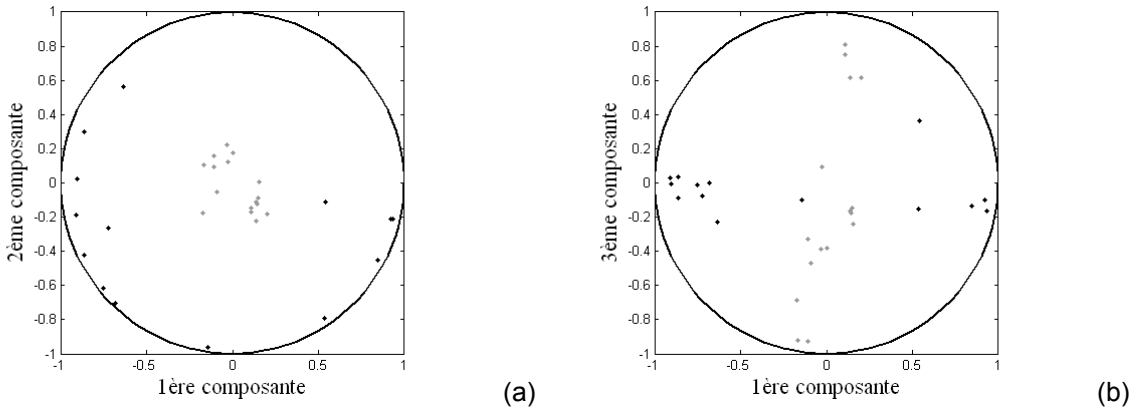
L'analyse en composantes principales, comme décrite précédemment, constitue un outil pour décrire et intégrer la redondance entre plusieurs mesures ou variables. Elle est souvent utilisée, et c'est le cas ici, pour représenter graphiquement et de manière synthétique les faits saillants d'un ensemble de données. Grâce à cette méthode descriptive, nos données articulatoires (nombreuses et variables !) peuvent être visualisées et résumées graphiquement.
Nous avons effectué deux séries d'analyses en composantes principales :

- La première série a été réalisée sur les degrés de liberté marqués du conduit vocal pour les 4043 trames de la séquence. Nous avons considéré dans un premier temps uniquement les 15 degrés de liberté de la langue, puis dans un deuxième temps ces mêmes 15 ddl avec les 16 ddl relatifs aux lèvres.
- La deuxième série a été réalisée sur les distances mesurées des sections médio-sagittales, à partir des contours géométriques. L'ensemble de données considéré est constitué de 4043 observations de 28 sections. L'ACP a été effectuée sur les distances sagittales ainsi que sur les aires calculées grâce à ces distances, par le modèle alpha-beta.

3.2.1. ACP sur les degrés de liberté géométriques

Les premières composantes de l'ACP sur les degrés de liberté des points marqués du conduit vocal permettent de représenter les données sur les graphes qui suivent. Dans le cas où seuls les 15 degrés de liberté de la langue sont pris en compte (Fig. 73a), ces 2 premières composantes expliquent respectivement 64% et 19% de la variance (soit 83% de la variance totale).
Lorsqu'on ajoute les 16 ddl des lèvres, la représentation dans le plan principal des 2 premières composantes est très semblable mais la variance expliquée par ces 2 composantes n'est plus que de 70% (54%+16%). Afin de voir comment les degrés de liberté sont distribués selon les axes principaux, nous traçons les cercles de corrélation, qui comme

leur nom l'indique, montre la corrélation entre les données de départ et les vecteurs principaux. Nous traçons figure 72 le cercle des corrélations pour les composantes 1 et 2 et celui pour les composantes 1 et 3. Les points noirs représentent les ddl de la langue, les points gris ceux des lèvres. Les ddl de la langue se placent sur le contour du cercle de la figure 72a, ils sont donc correctement représentés par les 2 premières composantes, ce qui n'est pas le cas des ddl des lèvres qui se concentrent au centre du cercle. La figure 72b tend à montrer que les lèvres sont représentées selon la 3ème direction puisque les ddl s'alignent sur cette dimension. Sur la figure 73b, nous avons représenté la projection des données (langue et lèvres) dans le plan des 1ère et 3ème composantes de l'ACP.



(a) (b)

Figure 72 : Cercles de corrélation mettant en évidence la distribution des degrés de liberté de la langue (en noir) et des lèvres (en gris) selon les 3 premières composantes principales de l'ACP.

L'observation des voyelles dans ces représentations (Fig. 73a et 73b) montre que pour chaque catégorie vocalique, les exemplaires sont à peu près concentrés au même endroit mais chacune couvre une surface importante du plan principal. Il y a recouvrement des classes. Les ellipses de dispersion à un écart-type des voyelles mettent en évidence ce recouvrement, qui est particulièrement important au niveau des voyelles antérieures dans le cas des 15 ddl de langue (Fig. 73a et 73c). Ce recouvrement est réduit lorsque l'on prend en compte en plus les 16 degrés de liberté des lèvres (Fig. 73b). La distinction entre le [i] et le [y] est mieux marquée.

Les voyelles couvrent l'espace complet de projection (et donc l'espace maximal). Si on se limite aux classes extrêmes [i], [u] et [a], on constate qu'elles occupent des zones extrêmes de l'espace géométrique, que l'on met en parallèle des positions extrêmes de la langue (position haute ou basse, avant ou arrière, Fig. 73d) : [u] en arrière, [i] en avant, [a] en bas.

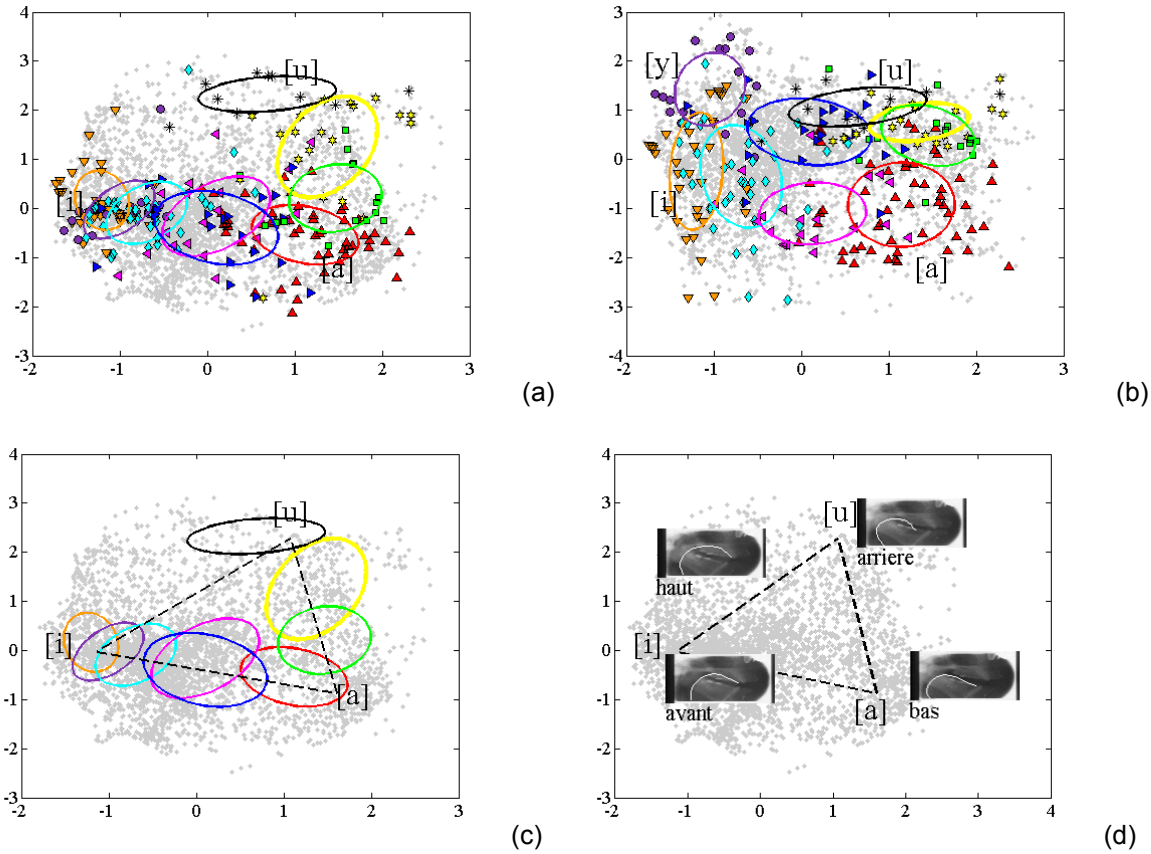


Figure 73 : ACP sur les degrés de liberté des points marqués du conduit vocal pour la séquence complète (silence compris) et observation des voyelles – (a) 15 ddl de la langue : 2 premières composantes – (b) 15 ddl de la langue + 16 ddl des lèvres : composantes 1 et 3 – (c) Ellipses de dispersion pour les 15 ddl de la langue – (d) Positions extrêmes de la langue.

Pour quantifier la confusion entre les classes vocaliques, mise en évidence avec les ellipses de dispersion sur les graphiques, nous considérons une mesure quantitative de capacité discriminante. Cette mesure sera à nouveau utilisée au chapitre 7, elle s'appuie sur l'élaboration de matrices de confusion à partir des données.

La mesure de capacité discriminante est basée sur le calcul des distances des voyelles aux centres des ellipses de dispersion figurées dans les plans principaux. Etant donnée par exemple une voyelle [a], on estime que cette voyelle est correctement placée si, en terme de distance euclidienne dans le plan principal, cette voyelle est plus proche du centre de l'ellipse des [a] que de n'importe quel autre centre d'ellipse. Sinon on estime qu'il y a confusion et cette voyelle est alors classée dans la catégorie vocalique correspondant à l'ellipse dont le centre est le plus proche. Ceci permet de caractériser le pourcentage de voyelles correctement classifiées et d'évaluer le nombre et le type de confusions qui apparaissent. Ces résultats sont présentés sous forme de tables de confusions (figures 74a et 74b). La diagonale indique, par voyelle, les pourcentages de « bonnes détections ». On définit la capacité discriminante totale comme la moyenne des éléments de la diagonale, ce qui

correspond à moyenner sur les 9 classes vocaliques le taux de bonne reconnaissance de chaque classe.

Cette capacité est égale à 45,5% lorsque l'on analyse uniquement les 15 ddl de la langue. Le [u], dont l'ellipse se dégage bien des autres, est discriminé à 90%. Si on ajoute les 16 ddl des lèvres, la capacité discriminante totale atteint 55%. Dans ce dernier cas, on observe, en particulier, une meilleure discrimination du [i] et surtout du [y] qui est reconnu à 87%.

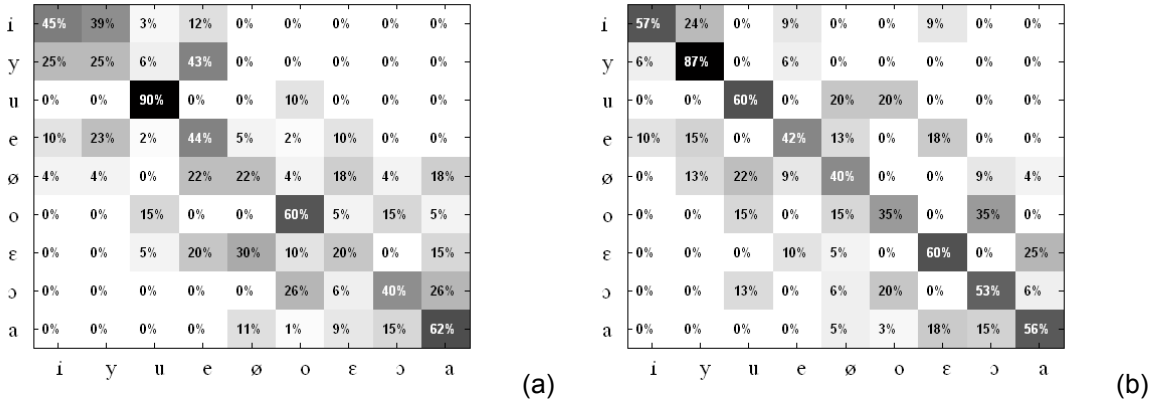


Figure 74 : Tables de confusion associées aux plans principaux obtenus à partir des données géométriques – (a) 15 ddl de la langue – (b) 15 ddl de la langue + 16 ddl des lèvres.

3.2.2. ACP sur les sections médio-sagittales et les fonctions d'aire

Nous présentons maintenant les projections dans le plan principal des paramètres géométriques représentés soit par les sections médio-sagittales (Fig. 75a), soit par les fonctions d'aire (Fig.75b).

Les variances expliquées sont respectivement 80% et 74%. Il semble que le recouvrement dans le plan principal des fonctions d'aire soit moins important, mis à part pour les voyelles postérieures [o] et [ɔ]. On met à nouveau en parallèle les voyelles extrêmes avec des formes particulières de la fonction d'aire (figures 75c et 75d) : une grande cavité arrière et une constriction avant pour [i], une constriction arrière pour [u] et enfin pour [a], une grande cavité avant, résultant d'une langue basse.

De même que précédemment, l'usage de la capacité discriminante permet d'évaluer ces projections. Cette capacité est de 44% lorsque l'on considère la projection des sections et de 52% lorsque ce sont les fonctions d'aire qui sont projetées.

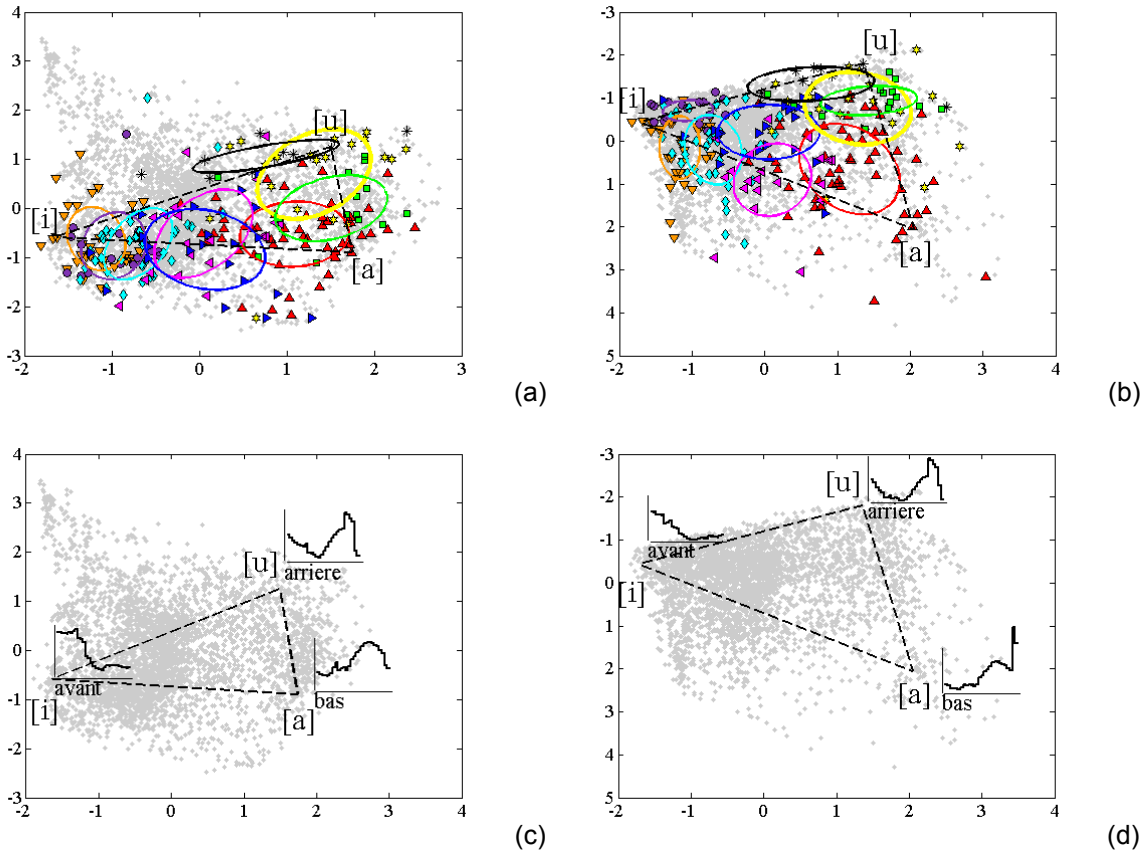


Figure 75 : (a) et (c) ACP sur les 28 distances sagittales du conduit vocal pour la séquence complète et observation des voyelles et ellipses de dispersion associées – mise en parallèle avec des allures de sections.
(b) et (d) ACP sur les fonctions d'aire du conduit vocal pour la séquence complète et observation des voyelles et ellipses de dispersion associées – mise en parallèle avec des allures de fonction d'aire.

Compte-tenu de ces distributions, dispersées, dans les plans principaux, nous avons observé quelques configurations géométriques, celles correspondant à des voyelles représentées par des points éloignés du centre de l'ellipse associée. Le contrôle visuel des contours du conduit vocal marqués pour ces images ne montre pas d'erreur importante de marquage géométrique. Plusieurs configurations géométriques ou formes articulatoires peuvent produire le même spectre de parole [ACM+78]. L'étude de ces spectres est donc indispensable et sera l'objet du chapitre 7.

En conclusion, l'observation des données permet de dire que la structure générale du triangle des voyelles est conservée. Nous retrouvons les observations classiques des configurations géométriques moyennes pour chaque classe vocalique. La variabilité dans chacune des ces classes selon le plan principal est grande avec un recouvrement plus ou moins important selon l'espace paramétrique adopté.

CHAPITRE 6 : ASSOCIATION LINÉAIRE

Nous avons vu au chapitre précédent que la forme du conduit vocal correspondant à chaque classe vocalique est bien retrouvée dans Laval43. Nous pouvons donc tenter d'associer linéairement les degrés de liberté géométriques et la position des formants.
Nous nous intéressons donc dans ce chapitre à l'étude d'un modèle d'association linéaire entre données audio et géométriques. Le modèle que nous allons présenter est proche de celui proposé par Yehia et al. [YRV98], mais nous considérerons les formants au lieu des LSP (Line Spectral Pair). En effet, il est établi que les trajectoires formantiques sont associées aux mouvements du conduit vocal. Nous proposons donc d'analyser l'association linéaire entre formants et données géométriques extraites depuis la cinéradiographie, dans l'optique de synthétiser des sons de parole à partir de ces données.
Cette étude, contrairement aux autres chapitres de cette seconde partie, sera basée sur les données articulatoires obtenues à partir de la séquence cinéradiographique Wioland, et non pas Laval43.

1. Modèle linéaire

Le modèle linéaire, type Yehia [YRV98], est utilisé dans l'idée d'associer linéairement des données acoustiques et articulatoires.

1.1. Données articulatoires et acoustiques

1.1.1. Données articulatoires

Les données articulatoires en question sont les configurations géométriques obtenues à partir de la méthode d'extraction de données. Largement détaillées dans la première partie, nous n'en dirons pas plus. Notons simplement que nous considérons les contours marqués pour la langue et les lèvres. Nous disposons ainsi de 8 points (ou 10 degrés de liberté) pour les lèvres et de 10 points (ou 12 degrés de liberté) pour la langue et ce, pour quelques 4200 trames (nous laissons de côté les images surexposées de la troisième partie de la séquence). Rappelons aussi que la séquence Wioland numérisée propose 25 images par seconde, soit une image toutes les 40ms.

1.1.2. Données acoustiques

Les données acoustiques sont les fréquences des formants F_1, F_2 et F_3.

Les formants sont des paramètres acoustiques très classiques, qui permettent d'observer le contenu phonétique d'un corpus. Lors de la phonation, certaines fréquences du son produit par les cordes vocales sont amplifiées par les cavités de résonance en fonction de la fréquence de résonance propre à chaque cavité. Ce sont ces fréquences que l'on nomme "formants". Ce sont les endroits du spectre sonore qui présentent les plus grandes accumulations de pression sonore. Les formants (abréviation F_1, F_2, F_3, F_4) permettent d'identifier le timbre des sons, chaque voyelle se caractérise par son timbre spécifique. Les valeurs en Hertz des quatre premiers formants des sons humains sont particulièrement significatives.

Depuis les travaux de Delattre et Joos en 1948 ([Del48], [Joo48]), on représente généralement l'espace vocalique sous la forme d'un triangle. Les formants sont tracés dans un plan F_1-F_2. En effet, Delattre démontre que la position fréquentielle des deux premiers formants est suffisante pour caractériser chaque voyelle du point de vue de son timbre.
On présente les formants F_1-F_2 de façon à ce qu'il y ait une relation entre F_1 et la hauteur de la langue, et également entre F_2 et l'avancement de la langue dans le conduit vocal. L'organisation des voyelles dans ce triangle acoustique rappelle celle des voyelles sur le triangle articulatoire de la phonétique classique. Ce triangle articulatoire représente très grossièrement la position moyenne de la langue dans la cavité buccale selon deux axes nommés « antérieur-postérieur » (ou « avant-arrière ») et « ouvert-fermé », selon que la langue est massée en avant et vers la zone dentale pour [i], basse et étalée loin du palais pour [a] (ouvert) ou massée postérieurement vers le voile pour [u].
Ceci renvoie à l'étude réalisée préalablement au chapitre 5.

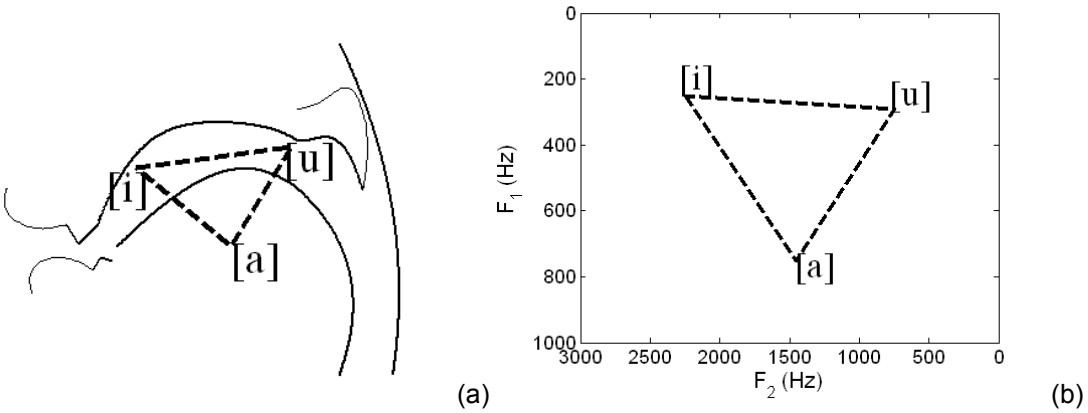


Figure 76 : Relation articulatoire-acoustique définie à partir (a) de la position de la langue (sans échelle) et (b) des formants.

Ce triangle entend représenter graphiquement les caractéristiques articulatoires des voyelles cardinales, à partir desquelles il est possible de définir tout autre type de voyelles. La pointe

du triangle présente la voyelle ouverte [a], alors que sa base présente les voyelles fermées [i] et [u]. Entre les extrémités du triangle se situent les autres voyelles.
On peut interpréter une augmentation de F_1 comme le résultat d'une ouverture articulatoire et une augmentation de F_2 comme une antériorisation de l'articulation.
Quant au F_3, non pris en compte dans cette représentation, il est corrélé avec la configuration des lèvres pour les voyelles antérieures.

Le modèle linéaire que nous allons proposer a essentiellement pour vocation de restaurer cette relation entre position de la langue et des lèvres et position des formants.

Les fréquences de formants ont été mesurées à partir du signal audio original à l'aide du logiciel *Praat*. *Praat* est un logiciel d'analyse et de transcription phonétique (spectre, intonation, intensité etc.). Le logiciel comporte aussi des fonctionnalités importantes pour l'enregistrement, pour la manipulation et pour la synthèse de sons, pour la création d'algorithmes d'apprentissage, pour l'analyse statistique, ainsi que pour diverses expériences auditives. Le logiciel *Praat* a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam [BW05]. Parmi toutes les fonctionnalités de ce logiciel, nous nous sommes limités à effectuer des analyses phonétiques et acoustiques de base (spectrogramme, analyse de formants, courbe de F0…), et en particulier, nous avons utilisé *Praat* pour extraire les valeurs des trois premiers formants (F_1, F_2, F_3) du signal audio Wioland, préalablement traité avec le logiciel *Adobe Audition* pour réduire le bruit.
Les trajectoires formantiques sont extraites toutes les 20 ms, une décimation d'ordre 2 permet d'obtenir une mesure toutes les 40 ms et ainsi d'associer à chaque configuration géométrique de la séquence un jeu de fréquences formantiques. Nous parlerons de formants mesurés ou formants du signal original.

1.2. Estimation acoustique linéaire

Un modèle statistique linéaire est construit pour permettre de prédire des données audio (fréquences de formants) à partir de données vidéos (degrés de liberté décrivant la forme du conduit vocal et ses mouvements). Notons qu'il serait envisageable de construire un modèle inverse, pour prédire la vidéo à partir de l'audio.
La transformation linéaire T_{yx} de passage des données vidéos Y aux données audio X est estimée à partir d'une base d'apprentissage correspondant aux trames de parole.
La distinction parole-silence a été effectuée, comme précédemment décrit, en tenant compte de l'énergie dans la première sous-bande, à partir d'un banc de 4 filtres (voir figure 65). Sur

les 4200 trames, plus de la moitié sont des trames de silence, nous ne conservons pour la base d'apprentissage que les quelques 2000 restantes.
Les données audio sont filtrées temporellement avant d'être prises en compte, pour réduire la variabilité du bruit de mesure. Le filtre considéré est un filtre passe-bas de Butterworth d'ordre 4 de fréquence de coupure 2,5 Hz.
Le calcul de la matrice de passage T_{yx} nécessite l'évaluation des moyennes des jeux de données. Ces moyennes μ_x et μ_y sont calculées sur la séquence complète (silence inclus).
L'expression de T_{yx} est donnée par la formule suivante :

$$T_{yx} = (X - \mu_x)(Y - \mu_y)^T \left((Y - \mu_y)(Y - \mu_y)^T\right)^{-1}$$

A partir de cette matrice de passage, il est alors possible d'estimer ou de prédire des paramètres audio associés à une nouvelle configuration géométrique, par une régression multilinéaire.

$$\tilde{X} = T_{yx}(Y - \mu_y) + \mu_x$$

Pour l'estimation des paramètres prédits $\tilde{X}$, on utilise une base de test.
Nous avons procédé par une technique de Jackknife pour l'obtention des bases d'apprentissage et de test, sur les trames sélectionnées.

Pour prédire les 2 premiers formants, nous prenons en compte, dans les données vidéos, uniquement les degrés de liberté correspondant à la langue, en accord avec l'idée que F_1 est associé à la hauteur de la langue et F_2 à sa rétraction (position avant/arrière).
Pour prédire le troisième formant, le précédent jeu de données géométriques est complété avec les 10 degrés de liberté des lèvres (6 ddl) et des dents (4 ddl).

1.3. Comparaison des formants d'origine et des formants estimés

1.3.1. Corrélations

L'objectif de la corrélation est de quantifier la relation entre deux mesures, i.e. de mesurer à quel point deux mesures varient simultanément. Une mesure de cette corrélation est obtenue par le calcul du coefficient de corrélation linéaire. Ce coefficient est égal au rapport de leur covariance et du produit non nul de leurs écarts types.
Des analyses de corrélation ont été menées systématiquement dans Yehia et al. [YRV98] entre les paramètres LSP mesurés et ceux estimés linéairement. Par exemple, pour les paramètres estimés à partir du conduit vocal, les coefficients de corrélation obtenus varient entre 0,7 et 0,84.

Nous présentons dans le tableau 10 les coefficients de corrélation que nous obtenons, sur la séquence privée des silences, entre les formants du signal d'origine (formants de référence) et les formants prédits.

$$c = \frac{\sqrt{\sum_{p=1}^{N} \widetilde{F}_i(p) * F_i(p)}}{\sqrt{\sum_{p=1}^{N} F_i(p)^2} * \sqrt{\sum_{p=1}^{N} \widetilde{F}_i(p)^2}}$$

où $F_i(p)$ est la valeur de référence du formant i à la trame p et $\widetilde{F}_i(p)$ la valeur prédite.

Une séquence de formants, d'origine et estimés, est observable sur la figure 77. L'estimation de F_1 et F_2 à partir des données géométriques de langue donne une corrélation de 0,7. Celle de F_3 est à peine de supérieure à 0,5 en prenant en compte la langue et les lèvres. La moyenne de ces valeurs signifie que 64% de la variance totale observée sur les formants peut être récupérée à partir des mouvements de la langue et des lèvres.
Les corrélations de 0,7 pour F_1 et F_2 sont meilleures que les 55% qu'obtenaient Barker et Berthommier [BB99b] avec les données labiales seules.

	Langue	**Langue et lèvres**
F_1	0,7	0,7
F_2	0,69	0,7
F_3	0,47	0,52

Table 10 : Corrélations entre formants d'origine et formants prédits pour la séquence (hors moments de silence) en fonction des degrés de liberté pris en compte.

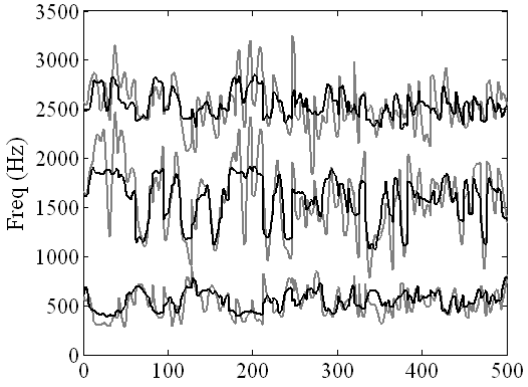


Figure 77 : Séquence de formants F_1, F_2, F_3 estimés à partir du modèle linéaire (en noir) superposés aux formants mesurés depuis le signal d'origine et filtrés (en gris).

1.3.2. Espaces formantiques

La superposition des formants estimés et des formants mesurés est aussi réalisée (Fig. 78) dans un plan F_1-F_2 et dans un plan F_2-F_3, qui sont des espaces de représentation classiques,

dont nous reparlerons plus loin. L'observation de ces espaces formantiques sur les figures qui suivent mettent en évidence un centrage des données estimées par rapport aux données d'origine. Cet effet est inhérent à l'usage d'un associateur linéaire. Il y a une perte de dynamique au cours de l'estimation linéaire. En superposant un triangle vocalique dans le plan F_1-F_2, on constate notamment que les voyelles extrêmes [i], [a] et [u] sont nettement moins bien représentées par les données estimées que par les données mesurées. Acoustiquement, l'ensemble des sons produits devraient être perçus comme des voyelles centrales de type [ø] ou [ɛ], comme on pourra le montrer avec un test de perception un peu plus loin.

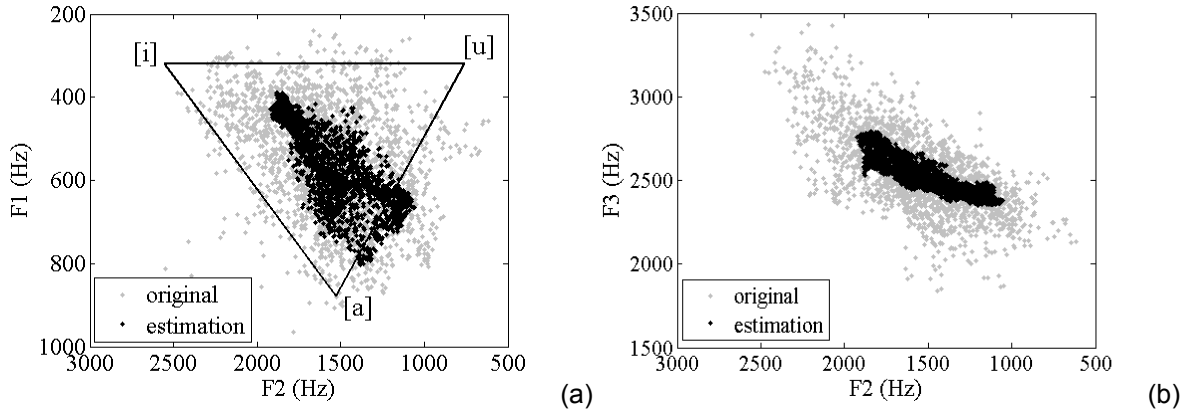


Figure 78 : Espaces (a) F_1-F_2 et (b) F_2-F_3 obtenus à partir données mesurées (en gris) et données estimées (en noir).

Si la corrélation temporelle des estimations avec l'original est assez élevée (comparable à celle observée par Yehia et al. [YRV98]), cela n'implique pas nécessairement que les estimations soient phonétiquement correctes.

1.3.3. Mesures de biais

Considérant les formants extraits du signal d'origine comme référence, on peut calculer le biais d'estimation, c'est-à-dire la différence moyenne μ entre les formants estimés et la référence. Cette mesure est normalisée par rapport aux valeurs d'origine, elle est décrite par la formule suivante :

$$\mu(F_i) = \frac{1}{N}\sum_{p=1}^{N}\frac{\left|\widetilde{F}_i(p) - F_i(p)\right|}{F_i(p)}$$

où $F_i(p)$ est la valeur de référence du formant i à la trame p et $\widetilde{F}_i(p)$ la valeur prédite.

La mesure effectuée se limite aux trames associées aux instants vocaliques. A partir du signal audio, nous avons étiqueté 385 voyelles du corpus. Le biais entre les formants

estimés et les formants d'origine pour ces trames est évalué à 12% pour le premier formant et 19% pour le second. Ces écarts ont des conséquences d'un point de vue perceptif.

1.4. Interaction production-perception

L'interaction production/perception (locuteur/auditeur) est supposée être basée sur deux principes majeurs qui sont :

- La simplicité articulatoire (appelée aussi critère du moindre effort ou principe de l'économie articulatoire),
- La distinctivité perceptive (principe du contraste).

Dans le processus de communication parlée, des moyens sont donc a priori mis en œuvre par le locuteur pour, d'une part, minimiser l'effort articulatoire et pour d'autre part, maintenir un contraste perceptif suffisant entre les phonèmes. Ce contraste est nécessaire si l'on souhaite éviter toute confusion dans l'interprétation.
Deux théories complémentaires ont alimenté le débat autour de ces hypothèses dans les années 1970. Il s'agit de la Théorie Quantique proposée par Stevens [Ste72] et de la Théorie de la Dispersion énoncée par Liljencrants et Lindblom [LL72].

Le postulat de base de la théorie quantique [Ste72, Ste89] est la non-linéarité du passage de l'articulatoire vers l'acoustique. On montre que pour certaines positions des articulateurs, un petit changement entraîne peu de modifications sur la perception acoustique Par contre, pour d'autres positions, un même changement provoque des modifications énormes. Ceci a amené Stevens à concevoir l'espace des possibilités articulatoires comme composé d'attracteurs (des points d'ancrage stables) et de régions de transition rapide. Selon lui, les traits distinctifs des phonèmes peuvent être prédits et expliqués par les positions de ces attracteurs et régions de transition. Un phonème est d'autant plus fréquent dans les langues du monde qu'il est stable. Stevens considère qu'un phonème acoustiquement stable possède une forme perceptive plus riche et est, par conséquent, plus aisément discriminable.

La théorie de la Dispersion cherche à expliquer la structure phonologique des systèmes vocaliques. Etant donné que ces systèmes ne présentent qu'un ensemble limité de combinaisons possibles de voyelles, Liljencrants et Lindblom [LL72] ont étudié les systèmes vocaliques d'un certain nombre de langues afin de prédire leur structure phonologique, et ont analysé la manière dont les voyelles sont distribuées dans les différents systèmes. Les typologies sur les systèmes de sons des langues naturelles montrent qu'il existe des voyelles universellement favorisées. Le postulat de la théorie est que les sons favorisés sont ceux qui maintiennent entre eux, à l'intérieur des systèmes, une distance permettant une

discrimination. A la différence de la théorie quantique, les propriétés du contraste sont déterminées par les relations que les voyelles entretiennent entre elles à l'intérieur du système, et non pas par les caractéristiques acoustiques et articulatoires intrinsèques à chacune d'elles. Le critère de dispersion vise à maximiser les distances perceptives entre les voyelles et constitue le principe du contraste maximal qui permet de générer, par simulation, un système où les voyelles s'organisent pour présenter un maximum de distinction globale.

Les régions périphériques ou extrêmes du triangle vocalique sont donc perceptuellement importantes du fait de cette préférence linguistique pour les voyelles extrêmes (voyelles dans les coins du triangle vocalique). Ces voyelles sont les plus présentes dans les langues du monde. La théorie Quantique de Stevens montre que ces voyelles sont les plus stables en terme de production. La théorie de la dispersion prend en considération le rôle de la perception et montre que les voyelles qui ont le plus de chances d'être discriminées par rapport aux autres sont celles qui sont le plus distinctes, soit les voyelles extrêmes.
Partant de cette idée de privilégier ces voyelles périphériques, nous proposons une méthode globale, indépendante des classes vocaliques, pour augmenter la couverture de l'espace vocalique par les données estimées. Cette méthode tend à améliorer l'estimation des formants estimés sans aucun a priori phonétique.

2. Transformation affine globale

Nous proposons de corriger les estimations formantiques obtenues par le modèle linéaire, en appliquant une transformation affine globale qui a pour but de disperser ces estimations et de les recadrer dans l'espace maximal. Il s'agit d'une contrainte phonétique susceptible de dégrader les mesures d'erreurs et de corrélation.
La méthode que nous proposons est une méthode statistique qui consiste en une transformation affine d'ajustement de contours, déduits à partir des distributions des formants mesurés et estimés.

2.1. Transformation sur F_1 et F_2

Expliquons le principe de la méthode avec les formants F_1 et F_2.
A partir des distributions des données estimées et d'origine dans l'espace formantique F_1-F_2, il est possible d'établir les contours de ces distributions. Les données sont préalablement traitées et normalisées (Fig. 79a), de telle sorte que la moyenne soit égale à 0 et l'écart-type à 1. Les contours des espaces F_1-F_2 normalisés sont obtenus à partir des histogrammes de

distribution (Fig. 79b et 79c) avec un seuil de 5%, c'est-à-dire que chaque contour est établi de façon à contenir 95% des données qu'il représente.

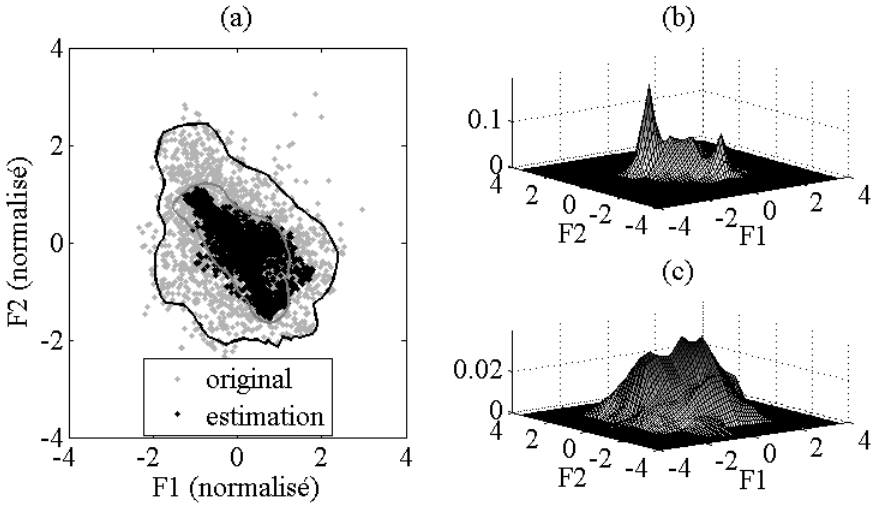


Figure 79 : A partir des distributions dans l'espace formantique F_1-F_2, 95% de chacun des deux jeux de données sont conservés pour établir les contours.
(a) Données acoustiques normalisées.
(b) Histogramme de distribution des données acoustiques estimées.
(c) Histogramme de distribution des données acoustiques d'origine.

On recherche ensuite la transformation affine, représentée par la matrice M et définie avec un paramètre de rotation θ et 2 paramètres de dilatation s_x et s_y, telle que le contour des données estimées ainsi transformé soit le plus proche du contour des données mesurées, au sens des moindres carrés. Il s'agit d'une régression linéaire appliquée sur les points du contour et répétée jusqu'à ce qu'il y ait convergence. Les paramètres θ, s_x et s_y sont initialisés respectivement à 1, 1 et 0. L'algorithme consiste ensuite à faire varier ces paramètres pour minimiser le critère d'erreur. Soit $x(f_1,f_2)$ le contour obtenu à partir des formants estimés et $y(f_1,f_2)$ celui des formants d'origine, où f_1 et f_2 sont les fréquences normalisées des formants, le critère d'erreur à minimiser est :

$$err = \sum_{f_1}\sum_{f_2}(y(f_1,f_2) - M * x(f_1,f_2))^2 \text{ , avec } M = \begin{bmatrix} s_x\cos(\theta) & -s_x\sin(\theta) \\ s_y\sin(\theta) & s_y\cos(\theta) \end{bmatrix}.$$

Une fois le critère minimisé, à partir des valeurs finales des paramètres θ, s_x et s_y introduits dans la matrice M, le nouveau contour de l'estimation est donné par $z(f_1,f_2)=M* x(f_1,f_2)$, il est représenté par le contour en pointillé sur le graphe 80.

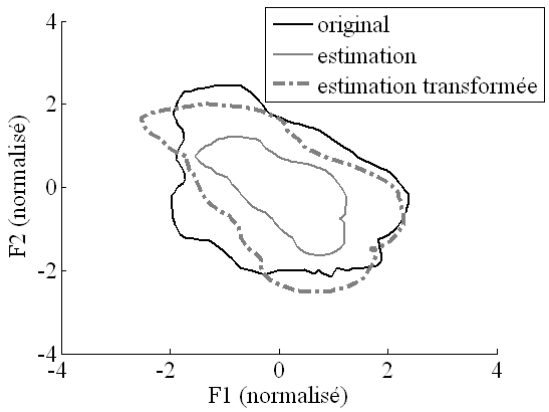


Figure 80 : Le contour (en trait gris plein) pour les paramètres audio estimés est dilaté et orienté (grâce à la transformation M) pour s'ajuster au contour des paramètres audio mesurés (en trait noir plein). Le contour obtenu est représenté en pointillés.

La transformation M est ensuite appliquée directement aux paramètres audio estimés F_1 et F_2 et la nouvelle estimation de formants ainsi obtenue est observable sur la figure 81, où elle est superposée aux données de départ. On constate que la perte de dynamique a été en partie compensée et que la couverture de l'espace par les formants nouvellement estimés se rapproche de celle des données mesurées.

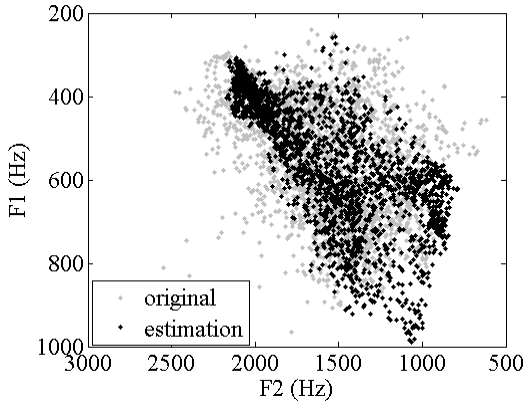


Figure 81 : Espace F_1-F_2 après transformation affine. La couverture de l'espace formantique est meilleure.

2.2. Transformation sur F_3

Une fois F_1 et F_2 ainsi traités avec la transformation affine, un traitement analogue est effectué sur le troisième formant. Comme précédemment, on recherche la transformation appropriée pour ajuster au mieux les contours des espaces F_2-F_3 estimés et mesurés. Le contour de l'espace F_2-F_3 pour les estimations prend en compte l'estimation de F_3 par le modèle linéaire appliqué avec les degrés de liberté de la langue et des lèvres et l'estimation transformée de F_2, qui vient d'être calculée avec la transformation affine décrite juste avant. Le seuil pris en compte pour l'évaluation des contours est cette fois de 10% (et non plus de

5%). La transformation affine alors obtenue sur les contours est appliquée aux formants estimés F_3.

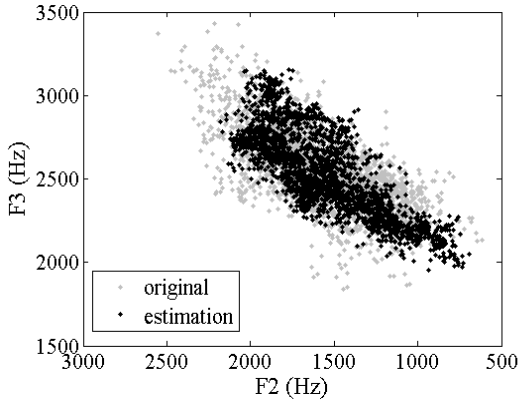


Figure 82 : Espace F_2-F_3 après transformation affine.

2.3. Apport de la transformation

A partir des formants, on re-synthétise des signaux et en particulier des voyelles. Pour cela, nous utilisons un filtrage inverse par LPC (Linear Predictive Coding) sur un seul segment voisé du signal d'origine de façon à estimer le signal de source. En effet, un signal de parole peut être vu comme un signal source filtré par une fonction de transfert représentative des résonances (ou formants) du conduit vocal. Nous ne nous attarderons pas ici sur cet aspect, qui sera abordé plus largement au chapitre de synthèse articulatoire.

L'étude va se limiter ici aux voyelles du corpus de Wioland et nous observons (Fig. 83) quelques exemples de voyelles stationnaires synthétisées à partir des différents jeux de formants.

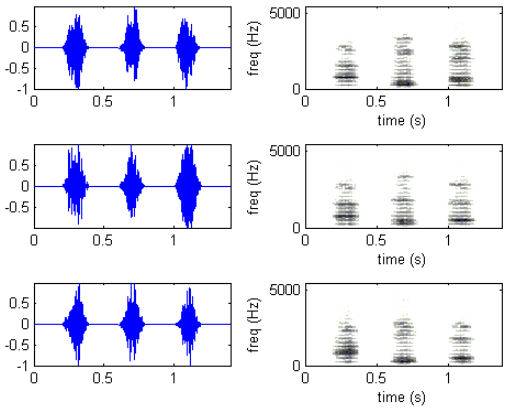


Figure 83 : Signaux synthétisés et spectrogrammes ([a], [i], [ɛ], de gauche à droite), pour les formants mesurés (en haut), les formants estimés (au milieu), et les formants estimés puis « transformés » (en bas).

A ce stade, nous disposons de plusieurs jeux de formants : les formants mesurés depuis le signal audio, les formants prédits par le modèle linéaire et les formants prédits puis traités

par la transformation affine. Le calcul de la corrélation entre ces derniers formants et les formants mesurés montre des résultats (0,67 pour F_1 et 0,68 pour F_2) à peine plus faibles que pour les formants estimés du modèle linéaire. Si l'on examine la répartition des erreurs en fonction de la position dans l'espace vocalique, on constate que les erreurs d'estimation sont maintenant plus réduites dans les zones périphériques des espaces formantiques, au détriment des voyelles centrales.
Afin d'évaluer la correction phonétique, nous réalisons un test de perception.

3. Test de perception

3.1. Préparation

Nous avons établi un corpus réduit de voyelles pour ce test de perception, à partir de voyelles présentes dans le corpus Wioland. Nous avons sélectionné, à partir du signal audio de la base de données et des valeurs de formants extraits, 90 trames correspondant à 9 voyelles du français ([a], [i], [e], [u], [y], [o], [ɛ], [ø], [ɔ]). Chaque classe vocalique est représentée par 10 items différents.
Pour chaque trame, nous disposons des valeurs de fréquences formantiques associées qui nous permettent de synthétiser les différents types de stimuli.

Le test de perception est conçu à partir de 4 types de signaux synthétisés, notés S_0 à S_3, pour les 90 items de voyelles. La synthèse est réalisée pour tous les stimuli, à partir d'une même source. Cette source a été obtenue, comme décrit au paragraphe 2.3. : un segment unique du signal de départ, voisé et de durée 200 ms, a été filtré par un filtre inverse LPC d'ordre 20. Ce signal ainsi blanchi est utilisé comme source pour tous les stimuli.

- S_0 est obtenu en filtrant la source par le spectre LPC (prédiction linéaire) d'ordre 20 du signal d'origine.
- S_1, S_2 et S_3 sont obtenus en convoluant la source par une somme de sinusoïdes aux fréquences des formants correspondants :
    - S_1 : F1, F_2 et F_3 mesurés à partir du signal audio d'origine
    - S_2 : F_1, F_2 et F_3 estimés avec le modèle linéaire
    - S_3 : F_1, F_2 et F_3 estimés avec le modèle linéaire et traités avec les transformations affines.

Chaque stimulus a une durée de 200 ms.

La tâche demandée aux sujets est une tâche de catégorisation ; la consigne consiste à choisir la voyelle qu'ils entendent parmi [a], [i], [e], [u], [y], [o], [ɛ], [ø], [ɔ] ou [ ] s'ils ne

distinguent pas du tout la voyelle. Les sujets peuvent écouter chaque stimulus autant de fois que nécessaire.

Nous présentons ici les résultats moyens de 6 sujets français.

3.2. Résultats

A partir des réponses des sujets, nous construisons des tables de confusion (Fig. 84). Chaque ligne d'une matrice correspond aux classes vocaliques des stimuli, chaque colonne correspond aux classes vocaliques perçues. La matrice est exprimée en pourcentage de reconnaissance des voyelles pour chaque catégorie vocalique. La moyenne des éléments de la diagonale est appelée taux de reconnaissance.

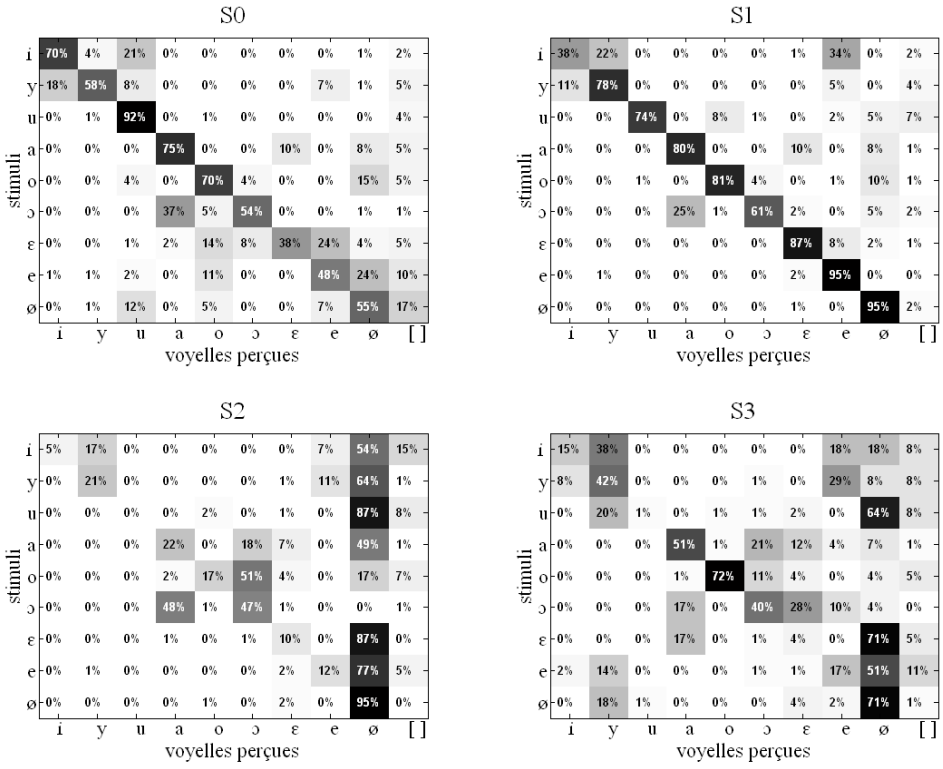


Figure 84 : Matrices de confusion obtenues à partir des réponses des sujets pour les 4 modèles de signaux testés.

L'analyse de ces matrices est complexe ; aussi, dans un premier temps, nous faisons simplement la distinction entre voyelles centrales ([e], [o], [ɛ], [ø], [ɔ]) et voyelles périphériques ([a], [i], [u], [y]) et nous calculons le taux de perception correcte de ces voyelles périphériques ou centrales. Nous évaluons le taux de catégorisation des voyelles entre le centre et la périphérie et comparons ce résultat (Fig. 85a) pour les 4 types de signaux synthétisés.

Les voyelles centrales et périphériques sont bien discernées pour les stimuli S_0 et S_1 (près de 85% de bonne reconnaissance). Avec les formants estimés à partir du modèle linéaire (S_2), ce taux tombe à 55% et il remonte à 65 % avec les formants estimés puis soumis aux transformations affines (S_3). La différence entre les taux obtenus pour S_2 et S_3 est significative (test de Student, p<0.03%).

Des tables de confusion centre/périphérie sont présentées figure 85b et montrent la distribution de l'erreur. La somme des lignes ne vaut pas 100%, ceci est dû au fait que nous ne prenons pas compte le cas des non-réponses par les sujets (dernière colonne des matrices de la figure 84). Pour S_2, la plupart des voyelles sont perçues comme centrales. Après la transformation (stimuli S_3), les voyelles périphériques sont mieux perçues avec un taux de bonne reconnaissance passant de 14% à 40%, sans affaiblir notablement la perception des voyelles centrales.

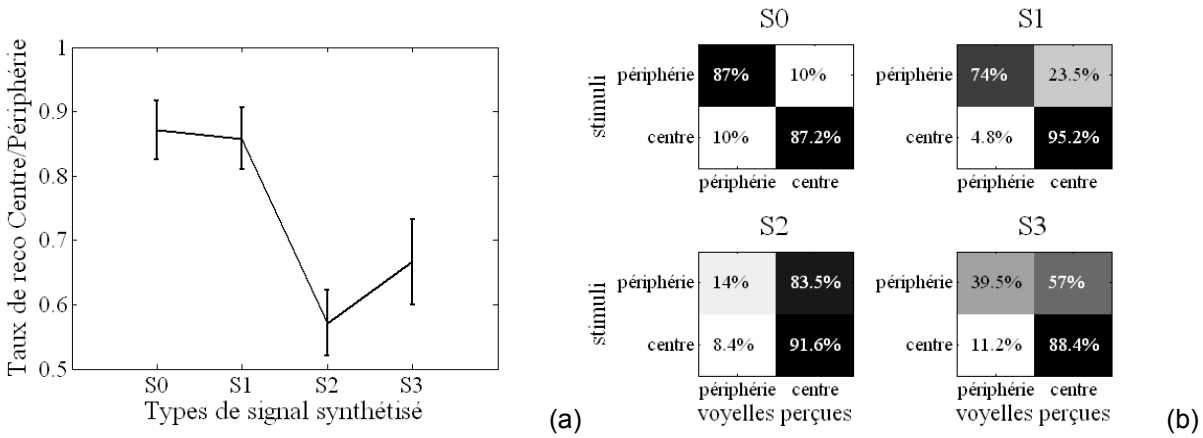


Figure 85 : (a) Résultats du test en terme de taux de reconnaissance entre voyelles périphériques et voyelles centrales.
(b) Tables de confusions entre voyelles centrales et périphériques.

Une étude plus fine présente ensuite la perception des catégories vocaliques en fonction de la catégorisation des stimuli. La figure 86 consiste en deux séries de triangles vocaliques. La première ligne correspond aux stimuli produits, la seconde aux réponses des sujets et donc à la catégorisation perçue des stimuli. Sur la première ligne, on a représenté dans un plan F_1-F_2 les ellipses de dispersion correspondant aux 9 classes vocaliques en jeu dans le test de perception. Pour chaque classe, à partir des 10 items, les ellipses à un écart-type sont centrées sur les valeurs moyennes des formants de la classe, pour les formants d'origine (à gauche), pour les formants estimés (au milieu) et pour ceux obtenus après transformation affine (à droite). La seconde ligne met en évidence les résultats du test. Pour chaque stimulus entendu, la paire correspondante de formants F_1-F_2 produits est « étiquetée » avec la classe vocalique de la réponse donnée par le sujet. Pour cette seconde ligne du graphique, chaque classe vocalique se compose des exemplaires perçus comme appartenant à cette classe et positionnés dans le plan F_1-F_2 aux valeurs formantiques des

stimuli. Les ellipses de dispersion représentées (à un écart-type) mettent en évidence la distribution des exemplaires de chacune de ces classes perçues.

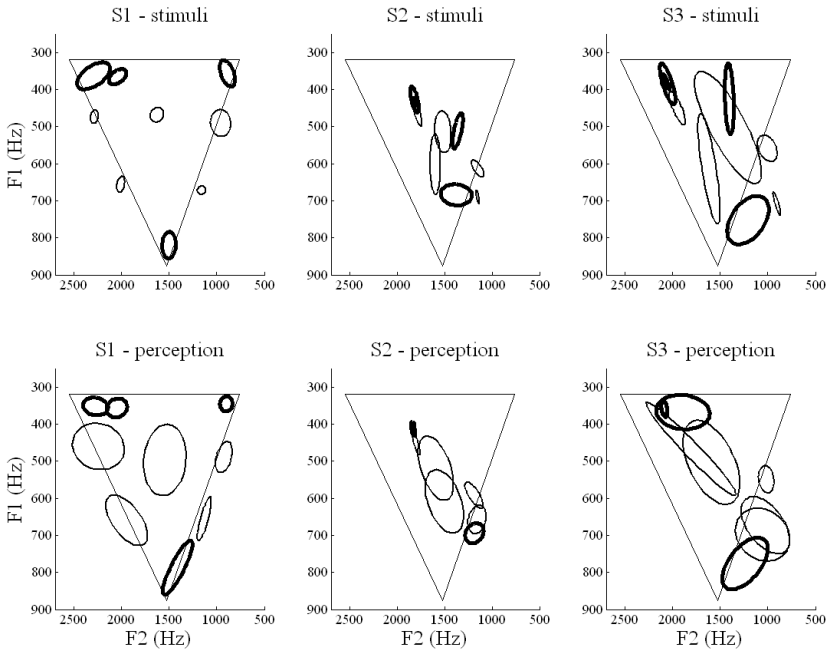


Figure 86 : Comparaisons sur des triangles vocaliques entre stimuli et perception pour les différentes catégories de voyelles (10 items pour chaque) : les voyelles périphériques ([a], [i], [u], [y]) sont en trait gras.

La perception des voyelles suit la position des formants des stimuli. Les voyelles estimées à partir du modèle linéaire sont centrées et les sujets ne sont pas capables de percevoir correctement les voyelles périphériques. Il y a un recouvrement des catégories vocaliques au centre du triangle vocalique. La transformation affine, comme post-traitement, permet de dilater l'espace des formants, les catégories de voyelles couvrent mieux le triangle. Les classes vocaliques extrêmes sont en partie repositionnées à des valeurs classiques, mais 57% des voyelles périphériques sont encore perçues comme centrales après la transformation affine. [u] n'est pas du tout perçue alors que [a] est perceptuellement discriminée à 51%. Si le [i] et le [y] sont mieux discriminées par rapport aux voyelles centrales après la transformation affine (S3), les confusions sont importantes entre ces 2 voyelles. Ce traitement postérieur rétablit en partie la position mais ne corrige pas le recouvrement. Le taux de reconnaissance en prenant en compte chaque catégorie de voyelles, et non plus centre vs périphérie, est faible (moyenne de 35% pour le modèle S_3).

Partant des configurations géométriques de la langue et des lèvres de la séquence Wioland, nous avons estimé, par le biais d'un modèle linéaire, la fréquence des formants associés.

Les résultats obtenus montre que le modèle linéaire n'est pas suffisant. La plupart des voyelles périphériques ne sont pas correctement perçues. Notre tentative de post-traitement basée sur une approche globale de dispersion maximale des distributions, permet en appliquant une transformation affine aux estimations de formants d'améliorer la distinction des voyelles périphériques, mais seulement dans une certaine mesure. Les performances ne sont pas pleinement satisfaisantes, les confusions entre catégories de voyelles sont encore importantes.

4. Approche linéaire sur éléments sélectionnés

L'approche linéaire qui vient d'être présentée est globale : aussi bien l'association linéaire que la transformation affine sont appliquées sur le jeu entier de formants de la séquence, sans considération des classes vocaliques. Ce n'est que pour le test de perception que nous prenons en compte cette catégorisation.

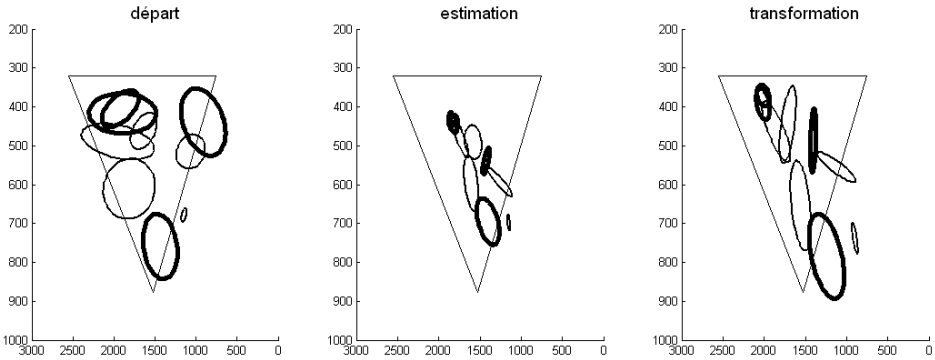
La démarche qui suit consiste en une approche limitée à une sélection de données pour lesquelles la relation articulatoire-acoustique est a priori correcte. La transformation linéaire T_{yx} est calculée sur une base d'apprentissage constituée seulement de 385 trames correspondant aux voyelles du corpus. Cette transformation est ensuite appliquée sur la base de test comprenant l'ensemble des trames de parole de la séquence. La mesure de biais entre les formants d'origine et les formants estimés est effectuée sur les 385 trames de voyelles, le biais est évalué à 11% pour F_1 et 17% pour F_2, il est donc un peu moins élevé que pour l'approche globale. La corrélation entre les formants estimés et les formants de départ est de 0,6 pour F_1 et 0,54 pour F_2.

Comme précédemment pour l'approche globale, une transformation affine est élaborée puis appliquée aux données pour tenter de mieux couvrir l'espace vocalique. Cette transformation est basée sur le même principe que celui, décrit au §2., qui s'appuie sur les contours des distributions des formants d'origine et des formants estimés dans les plans F_1-F_2 et F_2-F_3. Les transformations sont ici calculées à partir des contours correspondants aux éléments sélectionnés uniquement. Les matrices de transformation affine (rotation, dilatation) alors définies sont appliquées à ces seuls éléments sélectionnés.

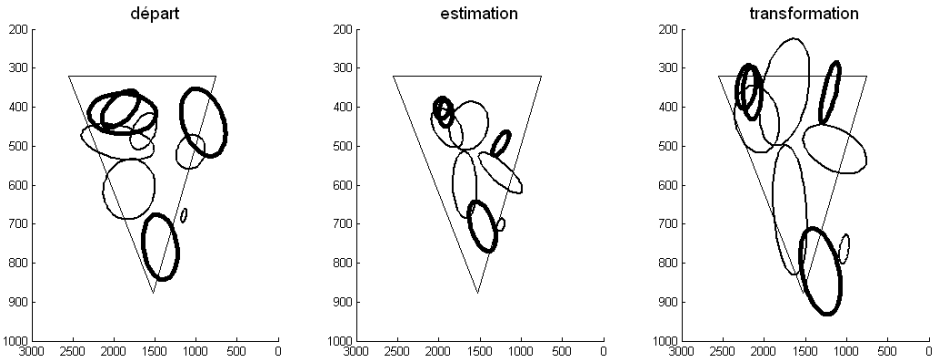
La figure 87 compare l'approche globale (Fig. 87a) et l'approche par éléments sélectionnés (Fig. 87b) en se basant sur les ellipses de dispersion des catégories vocaliques obtenues à partir des valeurs formantiques des 385 éléments sélectionnés.

De même qu'à la première ligne de la figure 86, nous représentons dans le plan F_1-F_2 les ellipses obtenues à partir des formants des exemplaires de chaque classe vocalique, du signal de départ (à gauche), estimées (au milieu) et après transformation affine (à droite).

Avec les formants d'origine, on constate déjà un certain recouvrement des ellipses. Pour les formants estimés par le modèle linéaire, l'approche par sélection présente un centrage moins grand que l'approche globale. Après transformation affine, on observe que pour l'approche par sélection, les ellipses couvrent mieux le triangle vocalique et notamment l'ellipse du [u] se place mieux dans le coin en haut à droite. Mais les ellipses sont très étendues, la dilatation est trop importante, elle surcompense l'effet de centrage plus faible et la perception des voyelles n'est pas améliorée.



(a)



(b)

Figure 87 : Classes vocaliques dans l'espace F_1-F_2 pour les données d'origine, les données estimées par modèle linéaire et les données estimées puis soumises à transformation affine. (a) méthode globale – (b) méthode par classe.

Que ce soit de façon globale ou sur une sélection d'éléments, le modèle linéaire ne suffit pas pour prédire correctement l'acoustique à partir des données géométriques. Nous supposons qu'un modèle linéaire appliqué directement aux degrés de liberté géométriques n'est pas suffisant, et que d'autres représentations et d'autres transformations de ces données sont nécessaires.

Nous avons alors orienté nos travaux vers le développement de synthèse articulatoire à partir des configurations acquises depuis les images cinéradiographiques. Cette approche utilise les représentations et les transformations physiques liées à la production de la parole.

CHAPITRE 7 : SYNTHÈSE ARTICULATOIRE

Comme mentionné au chapitre 4 avec la fonction d'aire, pour avoir un modèle complet capable de produire un signal de parole, le dernier module nécessaire est un modèle acoustique. Ce modèle va permettre de passer de la fonction d'aire à des paramètres acoustiques, et principalement aux formants. Nous utiliserons la méthode consistant à obtenir un analogue électrique en ligne de transmission et à en calculer la fonction de transfert [BF84]. Le conduit vocal est vu alors comme une concaténation de tubes uniformes assimilés chacun à un quadripôle électrique et à une fonction de transfert. La fonction de transfert totale présente des pôles qui correspondent aux formants produits par le conduit vocal.

L'étude qui suit se décompose en 2 parties. Une première s'intéresse aux formants ; elle consiste à estimer les formants produits par les configurations géométriques du conduit vocal, extraites grâce à la méthode d'extraction semi-automatique, et à les comparer avec les formants du signal d'origine.

Une seconde partie introduit une comparaison au niveau du signal complet. Elle s'attache à comparer spectralement le signal estimé synthétisé avec le signal d'origine, à partir d'une mesure de spectres LPC. Nous essayons de combiner nos estimations avec la partie non modélisable du signal afin de réaliser une resynthèse d'un signal de parole. Dans la décomposition source-filtre mise en place, la source est estimée à partir du signal audio d'origine et décomposée en 2 éléments (une source blanchie et une modulation d'amplitude en deux sous-bandes). Un test de perception met en évidence l'intelligibilité des signaux synthétisés.

1. Données

Nous présentons les données d'origine, qui serviront de référence. Il s'agit du signal audio Laval43 et des formants extraits de ce signal. Nous détaillons également dans ce paragraphe le modèle permettant le passage de la fonction d'aire à la fonction de transfert représentative du conduit vocal.

1.1. Formants et signal audio de référence

En complément des images de la séquence Laval43, nous disposons du signal audio correspondant (format .wav). Le signal audio enregistré au cours de la séance de rayons X est très bruité. Un traitement préliminaire a été réalisé sur ce signal de façon à réduire le

bruit. Ce traitement a été réalisé à l'aide du logiciel Adobe Audition et a consisté en une soustraction spectrale. Ce pré-traitement a bien diminué le bruit mais il a introduit une sorte d'écho extrêmement rapide ou d'effet « cathédrale », moins gênant cependant que le bruit initial. Ce signal échantillonné au départ à 44100Hz a été sous-échantillonné à 11025 Hz.

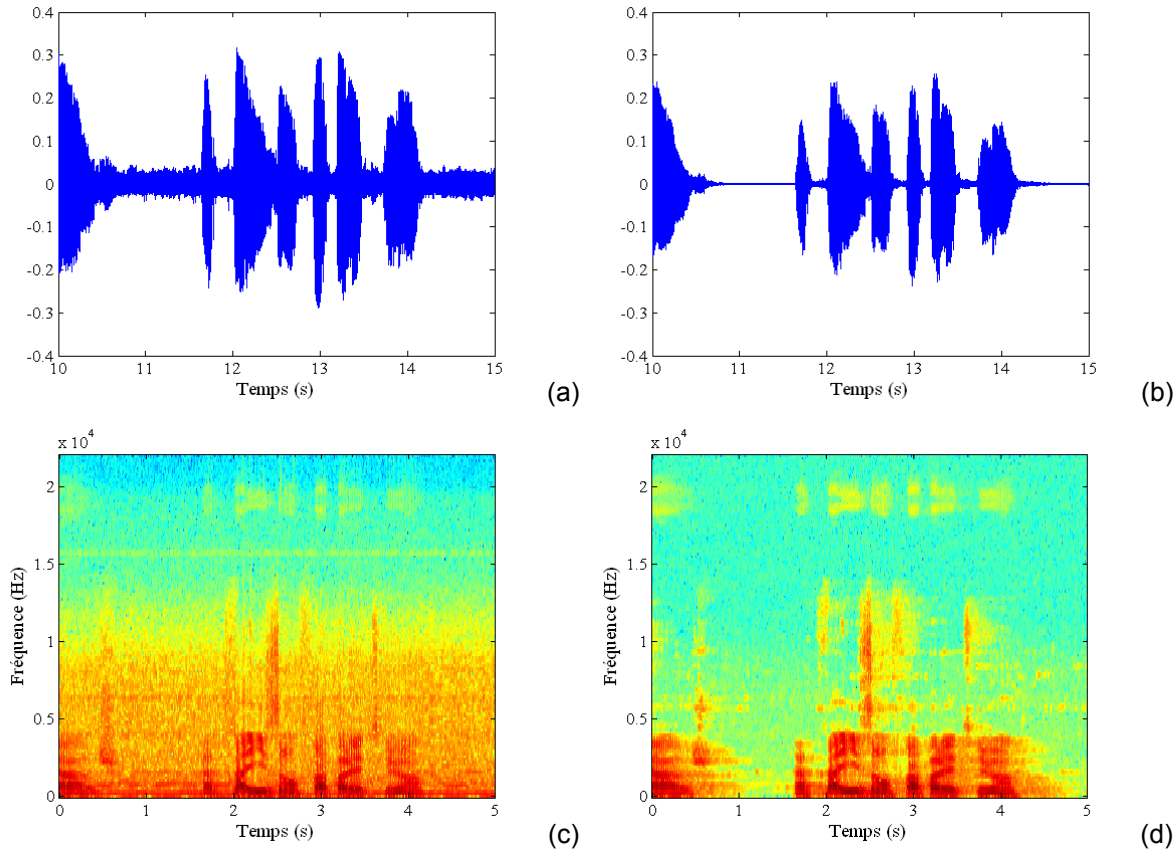


Figure 88 : (a) Signal audio Laval43 original - (b) Signal audio Laval43 après débruitage - (c) Spectrogramme du signal original – (d) Spectrogramme du signal après débruitage.

C'est dorénavant ce signal audio débruité et sous-échantillonné à 11,025 KHz que nous considérons comme signal original et qui nous servira de référence pour la suite.

Comme il l'a été mentionné précédemment, les formants sont des paramètres acoustiques qu'on utilise classiquement pour observer le contenu phonétique d'un corpus. Le F_1 est corrélé avec l'ouverture, le F_2 avec la position antérieure (F_2 élevé) et postérieure (F_2 bas) de la langue et F_3 avec la configuration des lèvres pour les voyelles antérieures.

Ce sont ces paramètres de fréquences formantiques que nous considérons pour relier l'articulatoire à l'acoustique.

Nous avons à nouveau utilisé le logiciel *Praat* pour extraire les valeurs des trois premiers formants (F_1, F_2, F_3) du signal audio Laval43. Ceci nous permet ainsi de disposer d'une valeur de fréquence de ces formants pour chaque image de la séquence. Ces valeurs seront

considérées comme les valeurs de référence, c'est-à-dire celles associées au signal d'origine. Nous parlerons par la suite de formants d'origine ou formants du signal original.

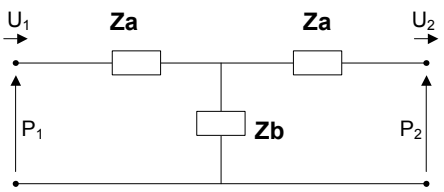
1.2. Estimation de la fonction de transfert du conduit vocal via la fonction d'aire et estimation de formants

Nous avons besoin d'un modèle acoustique pour passer de la fonction d'aire à l'acoustique. Comme mentionné en introduction, nous calculons la fonction de transfert du conduit vocal par l'implémentation d'un circuit analogue [BF84]. L'idée de cette méthode est de modéliser les résonances du conduit vocal. La transmission du son est modélisée en bénéficiant de l'analogie de comportement avec le courant électrique dans un circuit.

Cette analogie prend en compte un certain nombre d'hypothèses mécaniques comme la nature laminaire de l'écoulement et une propagation unidimensionnelle longitudinale (vérifiée jusqu'à 5kHz). Elles permettent de considérer que la forme du conduit vocal n'a pas d'influence et que seule compte l'aire de celui-ci, qui est variable suivant la région du conduit vocal, de la glotte aux lèvres. Ce modèle repose sur l'hypothèse que l'onde sonore qui se propage est parfaitement plane. Cette méthode constitue une assez bonne approximation dans les fréquences pas trop élevées (inférieures à 8 kHz) et les configurations du conduit vocal qui ne sont pas trop larges. En effet, dans les hautes fréquences où la longueur d'onde est très petite comme dans les configurations articulatoires où la section est large, on imagine qu'il peut se produire des résonances transversales, qui ne sont pas intégrées dans cette méthode de modélisation. Dans le cadre de notre étude, l'approximation classique des ondes planes est suffisante ; le but n'est pas d'obtenir un modèle physique mais des représentations et des transformations adéquates.

Le modèle de Badin et Fant [BF84] prend en compte les pertes visco-thermiques et les vibrations de parois, ce qui permet d'évaluer des largeurs de bandes associées aux formants. Ceci est accompli en introduisant dans l'analogie électrique des impédances pour chacune des pertes [Fla72].

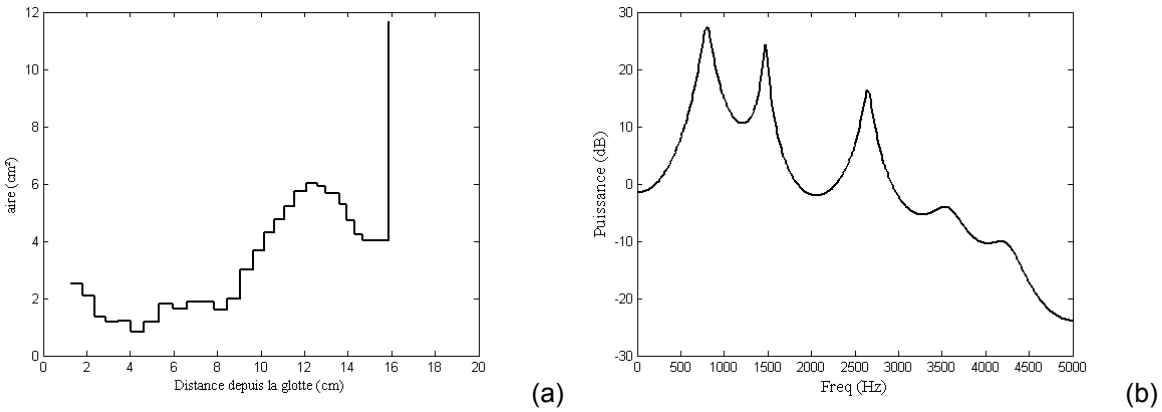
Un tube uniforme (de section constante) peut être assimilé à un quadripôle électrique. Aux grandeurs tension et courant correspondent respectivement pression (P) et débit d'air (U). Si les tubes élémentaires sont suffisamment courts, on fait une approximation en utilisant le quadripôle électrique en T suivant :



Les formules de Za et Zb sont complexes lorsque l'on modélise les pertes par vibration des parois et par viscosité de l'air (modélisation par des résistances ou conductances supplémentaires). Nous ne détaillerons pas ici. Notons simplement que Za et Zb dépendent de l'aire et de la section du tube. Chaque quadripôle est défini par sa fonction de transfert calculée en fonction des valeurs des impédances.

Le conduit vocal est décomposé en un certain nombre de sections, réparties le long du conduit vocal et modélisées par des tubes uniformes. Sa réponse en fréquence est calculée comme celle d'un ensemble de quadripôles en série, par multiplication des fonctions de transfert de chacun des quadripôles. Les résonances se cumulent tout au long du conduit. Les formants s'identifient aux pics de résonance de la fonction de transfert totale.
Un conduit vocal simple permettant de prononcer une voyelle orale, a une réponse qui ne présente que des pôles. La modélisation par une succession de tubes élémentaires s'identifie bien au conduit réel. Nous ne modéliserons pas ici le conduit nasal, avec toutes les imperfections que cela pose pour la synthèse articulatoire du signal complet. Les fonctions de transfert, que nous calculons pour toutes les trames de la séquence, ne présentent que des pôles.

La fonction de transfert (ou réponse en fréquence) associée à une fonction d'aire est une fonction qui décrit les propriétés du conduit vocal en tant que système (ou filtre) qui transforme un signal d'entrée (la source) en signal de sortie. Dans une perspective source-filtre, le spectre d'un son résulte de la multiplication du spectre de la source par la fonction de transfert. Nous détaillerons cet aspect après l'étude des formants, au paragraphe 3.
A partir d'une fonction de transfert calculée, nous sommes en mesure de déterminer la fréquence des formants, à l'aide d'un détecteur de pics. Les valeurs de fréquences des 3 premiers pics (correspondant aux 3 premiers formants) ont été estimées pour chaque trame de la séquence. Ces fréquences seront considérées comme les valeurs estimées, c'est-à-dire celles associées aux configurations géométriques extraites semi-automatiquement par la méthode. Nous parlerons de formants estimés via la fonction d'aire ou simplement de formants estimés.



(a) (b)

Figure 89 : Fonction d'aire et fonction de transfert simulée d'un [a].

2. Analyse de formants

Nous nous intéressons dans ce paragraphe à l'estimation des formants à partir des fonctions d'aire et nous comparons les fréquences obtenues avec celle des formants extraits du signal audio d'origine, et ceci pour l'ensemble de la séquence Laval43.

Cependant, les formants sont particulièrement caractéristiques pour les voyelles, qui sont des signaux quasi-stationnaires variant peu en amplitude et en fréquence. Ainsi, après un aperçu de l'estimation des formants globalement sur la séquence (avec ou sans silence), nous étudierons plus en détail ce qu'il se passe au niveau des voyelles du corpus, qui ont été précédemment détaillées au chapitre 5.

Nous avons étiqueté à la fois les voyelles orales et les voyelles nasales mais nous ne traiterons pas ces dernières pour l'instant. Leur étude est complexe et nécessiterait la modélisation du conduit nasal. Leur production sera néanmoins observée, un peu plus loin (Chapitre 9), en parallèle des mouvements du voile du palais.

Les voyelles concernées dans ce paragraphe-ci sont uniquement les quelques 200 voyelles orales ([a], [i], [e], [u], [y], [o], [ɛ], [ø], [ɔ]) étiquetées du corpus Laval43.

2.1. Représentations

Il est établi qu'il existe une relation articulatoire-acoustique entre les formants et la forme du conduit vocal et notamment entre les 2 premiers formants et la position de la langue.

Cette idée renvoie à l'étude réalisée préalablement au chapitre 5. Et le modèle linéaire proposé au chapitre 6 avait essentiellement pour vocation de restaurer cette relation entre position de la langue et position des formants.

Pour simplifier ce qui a déjà été dit, on considère que le formant F_1 est corrélé avec l'ouverture et le formant F_2 avec la position antérieure (F_2 élevé) et postérieure (F_2 bas) du son.

On a l'habitude de représenter les formants F_1-F_2, sous la forme d'un triangle, dit acoustique, dans lequel l'organisation des voyelles rappelle celle des voyelles sur le triangle articulatoire de la phonétique classique (figure 76 au chapitre 6).

Nous nous limitons ici à la représentation de F_1 et F_2, l'analyse du troisième formant sera traitée plus loin.

A partir des formants d'origine, nous représentons les trames de la séquence dans un plan F_1-F_2. Les figures 90a et 90b sont obtenues en portant, pour chaque trame de la séquence, en abscisses la fréquence du second formant et en ordonnées celle du premier formant, selon une échelle linéaire. La figure de gauche (90a) permet de voir en gris plus clair les trames dites « de parole », les points gris foncé représentent les trames de silence du corpus. On constate que pour ces trames, les résultats donnés par *Praat* sont très disparates et atteignent des valeurs de F_1 de plus de 2KHz. La détection de formants ne fonctionne pas dans les silences, il s'agit de bruit.

La distinction des trames de parole et de silence est réalisée à partir d'un seuil fixé pour l'amplitude dans la 1ère sous-bande, comme cela a été expliqué au chapitre 6.

Ainsi, nous laissons désormais pour un temps les trames de silence et nous représentons uniquement les trames de parole. La figure 90b est un zoom sur les trames de parole du signal, les signes de couleur correspondent aux voyelles sélectionnées à partir de l'audio. Le choix des signes est fonction de la classe vocalique et correspond à celui qui a été défini précédemment au chapitre 5, § 3., Table 9.

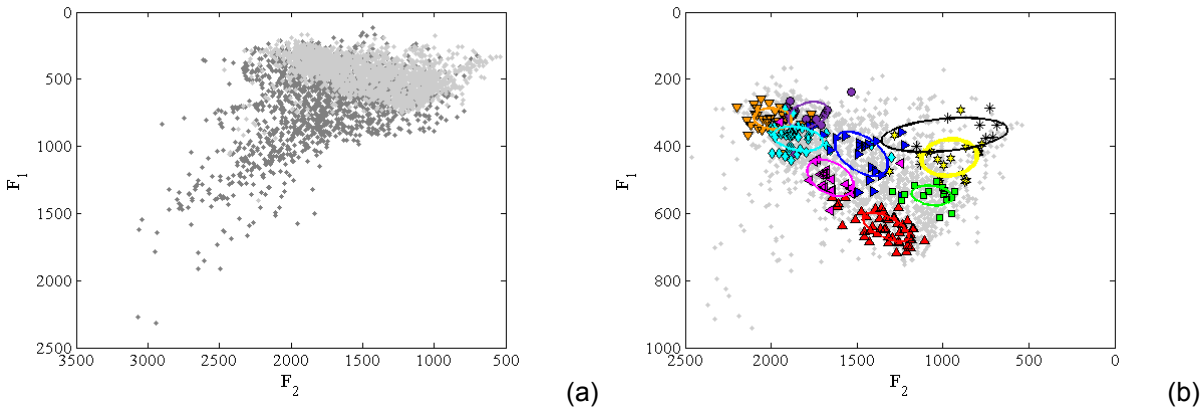


Figure 90 : (a) Triangle vocalique de la séquence Laval43 à partir des données d'origine : les points gris foncé sont des instants de silence.
(b) Zoom sur les trames de parole et les voyelles sélectionnées.

Les voyelles sélectionnées se placent correctement dans le triangle vocalique. Les voyelles [a] occupent le sommet inférieur du triangle, qui correspond à une langue abaissée, les [i] ont des valeurs de F_1 faibles et des valeurs de F_2 élevées, elles se placent dans le coin supérieur gauche, correspondant à des configurations de langue haute et en avant. Les

voyelles [u] se placent dans le dernier coin du triangle, en haut à droite, en accord avec une position haute et en arrière de la langue attendue pour ces voyelles. Les autres catégories de voyelles se placent par paquet dans le triangle, à des valeurs moyennes attendues et habituelles de fréquence.

De la même façon que les formants d'origine, les formants estimés à partir des fonctions d'aire sont représentés dans un plan F_1-F_2. Le triangle vocalique obtenu (Fig. 91a) est moins étendu que celui des données originales. De même que précédemment, les points gris clair représentent les trames de parole et les trames de silence sont en gris foncé. L'espace maximal vocalique [BPGS89] est couvert par toutes les configurations observables, même pendant les périodes de silence. Par rapport au modèle linéaire, on n'observe pas de centrage des données, il n'y a pas de problème de définition de l'espace maximal. On observe une véritable allure de triangle à des valeurs « classiques » de fréquences. La notion d'espace maximal triangulaire est issue de l'application du modèle acoustique. L'espace couvert par les données géométriques (par projection ou directement avec l'observation de la constriction) est plutôt un quadrilatère avec les caractéristiques haut/bas - avant/arrière.

Le zoom sur les instants de voyelles (Fig. 91b) permet d'observer que les différentes catégories de voyelles sont globalement bien placées ; une étude plus détaillée et quantitative sera réalisée un peu plus loin. On remarque que l'organisation générale est respectée. Les voyelles [a] se placent globalement dans le bas du triangle, de la même façon que les [i] sont représentés en haut à gauche et les [u] en haut à droite. Cependant on constate, par l'observation des ellipses de dispersion (à un écart-type), plus d'excursions que précédemment avec les formants du signal d'origine. Une analyse de la discrimination entre les classes vocaliques sera traitée au §2.2.3..

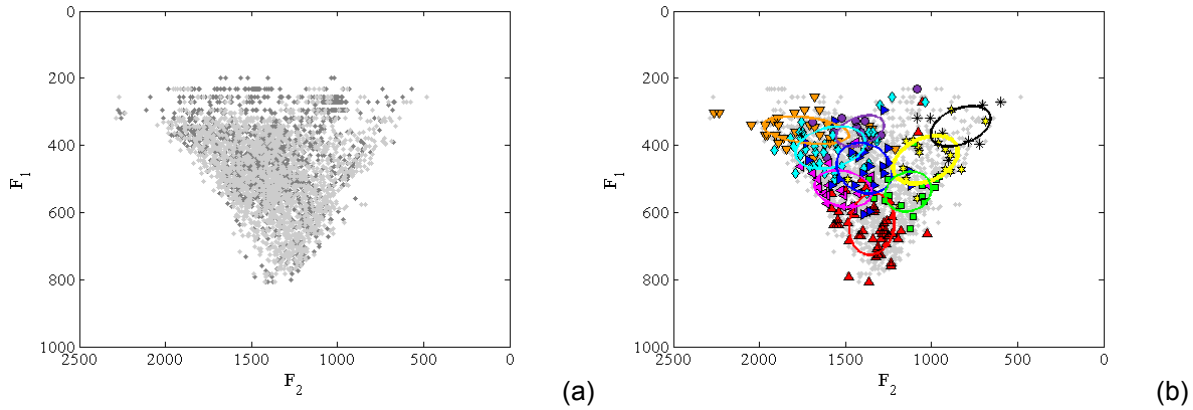


Figure 91 : (a) Triangle vocalique de la séquence Laval43 à partir des formants estimés des fonctions d'aire : les points gris foncé sont des instants de silence. (b) Zoom sur les trames de parole et les voyelles sélectionnées pour ces mêmes données estimées.

Si l'utilisation du modèle acoustique améliore la couverture de l'espace maximal par rapport au modèle linéaire, il n'améliore pas forcément la discrimination.

2.2. Comparaison entre les formants d'origine et les formants estimés

Comme au chapitre précédent, avec le modèle linéaire, le but de cette comparaison est de montrer une cohérence en terme de formants, entre les données d'origine, celles extraites du signal audio et les données estimées, celles obtenues à partir des contours géométriques extraits semi-automatiquement et exploités pour en prédire les fonctions d'aire associées.
Avant de synthétiser un signal à partir des données estimées et de parler en terme perceptif, plusieurs indicateurs quantitatifs nous permettent d'évaluer cette cohérence. L'analyse qui suit commence par l'observation de la corrélation puis celle du biais et de l'écart-type entre données estimées et d'origine. Elle se poursuit par une étude dans le plan F_1-F_2 avant d'observer le comportement du troisième formant.

2.2.1. Corrélation

Des mesures de corrélation ont déjà été faites au chapitre précédent avec le modèle linéaire, nous étudions ici la corrélation entre les formants d'origine et les formants estimés à partir du modèle acoustique.
Les coefficients de corrélation obtenus pour la séquence laval43 complète sont faibles : 0,08 pour F_1, 0,35 pour F_2. Ils n'indiquent aucune corrélation des formants estimés avec ceux extraits du signal audio.
Mais compte-tenu de l'allure du triangle vocalique des données d'origine et du nombre de valeurs extrêmes présentes, ce résultat peut s'expliquer : prenons comme exemple le premier formant, on observe de grandes variations parmi les données extraites, F_1 monte jusqu'à 2KHz, ce qui n'est pas observable pour les données estimées. Ce bruit important dans le jeu de données originales biaise le calcul du coefficient de corrélation.
Comme on a déjà pu le mentionner et l'observer figure 90a, ces grandes dispersions de valeurs sont principalement présentes dans les zones de silence.
Si on limite le calcul du coefficient de corrélation aux trames dites de « parole », la similitude entre les données extraites et estimées augmente, on atteint plus de 0,60 de corrélation pour F_1 et F_2.
En se limitant encore et en ne s'intéressant plus cette fois qu'aux trames étiquetées de voyelles, la corrélation dépasse 0,85 pour F_1 et 0,75 pour F_2.
On observe ces corrélations pour les trames de voyelles sur les graphes suivants.

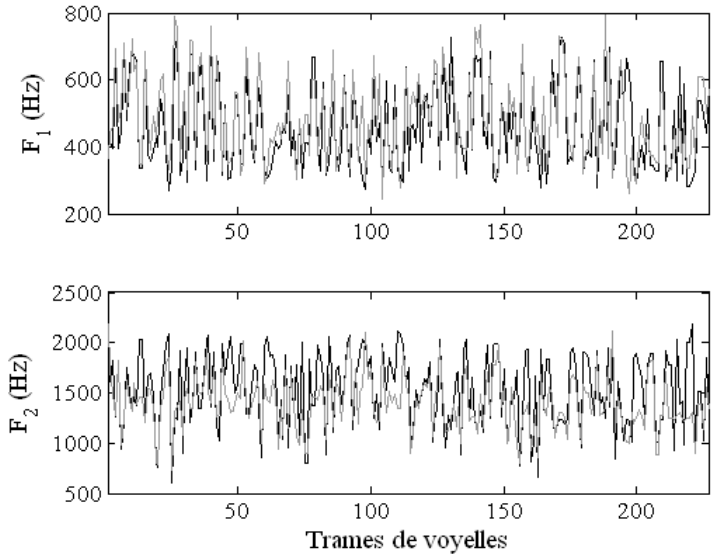


Figure 92 : F_1 et F_2 pour les trames étiquetées de voyelles. Le tracé noir correspond aux valeurs d'origine et le gris aux fréquences estimées via les fonctions d'aire.

L'observation des corrélations pour F_3 ne donne pas de résultats probants. Même en se limitant à la séquence sans silence ou simplement aux voyelles, aucune corrélation ne semble exister entre les données extraites du signal audio et celles estimées via le modèle géométrique. En théorie, F_3 est corrélé avec la configuration des lèvres pour les voyelles antérieures. Mais nos images radiographiques ne donnent pas toutes les informations sur la géométrie des lèvres, et en particulier sur leur arrondissement. Cette donnée peut expliquer les faibles résultats obtenus avec ce troisième formant.
On récapitule dans le tableau 11 les coefficients de corrélation obtenus.

	Séquence complète	Séquence sans silence	Voyelles
F_1	0,08	0,66	0,86
F_2	0,35	0,61	0,75
F_3	0,06	0,07	0,22

Table 11 : Coefficients de corrélation entre formants d'origine et formants estimés.

La corrélation fournie ici par l'approche acoustique pour les instants de parole est inférieure à 0,7, qui était le résultat obtenu avec le modèle linéaire.

Nous notons de plus que l'influence du rapport pixel/cm est faible. Les coefficients de corrélation varient peu suivant que la fonction d'aire utilisée pour l'estimation des formants a été compilée avec 38 ou 42 pixels pour 1 cm (les différences ne sont pas significatives). L'allure des trajectoires formantiques est peu perturbée par la variation de ce rapport, c'est le niveau de ces trajectoires qui est modifié.

Le calcul des coefficients de corrélation réalise un recentrage et annule le biais entre les valeurs moyennes des signaux considérés. Le paragraphe suivant s'attache alors à l'étude des signaux en terme de différence moyenne et d'écart-type.

2.2.2. Biais et écart-type

Considérant les formants extraits du signal d'origine comme référence, nous calculons le biais d'estimation (comme cela a été fait pour le modèle linéaire au chapitre 6), c'est-à-dire la différence moyenne μ entre les formants estimés et la référence ainsi que l'écart-type σ pour chacun des 3 premiers formants. Ces mesures sont normalisées par rapport aux valeurs d'origine, elles sont décrites par les formules suivantes :

$$\mu(F_i) = \frac{1}{N}\sum_{p=1}^{N}\frac{\left|\widetilde{F}_i(p) - F_i(p)\right|}{F_i(p)}$$

$$\sigma(F_i) = \sqrt{\frac{1}{N}\sum_{p=1}^{N}\left(\frac{\left|\widetilde{F}_i(p) - F_i(p)\right|}{F_i(p)}\right)^2}$$

où $\widetilde{F}_i(p)$ est la valeur estimée du formant i à la trame p.

2.2.2.1. Réglages de paramètres

Le calcul du biais est effectué pour la séquence hors moments de silence en faisant varier les paramètres dont nous avons parlé à la fin du chapitre 4 sur la fonction d'aire, à savoir la position de la glotte, le rapport pixels/cms et les paramètres α-β. Ceci nous permet ainsi de confirmer ou de modifier les choix qui ont été faits.

Nous observons (Fig. 93c) que faire varier la position de la glotte, de quelques pixels autour de la position définie, ne modifie quasiment pas la différence moyenne entre les formants d'origine et les formants estimés. Une ANOVA à 1 facteur sur ces différentes positions confirme statistiquement le résultat graphique (au risque α=5%, F<1), il n'y a pas de différence significative.

Au contraire, le rapport entre les pixels et les centimètres implique des différences significatives sur le biais entre original et estimation. Ce résultat obtenu par une ANOVA à 1 facteur sur les différents rapports est vérifié pour la séquence hors moments de silence et pour les 3 formants (resp. F=39,47/2,92/28,71 p<0.01). Sur le graphe 93a ci-dessous, on observe pour chaque formant, l'évolution du biais en fonction du rapport pixels/cms. Le comportement du biais en fonction de ce rapport est identique pour les formants 2 et 3, et présente un minimum autour de 38-39 pixels/cms, l'allure est différente pour le premier

formant pour qui, le biais croît avec le rapport. La moyenne des biais sur les 3 premiers formants pour la séquence hors silence (Fig. 93a en gras et Fig. 93b) atteint un minimum pour un rapport de 38 pixels pour 1 cm. Ceci est également vérifié pour la séquence complète et en se limitant aux seules trames de voyelles.

Compte-tenu de ces résultats, nous considérons que la position de la glotte et le rapport choisi de 38 pixels pour 1 cm sont corrects et permettent des résultats optimaux.

Enfin, nous testons (Fig. 93d) quelques jeux de paramètres α-β : nous comparons les résultats obtenus avec les paramètres de Soquet et al. [SLMD02] du locuteur mâle (h), ceux de la locutrice (f) et ceux obtenus en moyennant ces deux précédents jeux de paramètres (m). L'influence de ces paramètres sur le calcul du biais pour la séquence hors moments de silence dépend du formant considéré : pour F_3, il n'y a quasiment pas d'influence de ces paramètres, pour F_1 et F_2, les influences sont opposées, le biais sur F_2 est plus faible avec les paramètres du locuteur mâle, par contre avec ces mêmes paramètres le biais sur F_1 est plus important. Les paramètres « moyennés » proposent des résultats intermédiaires. En moyenne sur les 3 formants, on constate peu d'influence des paramètres α-β. Nous conservons le choix d'utiliser les paramètres du locuteur mâle.

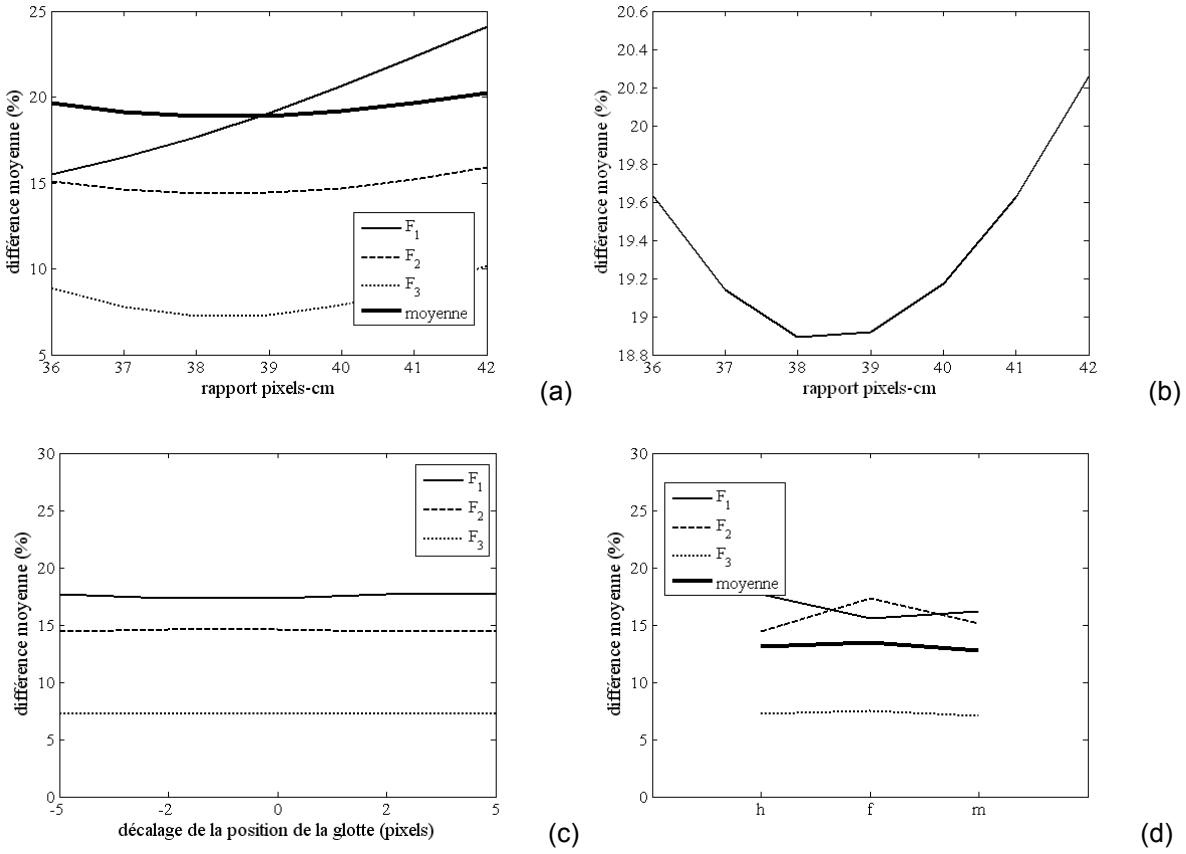


Figure 93 : Biais entre formants estimés et formants d'origine pour les 3 premiers formants, pour la séquence hors moments de silence et pour différentes valeurs de paramètres : (a) rapport pixels/cm - (b) zoom sur la moyenne des 3 formants - (c) position de la glotte - (d) paramètres α-β.

2.2.2.2. Analyse des résultats

Etant donnés ces paramètres réglés, nous analysons à présent les mesures obtenues en terme de biais et d'écart-type.

Le tableau 12 résume ces valeurs en distinguant les cas où l'on considère la séquence entière (les 4043 trames), la séquence sans les silences (les quelques 2000 trames de « parole ») ou les trames de voyelles. Comme pour la corrélation, les différences moyennes et les écart-types sont meilleurs si on ne considère pas les trames de silence, et encore plus si on se limite aux voyelles. Les valeurs de biais et d'écarts-types sont plus faibles pour F_3 en comparaison à F_1 et F_2, mais ce résultat est à nuancer compte-tenu des coefficients de corrélation. Les comportements de F_1 et F_2 en terme de biais sont assez similaires, les écarts-types pour F_1 sont supérieurs à ceux de F_2. On estime à 13% le biais d'estimation moyen des formants par rapport aux valeurs d'origine, pour les trames de parole et à 10% pour les voyelles.

	Différence moyenne (%)			Ecart type (%)		
	Séquence	Parole	Voyelles	Séquence	Parole	Voyelles
F_1	26	18	12	26	21	11
F_2	19	14	13	16	13	10
F_3	12	7	7	11	8	7

Table 12 : Différence moyenne et écart-type des formants estimés par rapport aux formants d'origine.

Une analyse plus fine est réalisée sur les voyelles sélectionnées du corpus, en s'intéressant aux différents types de voyelles.

Avant d'observer en détail les représentations F_1-F_2 et l'organisation des éléments sélectionnés dans les triangles vocaliques, nous examinons les valeurs moyennes des formants F_1, F_2 et F_3 pour les différentes voyelles étiquetées.

Il n'existe pas d'étude normative pour les voyelles françaises comparable aux données de Peterson et Barney [PB52] pour l'anglais américain. Néanmoins, nous comparons ici les valeurs d'origine et celles estimées à des valeurs moyennes pour des sujets masculins [Cal89], que nous considérons ici comme référence ; il s'agit non pas d'une référence représentant la norme du français mais plutôt d'une référence externe, pour comparer nos données à d'autres données de la littérature.

	Signal original			Estimation			Référence		
	F_1(Hz)	F_2(Hz)	F_3(Hz)	F_1(Hz)	F_2(Hz)	F_3(Hz)	F_1(Hz)	F_2(Hz)	F_3(Hz)
[a]	651	1333	2397	632	1356	2413	684	1256	2503
[i]	312	2006	2899	354	1730	2484	308	2064	2976
[y]	332	1801	2368	356	1416	2440	300	1750	2120
[u]	348	958	2261	342	831	2441	315	764	2027
[o]	430	984	2306	441	1048	2420	383	793	2283
[ɔ]	546	1071	2370	537	1135	2414	531	998	2399
[e]	383	1838	2564	406	1587	2410	365	1961	2644
[ɛ]	490	1673	2438	529	1512	2385	530	1718	2558
[ø]	417	1486	2313	467	1390	2426	381	1417	2235

Table 13 : Comparaison des valeurs moyennes des 3 premiers formants, pour chaque classe vocalique orale du corpus, entre les données d'origine et celles estimées. Mise en parallèle avec des valeurs de référence de ces mêmes formants [Cal89].

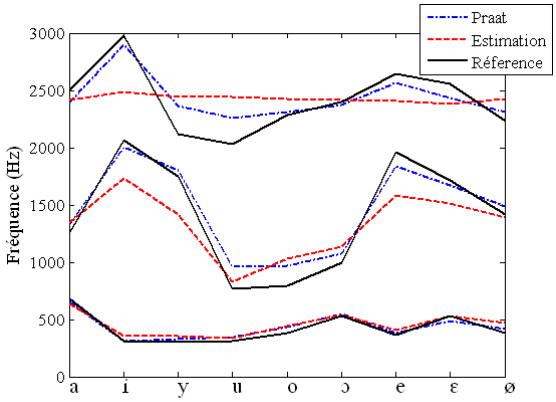


Figure 94 : Comparaison des valeurs moyennes des 3 premiers formants, pour chaque classe vocalique orale du corpus, entre les données d'origine et celles estimées. Mise en parallèle avec des valeurs de référence de ces mêmes formants [Cal89].

Le formant F_3 est mal estimé, il n'y a pratiquement pas de différence entre les 9 types de voyelles considérées. Les 2 autres formants présentent des comportements en accord avec la référence et les données d'origine, on observe une tendance à l'écrasement dynamique des formants estimés, ceci est particulièrement visible pour F_2 qui est soit sous-estimé pour les voyelles antérieures, soit sur-estimé pour les voyelles postérieures.

Nous analysons, pour les différentes classes vocaliques, le comportement du biais en fonction des 3 formants.

Dans cette mesure (représentée Fig. 95), pour les voyelles, nous comparons d'une part les formants estimés, et d'autre part les formants extraits de l'audio, avec les valeurs considérées comme référence dans le tableau 13. Les différences obtenues sont normalisées par rapport à la moyenne de cette référence sur les 9 classes. Les résultats sont présentés par formants.

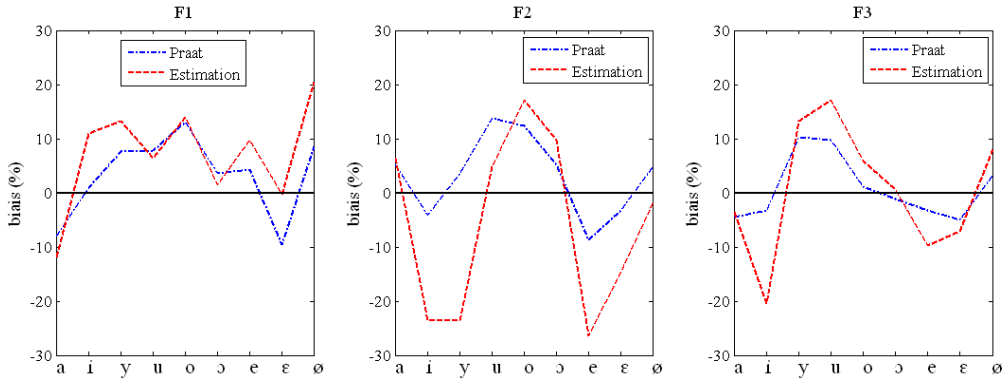


Figure 95 : Biais, pour les 3 premiers formants et pour les 9 classes vocaliques, entre les formants extraits de Praat et une référence [Cal89] et entre les formants estimés et cette même référence.

Ces écarts observés pour notre estimation sont plus importants que pour les formants extraits du signal, en particulier pour le second formant. Ils oscillent entre –10% et 20% pour F_1, entre –25% et 18% pour F_2 et entre –20% et 18% pour F_3. En moyenne, en valeur absolue, on retrouve les valeurs du tableau 12 pour les voyelles. En ayant toujours conscience que cette référence n'est pas une référence absolue, ces écarts sont rapprochés de résultats en perception, connus sous le nom de Difference Limen.

2.2.2.3. Seuil différentiel (Difference Limen)

Pour l'analyse et la synthèse de phonèmes, il n'est pas nécessaire d'avoir une précision des fréquences de formants excédant celle que l'oreille humaine peut détecter. Aussi, de nombreuses expériences ont été effectuées pour déterminer le seuil de perception des différences fréquentielles pour les formants des voyelles. En 1955, Flanagan détermine le minimum de précision nécessaire pour l'analyse et la synthèse des formants F_1 et F_2 des voyelles [Fla55]. Il entreprend des tests pour juger de la similitude de 2 sons voisins pour différentes valeurs de fréquences de chaque formant, séparément. Le seuil de la variation des fréquences des formants est atteint lorsque la différence $\Delta F_n = F_{ref,n} - F_n$ est perçue. F_n est la fréquence du formant testé. Le seuil, noté DL_n (Difference Limen) est calculé comme suit (et correspond à la notion de biais décrite juste avant).

$$DL_n = \frac{\Delta F_n}{F_{ref,n}}$$

Les résultats obtenus par ces tests indiquent la nécessité d'une bonne résolution. En effet, le tableau 14 montre des valeurs de DL_n inférieures à 6% pour les différentes fréquences du test. Ces valeurs dépendent de la distribution des formants dans le spectre. Elles sont d'autant plus faibles que les formants voisins de $F_{ref,n}$ sont proches.

Fréquence Fn (Hz)	300	500	700	1000	1500	2000
-DLn (%)	5,7	5	3,9	2	3	4,5
+DLn(%)	4	5,4	2,7	5	5	1

Table 14 : Seuils de perception pour différentes valeurs de fréquences de formants F_1 et F_2 d'après Flanagan (1955).

Ce seuil différentiel désigne la limite en dessous de laquelle un sujet ne parvient plus à distinguer deux stimuli. Cette différence aussi appelée JND (Just Noticeable Difference) en anglais, a été étudiée en contexte consonantique par Mermelstein en 1978 [Mer78], pour compléter l'étude de Flanagan limitée aux voyelles hors contexte. Pour les voyelles seules, il trouve des seuils différentiels plus grands que Flanagan, allant de 4% à 14%. Ils trouvent des différences de 14% et 5,5% pour le premier formant, un peu moins pour le second formant (4,2% et 7%). Les JNDs trouvées en réalisant un changement simultané des 2 formants sont plus faibles. En contexte consonantique, les seuils sont plus grands.

Si nous considérons nos mesures sur les formants estimés, les différences observées entre ces formants et les formants de référence sont supérieures aux seuils différentiels proposés pour des voyelles statiques. Perceptuellement, la différence sera marquée entre nos estimations et la référence. Mais, compte-tenu du contexte dynamique dans lequel nous nous plaçons, avec l'influence du contexte consonantique et de la coarticulation, et en accord avec les résultats de Mermelstein, ces différences apparaissent plus faibles. Nos valeurs moyennes sur les voyelles de 12% pour F_1 et 13% pour F_2, se rapprochent des seuils différentiels définis par [Mer78], elles sont comparables du point de vue du premier formant, supérieures pour le second formant. Ceci laisse supposer qu'un certain nombre de voyelles pourront être correctement perçues dans un contexte de synthèse dynamique, qui sera l'objet du paragraphe 3 de ce chapitre.

2.2.3. Ellipses de dispersion et tables de confusion

Le triangle vocalique constitue un moyen de visualiser la relation articulatoire-acoustique pour les voyelles. Nous présentons sur la figure 96a les ellipses de dispersion (un écart type) de chacune des classes vocaliques orales du corpus considéré. La figure 96b présente les ellipses des mêmes voyelles en utilisant les formants estimés des configurations articulatoires.

Nous constatons des zones communes entre les ellipses de différentes voyelles. Des combinaisons voisines de valeurs formantiques renvoient à des voyelles différentes. Ces zones de recouvrement vont diminuer la discrimination entre les voyelles.

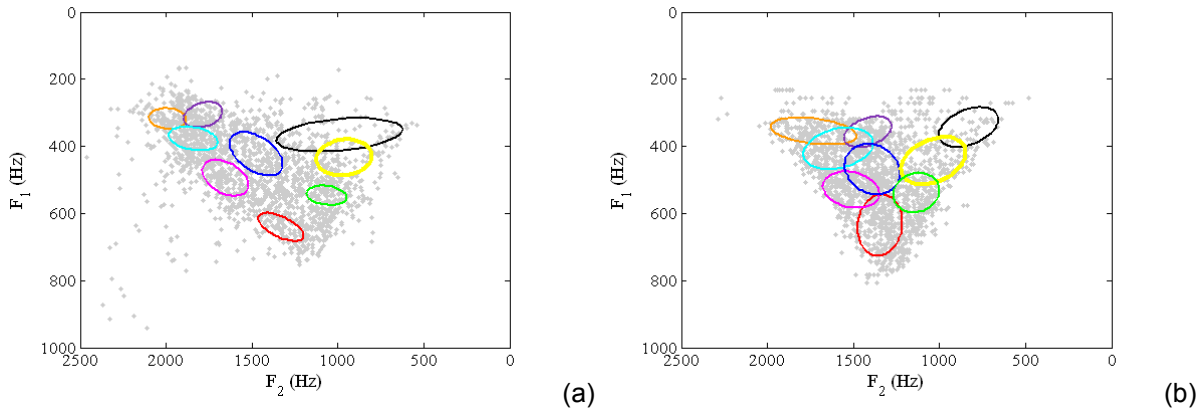


Figure 96 : (a) Plan F_1-F_2 des données extraites du signal audio et ellipses de dispersion (un écart-type) des voyelles sélectionnées.
(b) Plan F_1-F_2 des données estimées et ellipses de dispersion (un écart-type) des voyelles sélectionnées.

Les ellipses de dispersion permettent de visualiser le recouvrement mais pour quantifier cette confusion observée entre les classes vocaliques, nous analysons la capacité discriminante à partir des matrices de confusion, de façon similaire à l'analyse réalisée au chapitre 5. Les matrices de confusion sont établies à partir des données directes en s'appuyant sur le diagramme de Voronoï des données. Ces diagrammes sont représentés sur les figures 97a (données d'origine) et 97b (données estimées), ils sont calculés à partir des 9 centroïdes d'ellipses. Cette représentation sous forme de Voronoï permet une vision complémentaire pour appréhender la distribution des données dans l'espace formantique. Le partage de l'espace entre les 9 classes vocaliques considérées est globalement préservé entre les données d'origine et les données estimées, mais on observe quelques défauts locaux, et notamment des modifications de frontières, comme celle entre le [y] et le [i] ou le [u] ou le [o].

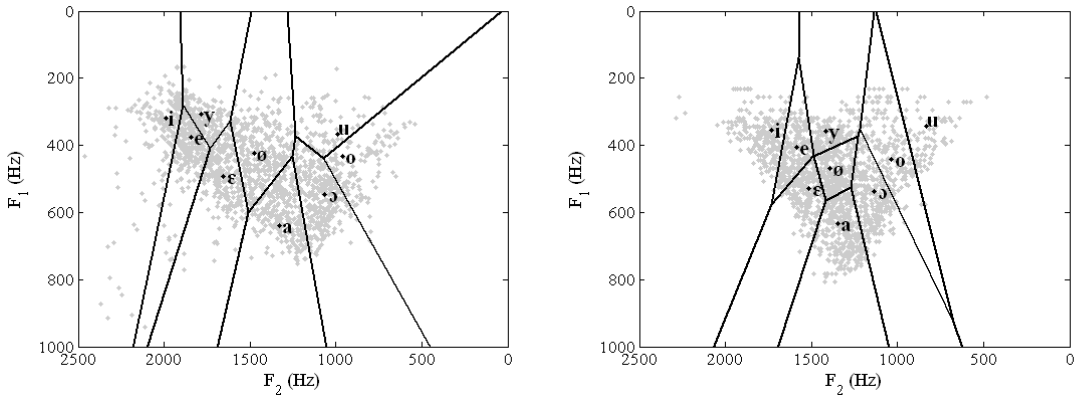


Figure 97 : (a) Plan F_1-F_2 des données extraites du signal audio et diagramme de Voronoï des classes vocaliques.
(b) Plan F_1-F_2 des données estimées et diagramme de Voronoï des classes vocaliques.

La capacité discriminante évalue la « probabilité » de correctement classifier les voyelles. Nous considérerons ici 2 méthodes de mesures :

- La première consiste en une classification au plus proche voisin en terme de distance euclidienne dans le plan F_1-F_2. Elle est basée sur le calcul des distances des voyelles aux centroïdes des ellipses dans le plan F_1-F_2 (ces centroïdes correspondent aux valeurs moyennes F_1 et F_2 pour chaque catégorie de voyelles). Etant donnée par exemple une voyelle [a], on estime que cette voyelle est correctement placée si, en terme de distance euclidienne dans le plan F_1-F_2, cette voyelle est plus proche du centre de l'ellipse des [a] que de n'importe quel autre centroïde d'ellipse. Sinon on décide que cette voyelle est classée dans la catégorie vocalique associée au centroïde le plus proche.
- La seconde est basée sur la méthode des plus proches voisins (ou méthode des k-voisins, qui est très courante en algorithmique). Elle consiste, étant donné un point *x,* à déterminer quels sont les *k* points de l'ensemble *A* considéré les plus proches de *x*. On parle alors de trouver un voisinage de taille *k* autour du point *x*. Ici, soit *x* une voyelle, il s'agit de trouver les *k* points de l'ensemble des voyelles les plus proches de *x*. La voyelle *x* sera classée dans la catégorie vocalique la plus représentée parmi ces *k* voisins.

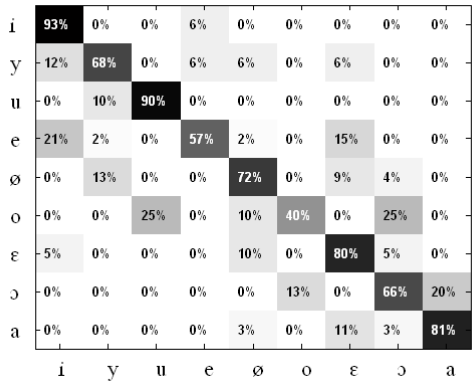
Ces méthodes permettent de caractériser le pourcentage de voyelles correctement détectées et d'évaluer le type de confusions qui apparaissent. Ces résultats sont présentés sous forme de tables de confusions (figures 98 et 99). La diagonale indique, par voyelle, les pourcentages de « bonnes détections ». On définit, pour chaque triangle vocalique, le taux de reconnaissance des voyelles, comme la moyenne des éléments de la diagonale. La première mesure avec la distance au plus proche voisin donne un taux évalué à 72% à partir des formants du signal d'origine et à 50% pour les estimations. La seconde mesure, basée sur 4 voisins, donne des taux à peu près équivalents, un tout petit peu plus faibles : 71% pour les formants d'origine et 47% pour les estimations. Les tables de confusion sont différentes dans le détail, mais le nombre d'exemplaires de chaque classe est trop faible pour que l'on puisse conclure sur des différences significatives (sur les 9 classes, 5 possèdent moins de 21 éléments).

Les ellipses des formants d'origine sur la figure 96a sont relativement disjointes, mis à part celles du [u] et du [o]. Cependant, on constate que le [u] est sous-représenté par rapport aux autres catégories de voyelles (seulement 10 éléments, alors que le [a], par exemple, en compte plus de 50). L'ellipse du [u] est large car très sensible à un seul exemplaire mal

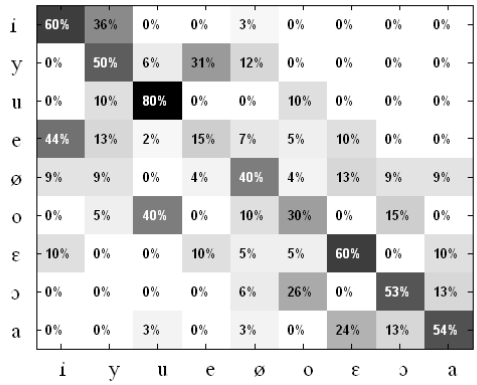
représenté. Plus d'un tiers des [o] vont être détectés comme des [u]. Les ellipses des [i], [y] et [e] sont proches.

Les ellipses des formants estimés (Fig. 96b) sont un peu plus superposées, comme le met aussi en évidence la table de confusion. La capacité discriminante est moins bonne. Néanmoins, on observe une répartition correcte des ellipses dans le triangle vocalique, les voyelles de chaque classe sont globalement à la bonne place (les moyennes sont proches des valeurs de référence), même s'il y a des exemplaires qui s'écartent dans chaque catégorie.

Les différences observées sur les ellipses entre les données d'origine et les données estimées sont homogènes pour les différentes classes, en dehors du [u] dont l'ellipse est moins étendue pour l'estimation.

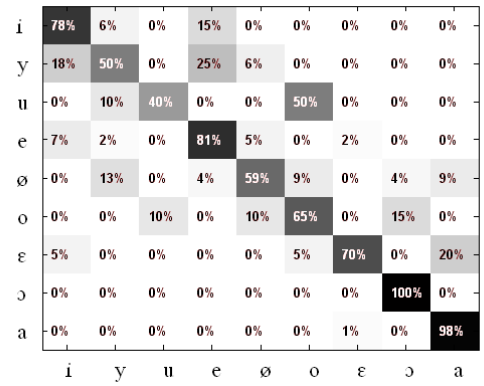


(a)

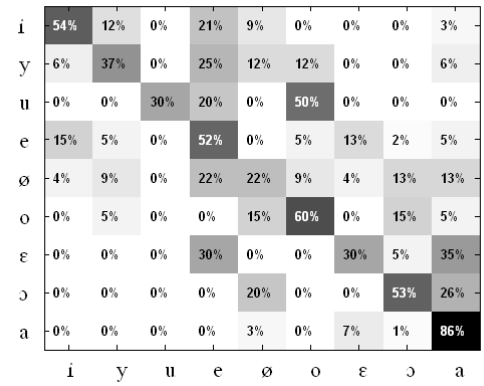


(b)

Figure 98 : (a) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 d'origine et sur les distances aux valeurs moyennes des classes vocaliques. (b) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 estimés à partir des fonctions d'aire et sur les distances aux valeurs moyennes des classes vocaliques.



(a)



(b)

Figure 99 : (a) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 d'origine et sur *k*-voisins dans l'espace vocalique. (b) Table de confusion des voyelles en se basant sur les formants F_1 et F_2 estimés à partir des fonctions d'aire et sur *k*-voisins dans l'espace vocalique.

Ces résultats mettent en évidence la variabilité dans les fréquences formantiques des voyelles ; cette variabilité, souvent abordée dans la littérature, est en partie attribuée à la coarticulation, qu'on peut définir comme l'influence qu'exerce un son sur un son contigu. La variabilité de la production des voyelles due aux phénomènes de coarticulation s'explique par des chevauchements de gestes articulatoires, elle est donc liée à une variabilité articulatoire sous-jacente.

2.2.4. Discussion sur F_3

Les triangles vocaliques en deux dimensions (F_1-F_2) ne fournissent pas une image complète des systèmes vocaliques. La prise en compte du troisième formant a une importance particulière pour les langues, comme le français, dans lesquelles le trait d'arrondissement labial a valeur distinctive. En effet, le troisième formant est utile pour faire ressortir clairement, par exemple, la distance qui sépare [i] et [y], voyelles qui paraissent extrêmement proches lorsqu'elles sont présentées dans un espace F_1-F_2. Par contre, ce troisième formant est moins nécessaire pour les voyelles postérieures. L'estimation précise de sa fréquence est en outre difficile dans le cas de ces voyelles-là du fait de sa faible amplitude.

F_3 joue donc un rôle appréciable dans l'identification des voyelles qui ont F_2 et F_3 très rapprochés, comme pour les voyelles antérieures [Del48]. La perception de leur somme équivaut à peu près à la perception d'un seul formant dont la fréquence sera intermédiaire entre F_2 et F_3 (travaux de Chistovich sur le F_2', [CSL79], [Chi80]).

Le tracé des éléments du corpus Laval43 dans un plan F_2-F_3 a été effectué sur les figures suivantes. On observe les données d'origine (Fig. 100a) et les données estimées à partir de la fonction de transfert (Fig. 100b) pour les trames de la séquence (hors silence), ainsi que les ellipses de dispersion associées aux catégories vocaliques. Le recouvrement des ellipses est important dans chacune des 2 représentations.

Nous distinguons 2 problèmes différents concernant le troisième formant. D'un côté, les valeurs très dispersées du F_3 de [u] (grande ellipse noire) pour les données d'origine (Fig. 100a) reflètent la difficulté de la mesure de ce formant (problème d'estimation de formants de Praat). D'un autre, nous constatons le manque de variabilité dans les mesures de ce formant pour les données estimées : les ellipses sont presque toutes superposées, le formant F_3 varie très peu entre les différentes classes vocaliques. Cette dégradation de l'estimation par rapport aux données d'origine s'explique par le mouvement réduit des lèvres au cours de la séquence.

L'analyse discriminante, élargie en prenant en compte les 3 formants, présente des résultats comparables à celle réalisée à partir de F1 et F2, en terme de discrimination des classes

vocaliques pour les données estimées. Toutefois, l'observation de la table de confusion montre quand même une meilleure discrimination du [i] (on passe de 60% avec F_1-F_2 à 72% en ajoutant F_3, pour le taux de discrimination de cette classe vocalique) et donc une confusion moins grande avec la classe du [y].

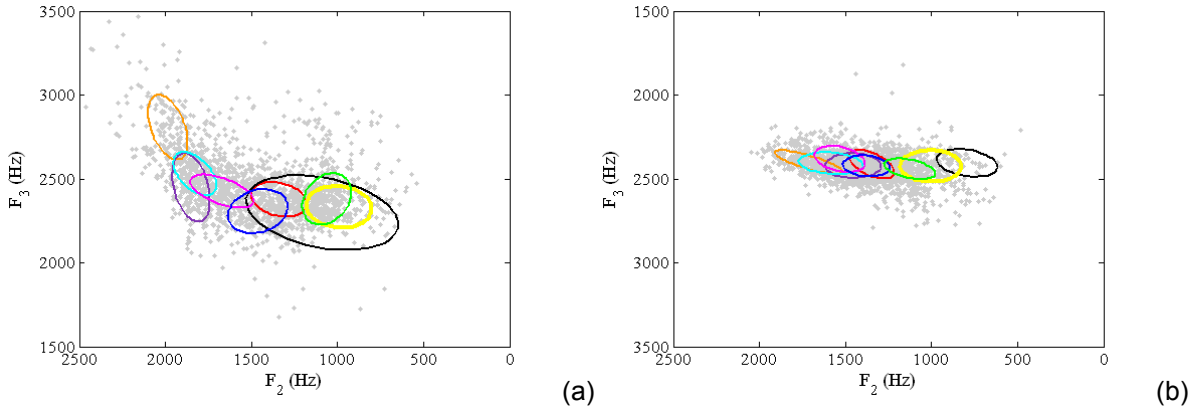


Figure 100 : Plan F2-F3 pour les trames de parole de Laval43 à partir des données (a) d'origine et (b) estimées.

Ces observations concordent avec ce qui a été constaté avec le calcul des coefficients de corrélation entre formants d'origine et formants estimés. L'estimation de F_3 à partir des données géométriques du conduit vocal n'est pas bonne. Même si, de manière générale, la hauteur fréquentielle du troisième formant varie peu pour la majorité des voyelles, on observe ici des valeurs de F_3 quasiment constantes pour toutes les trames de parole de la séquence. L'estimation des données souffre d'un manque d'information concernant les lèvres, et en particulier concernant le trait d'arrondissement.

3. Signal audio synthétisé

Après l'analyse spécifique aux formants que nous venons de réaliser, nous nous intéressons à présent à la synthèse d'un signal de parole à partir des données articulatoires extraites.

Les formants estimés sont le résultat d'une détection de pics réalisée sur les réponses en fréquence (fonctions de transfert) obtenues à partir des fonctions d'aire, grâce à l'analogue électrique-acoustique utilisé. Ces mêmes fonctions de transfert sont maintenant utilisées en tant que filtres, dans l'optique d'un système source-filtre qui permet la production d'un son. On considère que les sons de parole sont produits par une source (voisement ou bruit de friction) qui passe à travers le filtre qui est la fonction de transfert du conduit vocal.

L'extension de l'utilisation du modèle source-filtre aux consonnes nécessite une reconsidération des conditions d'application de ce modèle. En effet, ce dernier est adapté pour les voyelles, la modélisation de la source mise en jeu n'est pas conçue pour la synthèse

de consonnes. Nous proposons alors d'introduire une modulation d'amplitude en 2 sous-bandes.

3.1. Synthèse du signal

La synthèse du signal fait appel à 3 éléments : une source, une modulation d'amplitude et une modulation fréquentielle. Cette dernière correspond au filtrage de la source par la fonction de transfert du conduit vocal estimée depuis le marquage géométrique. Ni la source, ni les variations d'amplitude ne peuvent être extraites directement depuis les contours du conduit vocal : le modèle acoustique utilisé n'est pas, en effet, construit pour rendre compte des variations d'amplitude et par définition du modèle, la source est ajoutée. Par contre, aussi bien la source que les variations d'amplitude peuvent être dérivées du signal audio d'origine.

3.1.1. Source

Pour faire de la synthèse ou du codage articulatoire, un modèle de production de parole doit intégrer la source [SS87].

Les sources sont :

- la vibration des cordes vocales. Pour les sons voisés, la source quasi-périodique se situe au niveau de la glotte.
- le bruit de friction lorsque la section du tube est très faible. Dans ce cas la source se situe au niveau de la constriction la plus étroite dans le conduit.
- le bruit des occlusives. Il est dû au relâchement soudain de la pression occasionnée par la fermeture du conduit. L'occlusion complète du conduit vocal s'accompagne d'un bref silence. Puis le conduit vocal s'ouvre et l'air sous pression derrière la constriction s'échappe en créant un bruit turbulent.

Un des objectifs de l'étude menée dans cette partie est de valider la qualité du contenu acoustique estimé à partir des fonctions d'aire, c'est-à-dire montrer que les réponses en fréquence obtenues par analogie électrique sont valables. C'est pourquoi nous cherchons à utiliser une source la plus blanchie possible, pour évaluer l'apport du filtre.

A priori un filtrage LPC inverse seul ne suffit pas pour réaliser cette opération : le signal soumis au simple filtrage inverse est intelligible. Nous proposons d'ajouter une seconde étape en utilisant la transformée de Hilbert.

La source considérée est obtenue à partir du signal original comme suit : on décompose le signal par un filtrage de Hilbert puis on filtre l'information temporelle fine avec un filtrage LPC inverse. Le schéma de blanchiment de la source est présenté Figure 101.

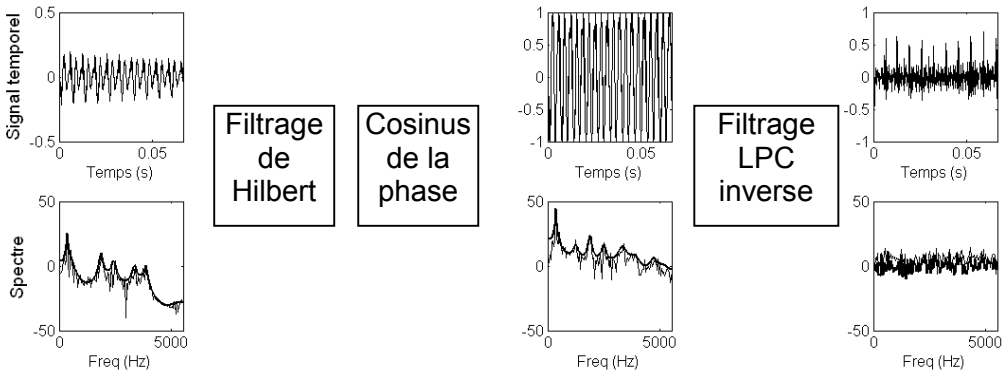


Figure 101 : Schéma de blanchiment de la source.

1/ La transformée de Hilbert correspond à un filtre passe-tout qui déphase toutes les composantes fréquentielles de –π/2. Sa fonction de transfert est $H(f) = -i * signe(f)$. C'est un filtre linéaire non causal de réponse impulsionnelle $h(t) = \frac{1}{\pi * t}$. Le filtre de Hilbert transforme un cosinus en sinus, c'est un quadrateur parfait.

La transformée de Hilbert est fondamentale en télécommunications et intervient largement dans les méthodes de modulation. En effet, appliquer cette transformée à un signal permet de décomposer le signal en son enveloppe (son amplitude, par le module du signal filtré) et en une partie plus fluctuante, la porteuse. En appliquant un filtre de Hilbert à notre signal audio d'origine et en ne conservant que le cosinus de la phase du signal ainsi filtré, nous éliminons l'enveloppe et ne conservons que la structure temporelle fine du signal.

2/ Cette porteuse de Hilbert est partiellement intelligible. C'est pourquoi nous la couplons avec un filtrage inverse LPC afin d'éliminer les pôles résiduels et se débarrasser de l'influence des formants.

La prédiction linéaire (LPC) est un modèle auto-régressif, c'est-à-dire qu'un échantillon de parole à un instant peut être estimé par une combinaison des p échantillons précédents. Les coefficients de cette combinaison (dits coefficients de prédiction et notés a_k), pour des courts segments de parole, sont tels qu'ils minimisent l'erreur quadratique moyenne de prédiction entre le signal prédit et le signal réel (i.e. la puissance moyenne du résidu). Nous ne détaillons pas les méthodes de calcul de ces coefficients. Plusieurs méthodes existent, par exemple, celle de l'autocorrélation qui fait appel à l'algorithme de Levinson-Durbin.

Le filtre LPC ainsi défini est un filtre tout pôle dont la fonction de transfert est :

$$H(z) = \frac{1}{A_p(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

$A_p(z)$ est le filtre inverse de la fonction de transfert.

Les coefficients du filtre $H(z)$ correspondent aux formants du signal, et si l'on inverse ce filtre, on annule l'effet des formants. Le filtre LPC que nous calculons est d'ordre 20.

La source que nous obtenons finalement après filtrage de Hilbert et filtrage inverse est écoutée et n'est pas intelligible. Nous en reparlerons plus loin avec la mise en place d'un test de perception. L'observation du spectre (en fin de chaîne, Fig. 101) et du spectrogramme de cette source blanchie (Fig. 102) montre l'absence de formants.

Par la suite, nous notons sb les signaux synthétisés à partir de cette source blanchie.

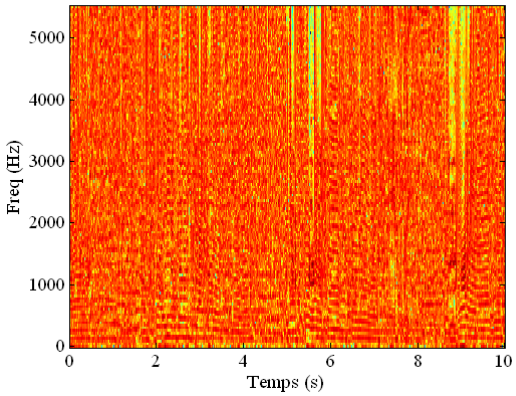


Figure 102 : Spectrogramme d'un segment de la source obtenue par filtrage de Hilbert et filtrage inverse LPC. La structure harmonique est préservée dans les périodes voisées.

3.1.2. Amplitude

Les signaux synthétisés sont obtenus en filtrant la source par la fonction de transfert du conduit vocal. Ce modèle source-filtre est appliqué trame par trame, chaque trame ayant une durée de 33,4 ms et étant associée à une fonction de transfert.

Le signal ainsi synthétisé est difficilement intelligible, comme nous en reparlerons plus loin. La fonction de transfert obtenue à partir de la fonction d'aire grâce au modèle acoustique n'estime pas l'amplitude du signal.

La prise en compte de la modulation d'amplitude est maintenant classique en perception, depuis les travaux de Shannon et al. [SZK+95], mais elle est peu utilisée en production.

Nous introduisons une modulation d'amplitude à 29,97 Hz, c'est-à-dire synchrone avec les trames vidéo. Les modulations d'amplitude correspondent aux fluctuations d'amplitude de l'enveloppe temporelle.

Le paradigme dit "d'appauvrissement" de la parole ou de parole réduite spectralement est utilisé, en perception, dans l'investigation du rôle de l'information temporelle, puisqu'il oblige le sujet à n'utiliser que l'information temporelle présente dans le signal. L'information spectrale est supprimée par des techniques de traitement du signal. Une des méthodes consiste à extraire l'enveloppe à travers un banc de filtres, au lieu de l'extraire en pleine bande. Ce banc de filtres passe-bande se recouvrant se rapproche du comportement du

système auditif périphérique qui est capable de décomposer les sons complexes en éléments plus simples et de les séparer partiellement. Le signal est découpé en plusieurs bandes de fréquences, l'enveloppe temporelle de chaque bande est extraite.

Dans notre étude, une décomposition du signal en 2 sous-bandes est réalisée. On utilise pour cela un banc de 2 filtres quasi-rectangulaires à gain unitaire (Fig. 103), définis en échelle Bark. Ils sont construits à partir du regroupement et de la sommation de filtres initiaux pondérés par une fenêtre de Hanning répartis aussi selon une échelle Bark. La coupure entre les 2 filtres est définie à 1000 Hz, elle divise le spectre entre le 1er et le 2ème formant, qu'on peut voir comme un partage BF (basses fréquences) - HF (hautes fréquences).

La décomposition en 2 sous-bandes a la propriété de prendre en compte les modes de production des consonnes [Gro05]. Dans la sous-bande HF, les modulations d'amplitude liées aux constrictions du conduit vocal (frications, bursts) sont présentes. En BF, nous observons les modulations d'amplitude produites par la glotte (barre de prévoisement des consonnes voisées). Pour les noyaux vocaliques, la modulation HF et BF est simultanée et une mesure de l'amplitude en pleine bande suffirait.

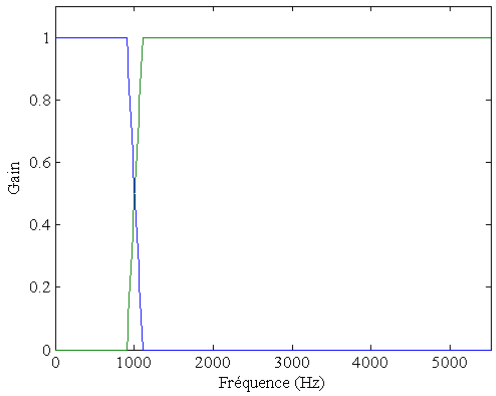


Figure 103 : Banc de filtres pour le calcul des amplitudes en 2 sous-bandes.

Le filtrage par ce banc de filtres du signal d'origine et du signal synthétisé permet de calculer les amplitudes en sous-bandes des 2 signaux. Comme on le voit dans la première ligne du tableau 15, les amplitudes des 2 signaux sont très peu corrélées (moins de 0,4 pour chacune des 2 sous-bandes).

L'amplitude est nécessaire à l'intelligibilité d'un signal, ainsi pour améliorer celle du signal synthétisé, on se propose de ré-affecter, dans la source, l'amplitude du signal d'origine, c'est-à-dire réintroduire dans la source la part d'information du signal d'origine relative aux variations d'amplitude des 2 sous-bandes. La source blanchie est alors modulée en amplitude, comme le signal d'origine.

Au final le signal synthétisé est obtenu en filtrant par la fonction de transfert du conduit la source modulée en amplitude en 2 sous-bandes. La corrélation temporelle des amplitudes en sous-bandes entre ce nouveau signal synthétisé et le signal de départ est alors très forte (tableau 15).

	1ère sous-bande	2ème sous-bande
Synthèse par filtrage par la fonction de transfert de la source blanchie	0,391	0,37
Synthèse par filtrage par la fonction de transfert de la source blanchie et modulée en amplitude	0,973	0,969

Table 15 : Coefficient de corrélation entre l'amplitude du signal original et celle du signal synthétisé, sans et avec modulation d'amplitude en 2 sous-bandes.

Pour résumer, nous avons ainsi un bon découplage de ce que nous pouvons estimer avec notre modèle acoustique et de ce que nous n'estimons pas. Du côté de la source, nous avons une porteuse modulée en amplitude en deux sous-bandes et du côté du conduit une modulation spectrale. Nous avons aussi montré que la porteuse non modulée ne contient pas d'information phonétique.

Avec cette décomposition, nous tenons compte du fait qu'une partie de l'intelligibilité résulte de la ré-affectation d'amplitude, soit des deux paramètres de modulation d'amplitude en sous-bandes à 30 Hz, et non pas de la source blanchie.

Le taux d'intelligibilité attribuable à la modulation d'amplitude en 2 sous-bandes est non négligeable [SZK+95], nous le verrons plus loin avec le test de perception.

Ceci met en évidence que la parole n'est pas uniquement formantique (spectrale). La synthèse articulatoire à partir du conduit vocal rend bien compte de la structure formantique, mais le modèle acoustique n'est pas conçu pour rendre compte de la modulation d'amplitude. D'où la nécessité de remettre une part de cette information pour l'intelligibilité du signal.

Par la suite, nous noterons avec un indice m les signaux synthétisés avec modulation d'amplitude (sb_m).

3.2. Analyse spectrale

L'évaluation de cette synthèse de signaux sera analysée de manière qualitative avec une notion d'intelligibilité, au §3.3.. Elle peut bien évidemment s'intéresser aux formants, comme nous l'avons déjà vu. Nous définissons ici un critère d'évaluation spectrale qui tient compte, au moins partiellement, de la décomposition que nous avons réalisée, dans le but de comparer quantitativement le signal synthétisé estimé au signal d'origine.

3.2.1. Paramétrisation

L'étude qui suit porte essentiellement sur une paramétrisation basée sur le codage prédictif linéaire ou LPC. Nous ne quantifierons pas les résultats directement dans le domaine fréquentiel. Pour chaque trame, nous considérons et comparons les spectres LPC des signaux (d'origine ou synthétisés). En effet, l'ajustement des paramètres du modèle LPC permet de déterminer à tout instant la fonction de transfert associée à la trame considérée. Cette fonction fournit une approximation de l'enveloppe du spectre du signal à l'instant d'analyse. La prédiction linéaire (LPC) a été décrite au §3.1.1..

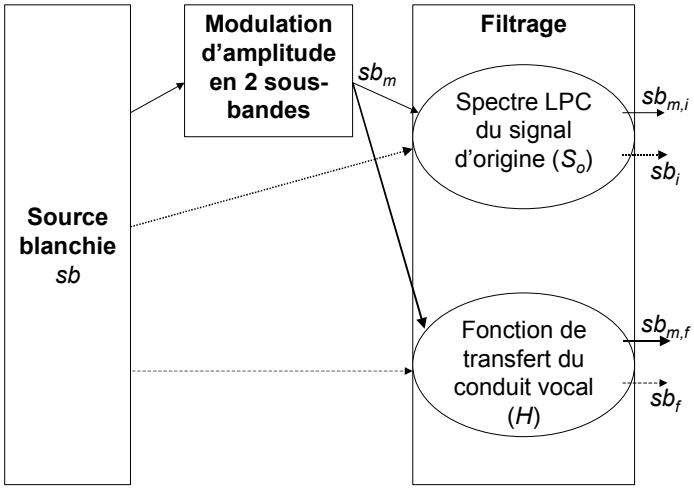


Figure 104 : Schéma-Bloc de synthèse des signaux.

Pour chaque trame, nous pouvons déterminer les spectres LPC d'ordre 15 :

- du signal d'origine (signal s_o, spectre LPC S_o)
- du signal d'origine synthétisé par filtrage, par le spectre LPC (S_o) du signal d'origine, de la source, avec modulation d'amplitude (signal $sb_{m,i}$, spectre LPC $Sb_{m,i}$) ou sans (signal sb_i, spectre LPC Sb_i).
- du signal estimé synthétisé par filtrage, par la fonction de transfert associée à la forme du conduit vocal (*H*), de la source, avec modulation d'amplitude (signal $sb_{m,f}$, spectre LPC $Sb_{m,f}$) ou sans (signal sb_f, spectre LPC Sb_f).

3.2.2. Distance spectrale

La comparaison entre les spectres LPC définis ci-dessus s'appuie sur une notion de distance spectrale, qui permet de montrer dans quelle mesure 2 spectres LPC s'ajustent entre eux. C'est une mesure de similarité entre 2 spectres. Elle permet de quantifier l'écart entre l'estimation et la référence en tenant compte des facteurs de fréquence et d'amplitude des

formants (ainsi que, dans une moindre mesure, de largeur de bande). Cette mesure ne se limite donc pas uniquement à la fréquence des pics du spectre LPC.

La mesure utilisée est la somme sur les échantillons de la différence de 2 spectres ([RJ93], page 158). Soit S_n et S_n' 2 spectres LPC pour une trame n, la distance spectrale considérée pour cette trame est :

$$d(S_n, S_n') = \frac{1}{p}\sum_{i=1}^{p}\left|S_n(i) - S_n'(i)\right|$$

Cette différence est appliquée sur les spectres exprimés en dB.

Les fonctions de transfert estimées à partir du conduit vocal présentent des pics, généralement 5. Lors de la comparaison des spectres d'origine et synthétisé, il est fréquent d'observer que les 3 premiers pics (formants) s'ajustent alors que les formants suivants ne sont plus similaires, en amplitude ou en fréquence. Sur l'exemple de la figure 105, l'ajustement est en fréquence mais pas en amplitude. Nous nous intéressons à la similitude des spectres essentiellement au niveau de la bande de fréquence propre aux 3 premiers formants et ne prenons donc en compte dans le calcul de d que la première partie du spectre, couvrant les fréquences de 0 à 3,5 KHz. En effet la fiabilité des mesures de formants au-delà de 4-5 KHz n'est pas très bonne, l'usage que nous faisons du modèle acoustique n'est pas adapté pour les formants 4 et 5.

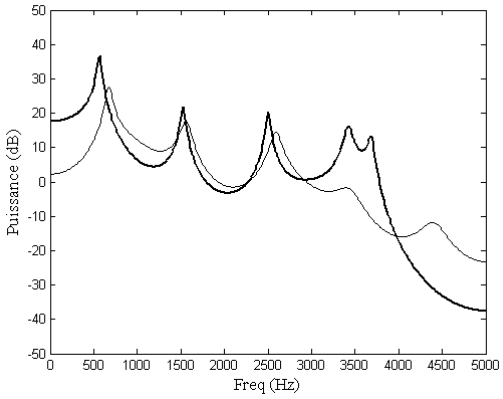


Figure 105 : Spectres LPC du signal original (trait gras) et du signal synthétisé à partir de la fonction de transfert estimée (trait fin) : les 2 spectres sont assez similaires jusqu'à 3KHz puis plus du tout ensuite.

En moyennant cette distance d sur l'ensemble des trames de la séquence, on obtient la grandeur $D(S,S')$, représentant la similitude spectrale des 2 signaux s et s' qu'on voulait comparer. Cette grandeur est exprimée en dB.

Plus cette grandeur est petite, plus les spectres comparés sont considérés comme proches ou similaires. A l'opposé, plus D est grande, plus les spectres sont éloignés.

Cette distance évalue de façon globale le décalage qu'il peut y avoir entre les fréquences des formants mais également les différences d'amplitude. Ces deux aspects combinés dans

une même grandeur rendent la distance D assez peu sensible. Cette distance est utilisée pour comparer des signaux, on s'intéresse à l'analyse des distances relativement entre elles plutôt qu'à celle de ces valeurs dans l'absolu.

Ces différentes mesures ont l'avantage d'être complémentaires par rapport aux 2 facteurs en présence (fréquence et amplitude). Elles permettent d'une part d'évaluer quantitativement l'apport de la modulation d'amplitude en 2 sous-bandes en comparant des signaux synthétisés par filtrage de la source par la même fonction de transfert, mais modulés ou non en amplitude, c'est le cas de la figure 106a. Les fréquences des 3 premiers formants des 2 signaux sont identiques, mais l'amplitude est différente. D'autre part, ces mesures permettent de quantifier la précision des formants estimés : sur la figure 106b, on observe le spectre d'un signal synthétisé en filtrant la source blanchie modulée en amplitude par la fonction de transfert du conduit vocal et celui d'un signal synthétisé en filtrant la source blanchie modulée par le spectre LPC du signal d'origine. On constate des écarts entre les fréquences formantiques.

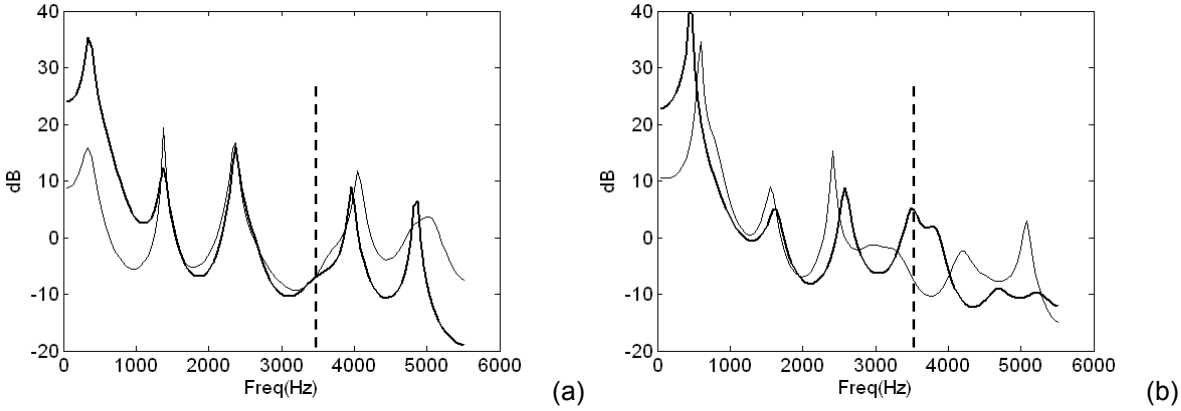


Figure 106 : Comparaison de spectres LPC (on se limite à la bande [0-3.5] Hz).
(a) Les 2 signaux ont été synthétisés à partir de la source blanchie par filtrage par la fonction de transfert du conduit vocal, le signal (en gras) a été modulé en amplitude en 2 sous-bandes, pas l'autre.
(b) Les 2 signaux ont été synthétisés à partir de la source blanchie modulée en amplitude par filtrage par la fonction de transfert du conduit vocal (en trait fin) ou par filtrage par le spectre LPC du signal d'origine (en trait gras).

En parallèle de cette distance spectrale, on s'intéresse à la corrélation entre les différents spectres LPC définis. Le coefficient de corrélation permet aussi de montrer la similitude, mais le calcul passe par une étape de centrage par rapport aux valeurs moyennes, qui a pour effet d'atténuer le facteur amplitude. C'est pourquoi cette mesure n'est pas suffisante, même si elle permet d'apporter quelques informations complémentaires. Cette mesure est elle aussi réalisée entre 0 et 3,5KHz sur les spectres LPC exprimés en dB.

3.2.3. Réglages des paramètres

La mise au point de la distance spectrale, comme mesure quantitative sur les spectres, permet de confirmer, à nouveau (après l'étude sur les formants), les choix que nous avons fait pour les paramètres α et β, le rapport pixels/cms et la position de la glotte. Pour cela, nous considérons les distances spectrales en dB entre le signal d'origine et le signal synthétisé en filtrant la source blanchie et modulée en amplitude par les fonctions de transfert du conduit vocal.

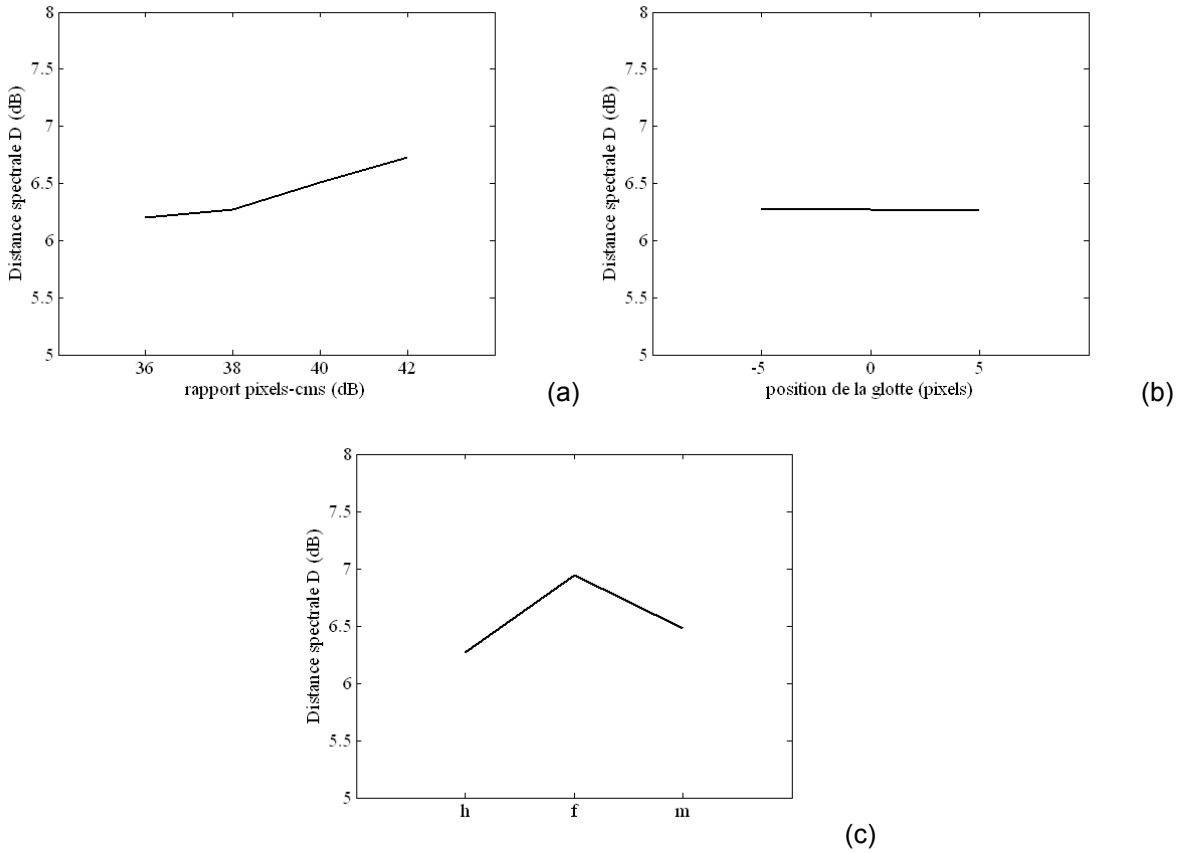


Figure 107 : Distances spectrales, sur la séquence complète pour différentes valeurs de paramètres, entre le signal d'origine et le signal synthétisé (à partir d'une source blanchie et modulée en amplitude en 2 sous-bandes, et filtrée par la fonction de transfert du conduit vocal) : (a) rapport pixels/cms – (b) position de la glotte – (c) paramètres α-β.

Sur le graphe 107a, on observe l'influence du rapport pixels-cms des images. La distance spectrale moyenne augmente lorsque ce rapport augmente. Cette différence graphique est confirmée statistiquement par une ANOVA à un facteur ($F=50,48$, $p<0.01$). 38 pixels/cm permet un écart spectral entre l'estimation et l'original plus faible, tout en conservant une longueur moyenne de conduit vocal acceptable.

Le graphe 107b montre l'influence de la position de la glotte sur la distance spectrale entre le signal synthétisé et le signal original. En faisant varier la position de la glotte de quelques pixels par rapport à la position définie, il n'y a que très peu de variations en terme de

distance spectrale. La différence n'est pas significative (F<1, au risque α=5%), la position de la glotte n'a pas besoin d'être modifiée.

Enfin, nous testons (Fig. 107c) les jeux de paramètres α-β définis précédemment avec la mesure de biais : nous comparons les résultats obtenus avec les paramètres du locuteur mâle (h) de Soquet et al. [SLMD02], ceux de la locutrice (f) et ceux obtenus en moyennant ces deux précédents jeux de paramètres (m). Les distances spectrales moyennes obtenues sont significativement différentes, comme le confirme une ANOVA à un facteur (F=91,91, p<0.01). L'utilisation des paramètres α-β de la locutrice augmente l'écart entre les spectres. Le choix fait d'utiliser les paramètres du locuteur mâle est confirmé.

3.2.4. Analyse des distances spectrales

Plusieurs mesures de distances spectrales sont menées de façon à comparer les signaux suivants :

Signaux de référence	s_o, S_o	Signal d'origine
	sb_i, Sb_i	Source blanchie et filtrée par le spectre LPC du signal d'origine par le spectre LPC du signal d'origine (S_o)
	$sb_{m,i}, Sb_{m,i}$	Source blanchie, modulée en amplitude en 2 sous-bandes et filtrée par le spectre LPC du signal d'origine (S_o)
Signaux estimés	sb_f, Sb_f	Source blanchie et filtrée par la fonction de transfert estimée du conduit vocal (H)
	$sb_{m,f}, Sb_{m,f}$	Source blanchie, modulée en amplitude et filtrée par la fonction de transfert estimée du conduit vocal (H)
	$sb_{m,fd}, Sb_{m,fd}$	Source blanchie, modulée en amplitude et filtrée par une fonction de transfert (H'), estimée du conduit vocal décalée de 30 trames

Nous comparons des signaux « de référence » avec des signaux « estimés ». Outre le signal de départ Laval43 (s_o), les signaux de référence sont obtenus par synthèse de la source blanchie, modulée ($sb_{m,i}$) ou non (sb_i) en amplitude en 2 sous-bandes, et filtrée par les spectres LPC du signal d'origine. Les signaux estimés sont synthétisés à partir de la même source blanchie, qui est modulée ($sb_{m,f}$, $sb_{m,fd}$) ou non (sb_f), et filtrée par la fonction de transfert du conduit vocal. L'indice d du signal $sb_{m,fd}$ signifie que la source blanchie et modulée a été filtrée par des fonctions de transfert, estimées à partir des fonctions d'aire mais avec un décalage ; le filtrage est réalisé avec des fonctions de transfert décalées de 30 trames (en avance) par rapport à la modulation d'amplitude qui est synchrone.

La première ligne du tableau 16 évalue l'écart spectral entre le spectre d'origine et les spectres des 3 signaux estimés.

Les 2 lignes suivantes du tableau comparent ces mêmes signaux synthétisés depuis le conduit vocal avec les signaux de référence synthétisés à partir des spectres LPC du signal

d'origine. Ceci permet de comparer les signaux en s'affranchissant de la source, dans la mesure où cette fois, la référence n'est pas directement le signal d'origine mais une version synthétisée du signal d'origine. Les distances mesurées sont alors plus faibles.

estimation / référence	Sb_f	$Sb_{m,f}$	$Sb_{m,fd}$
S_o	8,44 dB	6,27 dB	6,84 dB
$Sb_{m,i}$	-	5,27 dB	5,99 dB
Sb_i	7,29 dB	-	-

Table 16 : Distances spectrales D entre spectres LPC sur la séquence complète.

Ces mesures de distances spectrales ont été réalisées sur la séquence complète, sur la séquence sans silence ou sur les voyelles seules mais nous avons constaté que les résultats diffèrent de façon non significative. Les résultats présentés ici sont obtenus à partir de la séquence complète.

La distance spectrale D est plus grande lorsque la synthèse est réalisée sans modulation d'amplitude : on trouve une valeur de 8,44dB sans modulation et 6,27dB avec modulation d'amplitude en 2 sous-bandes. Cela complète ce qui a été dit précédemment concernant l'amplitude, celle-ci est nécessaire à la synthèse du signal. En décalant la fonction de transfert de quelques trames, on peut considérer que la modulation fréquentielle ou les formants ont été mal estimés et on constate que la distance D augmente un peu (comparaison entre les 2 dernières colonnes du tableau 16). On passe de 6,27dB à 6,84dB par rapport au signal d'origine et de 5,27dB à 5,99dB par rapport au signal synthétisé à partir du spectre LPC d'origine. Ces différences sont assez faibles mais nous verrons un peu plus loin que ce décalage implique une perte notable en terme d'intelligibilité.

Ces mêmes résultats sont confirmés avec les coefficients de corrélation calculés entre les spectres LPC du signal d'origine et des signaux estimés. La corrélation est bien meilleure avec la modulation d'amplitude en 2 sous-bandes, elle dépasse 0,7 alors qu'elle ne vaut que 0,34 sans modulation. En décalant la fonction de transfert de filtrage, on perd un peu en corrélation (0,68 au lieu de 0,73).

	Sb_f	$Sb_{m,f}$	$Sb_{m,fd}$
S_o	0,34	0,73	0,68

Table 17 : Coefficients de corrélation entre spectres LPC synthétisés et d'origine.

Les deux figures suivantes illustrent les deux analyses qui ont été menées, la première sur les formants, la seconde sur le signal complet par l'étude des spectres LPC.

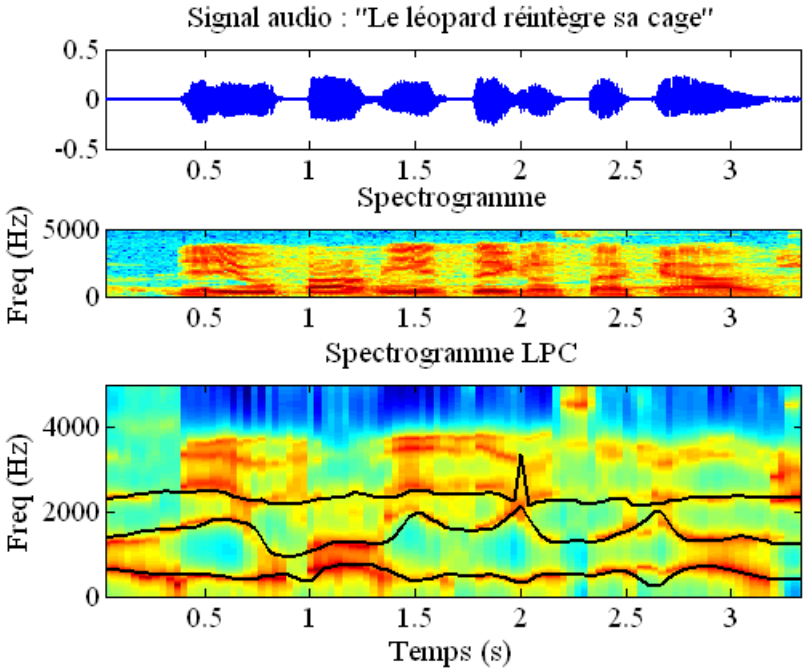


Figure 108 : Signal d'origine, spectrogramme, spectrogramme LPC et formants estimés (en traits noirs épais) à partir des fonctions d'aire extraites.

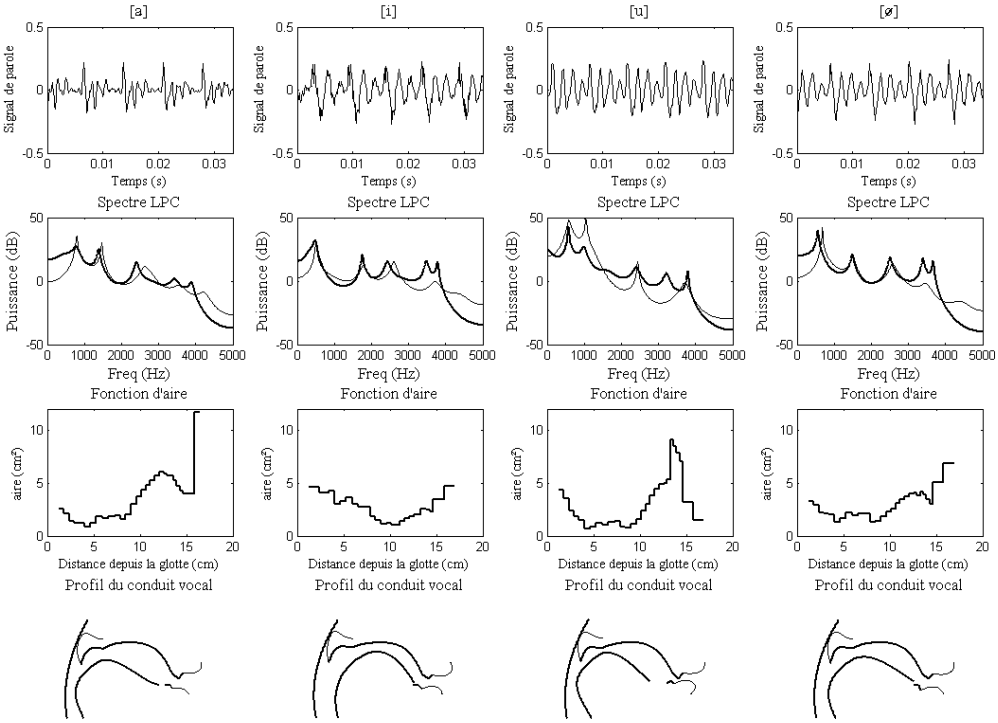


Figure 109 : Le spectre LPC (trait gras) obtenu à partir du signal original est comparé au spectre LPC du signal synthétisé depuis la fonction d'aire (trait fin). Le signal de parole, les fonctions d'aire et les profils de conduit vocal sont également donnés, à titre d'exemple, pour 4 trames étiquetées de voyelles du corpus, de gauche à droite [a], [i], [u] et [ø].

La figure 108 montre pour une phrase du corpus le spectrogramme LPC sur lequel on a superposé les trajectoires formantiques estimées. Mis à part quelques décrochages (observables par exemple en t=2s), on constate que ces trajectoires sont globalement bien synthétisées : les estimations suivent à peu près correctement les trajectoires d'origine.
La figure 109 montre 4 trames particulières de la séquence, il s'agit de 4 voyelles étiquetées du corpus. Pour chacune d'elles, on peut observer le contour géométrique extrait, la fonction d'aire alors calculée et le spectre LPC synthétisé superposé à celui d'origine.

3.3. Test de perception

Après cette analyse quantitative, un test perceptif a été mené pour évaluer globalement l'intelligibilité des signaux synthétisés.
Le corpus dont nous disposons (38 phrases) ne permet pas d'évaluer l'intelligibilité phonétique ou au niveau du mot. Le test mis en place est subjectif (on demande au sujet de juger s'il entend, plus ou moins, des phrases qui lui sont présentées sous forme de texte), il se rapproche de tests effectués en multimédia pour évaluer les qualités perçues de média compressés et transmis, connus sous le nom de MOS.
Le « Mean Opinion Score » (MOS), ou note d'opinion moyenne, exprime par une simple note entre 0 et 5 la qualité de la restitution sonore d'un codec audio (0 pour la plus mauvaise et 5 pour la meilleure, comparable à la version d'origine).
La méthode du score MOS a été envisagée pour évaluer la qualité perçue d'un codec audio et pour comparer les qualités respectives de différents codecs[8] (souvent basés sur des algorithmes de compression radicalement distincts). En effet, face à cette variété d'algorithmes et de résultats produits et avec la remise en cause de l'utilisation de l'erreur quadratique entre signal original et signal codé-décodé, le besoin d'une mesure prenant en compte les effets psycho-acoustiques a été ressenti. Le principe de calcul du MOS est basé sur le sondage d'un échantillon supposé représentatif de la population des utilisateurs. Ces personnes écoutent un signal (souvent de la voix), puis son équivalent codé-décodé. Après chaque écoute, l'auditeur donne une note correspondant à la qualité qu'il a perçue. La moyenne des notes fournies constitue le MOS.
Sur ce même principe, nous voulons comparer, d'un point de vue perceptif, plusieurs types de nos signaux synthétisés. L'objectif est de quantifier la qualité de la synthèse effectivement perçue par un échantillon d'auditeurs. L'expérience perceptive mise en place a pour but d'analyser l'intelligibilité de ces signaux selon une échelle à 8 niveaux définie comme suit (Table 18). Tout en restant subjective, cette échelle diffère de celle utilisée pour un score

[8] Il a aussi été proposé une relation entre le MOS et une mesure de distance spectrale pour le codage GSM [RFR03].

MOS puisqu'elle repose, d'une certaine façon, sur un critère plus stable. La variabilité entre sujets est moins importante lorsqu'il s'agit de choisir entre « je comprends quelques mots » ou « je comprends la phrase » que lorsqu'il faut attribuer une note qui juge de la qualité de l'échantillon perçu. Ceci permet, pour notre expérience, de limiter le nombre de sujets auditeurs.

0	je n'entends que du bruit
1	j'entends des voyelles
2	j'entends quelques syllabes
3	je comprends quelques mots
4	je comprends la phrase avec effort
5	la phrase est intelligible mais je suis gêné par des distorsions
6	je comprends la phrase sans effort et sans distorsions gênantes
7	écoute parfaite sans distorsion

Table 18 : Echelle utilisée pour évaluer le niveau d'intelligibilité des stimuli.

3.3.1. Stimuli

Les stimuli utilisés pour ce test sont au nombre de 99. Il s'agit de 11 phrases extraites uniformément du corpus Laval43, sous 9 formes différentes de synthèse.

Les phrases et les 9 types de signaux considérés et évalués au cours de ce test de perception sont présentés dans les tableaux qui suivent (Tables 19 et 20). La plupart de ces signaux ont déjà été présentés, lors de l'analyse de la distance spectrale. Nous ajoutons à l'analyse l'utilisation d'un bruit blanc gaussien, échantillonné à 11025Hz, en tant que source. Le bruit blanc est utilisé de façon analogue à la source blanchie, il peut être modulé ou non en amplitude en 2 sous-bandes et filtré par la fonction de transfert estimée du conduit vocal. Les signaux synthétisés à partir du bruit blanc sont notés b.

1	Le léopard réintègre sa cage
2	La trahison frappa les truands
3	Le co-auteur est anéanti
4	C'est une clarté extraordinaire
5	Il a obtenu un mets infecte
6	Les documents indiens coïncident
7	En coopération on est frères
8	Louis est un enfant orgueilleux
9	Il est dans un état euphorique
10	Le cadet emporta un ballon
11	J'aimais obéir à mes parents

Table 19 : Corpus de phrases utilisées pour le test d'intelligibilité.

b_m	Source bruit blanc modulée en amplitude
b_f	Source bruit blanc filtrée par la fonction de transfert estimée du conduit vocal (H)
$b_{m,f}$	Source bruit blanc modulée en amplitude et filtrée par la fonction de transfert estimée du conduit vocal (H)
sb	Source blanchie
sb_f	Source blanchie et filtrée par la fonction de transfert estimée du conduit vocal (H)
sb_m	Source blanchie et modulée en amplitude
$sb_{m,f}$	Source blanchie, modulée en amplitude et filtrée par la fonction de transfert estimée du conduit vocal (H)
$sb_{m,i}$	Source blanchie, modulée en amplitude et filtrée par le spectre LPC du signal d'origine (S_o)
$sb_{m,fd}$	Source blanchie, modulée en amplitude et filtrée par une fonction de transfert (H'), estimée du conduit vocal mais correspondant à une configuration d'une autre trame (décalage de 30 trames)

Table 20 : Signaux considérés pour le test d'intelligibilité.

L'idée est de montrer l'influence ou non sur l'intelligibilité de plusieurs facteurs :

- Source : bruit blanc gaussien ou source blanchie
- Modulation d'amplitude en 2 sous-bandes à 29,97 Hz : avec ou sans
- Filtrage : sans, avec fonction de transfert estimée du conduit vocal ou avec spectre d'origine.

Au cours de cette expérience, les stimuli des phrases, claires, directement extraite du signal d'origine ne sont pas présentés. Tous les stimuli ont été préalablement égalisés en amplitude.

Chaque phrase est entendue une seule fois, il n'est pas possible de la ré-écouter ni de revenir en arrière. Le sujet décide lui-même quand il veut passer à la phrase suivante. Au moment de l'écoute, la phrase entendue s'affiche à l'écran. Les sujets doivent apprécier le degré d'intelligibilité de chaque phrase présentée et donner leur avis selon l'échelle présentée au tableau 18, en cochant les cases d'un tableau.

Avant de commencer le test, chaque sujet écoute quelques phrases significatives, de façon à se familiariser avec le style de stimuli.

3.3.2. Résultats

Cinq sujets, de langue française maternelle, ont passé le test.

Le graphe 110 présente les résultats moyens obtenus (ainsi que l'écart-type suivant les sujets) pour les différents types de signaux testés.

On constate qu'avec la modulation d'amplitude, l'utilisation d'une source blanchie permet une meilleure intelligibilité que le bruit blanc. Un bruit blanc modulé (b_m) est perçu comme du

bruit (niveau moyen d'intelligibilité de 0,5) alors qu'une source blanchie modulée en amplitude (sb_m) permet de comprendre la phrase, même si cela nécessite des efforts.

La donnée temporelle d'amplitude, comme on l'espérait, augmente l'intelligibilité. Alors que la simple source blanchie (sb) ne permet de reconnaître que quelques syllabes (niveau moyen de 2,4) dans le signal, cette même source modulée en amplitude (sb_m) devient intelligible (score de 4).

La modulation fréquentielle, introduite grâce au filtrage, améliore l'intelligibilité ; elle permet de comprendre quelques mots (niveau 3) dans le cas d'une source bruit blanc modulée en amplitude ($Bb_{m,f}$). On constate aussi son importance avec le cas du signal filtré par une fonction de transfert décalée, c'est-à-dire une fonction de transfert estimée mais pour une autre trame, éloignée, du signal. Dans ce cas, l'intelligibilité est fortement réduite, à quelques voyelles ou syllabes seulement (score inférieur à 2).

Le filtrage de la source blanchie modulée par le filtre LPC du signal d'origine permet une compréhension quasi-parfaite des phrases (niveau d'intelligibilité moyen de 5,78).

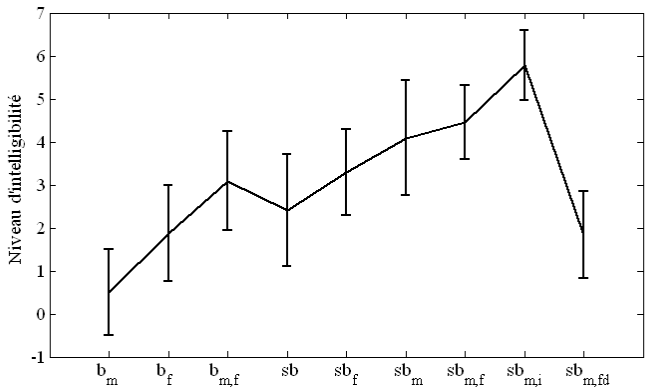


Figure 110 : Niveaux d'intelligibilité (MOS) moyens pour les 5 sujets pour les différents types de signaux testés et écart-type suivant les sujets.

L'observation graphique des résultats a été complétée d'analyses statistiques et notamment d'ANOVA.

Pour les ANOVA, nous laissons de côté les stimuli correspondant au signal $sb_{m,fd}$, c'est-à-dire la source blanchie, modulée en amplitude et filtrée par une fonction de transfert correspondant à une configuration décalée (de 30 trames). L'analyse de cette condition se fera séparément avec un test T et une correction de Bonferroni.

Nous cherchons à montrer l'influence relative de plusieurs facteurs (source, amplitude, filtrage) auxquels s'ajoutent le facteur sujets et le facteur phrases, sur les réponses au test de perception. Le facteur source comporte 2 conditions, source blanchie ou bruit blanc. Le facteur modulation comporte aussi 2 conditions: avec modulation ou sans. Le facteur filtrage en comporte 3 : sans filtrage, avec filtrage estimé, avec filtrage LPC d'origine.

A cause du facteur filtrage, l'analyse n'est pas équilibrée car toutes les combinaisons ne sont pas représentées. Par exemple, nous n'avons pas pris en compte la condition de filtrage par le spectre LPC du signal d'origine du bruit blanc modulé en amplitude, $b_{m,i}$. Ces conditions manquantes ne permettent donc pas de réaliser une ANOVA à 5 facteurs incluant le facteur filtrage. Nous effectuons une ANOVA à 4 facteurs (sujets/phrases/source/modulation). Le facteur filtrage sera traité indépendamment. Les résultats de l'ANOVA à 4 facteurs sont résumés dans le tableau 21.

Facteurs	ddl	F	p	Significatif
Sujets	4	6,16	0,0001	**
Phrases	10	1,46	0,1553	
Source	1	177,75	0,000	**
Modulation	1	39,34	0,000	**
Sujets*Phrases	40	0,33	1	
Sujets*Source	4	2,4	0,0514	*
Sujets*Modulation	4	3,18	0,0145	*
Phrases*Source	10	0,68	0,7442	
Phrases*Modulation	10	0,73	0,6932	
Source*Modulation	1	45,74	0,000	**
Sujets*Phrases*Source	40	0,3	1	
Sujets*Phrases*Modulation	40	0,39	0,9997	
Sujets*Source*Modulation	4	5,21	0,0005	**
Phrases*Source*Modulation	10	0,32	0,9756	
Sujets*Phrases*Source*Modulation	40	0,33	1	

Table 21 : ANOVA réalisée sur les réponses au test d'intelligibilité. Les facteurs pris en compte sont les phrases, les sujets, le type de source (bruit blanc ou source blanchie) et la présence ou non de la modulation d'amplitude.

Nous vérifions dans un premier temps que le facteur phrases n'est pas significatif. Que ce facteur soit considéré seul ou en interaction avec les autres, la valeur de F est petite, la valeur de p est toujours supérieure à 0,15.

Le facteur sujet est significatif ($F=6,16$ $p<10^{-4}$). Mais en combinant le facteur sujet avec les autres facteurs, on constate que les interactions sujet/source et sujet/modulation sont faiblement significatives (respectivement p=5% et p=1%). Dans le cas des stimuli de la source bruit blanc modulée en amplitude, 4 sujets sur 5 ont considéré qu'ils n'entendaient que du bruit (niveau 0) alors que le dernier a compris quelques syllabes, voire quelques mots (niveau 2,5). Ces signaux sont comparables à ceux utilisés par Shannon [SZK+95], mais ils ne sont pas ou peu intelligibles, dans un contexte expérimental où il sont présentés avec d'autres signaux synthétisés avec la source d'origine. Sur tous les autres stimuli, les différences entre les sujets sont moins importantes.

Les facteurs source et modulation donnent des différences très significatives ainsi que leurs interactions, avec des effets de potentialisation entre la source et la modulation d'amplitude. On observe par exemple une grosse différence sur le graphe 110 entre sb_m et b_m. Leurs

interactions avec le facteur phrases ne sont pas du tout significatives et très peu avec le facteur sujets.

Concernant le filtrage, il n'est pas possible de réaliser des ANOVA à 2 facteurs en faisant interagir le filtrage avec les facteurs modulation ou source, par manque de stimuli testés au cours du test (par exemple, $b_{m,i}$).

Une ANOVA à 3 facteurs (sujets, phrases, filtrage), ne tenant ainsi compte que de la condition filtrage, est réalisée (Table 22) et permet de mettre à nouveau en évidence que les différences observées sur les phrases ne sont pas significatives. Le facteur sujets, seul, est significatif mais pas l'interaction entre les facteurs sujets et filtrage. Le facteur filtrage est très significatif (F=78,93, p<<0.1%).

Facteurs	ddl	F	p	Significatif
Sujets	4	3,69	0,0061	**
Phrases	10	0,76	0,6669	
Filtrage	2	78,93	0,000	**
Sujets*Phrases	40	0,13	1	
Sujets*Filtrage	8	0,8	0,6042	
Phrases*Filtrage	20	0,18	1	
Sujets*Phrases*Filtrage	80	0,15	1	

Table 22 : ANOVA réalisée sur les réponses au test d'intelligibilité. Les facteurs pris en compte sont les phrases, les sujets et le filtrage.

Des tests de Student, avec correction de Bonferroni[9] sont également effectués pour comparer les différents filtrages.

Nous comparons les stimuli $sb_{m,i}$ et $sb_{m,f}$. La différence est très significative ($p\sim10^{-13}$). Les stimuli avec filtrage LPC d'origine sont perçus par les sujets sans effort et sans distorsions gênantes (niveau moyen d'intelligibilité de 5,78) alors que le filtrage par la fonction de transfert est perçu par les sujets avec des distorsions gênantes, même si les phrases sont intelligibles (niveau moyen de 4,43).

Nous comparons également les stimuli $sb_{m,f}$ et $sb_{m,fd}$. La différence est encore plus significative ($p\sim10^{-27}$). En décalant de quelques trames la fonction de transfert utilisée pour filtrer la source, on filtre avec de « fausses » estimations de formants, et on décorrèle la modulation spectrale et la modulation d'amplitude qui, elle, est réalisée de manière synchrone. Les stimuli ne sont pas intelligibles : seules quelques syllabes sont entendues par les sujets, le niveau moyen d'intelligibilité est de 1,8. Il s'agit en quelque sorte d'un témoin négatif (qui amène le résultat attendu) pour montrer que l'estimation réalisée de la modulation spectrale à partir du conduit vocal contient une information importante. Même si

[9] Cette correction revient à diviser le seuil d'erreur par le nombre de tests effectués. Ainsi, si on considère un seuil de 5%, la correction revient à comparer la valeur *p* du test au seuil 0,05/55 (le nombre de tests correspond au nombre de phrases (11) multiplié par le nombre de sujets (5)).

cette estimation n'est pas semblable à l'original, elle s'en approche et permet de restituer l'intelligibilité du signal. L'estimation fréquentielle des formants peut ainsi être validée, d'un point de vue perceptif.

Les résultats obtenus et présentés sur le graphe 110 mettent en évidence l'apport des 2 composantes, d'amplitude et de fréquence. Les niveaux d'intelligibilité moyens sont proches pour (sb_f) et (sb_m), en notant que ces niveaux seraient quasiment identiques si on comparait (sb_m) et (sb_i), obtenu par filtrage avec le spectre LPC du signal d'origine. Ces 2 composantes apportent une part similaire d'information.

3.4. Conclusion

La synthèse articulatoire consiste à décrire le phénomène de production de la parole, en partant d'une connaissance sur le mouvement des articulateurs du conduit vocal et d'un modèle acoustique. Nous avons intégré ici les informations géométriques dont nous disposions, avec un modèle de source, pour tenter de valider nos mesures.

Par des techniques de traitement du signal, nous extrapolons le modèle source-filtre [BF84] adapté pour les voyelles. En introduisant la modulation d'amplitude à 2 sous-bandes à 29,97 Hz, nous complétons le modèle de source pour aboutir à la synthèse d'un signal de parole.

La comparaison des spectres de fréquences du signal ainsi synthétisé et celui du signal original, l'observation des trajectoires formantiques et le test d'intelligibilité permettent de montrer qu'à partir des données géométriques, dans un contexte dynamique, nous sommes en mesure de synthétiser un signal de parole intelligible et proche de l'original, malgré le biais observé sur l'estimation des formants.

D'un point de vue temporel maintenant, la cinéradiographie et l'extraction des mouvements de la langue, et de la pointe en particulier, permettent la détection des moments de contact de la langue, que l'on met en parallèle de la production des consonnes.

CHAPITRE 8 : ETUDE DES CONSONNES DE LAVAL43

Etant donnée la nature dynamique de nos données, nous nous intéressons aux consonnes du corpus Laval43. L'étude présentée ici est une observation directe et limitée sur les données géométriques. Elle consiste à mettre en correspondance les données géométriques avec les moments consonantiques, en s'appuyant sur le lieu d'articulation des consonnes.
Le point d'articulation (ou lieu d'articulation) désigne l'endroit où s'effectue l'obstruction au passage de l'air (occlusion ou constriction). On distingue plusieurs points d'articulation, chacun pouvant être subdivisé. Nous avons essentiellement utilisé la classification suivante pour les consonnes de notre corpus :

- des consonnes labiales, ce sont les lèvres qui bloquent le passage de l'air : on distingue les bilabiales (rapprochement entre les 2 lèvres) et les labio-dentales (rapprochement entre la lèvre inférieure et les dents de la mâchoire supérieure)
- des consonnes alvéolaires, la langue vient appuyer sur la partie avant du palais
- des consonnes dorsales (palatales, vélaires ou uvulaires), qui sont articulées avec le dos de la langue. Le point d'articulation est situé vers l'arrière de la bouche.

Tout comme les voyelles, un étiquetage du signal audio Laval43 a été réalisé de façon à disposer d'instants consonantiques associés à des images de la base.
L'étiquetage réalisé sur les consonnes consiste à signaler les instants de production des diverses consonnes, en utilisant l'interface de marquage audio, identique à celle utilisée pour les voyelles et présentée en annexe A4. Les images de la séquence Laval43 sont disponibles à la cadence de 29,97 images par seconde ; aussi, dans la mesure où nous associons un instant consonantique à une image, chaque instant étiqueté a une durée de 33,4 ms. Les analyses qui vont suivre pourront allonger ces instants avec des fenêtres de 5 trames (5 images), centrées sur l'image étiquetée.
Tous les instants consonantiques n'ont pas été étiquetés, mais une grande partie l'a été de façon à avoir plusieurs exemplaires de chaque consonne. Nous avons ainsi à disposition, suite à cet étiquetage, le corpus de consonnes décrit dans le tableau 23.

Lieu d'articulation	Mode d'articulation	Consonnes	Nombre	
Bilabiales	Nasales	[m]	20	
	Plosives	[p]	20	30
		[b]	10	
Labio-dentales	Fricatives	[f]	16	30
		[v]	14	
Alvéolaires	Plosives	[t]	34	54
		[d]	20	
	Nasales	[n]	16	
	Fricatives	[s]	20	38
		[z]	10	
		[ʃ]	3	
		[ʒ]	5	
	Liquides	[l]	51	
Dorsales (palatales et vélaires)	Plosives	[k]	24	28
		[g]	4	
	Nasales	[ɲ]	1	
Uvulaires	Fricatives	[ʁ]	58	

Table 23 : Consonnes sélectionnées dans le corpus de Laval43 et classées par lieux et modes d'articulation.

Toutes les consonnes n'ont pas été étudiées au même titre. Les consonnes alvéolaires et palatales (au sens large) ont été analysées conjointement en terme de minimum de constriction entre la langue et le palais. Comme on le détaillera plus loin, cette observation minutieuse est rendue possible grâce notamment au suivi de la pointe de la langue, qui a été réalisé avec soin. Ensuite, une étude détaillée a été réalisée avec les consonnes bilabiales, permettant la mise en évidence de leur réalisation pour des minima d'ouverture aux lèvres.

Certaines consonnes n'ont pas été traitées, comme les [f] et les [v], qui sont des consonnes labio-dentales. Les consonnes [ʁ] n'ont pas du tout été étudiées car elles apparaissent le plus souvent dans des structures syllabiques de type CCV, plus complexes à étudier.

Une analyse propre aux nasales sera décrite plus tard : productions de voyelles et consonnes nasales seront mises en correspondance avec la position du voile du palais.

Les analyses qui suivent n'ont pas pour but de montrer de nouveaux résultats en phonétique, mais de valider le marquage géométrique qui a été réalisé à l'aide de notre méthode semi-automatique. En particulier, l'attention spécifique portée à la pointe pour la séquence Laval43 est ici mise en avant pour une meilleure précision de ces résultats. Nous cherchons à vérifier que pour un instant consonantique connu et étiqueté, nous observons une configuration géométrique plausible.

1. Consonnes alvéolaires et palatales - Etude des points de contact

Les consonnes alvéolaires désignent des consonnes apicales, c'est-à-dire articulée avec la pointe de la langue. La constriction est obtenue avec la langue. Leur lieu d'articulation se situe au niveau des alvéoles des dents de la mâchoire supérieure ou entre les alvéoles de la mâchoire supérieure et le palais dur pour les post-alvéolaires (nous ne ferons pas de distinction ici). Le français comporte les alvéolaires (ou post-alvéolaires) fricatives [s], [z], [ʃ] et [ʒ], plosives [t] et [d], nasales [n] et liquides [l]. Nous disposons de toutes ces consonnes dans notre corpus Laval43.

Une consonne dorsale est une consonne articulée avec le dos de la langue. Il s'agit des consonnes dont le lieu d'articulation est situé vers l'arrière de la bouche. Nous considérons les consonnes palatales et les consonnes vélaires. Le lieu d'articulation des premières est situé sur la partie supérieure du palais dite « palais dur ». En français, il n'y a que 2 consonnes palatales [j] et [ɲ]. Nous n'avons pas étiqueté la première et une seule fois la seconde. Les consonnes vélaires sont réalisées par un bombement de la partie postérieure de la langue qui se rapproche du palais mou. Le français et le corpus comportent les vélaires [k] et [g].

Dans ce paragraphe, nous étudions conjointement les consonnes dorsales et alvéolaires, car ce sont des consonnes constrictives, c'est-à-dire que leur articulation implique une obstruction plus ou moins complète du chenal respiratoire en un point d'articulation donné.

Lorsque l'obstruction est complète, la consonne est momentanée et se manifeste par un relâchement subit de l'air : c'est une plosive. Lorsque, l'obstruction est incomplète et si elle fait intervenir un écoulement turbulent, c'est une fricative.

Aussi la mise en correspondance des données articulatoires extraites des images radiographiques et de ces instants consonantiques s'appuie sur l'observation de la constriction produite par la langue contre le palais. Nous nous intéressons aux points de contact de la langue (pointe ou dos) avec la ligne du palais.

1.1. Méthode d'analyse

Les données géométriques extraites permettent pour chaque image de disposer du contour du conduit vocal, et en particulier des contours de la langue et du palais. Ces contours ont permis, comme cela a été présenté précédemment, de calculer des sections dans le plan médio-sagittal. Comme décrit au chapitre 4, nous avons fait les améliorations nécessaires pour considérer qu'à ce niveau de l'étude, nous avons à notre disposition pour chaque image de la séquence un jeu de 28 sections décrivant correctement la configuration du conduit vocal. Nous nous appuyons sur ce découpage en sections pour définir le lieu d'articulation, ou lieu de constriction, qui sera défini alors pour chaque image par un numéro de section.
Sur ces 28 sections, une représente l'ouverture aux incisives et une autre l'ouverture aux lèvres, nous les laissons de côté. Cette partie s'intéresse aux points de contact de la langue avec le palais, nous n'avons donc pas besoin de l'ensemble des sections, mais uniquement de celles qui décrivent le palais. Seules 14 sections nous intéressent, celles représentées en blanc sur la figure qui suit. Nous les numérotons de 1 à 14, en partant de la section la plus proche de la pointe.

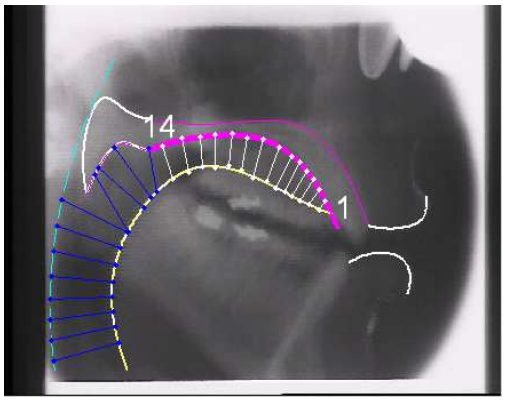


Figure 111 : Sections (en blanc) prises en compte pour le calcul du minimum de constriction entre la langue et le palais.

De ces 14 sections, il est alors possible d'extraire pour chaque image, la position et la taille du minimum de constriction entre la langue et le palais. Il s'agit de la section du conduit vocal, qui a la plus petite distance sagittale. Cette distance correspond à la taille de la constriction et la position de la constriction est définie par le numéro de la section en question.

Chercher à définir des seuils pour la taille ou le point de constriction, de façon à prédire le type de consonnes, n'est pas envisageable sur la base complète. Les voyelles sont aussi concernées par ce minimum de constriction, comme les [i] et [y], et ces seuils ne peuvent donc servir de critère de discrimination des consonnes sur la séquence. C'est pourquoi la mise en parallèle des données acoustiques et articulatoires se limite ici à observer le lieu et

la taille de la constriction uniquement pour les instants consonantiques des alvéolaires et des dorsales. Etant donnée la position connue des instants de ces consonnes, nous cherchons à mettre en évidence que pour chacun de ces instants, le lieu de constriction, correspondant au lieu d'articulation, est en accord avec la consonne considérée.

Les sections calculées (Fig. 111) ne permettent pas directement d'avoir la distance exacte entre la pointe de la langue et du palais. En effet, comme on l'a montré, ni la grille fixe pour le calcul des sections, ni les corrections apportées image par image ne rendent compte directement de la position exacte de la pointe. Cette donnée est cependant importante pour le calcul de la constriction. On le voit bien sur la figure précédente : il s'agit de la configuration d'un [t], le contact est correctement marqué entre la pointe et le palais, mais les sections sont positionnées trop en arrière et ne parviennent pas à mettre en évidence cette constriction complète. Une étape supplémentaire a donc été implémentée pour extraire une nouvelle information des contours géométriques estimés par notre méthode de rétro-marquage. Il s'agit pour chaque image de calculer la distance de la pointe de la langue au palais en partant du point à 2 degrés de liberté marqué et estimé pour la pointe. Le calcul revient à définir la droite normale au palais passant par ce point et à mesurer la distance entre la pointe et le palais le long de cette droite. Deux exemples sont présentés sur la figure 112, cette mesure complémentaire à celle réalisée à partir des 14 sections précédemment définies améliore la précision de la recherche du minimum de constriction entre la langue et le palais.

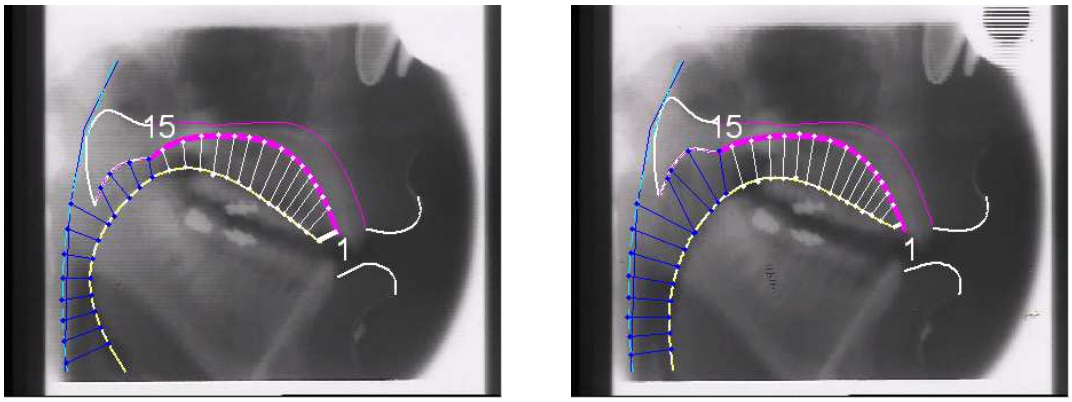


Figure 112 : Section supplémentaire (en trait gras et blanc) prise en compte pour le calcul du minimum de constriction entre la pointe de la langue et le palais. Cette mesure supplémentaire permet d'affiner la mesure de la taille de constriction entre la pointe et le palais.

Pour chaque image, le minimum de constriction est désormais évalué parmi un jeu de sections constitué de la distance entre la pointe de la langue et le palais et des 14 sections précédemment mentionnées. Nous numérotons ces sections de 1 à 15, la section 1

correspondant à la pointe de la langue et la section 15 correspondant à la section la plus en arrière, comme on le voit sur la figure 112.

1.2. Résultats

A partir de cette méthode d'analyse, nous obtenons les résultats qui suivent.

Les histogrammes de la figure 113 présentent les lieux d'articulation détectés pour les consonnes dorsales (à gauche) et alvéolaires (à droite).

Les consonnes dorsales sont toutes détectées comme postérieures, puisque leurs lieux de constriction sont détectés entre les sections 11 et 15. Le marquage géométrique indique qu'elles sont articulées vers l'arrière, ce qui est cohérent avec ce qui a été étiqueté d'après le signal audio.

Les consonnes alvéolaires sont très majoritairement détectées comme apicales ; c'est au niveau de la pointe de la langue (section numéro 1) que la constriction est détectée, mais on constate aussi quelques erreurs de détection. En effet, un petit nombre de consonnes étiquetées comme alvéolaires d'après l'audio sont détectées comme postérieures à partir des données géométriques, il s'agit du « résidu » aux sections 11 à 15 que l'on observe sur la figure 113b.

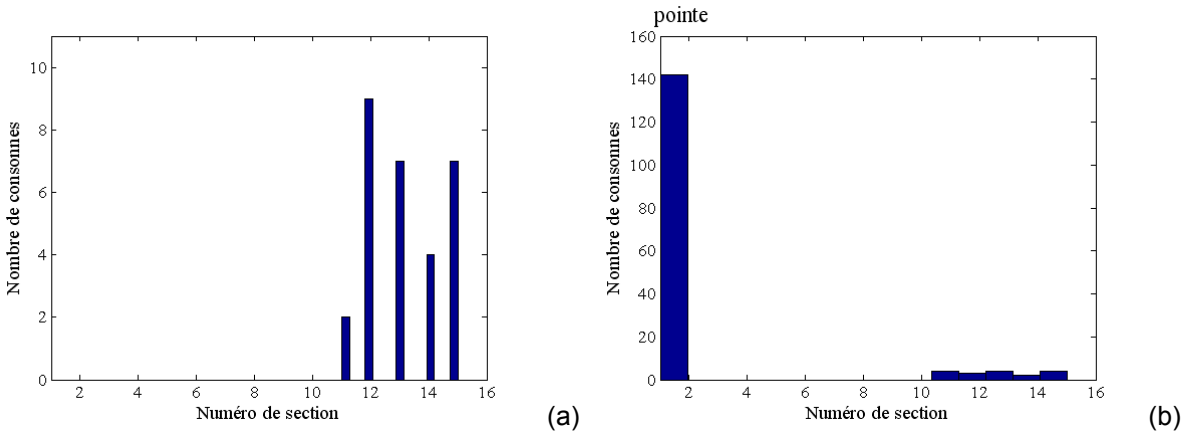
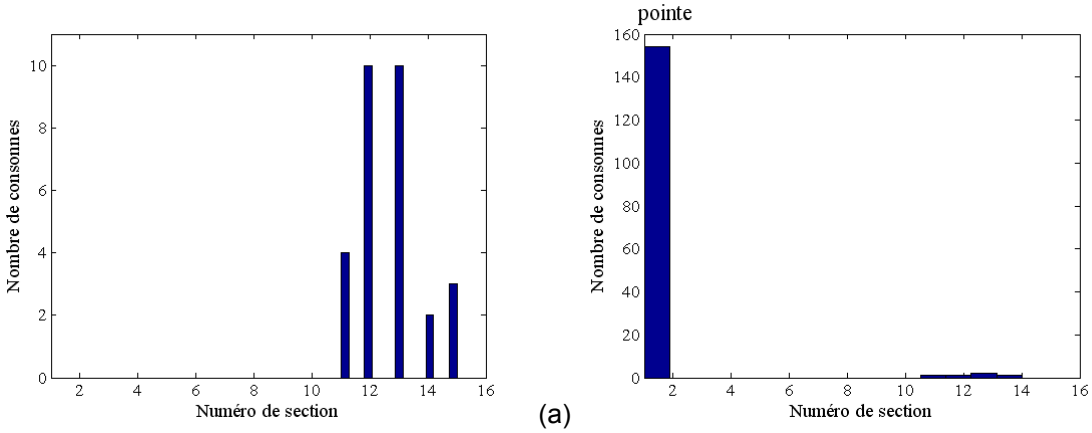


Figure 113 : (a) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes dorsales du corpus Laval43.
(b) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes alvéolaires du corpus Laval43.

Ce premier résultat a été obtenu en considérant les instants étiquetés sans opération de fenêtrage. Pour compenser d'éventuelles erreurs d'étiquetage et pour prendre en compte une certaine durée pour la consonne, il est intéressant de considérer, non pas la trame étiquetée, mais une fenêtre de 5 images autour de cette trame, c'est-à-dire 2 images de part et d'autre de l'image en question. Ainsi, pour chaque instant consonantique étiqueté comme alvéolaire ou dorsal, on recherche le minimum de constriction parmi les sections et parmi les 5 trames de la fenêtre.

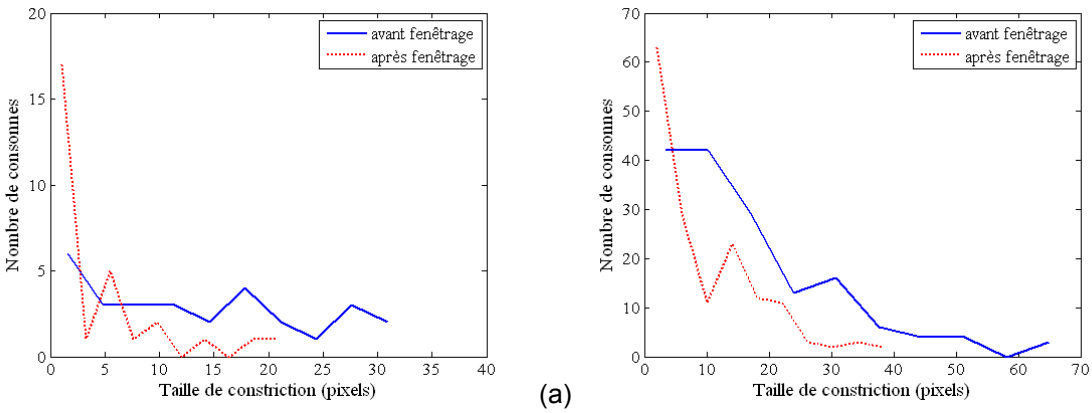
Les histogrammes présentés ci-dessus sont alors légèrement modifiés (Fig. 114). Le nombre d'erreurs de lieu d'articulation observé pour les consonnes alvéolaires diminue (le « résidu » des sections 11 à 15 a été réduit) ; seules 5 consonnes sont détectées comme postérieures, au lieu de 17 avant la prise en compte du fenêtrage. Il s'agit de consonnes [l], nous en reparlons juste après.



(a) (b)

Figure 114 : (a) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes dorsales du corpus Laval43 après fenêtrage.
(b) Lieux d'articulations détectés à partir des sections géométriques pour les consonnes alvéolaires du corpus Laval43 après fenêtrage.

Le bénéfice dû aux détections du minimum dans les fenêtres de 5 trames est également visible au niveau de la taille de la constriction. Rechercher le minimum de constriction sur les trames adjacentes à l'instant étiqueté permet de définir le véritable instant de constriction. Les courbes ci-dessous représentent la distribution de la taille de constriction pour les consonnes postérieures et les consonnes alvéolaires. Avec le fenêtrage, la taille moyenne de constriction passe de 13,7 à 4 pixels pour les consonnes dorsales et de 16,4 à 9,4 pixels pour les consonnes alvéolaires.



(a) (b)

Figure 115 : (a) Répartition des consonnes dorsales en fonction de la taille de constriction, avant et après fenêtrage.
(b) Répartition des consonnes alvéolaires en fonction de la taille de constriction.

L'observation plus précise des erreurs de lieu d'articulation pour les consonnes alvéolaires montre qu'il s'agit de consonnes [l]. On justifie ces erreurs par le fait que le [l] est une consonne latérale, ce qui signifie qu'elle est produite en laissant l'air passer sur les deux côtés de la langue, plutôt qu'au-dessus du milieu, comme le sont les autres consonnes. Ce sont des consonnes qui nécessitent pour leur réalisation l'écoulement de l'air via un canal latéral formé par l'affaissement de l'avant de la langue et le contact de son dos avec le palais. La connaissance de la forme 3D est nécessaire pour ce type de consonnes. Ceci peut expliquer la confusion possible de lieux d'articulation sur les images concernées.

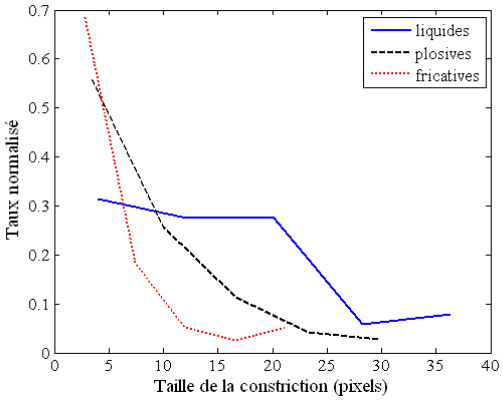
Certaines consonnes alvéolaires sont détectées avec des tailles de constriction supérieures à 25 pixels. Il s'agit encore de consonnes [l], auxquelles viennent s'ajouter quelques [n].

On analyse plus en détail (Fig. 116) le comportement de la constriction pour les consonnes alvéolaires (ou post-alvéolaires) en fonction du mode d'articulation ou de la consonne. Les fricatives sont les consonnes qui sont produites avec les constrictions les plus étroites (valeur moyenne de 5 pixels), suivies par les plosives avec une valeur moyenne de constriction de 8 pixels. Les liquides ont une valeur moyenne de constriction de près de 15 pixels pour un écart-type de 10. Pour ces consonnes, il semble que les données géométriques sagittales posent ici un problème. On se trouve sans doute confrontés à une difficulté au niveau du marquage, à cause du double contour (dû à la projection du 3D) qu'on peut observer pour la langue sur les images radiographiques correspondantes. Il y a peut-être aussi plus de variabilité en terme de production.

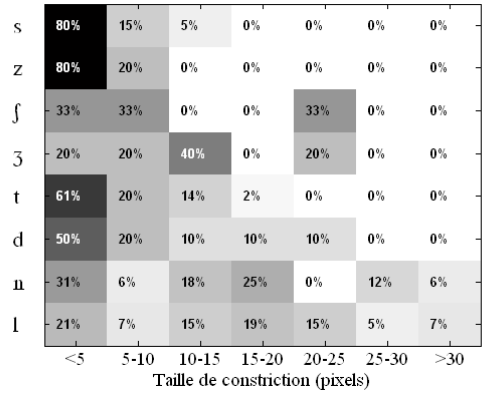
La table (Fig. 116b) détaille le pourcentage de consonnes détectées avec différentes tailles de constriction, et cela, pour les 8 types de classes consonantiques dites alvéolaires dans notre étude. On retrouve bien ce qui vient d'être dit, les consonnes liquides présentent le plus de variabilité en terme de taille de constriction. Au contraire, les fricatives [s] et [z] sont détectées avec de faibles tailles de constriction, les fricatives [ʃ] et [ʒ] montrent plus de variations. Les plosives [t] et [d] se comportent de façon similaire : pour chacune, plus de la moitié des exemplaires sont détectés avec une très faible constriction, les autres exemplaires présentent des constrictions de taille plus variable.

Une explication probable de ces différences est liée à la durée des consonnes et au rapport entre cette durée et la résolution temporelle de la cinéradiographie. Les consonnes fricatives [s] et [z] ont une durée longue, elles sont bien marquées, la taille de constriction minimale est très faible, la détection est bonne. Pour [t] et [d], la taille de la constriction est supérieure à 5 pixels dans 50% des cas. Ce sont des consonnes plosives, courtes, dont la durée de contact est de l'ordre de 15-20 ms ; la durée d'échantillonnage temporel disponible n'est pas suffisante, on manque l'instant précis de contact. Avec une fréquence d'échantillonnage de

50 Hz, la résolution est insuffisante pour une étude vraiment fine de l'organisation temporelle des consonnes et en particulier pour celle des plosives.



(a)

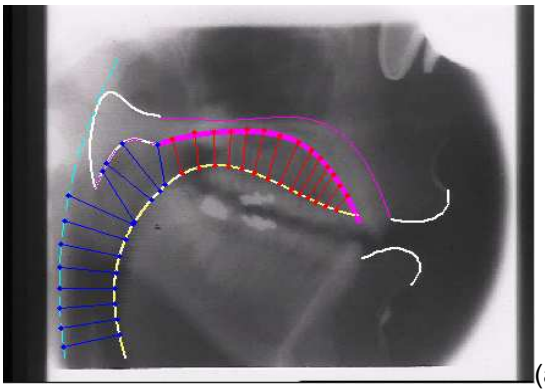


(b)

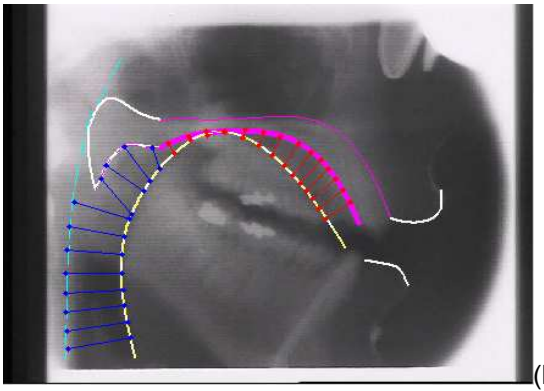
Figure 116 : (a) Distribution des tailles de constriction suivant les modes d'articulation, pour les consonnes alvéolaires. Les courbes ont été normalisées par le nombre de consonnes de chaque catégorie.
(b) Tailles de constriction suivant les 8 consonnes alvéolaires.

Les consonnes dorsales du corpus sont essentiellement des [k], nous ne disposons que d'un seul [ɲ] et de 4 [g], ceci limite l'étude détaillée par consonne que l'on pourrait faire.

Pour finir, les images suivantes montrent 2 exemples de configurations géométriques. L'image de gauche correspond à la production d'un [t], il y a contact entre la pointe de la langue et l'avant du palais. L'image de droite représente la production d'un [k], la langue est bombée vers l'arrière, son dos est en contact avec l'arrière du palais dur, juste avant le palais mou.



(a)



(b)

Figure 117 : (a) Exemple de configuration d'un [t].
(b) Exemple de configuration d'un [k].

2. Consonnes bilabiales

Une consonne bilabiale désigne, en phonétique articulatoire, une consonne labiale dont le lieu d'articulation est situé au niveau des lèvres ; elle est réalisée par rapprochement des

lèvres inférieure et supérieure. Le français comporte 3 bilabiales [b], [p] et [m], c'est aussi le cas de notre corpus.

A partir du signal acoustique et de l'observation des bursts, nous avons été en mesure d'étiqueter 50 consonnes bilabiales.

D'un point de vue articulatoire, nous disposons pour la base de données des marquages géométriques des lèvres supérieure et inférieure. C'est l'observation du marquage des lèvres pour ces consonnes qui a permis au chapitre 3 (§2.1.4.) de corriger le choix du cadre d'indexation pour les lèvres, en supprimant l'influence des incisives grâce à un masque noir.

A partir des contours estimés pour chacune des 2 lèvres, il est possible de mesurer l'écart entre les lèvres pour chacune des images de la base. Cet écart est calculé comme la distance minimale entre la lèvre supérieure et la lèvre inférieure. La mesure est réalisée à partir des points de splines (polynômes d'ordre 3) de chacune des lèvres.

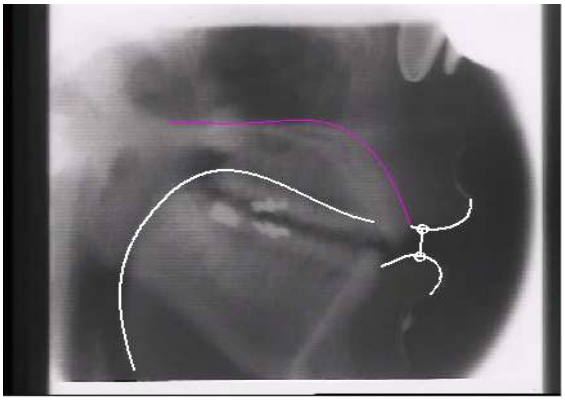


Figure 118 : Mesure de l'écart entre les lèvres supérieure et inférieure pour une image de la séquence Laval43.

Comme précédemment pour les consonnes alvéolaires et palatales, le fenêtrage de 5 trames autour des instants étiquetés est pris en compte pour l'analyse des consonnes bilabiales ; pour chaque instant consonantique associé aux labiales, on considère en réalité l'instant correspondant au minimum d'écart aux lèvres parmi les 5 trames comprises dans la fenêtre.

Ceci permet de mettre en évidence que les consonnes bilabiales étiquetées se situent bien à des minima locaux de l'écart aux lèvres, comme on peut le voir sur la figure 119a. Les étoiles rouges représentent les instants consonantiques associés aux consonnes [b], [p] et [m]. On remarque qu'un certain nombre des autres minima locaux sont associés à des moments de silence : le locuteur ferme les lèvres quand il ne parle pas. Sur l'histogramme qui suit (Fig. 119b), la différence entre les courbes pointillées (noire et rouge) pour de faibles valeurs d'écart aux lèvres montre que pour 15% des moments de silence, l'écart aux lèvres est inférieur à 20 pixels.

Pour les consonnes bilabiales, l'écart moyen entre les lèvres supérieure et inférieure est de 15 pixels alors qu'il est de 38 pixels pour la séquence complète.

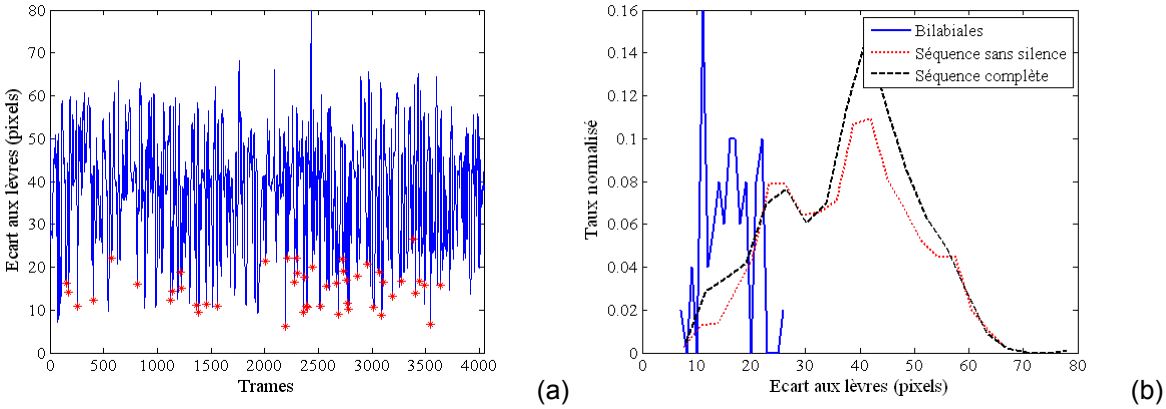


Figure 119 : (a) Ecart mesuré entre les lèvres inférieure et supérieure, à partir du marquage géométrique extrait pour la séquence Laval43 et mise en évidence des instants de production des consonnes bilabiales étiquetées du corpus.
(b) Répartition des trames en fonction de l'écart aux lèvres et mise en évidence d'un minimum d'ouverture pour les bilabiales.

Ces mesures ont donc mis en évidence une fermeture des lèvres pour les consonnes bilabiales ; en réalité, compte-tenu des valeurs obtenues, on ne peut pas vraiment parler de fermeture, mais plutôt de minimum d'ouverture. Il semble y avoir un biais de quelques pixels. Nous n'obtenons jamais une fermeture complète des lèvres au cours de la séquence. Un nouvel effort concernant le marquage des lèvres serait probablement nécessaire pour corriger ce biais.

La représentation qui suit, proposée par Dusan dans sa thèse [Dus00] sous le terme d'aérogramme, permet une vue d'ensemble de la fonction d'aire en fonction du temps, pour une phrase du corpus. Par analogie avec le spectrogramme (où les spectres sont affichés en fonction du temps), cette représentation 3D de la fonction d'aire est obtenue en traçant en abscisse le temps, en ordonnée les sections médio-sagittales et en représentant à chaque instant et pour chaque section les valeurs de la fonction d'aire par des niveaux de gris différents.

Nous considérons les 28 sections des lèvres jusqu'à la glotte, définies au chapitre 4, et les fonctions d'aire associées. De plus, nous intercalons la mesure précise réalisée pour la pointe, pour l'étude des consonnes dans ce chapitre : à partir de la distance sagittale calculée entre la pointe de la langue et la palais, l'aire est obtenue grâce au modèle alpha-beta et les valeurs de paramètres α et β de la région 7 et du locuteur mâle du modèle de Soquet et al. [SLMD02]. Nous disposons alors de 29 sections, numérotées sur l'aérogramme de 1 (pour les lèvres) à 29 (pour la section la plus proche de la glotte) et la section 3 correspond à la mesure pour la pointe de la langue.

Plus le niveau tend vers le blanc, plus l'aire est petite (inversement, plus le niveau est foncé, plus l'aire est grande). Les zones les plus claires de la figure 120 correspondent à des points de contacts. Par exemple, au temps t=1s, la tâche blanche située en haut de l'aérogramme illustre la fermeture des lèvres pour la plosive bilabiale [p]. Les tâches blanches autour de la section 15, peu avant t=2s et après t=2,5s, correspondent aux contacts palataux du [g] et du [k]. Le point de contact de la langue avec le palais pour le [s] est représenté par la tâche banche au niveau de la section 3 autour de t=2,3s. L'ouverture et l'abaissement de la langue pour les [a] de la phrase sont visualisés avec les tâches plus sombres à l'avant du conduit vocal, correspondant à une aire plus grande de la cavité avant. La tâche la plus sombre (t≈0,8s) met en évidence la prise en compte de la cavité sublinguale.

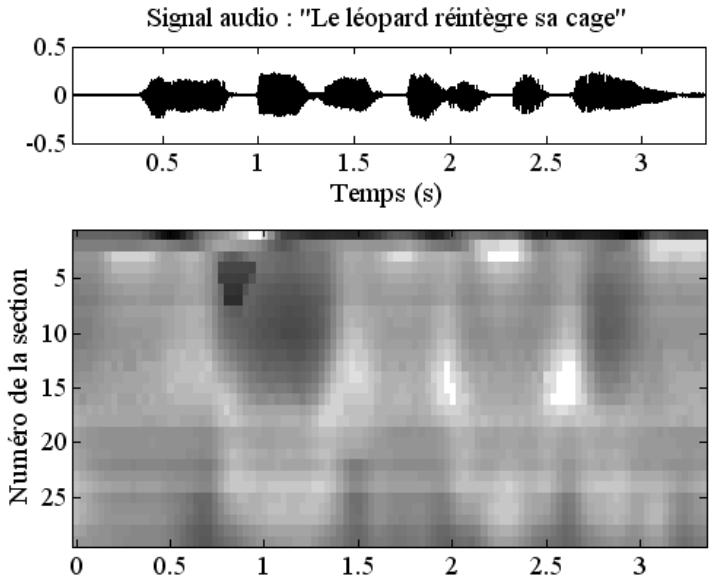


Figure 120 : Aérogramme pour la phrase du corpus « Le léopard réintègre sa cage ». Les constrictions sont représentées par les tâches blanches.

L'étude des consonnes à partir de la cinéradiographie reste en partie biaisée par la résolution temporelle limitée de cette technique d'imagerie, comme on a pu l'observer en particulier pour les plosives. L'échantillonnage temporel de l'ordre de 50 Hz pour les séquences dont nous disposons n'est pas suffisant pour une étude vraiment fine de l'organisation temporelle des consonnes. En effet, même s'il n'est pas possible de donner une durée moyenne d'une consonne en français (trop de variabilité à cause de la position dans la syllabe, du contexte vocalique ou du locuteur), on trouve dans la littérature ([Osh84], [KZ96]) des durées de moins de 100 ms, voire de quelques dizaines de ms (15-20 ms) pour les consonnes occlusives alvéolaires (comme [t], [d]). La cinéradiographie (et sa résolution temporelle) est alors susceptible de ne pas capturer l'instant exact du contact entre les lèvres ou entre la langue et le palais.

CHAPITRE 9 : UNE NOUVELLE SOURCE D'INFORMATIONS POUR L'ÉTUDE DES MOUVEMENTS DU VÉLUM

La nasalité est un phénomène complexe, tant du point de vue articulatoire, qu'acoustique ou aérodynamique. Elle est introduite par un articulateur, le voile du palais, dont l'abaissement permet de connecter les fosses nasales au conduit oral.

La modélisation de la nasalité est, de par sa complexité, l'objet d'études spécifiques. Elles sont nombreuses, nous n'en ferons pas l'état de l'art. Mais notons les diverses techniques d'imagerie utilisées pour enregistrer des données articulatoires sur le vélum.

On trouve un certain nombre d'études reposant sur l'utilisation de la cinéradiographie, notamment les travaux de Flament [Fla84] ou Zerling [Zer84] à Strasbourg. Les informations radiologiques concernant les régions velo et rhyno-pharyngales sont parfois de qualité discutable : lorsque le voile est en position haute, il peut se produire un phénomène de diffraction qui en rend le contour flou. C'est ce qu'on a pu observer sur certaines images de la séquence Wioland, comme par exemple sur la figure 121.

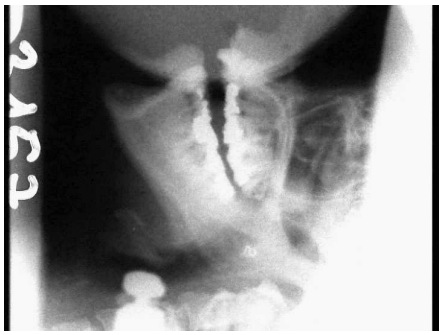


Figure 121 : Image de la séquence Wioland sur laquelle le contour du voile du palais, en position haute, n'est pas net.

Ces études cinéradiographiques s'intéressent aux mesures de positions du vélum ([SFK80], [ZDB$^+$87]), aux phénomènes d'articulation et de coarticulation (avec l'enregistrement de logatomes [Zer84]), à l'étude de pathologies suite à des becs de lièvre (enregistrement de courtes phrases [HI91]).

La problématique du marquage des articulateurs sur des séquences cinéradiographiques a été illustrée par Shaw et al. [SFK80] avec la mesure des positions du vélum. Différentes options pour suivre un articulateur y sont traitées :

(1) La première alternative consiste à poser un marqueur a priori sur l'articulateur et à enregistrer l'évolution de la position de cette marque au cours de la séquence. Il s'agit

alors de suivre un point à 2 degrés de liberté. Dans le cas de la cinéradiographie, par exemple pour le vélum, un marqueur opaque aux rayons X est fixé sur la partie orale du voile du palais avant l'enregistrement d'une courte séquence. Le traitement des données consiste ensuite à mesurer les coordonnées x-y de ce marqueur pour chaque image de la séquence.

(2) Une autre technique est basée sur le choix de points à 1 ddl, correspondant à l'intersection entre le contour de l'articulateur et des lignes de référence.

(3) Le suivi de l'articulateur peut être effectué à partir de points à 2ddl marqués sur le contour pour chaque image de la séquence enregistrée. Ces points, qu'on peut voir comme des points d'appui anatomiques, sont par exemple déterminés à partir d'une propriété de courbure du contour. Cette stratégie a l'inconvénient qu'il n'est pas toujours possible de déterminer un tel point avec précision.

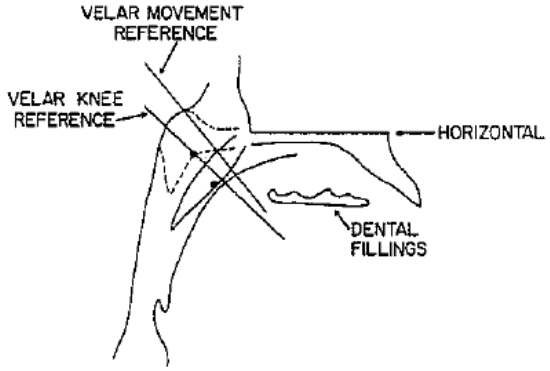


Figure 122 : Marquage manuel de points d'appui anatomiques sur le voile du palais, d'après [SFK80].

Notre méthode regroupe les deux dernières possibilités afin d'inférer un contour complet. Le marquage de l'articulateur est réalisé à partir de points à 1ddl (intersections avec les lignes d'une grille de marquage) et de points à 2ddl (comme la pointe de la langue ou celle du vélum). A partir du marquage de ces points, on établit le contour complet de l'articulateur par interpolation polynomiale.

Ce contour permet ensuite de simuler a posteriori les différentes options proposées ci-dessus. Il est possible de poser une marque a posteriori à 2 ddl ou de suivre des points caractéristiques en fonction de leurs propriétés. C'est le cas par exemple d'un point de courbure du contour ou d'un point de constriction.

Les études sur la nasalité font aussi appel à des données fibroscopiques [ARCM04] avec les problèmes de confort du locuteur et de fait les artéfacts que cela implique, à des données IRM ([SB05], [DMS02]) ou obtenues à partir de l'EMA [RBF00] en fixant une bobine sur le vélum. Cette dernière technique pose le problème de ne pouvoir enregistrer qu'une seule

position au cours du temps. Et l'IRM a des limites en termes d'enregistrement dynamique, les études actuelles se limitent à des données statiques.
Les études sur le vélum et le conduit nasal souffrent d'un manque de données dynamiques. La méthode de marquage semi-automatique, présentée dans ce manuscrit et qui a pu être utilisée avec succès sur le voile du palais sur de longues séquences, a l'avantage, par rapport aux méthodes à capteurs, de permettre l'observation directe du conduit en mouvement. Elle pourrait être en mesure de fournir des nouvelles données sur le vélum, difficilement accessibles autrement. Diverses analyses sont ensuite envisageables à partir des données géométriques extraites (comparaison de production entre locuteurs, études des phénomènes de coarticulation, d'anticipation…).

L'étude qui suit est une étude tout à fait préliminaire. Elle n'a pas pour ambition d'apporter de nouvelles informations sur le voile du palais, mais encore une fois, de valider le marquage qui a été réalisé sur cet articulateur avec la méthode semi-automatique mise en place.
Cette validation passe à nouveau par une mise en correspondance des données géométriques estimées avec les données étiquetées du signal, à savoir ici les voyelles et consonnes nasales du corpus.
Les voyelles nasales sont produites par le passage de l'air dans les fosses nasales grâce à l'abaissement du voile du palais. Le flux d'air continue en même temps de passer par le conduit oral. Les voyelles nasales du français sont [ɛ̃], [œ̃], [ɑ̃] et [ɔ̃].
Les consonnes nasales sont produites en abaissant le voile du palais, le français et notre corpus en comporte trois : [m], [n] et [ɲ].

D'un point de vue géométrique, nous disposons du marquage du vélum pour chaque image de la base, ainsi que du marquage de la paroi pharyngale. Le vélum a été marqué en appliquant la méthode semi-automatique avec 100 images clefs, 13 points (14 ddl) et un cadre focalisé sur cet articulateur. La paroi pharyngale a également été marquée pour toute la séquence à partir de la méthode (avec 50 images clefs et 5 points à 1ddl).
Une représentation analogue à celle réalisée précédemment au chapitre 5 (§3.1.1 et 3.1.2.) pour la langue et les lèvres est obtenue ici pour le voile du palais. L'idée de cette représentation est d'observer l'espace couvert au cours de la séquence par un point particulier de l'articulateur, de façon à mettre en évidence une organisation. Pour la langue par exemple, l'analyse du point le plus élevé du dos a permis la distinction des classes vocaliques extrêmes.
Etant donnés les degrés de liberté estimés sur la partie supérieure du vélum, nous moyennons la position des 3 points (à 1 ddl) les plus en arrière (représentés par des étoiles

sur la figure 123). Nous observons sur cette figure la position moyenne alors obtenue, en la superposant sur le conduit vocal, focalisé dans la région du vélum. Les positions associées aux instants de voyelles orales sont figurées en gris et sont proches de la paroi pharyngale, en comparaison de celles associées aux voyelles nasales du corpus (en noir) qui s'écartent du pharynx. Nous constatons en outre qu'il n'y a qu'un seul degré de liberté. Ce constat est en faveur des travaux menés à partir de capteurs, fixés sur la partie inférieure du vélum (comme par exemple le point noir sur la figure 123) : l'enregistrement des mouvements de ce capteur est censé représenter ceux du voile du palais. La corrélation entre le point noir fixé et la position moyenne obtenue pour la partie supérieure du vélum est proche de 0,9 ; cette mesure est comparable aux coefficients de corrélation calculés par Shaw et al. [SFK80].

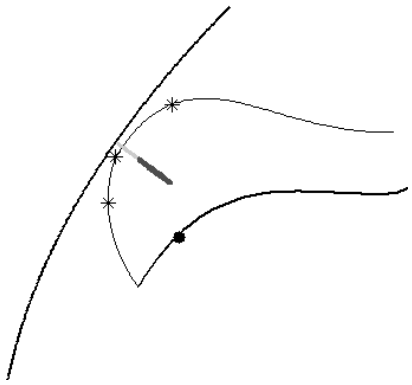


Figure 123 : Représentation graphique de la position moyenne de la partie supérieure du vélum (moyenne des 3 ddl indiqués par les étoiles) et distinction entre voyelles orales (en gris) et voyelles nasales (en noir).

De manière plus précise, à partir du marquage géométrique, il est aisé de calculer image par image l'écart minimal entre le pharynx et le voile du palais. Il s'agit d'estimer la constriction vélum-pharynx. Plus le diamètre de cette constriction est grand, plus le vélum est abaissé, c'est la position attendue pour les nasales. La recherche de cette constriction est effectuée sur les points des polynômes représentant respectivement le pharynx et la partie supérieure du vélum.

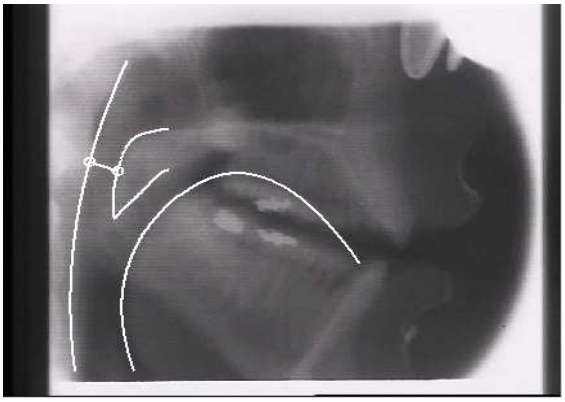


Figure 124 : Mesure de l'écart entre la paroi pharyngale et le voile du palais pour une image de la séquence Laval43.

Le fenêtrage de 5 trames autour des instants étiquetés, décrit au chapitre précédent, est à nouveau pris en compte dans cette analyse. Etant donné un instant étiqueté correspondant à une voyelle ou une consonnes nasale, on considère des fenêtres de 5 trames autour de cet instant. Pour chacune de ces fenêtres, on recherche, parmi les 5 trames, la trame associée au maximum d'écart entre la paroi pharyngale et le vélum.

La figure 125a représente l'écart mesuré entre le pharynx et le voile du palais pour toute la séquence et met en évidence 2 types de maxima, associés ou non à des phonèmes nasalisés. Les consonnes et voyelles nasales étiquetées se situent pour la plupart à des maxima locaux de l'écart entre le pharynx et le vélum, de moins de 35 pixels, comme le montrent les croix rouges correspondant aux voyelles nasales et les points noirs correspondant aux consonnes nasales. Il existe de nombreux autres maxima qui sont supérieurs à 35 pixels. Ces maxima sont majoritairement associés à des moments de silence : la position au repos est généralement un vélum abaissé, qui permet la respiration par le nez. La figure 125b montre la séquence privée des instants de silence. On constate qu'en excluant les silences, on réduit la proportion des ouvertures larges entre le vélum et le pharynx. Les consonnes et voyelles nasales occupent désormais un grand nombre des maxima locaux. On peut supposer que les maxima restants, non étiquetés voyelle ou consonne nasale, correspondent à des instants de respiration nasale.

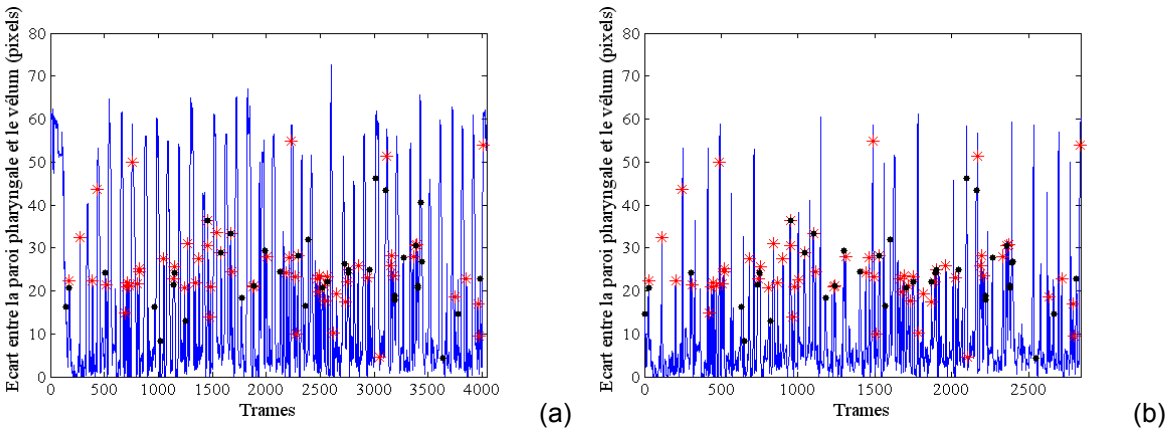


Figure 125 : (a) Ecart minimal entre le pharynx et le voile du palais pour la séquence complète Laval43. Les consonnes nasales sont représentées par les points noirs, les voyelles nasales par les croix rouges.
(b) Ecart minimal entre le pharynx et le voile du palais pour la séquence Laval43 privée des moments de silence.

L'ouverture de la cavité nasale pour les moments de silence a une valeur moyenne de 39 pixels. L'histogramme de répartition des trames de la séquence en fonction de l'écart entre le pharynx et le voile du palais a une forme tri-modale, comme on l'observe sur la figure 126a. Le mode étiqueté 1 correspond aux instants de silence.

Si on s'intéresse plus spécifiquement aux voyelles, on peut comparer les voyelles nasales avec les voyelles orales du corpus, en terme d'abaissement du vélum (on observe toujours l'écart entre le voile et le pharynx). Les histogrammes de la figure 126b montrent la distribution de l'ouverture de la cavité nasale pour les voyelles nasales (en trait plein bleu) et pour les voyelles orales (en pointillés rouges). Chaque courbe a été respectivement normalisée par le nombre de voyelles, orales ou nasales, du corpus. On obtient 2 distributions distinctes avec 2 pics très séparés. L'air ne passe pas par la cavité nasale pour les voyelles orales, le vélum est relevé. L'écart entre la paroi pharyngale et le vélum est faible, l'écart moyen pour ces voyelles est de 5 pixels alors qu'il est de 25 pixels pour les voyelles nasales. La distribution observée pour les voyelles nasales correspond au mode 2 de la figure 126a.

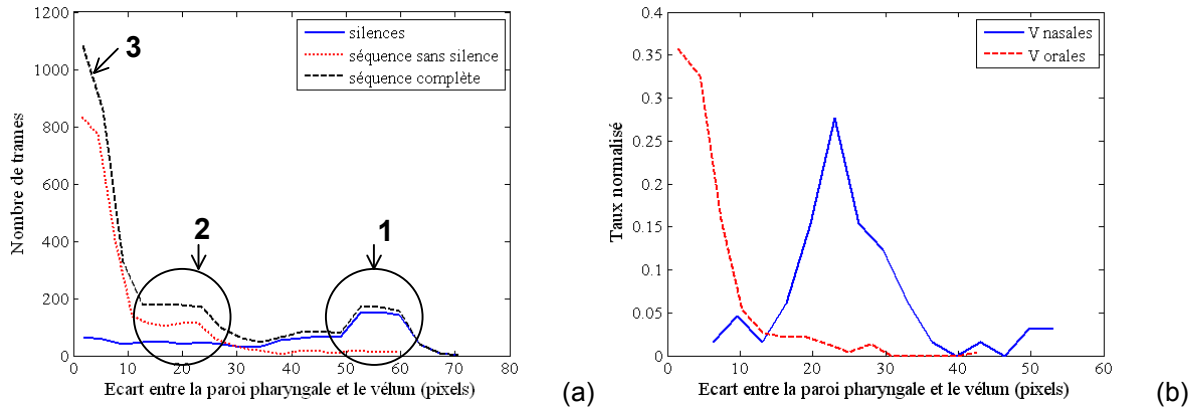


Figure 126 : (a) Répartition des trames en fonction de l'écart minimal entre le pharynx et le vélum pour la séquence, avec ou sans silence.
(b) Répartition de l'écart minimal entre le pharynx et le vélum pour les voyelles nasales et orales du corpus Laval43. Les ordonnées ont été normalisées par rapport au nombre de voyelles.

Ces observations sur le voile du palais permettent de retrouver des notions très classiques et pourront être complétées par des études plus approfondies de la relation entre acoustique et articulatoire.

DISCUSSION ET PERSPECTIVES

La cinéradiographie a, au cours des dernières décennies, largement fait les preuves de son utilité dans les études en parole. Cette technique est à l'origine d'un nombre important de travaux de référence en production de parole et en modélisation articulatoire. En effet, elle permet une visualisation complète des articulateurs du conduit vocal de la glotte jusqu'aux lèvres, à la fois statique (image par image) et dynamique avec des résolutions temporelles importantes (de l'ordre de 60 images par seconde). De grandes bases de données ont récemment été constituées par numérisation de séquences existantes, qui sont disponibles pour la communauté de recherche en parole.

Notre travail a pour objectif d'exploiter ces grandes quantités de données 2D+t du conduit vocal avec la mise en place d'une méthode semi-automatique d'extraction des contours et du mouvement des articulateurs, visant à minimiser l'intervention manuelle de marquage des images. Cette méthode est basée sur l'utilisation du rétro-marquage.

C'est l'élaboration de cette méthode qui est le cœur de cette thèse. L'étude qui a été menée en seconde partie ne cherche pas à apporter de nouveaux résultats concernant les relations articulatori-acoustiques, mais plutôt à « valider » la méthode et les données géométriques qui en résultent afin d'en montrer le potentiel. L'enjeu est d'analyser dans quelle mesure la précision obtenue au niveau du marquage des articulateurs est suffisante pour permettre l'utilisation de ces données dans des applications articulatori-acoustiques.

Par la suite et à partir de cette méthode, il sera envisageable d'analyser de plus larges corpus (plusieurs dizaines de phrases), ce qui n'a jusqu'à aujourd'hui que peu été réalisé, compte-tenu du travail long qu'implique le tracé manuel. Ces analyses pourront être effectuées à des fins diverses, dont nous donnerons quelques exemples un peu plus loin.

Ce chapitre est consacré à la discussion générale dont le but est de faire une synthèse des résultats obtenus et présentés aux chapitres précédents. Nous reviendrons sur certaines de nos questions et éventuellement, certaines hypothèses. Ce chapitre se présente en trois parties, d'abord un bilan du travail effectué, puis des conclusions et des réponses à des questions soulevées au cours des précédents chapitres, enfin certaines questions sont laissées ouvertes à titre de perspectives.

1. Bilan du travail effectué

Les grandes bases cinéradiographiques mentionnées sont constituées de plusieurs séquences, correspondant chacune à quelques minutes de parole lue. Enregistrées à des cadences de 50-60 images par seconde, chaque séquence se compose alors de plusieurs milliers d'images, représentant des configurations géométriques différentes du conduit vocal. Si l'on veut analyser complètement une de ces séquences et en extraire l'information géométrique, il serait trop laborieux de marquer une à une chacune des images.

Pour pallier aux difficultés rencontrées par les méthodes totalement automatiques, nous proposons de conserver une part d'expertise humaine, sur un nombre limité d'images, les images clefs, avant d'appliquer un traitement automatique. L'étape manuelle consiste à définir un petit jeu de degrés de liberté, associés à des points caractéristiques, décrivant la forme de l'articulateur considéré puis à les marquer sur chacune des images clefs. L'étape automatique permet ensuite de marquer toutes les images de la séquence, à l'aide d'un processus d'indexation basé sur une mesure de similarité entre les coefficients DCT basses fréquences des images. Il s'agit alors d'associer à chaque image, successivement, la configuration géométrique de l'image clef marquée la plus proche. Pour restaurer une meilleure continuité, des traitements postérieurs sont appliqués, tels qu'un filtrage temporel et un moyennage de configurations géométriques voisines, suivis d'une reconstruction des contours par interpolation polynomiale des points caractéristiques.
La méthode mise en place s'applique séquence par séquence et articulateur par articulateur. Pour une séquence donnée, les contours du conduit vocal complet sont reconstruits en combinant les configurations géométriques obtenues indépendamment pour chaque articulateur. Les articulateurs ont donc été préalablement soumis à des estimations spécifiques, correspondant à l'application du principe de base de la méthode avec des paramètres adaptés à chacun de ces éléments du tractus vocal.
La méthode a été élaborée et évaluée géométriquement pour un articulateur, la langue, sur la séquence Wioland de la base de données cinéradiographiques du français constituée par l'IPS et l'ICP [ABB+00]. Le réglage des paramètres a été effectué pour permettre un compromis entre l'erreur de reconstruction RMS et le coût (en temps de marquage manuel). La méthode a pu être assez facilement étendue à d'autres articulateurs et à d'autres séquences : la séquence Flament extraite de la même base de données que Wioland et la séquence Laval43 extraite de la base de données d'ATR [MVT95]. Sur la séquence Laval43, les articulateurs (langue, pointe de la langue, vélum, lèvres, mandibule) ont été traités, de manière indépendante, un par un et avec des images clefs différentes, avant d'être re-

combinés pour permettre la reconstruction du conduit vocal complet pour l'ensemble de la séquence.

Pour chaque articulateur, des vidéos ont été reconstruites en superposant sur les images d'origine les contours géométriques estimés. Ceci permet de suivre le mouvement de l'articulateur et de constater que les contours et leur mouvement sont bien reconstruits.

Notons que même si le traitement est quasiment analogue pour chaque articulateur, nous remarquons quelques particularités.

- Dans Laval43, les lèvres souffrent d'un manque de contraste, le marquage est plus difficile à réaliser et l'indexation présente plus facilement des erreurs, dans la mesure où les coefficients DCT se basent sur les différences de contrastes.
- La pointe de la langue demande une précision de marquage importante et pour cela une attention très particulière au moment de l'étape manuelle, ce qui augmente le temps de traitement. Sur Wioland, elle est peu visible et son suivi est moins bien réalisé que sur Laval43. Sur cette séquence, le traitement spécifique réalisé sur la pointe permet une détection des points de contact entre la langue et le palais et une mise en correspondance avec les événements consonantiques du corpus.
- De son côté, le vélum peut être considéré comme un articulateur « facile » à estimer. Sur les séquences Laval43 et Flament, il est bien visible et facilement marquable. Dès 50 images clefs marquées, l'application de la méthode permet un suivi très correct de cet articulateur. Ces résultats nous semblent utiles pour l'étude de la nasalité, en permettant l'exploitation de nombreuses données. A titre tout à fait préliminaire, des observations de l'abaissement du vélum ont été mis en parallèle avec la production de consonnes et voyelles nasales.

Une partie des travaux présentés a été menée pour valider les configurations géométriques obtenues. Pour commencer, nous nous sommes particulièrement intéressés à l'étude des voyelles. L'observation des contours géométriques du conduit vocal extraits par la méthode permet de retrouver des résultats classiques en phonétique en fonction des différentes classes vocaliques, notamment les positions haut-bas et avant-arrière de la langue.
Ensuite, afin d'évaluer la précision des contours estimés d'un point de vue acoustique, nous nous sommes basés sur les représentations classiques que sont les formants.
Dans un premier temps, nous avons tenté l'application d'un modèle linéaire pour estimer la position des formants en fonction des degrés de liberté géométriques. Si la corrélation temporelle entre les formants estimés et les formants d'origine est assez élevée (comparable à celle de Yehia [Yeh02]), le modèle linéaire n'est cependant pas suffisant pour prédire

correctement l'acoustique. Le triangle vocalique n'est pas bien couvert par les estimations formantiques.

Dans un deuxième temps, en partant des contours géométriques extraits par la méthode, une représentation du tractus vocal sous forme de fonctions sagittales a été développée à l'aide d'une grille composée de 27 lignes définies orthogonales au palais et à la langue en moyenne. Des corrections image par image ont ensuite été effectuées, en suivant la procédure proposée par Yehia dans sa thèse [Yeh02]. La transformation de la fonction sagittale en fonction d'aire a été réalisée en faisant appel au modèle alpha-beta de Heinz et Stevens [HS65], que nous avons appliqué avec les paramètres de Soquet et al. [SLMD02] pour un locuteur mâle.

Les fonctions d'aire ainsi calculées pour chaque image de la séquence ont été utilisées en entrée d'un modèle analogue électrique-acoustique, pour obtenir les fonctions de transfert associées à chaque configuration estimée du conduit vocal. Le modèle analogue considéré est celui de Badin et Fant [BF84]. Par un algorithme de détection de pics, les trois premiers formants ont été estimés pour chaque trame de la séquence. En considérant les trames de voyelles, un biais moyen de 10% a été évalué pour les trois formants, entre les estimations et les formants du signal original (extraits par Praat). L'étude des représentations F_1-F_2 a permis de retrouver l'organisation classique des classes vocaliques et de mettre en évidence une délimitation correcte du triangle vocalique.

Ce modèle, adapté pour les voyelles, a été extrapolé pour mettre à profit les données dynamiques dont nous disposons, et estimer un signal de parole à partir des données géométriques extraites (et notamment les consonnes). Sur le principe d'un modèle source-filtre, les fonctions de transfert estimées du conduit vocal sont utilisées en tant que filtre. Quant à la source, elle est définie en deux composantes à partir du signal : d'une part, une source blanchie par décomposition de Hilbert et filtrage inverse LPC de l'information temporelle fine et d'autre part une modulation d'amplitude en deux sous-bandes. De cette synthèse, une analyse spectrale a été menée ainsi qu'un test de perception qui a mis en évidence l'intelligibilité du signal resynthétisé à partir de l'estimation.

2. Conclusions

Les travaux menés au cours de cette thèse et résumés ci-dessus amènent à quelques conclusions et remarques.

La méthode semi-automatique décrite n'a pas pour objectif de concurrencer le marquage manuel en terme de qualité et de précision de mesures. L'étude se place dans un contexte d'exploitation de très larges bases de données cinéradiographiques, pour permettre l'analyse de longs corpus de parole naturelle. Les études existantes qui utilisent le marquage manuel se limitent généralement à des corpus de quelques phrases, comme pour les travaux de Yehia [Yeh02], ou des corpus de logatomes ([VAB98], [BGB+95], [BBB01], [VSR+03], [AVFG03]…). L'exploitation de plusieurs minutes de parole naturelle offre la possibilité d'étudier des phénomènes dynamiques et d'affronter la variabilité du signal de parole (articulation, coarticulation, anticipation, influence du contexte consonantique, variabilité de production inter et intra locuteurs...). Nous n'atteignons pas une précision comparable au marquage manuel, puisque les erreurs sont bien visibles quand elles sont présentes, mais nous proposons par cette méthode d'extraire en quelques jours les contours complets du conduit vocal pour une séquence de plusieurs milliers d'images.

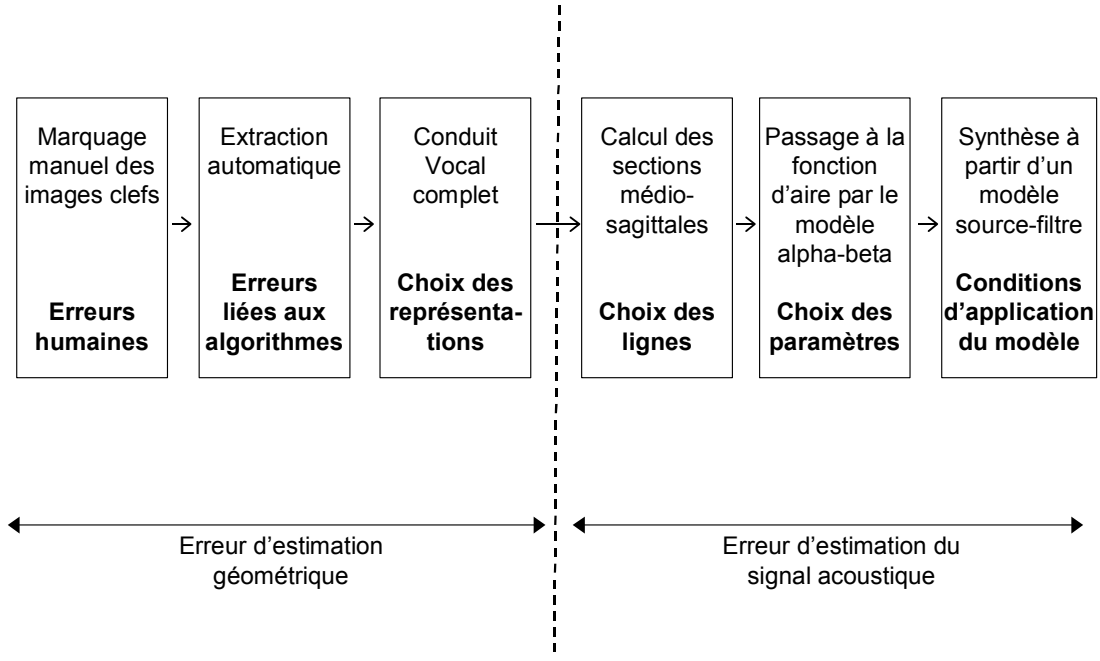
Les contours obtenus permettent de suivre les mouvements du tractus vocal et de synthétiser un signal de parole intelligible. Ceci souligne l'importance de l'aspect dynamique qui vient en fait compenser les erreurs mesurées trame par trame. Les 10% de biais observés sur les formants pour les voyelles en statique n'empêchent pas l'intelligibilité. Les données 2D seules ne sont pas suffisantes pour comprendre complètement la production des sons de parole. La nécessité de considérer le conduit vocal comme une structure 3D est au cœur de nombreux travaux, mais on voit ici que disposer de données 2D+t (données 2D dynamiques, en fonction du temps) ouvre des voies pour l'étude des phénomènes de production de parole.

Nos travaux n'ont pas eu pour objet de définir un modèle articulatoire. Le traitement effectué sur une séquence aboutit à des représentations articulatoires brutes. Les données dont nous disposons après traitement sont des coordonnées de points marqués sur le contour du conduit vocal pour chaque image de la séquence. Il ne s'agit pas de paramètres articulatoires de commande, définissant des mouvements d'ensemble des divers articulateurs. L'élaboration d'un tel modèle nécessite l'utilisation de données géométriques très précises. Aussi nos données nous semblent plus propices à des analyses directes plutôt qu'à la conception d'un modèle articulatoire de commande.

L'utilisation de ces données brutes intègre une plus grande variabilité que dans les études plus classiques utilisant des données régularisées en paramètres articulatoires.
Nos données sont complexes et leur étude montre une grande variabilité à divers niveaux. Nous avons d'abord pu constater un bruit important dans la mise en correspondance entre

les données vidéos (coefficients DCT) et les données géométriques et une régularisation a été réalisée par moyennage. L'observation des ellipses de dispersion des différentes classes vocaliques a permis de mettre en évidence d'une part la grande variabilité articulatoire (avec l'analyse de la constriction par exemple) et d'autre part la variabilité acoustique (avec la représentation formantique F_1-F_2) de nos données. Nous pouvons, certes, attribuer une part de cette variabilité à l'imperfection de notre modèle. En 1991, Maeda montre [Mae91], à partir de son modèle basé sur un marquage manuel plus précis, que la variabilité articulatoire est plus importante que la variabilité acoustique et suggère que la coordination entre les articulateurs permet de réduire le nombre de réalisations acoustiques possibles. Nos données issues d'un corpus non dédié à l'étude spécifique des voyelles permettent d'aborder, avec plus de données, le problème de la variabilité articulatoire de la parole continue.

Les étapes qui ont été suivies pour l'élaboration de la méthode et sa validation en terme acoustique sont schématisées ci-dessous. Ceci permet d'expliciter les différents niveaux d'erreurs auxquels nous sommes confrontés, le choix et le réglage des paramètres à chaque étape du processus étant, à chaque fois, une affaire de compromis.



- Insistons sur le fait que la méthode n'est pas purement automatique, c'est-à-dire que la part d'expertise humaine est importante. En effet, l'étape de marquage manuel conditionne la suite de la méthode. Il est essentiel de réaliser le marquage des images clefs

avec la plus grande attention et précision possible, de manière à ne pas propager d'erreurs à l'ensemble des images de la séquence lors de l'étape d'indexation automatique. Cette étape manuelle est donc primordiale, il faut garder en mémoire que le marquage de chaque image clef est difficile : par exemple, lorsque la langue est occluse par la mâchoire supérieure ou que la base de la langue apparaît avec un double contour à cause de la projection 2D. L'expert doit faire des choix et les suivre tout au long de l'étape de marquage. Cette étape prend du temps, plus ou moins long suivant l'articulateur considéré. Les erreurs humaines de marquage manuel n'ont pas pu être évaluées sur les séquences traitées, dans la mesure où un unique expert a marqué les images clefs.

- L'extraction automatique des contours via l'indexation à partir des images clefs est le cœur de la méthode. Tout au long des travaux menés, c'est l'erreur d'estimation liée à cette étape que nous avons cherché à évaluer et à minimiser en priorité, pour valider l'utilisation de la méthode. En effet, il s'agit d'une erreur qu'on qualifie d'algorithmique, puisqu'en jouant sur les paramètres des algorithmes utilisés, on peut la réduire. C'est à cette étape que la question majeure du nombre d'images clefs est posée, le choix de ce nombre est un compromis entre le temps de marquage manuel et l'erreur d'estimation. Dans un second temps, nous cherchons à évaluer l'effet de cette erreur sur l'estimation de paramètres acoustiques.

Un point d'attention essentiel de cette étape d'indexation est le choix du cadre spécifique à chaque articulateur. Chacun des articulateurs est estimé de manière la plus indépendante possible et cette notion d'indépendance est concrétisée par le choix du cadre qui sert au calcul des coefficients DCT. Il est donc indispensable de choisir avec précaution et attention le cadre pour ne prendre en compte que l'articulateur concerné et éviter les interférences.
Dans le cas de la langue, il n'est cependant pas possible d'éliminer, par le choix d'un cadre, l'interférence qui existe avec la mâchoire inférieure. Les mouvements de la langue et de la mâchoire se composent. Malgré la présence d'un mouvement spécifique de la mâchoire [BBV98], paramètre explicite dans de nombreux modèles articulatoires, nous assumons que le nombre d'images clefs considéré (et donc le nombre de configurations géométriques) est suffisant pour prendre en compte le mouvement de la langue indépendamment de celui de la mâchoire.

- Nous continuons l'analyse des sources d'erreurs avec celles liées à l'élaboration d'une représentation du conduit vocal complet. Celle-ci est obtenue, à partir des degrés de liberté estimés après l'étape automatique, en combinant les contours des divers articulateurs estimés. A cette étape, nous avons complété le conduit vocal avec le marquage des parties

fixes et considéré des interpolations polynomiales entre les points pour reconstruire les contours du tractus.

- Dans l'optique d'analyser l'erreur d'estimation de paramètres acoustiques évalués à partir des données géométriques, un modèle de synthèse articulatoire a été mis en place, en commençant par le calcul des sections médio-sagittales à partir des contours du conduit vocal. Nous avons proposé une méthode, alternative à la grille semi-polaire, pour le placement de la grille, le choix des lignes et de leur orthogonalité par rapport au conduit. Cette méthode est potentiellement porteuse d'erreurs.

- A cette liste viennent s'ajouter les erreurs liées à l'utilisation du modèle αβ, proposé par Heinz et Stevens [HS65], pour passer des fonctions médio-sagittales aux fonctions d'aire, i.e. du 2D au 3D. Ce modèle fait appel à des paramètres qui sont spécifiques à un locuteur. A partir de données cinéradiographiques seules, nous ne sommes pas capables d'estimer ces paramètres et nous avons été contraints d'utiliser des paramètres réglés pour un autre locuteur [SLMD02]. Des erreurs en résultent forcément, que nous ne pouvons pas directement quantifier. Un ajustement de ces paramètres pourrait être envisagé en optimisant les formants estimés dans le cas par exemple de configurations vocaliques extrêmes ([i], [a], [u]). A cette étape, deux estimations supplémentaires ont été nécessaires pour l'application du modèle : le rapport pixels/cms et la position de la glotte, non disponibles sur la séquence.

- Enfin, le modèle source-filtre considéré fait appel à l'analogue électrique-acoustique de Badin et Fant [BF84]. Ce modèle n'est valide que pour les voyelles dans la mesure où il s'agit d'un modèle de propagation acoustique avec lequel la source doit être d'origine glottique. Nous élargissons son application à la parole continue, ce qui implique des distorsions entre le signal synthétisé et le signal d'origine. En introduisant une modulation d'amplitude à 2 sous-bandes estimée à partir du signal d'origine, nous parvenons à synthétiser des consonnes et à une évaluation perceptive.

Les sources d'erreur pointées ici mériteraient pour la plupart des compléments d'analyse, qui viennent s'ajouter aux perspectives que laissent entrevoir nos travaux et que nous présentons maintenant.

3. Perspectives

Tout d'abord, nous présentons des « perspectives à court terme », correspondant à des pistes laissées en suspens ou de côté et qui mériteraient d'être traitées rapidement pour améliorer ou compléter la méthode proposée. Puis, dans un cadre plus général, nous pensons que nos recherches peuvent constituer une ressource pour aborder des questions anciennes, avant de soulever peut-être de nouvelles questions.

3.1. Perspectives à court terme

- Stratégie de choix des images clefs

La question du choix des images clefs s'est posée dès le début de l'étude. Les premières simulations réalisées ont cherché à comparer l'intérêt d'un choix uniforme par rapport à un choix aléatoire à partir de l'espace vidéo.

Le choix aléatoire, simple, propose une couverture de l'espace qui dépend de la fréquence des configurations. Aussi avec cette stratégie, on aura tendance à augmenter la densité d'images clefs dans la zone de l'espace vidéo qui est la plus dense, cette zone correspondant généralement aux positions moyennes de l'articulateur. Par contre, les positions extrêmes, moins nombreuses, seront sous-représentées.

La répartition uniforme est indépendante de la fréquence des images et comme son nom l'indique couvre l'espace de façon uniforme. Les extrêmes sont mieux représentées mais nous avons évalué que l'erreur de reconstruction RMS sur la séquence était au final plus importante. C'est pourquoi, et du fait de notre approche vidéo, sans a priori acoustiques, le choix de la stratégie aléatoire a été validé et ensuite utilisé comme une des conditions d'application de la méthode.

Il existe une troisième possibilité avec les « k-means[10] », que nous n'avons pas étudiée.

Aux termes de nos travaux, il semble à présent nécessaire de se poser à nouveau la question du choix des images clefs, d'autant plus que nous avons désormais une possibilité d'évaluation supplémentaire en considérant l'estimation de formants à partir des contours géométriques extraits.

D'après le signal audio, une distinction parole-silence a pu être réalisée. Est-ce que choisir les images clefs en majorité parmi les trames de parole serait une meilleure stratégie ? Et dans ce cas, pourrait-on alors envisager de réduire le nombre d'images clefs à marquer ?

Des premiers résultats concernant cette étude s'intéressent à l'estimation de la langue sur la séquence Laval43 et montrent que ce changement de stratégie a peu d'influence sur

[10] La méthode des *k*-means est un outil de classification classique qui permet de répartir un ensemble de données en *k* classes homogènes [McQ67].

l'estimation des formants. De même, en réduisant le nombre d'images clefs, on obtient encore des résultats comparables sur les formants estimés.

Au vu de ces résultats, on pourrait alors diminuer le nombre d'images clefs pris en compte dans l'estimation de la langue. Actuellement, cette estimation repose sur le marquage de 200 clefs, c'est beaucoup car le marquage de la langue est une tâche difficile et coûteuse en temps. Mais avec le choix aléatoire d'images clefs, 200 images nous paraissent nécessaires pour une estimation de qualité de la pointe.

Pour tenter de réduire ce nombre, il semble que ce soit au niveau de l'estimation spécifique de la pointe de la langue que d'autres stratégies de choix des images clefs soient à imaginer. Choisir directement des configurations associées à des voyelles ou de consonnes en tant qu'images clefs est une stratégie à ne pas négliger. Par exemple, [BBRS98] montre que la sélection limitée de voyelles ou de cibles consonantiques du corpus de départ mène à un modèle articulatoire qui représente les données avec une précision équivalente à celle obtenue avec un modèle articulatoire basé sur le corpus complet.

Concernant les voyelles, nous ne sommes pas certains que choisir spécifiquement à partir de l'audio des configurations vocaliques plutôt qu'un choix aléatoire réduise beaucoup l'erreur, à cause de l'occlusion de la langue par la mâchoire supérieure pour les [i] notamment. L'étude mérite néanmoins d'être faite.

Mais c'est plutôt sur les consonnes que nous espérons un gain possible grâce à cette stratégie de choix spécifique. En effet, les mouvements de contacts de la pointe sont rapides et donc rares par rapport à l'ensemble des trames de la séquence, leur proportion est trop faible pour qu'ils puissent être capturés par un choix aléatoire d'un petit nombre d'images clefs. Mieux représenter ces instants de contact parmi les images clefs apporterait un gain pour l'estimation du mouvement de la langue. Notre objectif premier en mesurant indépendamment la pointe était d'améliorer l'estimation des instants de contact de la langue. Aussi si nous étions capables de connaître à l'avance certains de ces instants, c'est-à-dire savoir le numéro de quelques trames correspondant à ces contacts, nous pourrions choisir de marquer ces images-là et ainsi espérer améliorer l'estimation réalisée de la pointe de la langue. A titre d'exemple, qu'il serait intéressant de tester, on pourrait considérer le marquage de 100 images clefs choisies aléatoirement pour l'estimation globale de la langue et celui de ces 100 mêmes images complétées de 50 correspondant à des trames de contact pour l'estimation spécifique de la pointe. On aurait ainsi 50 images de moins à marquer pour une précision qu'on pressent équivalente.

Cette question des images clefs est centrale et mériterait d'être encore analysée.

- Extensions à d'autres séquences cinéradiographiques

La méthode d'extraction proposée est dans l'état actuel des choses applicable film par film. Le traitement d'une séquence s'effectue indépendamment de celui d'autres séquences cinéradiographiques. Pour un articulateur donné, l'étape de marquage manuel des images clefs est propre à chaque film et doit être renouvelée pour chaque nouvelle séquence. En effet, des problèmes d'orientation ou de morphologie des locuteurs empêchent pour l'instant de réaliser le traitement d'une séquence à partir d'images clefs extraites d'une autre séquence. De plus, le cadrage des séquences n'est pas toujours identique d'une séquence à une autre et est pourtant essentiel pour l'application de la DCT. Il n'est donc pas possible aujourd'hui de réaliser directement le traitement complet de toute la base de données d'ATR (25 films). On peut cependant espérer trouver une solution à ces problèmes en se tournant vers des techniques de traitement d'images pour tenter de normaliser les séquences et permettre un traitement conjoint de plusieurs d'entre elles.

En attendant, pour faciliter l'utilisation de la méthode et son accessibilité, une interface est actuellement en cours de développement pour permettre le traitement de nouvelles séquences cinéradiographiques. Cette interface donnera la possibilité en partant des images d'origine d'une séquence quelconque d'effectuer le traitement complet de cette dernière. D'abord articulateur par articulateur, l'utilisateur pourra choisir le cadre, calculer les coefficients DCT puis définir les degrés de liberté avant de les marquer sur un nombre d'images clefs qu'il choisira. Ensuite chaque articulateur ayant été estimé sur l'ensemble de la séquence par application de la méthode à partir des paramètres réglés, un calcul de sections sagittales pourra être réalisé après élaboration d'une grille.

L'utilisation de cette interface offre la possibilité de traiter, à terme, la totalité des séquences de la base ATR ainsi que d'autres séquences, comme celle de Zerling [ABB+00]. Le volume de données articulatoires alors disponible permettra l'analyse de plus de 100000 configurations du conduit vocal, dans l'optique de nouvelles voies d'investigations.

3.2. Vers de nouvelles voies de recherche

L'objectif de nos travaux était donc de proposer une méthode d'extraction des mouvements du tractus vocal pour permettre ensuite l'exploitation de ces grandes bases de données cinéradiographiques existantes.

La question est maintenant ouverte de savoir comment exploiter ces données extraites. En quoi des contours géométriques du conduit vocal en contexte dynamique de parole naturelle peuvent-ils intéresser ? Quelles applications vont pouvoir se baser sur de telles données ?

- Modélisation articulatoire

A partir des données brutes extraites, on peut extraire des paramètres de commande afin d'élaborer un modèle articulatoire qui décrirait toutes les formes possibles du conduit vocal à partir d'un petit jeu de paramètres. Nous en avons déjà parlé.
Des travaux inter-locuteurs peuvent ensuite être envisagés à partir de données provenant de différentes séquences.

- Coarticulation

Il est aussi possible d'entrevoir avec cette direction de recherche de nouvelles études sur la modélisation des stratégies d'anticipation, d'articulation et de coarticulation. Ces études sont très souvent basées sur des corpus de logatomes et des données cinéradiographiques ([VSR+03], [CBRH03]). On peut espérer que des données de parole continue contribuent à de nouveaux résultats.

- Etude des consonnes

Nous avons montré que les données extraites de ces séquences cinéradiographiques permettent d'appréhender la production des consonnes. Les modèles articulatoires généralement bien adaptés pour les voyelles se trouvent confrontés aux problèmes spécifiques de la modélisation des consonnes (extraction de mouvements rapides, forme précise du conduit vocal). Nous pensons que des données dynamiques, dans un contexte de parole continue avec plus de variabilité que dans le cas de corpus de logatomes, favorisent l'étude des consonnes. Nous avons aussi pu constater que la synthèse de ces consonnes est encore un domaine de recherche à explorer.

- Etude des nasales

Les résultats préliminaires obtenus à partir du marquage du vélum laisse entrevoir la possibilité d'utiliser ces données directes et dynamiques pour retrouver des résultats et analyser d'une manière plus approfondie la relation entre articulatoire et acoustique.
Ceci sera facilité par le fait que le voile du palais est un des articulateurs les plus rapides à traiter par la méthode semi-automatique.

- Inversion acoustique-articulatoire

L'inversion cherche à récupérer les formes articulatoires à partir du signal acoustique de parole. La difficulté est liée au fait qu'une infinité de formes du conduit vocal peuvent produire le même spectre de parole. Les travaux sur l'inversion acoustique articulatoire reposent largement sur une approche d'analyse par synthèse articulatoire ([PL05], [OL05]) et ont donc généralement besoin d'exemplaires d'association entre des formes de tractus réel et le son correspondant.

- Extension de la méthode à d'autres techniques d'imagerie

Parmi les techniques d'imagerie, l'IRM semble très prometteuse pour l'avenir, avec le développement de l'IRM dynamique. Pourquoi ne pas transposer alors la méthode ici proposée ? Développée dans le cadre de la cinéradiographie, où l'extraction automatique de contours géométriques ne fonctionne pas, la question reste ouverte de savoir si la méthode est intéressante avec des images où les contours sont beaucoup mieux marqués. Partant du principe proposé par notre méthode, nous pouvons envisager de mettre en place une automatisation complète du procédé d'extraction, en rendant automatique l'étape initiale de marquage des images clefs, par application d'algorithmes de suivi de contours sur les images clefs elles-mêmes.

BIBLIOGRAPHIE

[AB86] Abry, C. & Boë, L.-J. (1986), “Laws for lips”, *Speech Communication*, 5, 97-104.

[ABB+00] Arnal, A., Badin, P., Brock, G., Connan, P.-Y., Florig, E., Perez, N., Perrier, P. Simon, P., Sock, R., Varin, L., Vaxelaire, B. & Zerling, J.-P. (2000), “Une base de données cinéradiographiques du français”, *XXIIIèmes Journées d'Etude sur la Parole*, Aussois, France, 425-428.

[ABS+94] Abry, C., Badin, P., Scully, C. & al. (1994), “Sound-to-gesture inversion in speech: The Speech Maps approach”, *Advanced speech applications*, Varghese, K., Pfleger, S. & Lefèvre, J.-P. (eds), Springer Verlag, 182-196.

[ACM+78] Atal, B.S., Chang, J.J., Mathews, M.V. & Tukey, J. W. (1978), “Inversion of articulatory-to-acoustic transformation in the vocal-tract by a computer-sorting technique”, *Journal of the Acoustical Society of America*, 63(5), 1535-1555.

[AHC93] Arman, F., Hsu, A. & Chiu, M.-Y. (1993), “Image processing on compressed data for large video databases”, *Proceedings ACM Multimedia*, Anaheim, Canada, 267-272.

[AKS99] Akgul, Y. S., Kambhamettu, C., & Stone, M. (1999), “Automatic extraction and tracking of the tongue contours”, *IEEE Transactions on Medical Imaging*, 18, 1035-1045.

[ARCM04] Amelot, A., Roubeau, B., Crevier-Buchman, L. & Maeda, S. (2004), “Prise de données simultanées aérodynamiques et fibroscopiques durant la production des voyelles nasales : Comparaison avec des données prises séparément”, *XXVèmes Journées d’Étude sur la Parole*, Fès, Maroc.

[AVFG03] Asci, A., Vaxelaire, B., Ferbach-Hecker, V. & Guedet, M. (2003), “Anticipatory and Carryover Coarticulation in Turkish”, *Proceedings of the International Congress on Phonetics Sciences*, Barcelona, Spain, 419-422.

[BA96] Badin, P. & Abry, C. (1996), “Articulatory synthesis from X-rays and inversion for an adaptive speech robot”, *Proceedings of the International Conference on Spoken Language Processing*, 1125-1128, Philadelphia, PA, USA.

[Bad91] Badin, P. (1991), “Fricative consonants: acoustic and X-ray measurements”, *Journal of Phonetics*, 19, 397-408.

[Bas78] Bassili, J. N. (1978), “Facial motion in the perception of faces and of emotional expressions”, *Journal of Experimental Psychology : Human Perception and Performance*, 4, 373–379.

[BB99a] Barker, J.P. & Berthommier, F. (1999), “Estimation of speech acoustics from visual speech features: a comparison of linear and non-linear models”, *Proceedings of Audio Visual Speech Processing*, Santa Cruz, California, USA, 112-117.

[BB99b] Barker, J.P. & Berthommier, F. (1999), “Evidence of correlation between acoustic and visual features of speech”, *Proceedings of International Congress of Phonetic Science*, San Francisco, California, USA.

[BB02] Bailly, G. & Badin, P. (2002), “Seeing tongue movements from outside”, *Proceedings of the International Conference on Speech and Language Processing*, Boulder, Colorado, 1913-1916.

[BBB01] Beautemps, D., Badin, P. & Bailly, G. (2001), “Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling”, *Journal of the Acoustical Society of America,* 109(5).

[BBB+96] Beautemps, D., Badin, P., Bailly, G., Galvan, A. & Laboissière, R. (1996), “Evaluation of an articulatory-acoustic model based on a reference subject”, *1st ESCA Tutorial and research workshop on speech production – 4th Speech Production Seminar*, Autrans, France.

[BBB+00] Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M. & Segebarth, C. (2000), “Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images”, *5th International Seminar on Speech Production : Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, 261-264.

[BBRS98] Badin, P., Bailly, G., Raybaudi, M. & Segebarth, C. (1998), “A Three-dimensional linear articulatory model based on MRI data”, *Proceedings of the International Conference on Spoken Language Processing*, Sidney, Australia.

[BBL95] Beautemps, D., Badin, P. & Laboissière, R. (1995), “Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data”, *Speech Communication*, 16, 27-47.

[BBV98] Bailly, G., Badin, P. & Vilain, A. (1998), “Synergy between jaw and lips/tongue movements: Consequences in articulatory modelling“, *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 5, 1859-1862.

[BEG03] Beskow, J., Engwall, O. & Granström, B. (2003), “Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements”, *Proceedings of International Congress of Phonetic Science*, Barcelona, Spain.

[Ber04] Berthommier, F. (2004), “Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Montreal, Québec, Canada.

[BF84] Badin, P. & Fant, G. (1984), “Notes on vocal tract computation”, *Speech Transmission Laboratory - Quarterly Progress Status Report* - Stockholm*,* 2-3, 53-108.

[BGB+95] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.-P. & Brock, G. (1995), “Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model”, *Proceedings of the International Conference on Acoustics*, Trondheim, Norway, 4, 349-352.

[BGGN91] Baer, T., Gore, J.C., Gracco, L.C. & Nye, P.W. (1991), “Analysis vocal tract shape and dimensions using magnetic resonance imaging : Vowels”, *Journal of the Acoustical Society of America*, 90(2), 799-828.

[BGO02] Bailly, G., Gibert, G. & Odisio, M. (2002), “Evaluation of Movement Generation Systems Using the Point-Light Technique”, *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA.

[BGP+95] Boë, L.-J., Gabioud, B., Perrier, P., Schwartz, J.-L. & Vallée, N. (1995), “Vers une unification des espaces vocaliques”, *Levels in Speech Communication: Relations and Interactions,* C. Sorin et al. (eds.), Elsevier B.V., 63-71.

[BJ03] Birkholz, P. & Jackel, D. (2003), “A three-dimensional model of the vocal tract for speech synthesis”, *Proceedings of the International Congress of Phonetic Science*, Barcelona, Spain.

[BL96] Berger, M.-O. & Laprie, Y. (1996), “Tracking articulators in x-ray images with minimal user interaction: example of the tongue extraction”, *Proceedings of IEEE International Conference on Image Processing*, Lausanne, Switzerland.

[BMP94] Boë, L.-J., Maeda, S. & Perrier, P. (1994), “La modélisation articulatoire : un demi siècle d'évolution entre fonctionnel, physique et biomécanique”, *XXèmes Journées d'Etude sur la Parole*, Trégastel, France, 41-54.

[BPB92] Boë, L.-J., Perrier, P. & Bailly, G. (1992), “The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion”, *Journal of Phonetics*, 20, 27-38.

[BPC84] Bertenthal, B.I., Proffitt, D.R. & Cutting, J.E. (1984), “Infant sensitivity to figural coherence in biomechanical motions”, *Journal of Experimental Child Psychology*, 37, 213-230.

[BPGS89] Boë, L.-J., Perrier, P., Guérin, B. & Schwartz, J.-L. (1989), “Maximal Vowel Space”, *Proceedings of the European Conference on Speech Communication and Technology,* Paris, France, 281-284.

[BPK87] Bertenthal, B.I., Proffitt, D.R. & Kramer, S.J. (1987), “Perception of biomechanical motions by infants: Implementation of various processing constraints. Special Issue: The ontogenesis of perception”, *Journal of Experimental Psychology: Human Perception and Performance*, 13, 577-585.

[BPR03] Bergeson, T. R., Pisoni, D. B. & Reynolds, J. T. (2003), “Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants”, *Proceedings of the Auditory-Visual Speech Processing Workshop*, St Jorioz, France, 55–60.

[BW05] Boersma, P. & Weenink, D. (2005), “Praat : doing phonetics by computer (Version 4.3.14)” [Computer program]. Retrieved May 6, 2005, from http://www.praat.org/.

[BSWZ86] Bothorel, A., Simon, P., Wioland, F. & Zerling, J.P. (1986), ”Cinéradiographie des voyelles et consonnes du français”, *Travaux de l’Institut de Phonétique de Strasbourg*, 296 pages.

[Cal89] Calliope (1989), “La Parole et son traitement automatique”, *Collection technique et scientifique des télécommunications*, Masson, Paris.

[Cat94] Cathiard, M.-A. (1994), “La perception visuelle de l'anticipation des gestes vocaliques: cohérence des évènements audibles et visibles dans le flux de la parole”, *Doctorat de Psychologie Cognitive*, Université Grenoble 2.

[CBRH03] Connan, P.-Y., Brock, G., Roy, J.-P. & Hirsch, F. (2003), “Using Digital Cine-Radiography to Study Anticipatory Labial Activity in French”, *Proceedings of the International Congress on Phonetics Sciences*, Barcelona, Spain, 3153-3156.

[CF66] Coker, C. H & Fujimura, O. (1966), “Model for specification of the vocal-tract area function”, *Journal of the Acoustical Society of America*, 40, 1271.

[Cha95] Chang, S.-F. (1995), “Compressed domain techniques for image/video indexing and manipulation”, *Proceedings of the IEEE International Conference on Image Processing*, 1, 314-317.

[Chi80] Chistovich, L.A. (1980), “Auditory processing of speech”, *Language and Speech,* 23(1), 67-73.

[CK41] Chiba, T. & Kajiyama, M. (1941), “The vowel: Its nature and structure”, *The Phonetic Society of Japan*, Tokyo.

[Cok76] Coker, C.H (1976), “A model of articulary dynamics and control”, *Proceedings of the IEEE*, 64(4), 452-460.

[CPK78] Cutting, J.E., Proffitt, D.R. & Kozlowski, L.T. (1978), “A biomechanical invariant for gait perception”, *Journal of Experimental Psychology: Human Perception and Performance*, 4**,** 357-372.

[CR99] Cham, T. & Rehg, J. (1999), “A multiple hypothesis approach to figure tracking”, *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 239–245.

[CSL79] Chistovich, L.A., Sheikin, R.L. & Lublinskaya, V.V. (1979), “« Centres of Gravity » and Spectral Peaks as the Determinants of Vowel Quality”, *Frontiers of Speech Communication Research*, Lindblom, B. & Öhman, S. (eds), London/New York/San Francisco, Academic Press.

[Dar87] Dart, S.N. (1987), “A bibliography of X-ray studies of speech”, *UCLA Working Papers in Phonetics*, 66, 1-97.

[DDS96] Davis, E.P., Douglas, A.S. & Stone, M. (1996), “A Continuum Mechanics Representation of Tongue Deformation”, *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, 788-792.

[Del48] Delattre, P. (1948), “Un triangle acoustique des voyelles orales du français”, *French Review*, 21, 477-484.

[DMS02] Delvaux, V., Metens, T. & Soquet, A. (2002), “Propriétés acoustiques et articulatoires des voyelles nasales du français”*, XXIV^{èmes} Journées d'Étude sur la Parole*, Nancy, France.

[Dix99] Dixit, P. (1999), “Palatometric investigation of selected coronal consonants of Hindi”, *Proceedings of the International Congress on Phonetic Sciences*, San Francisco, USA, 1, 439-441.

[DMS00] Demolin, D., Metens, T. & Soquet, A. (2000), “Real time MRI and articulatory coordinations in vowels”, *Proceedings of the 5th Seminar on Speech Production*, Kloster Seeon, Germany, 86-93.

[DODS06] Denby, B., Oussar, Y., Dreyfus, G. & Stone, M. (2006), “Prospects for a silent speech interface using ultrasound imaging”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France.

[DS04] Denby, B. & Stone, M. (2004), “Speech synthesis from real ultrasound images of the tongue”, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Montreal, Québec, Canada, 685-688.

[Dus00] Dusan, S.V. (2000), “Statistical Estimation of Articulatory Trajectories from the Speech Signal Using Dynamical and Phonological Constraints”, *PhD Thesis*, University of Waterloo, Ontario, Canada.

[Eng99] Engwall, O. (1999), “Modeling of the Vocal Tract in three Dimensions”, *Proceedings of the European Conference on Speech Communication and Technology*, Budapest, Hungary, 113-116.

[Eng00a] Engwal, O. (2000), “Dynamical aspects of coarticulation in Swedish fricatives – a combined EMA & EPG study”, *Speech Transmission Laboratory - Quarterly Progress Status Report* - Stockholm, 4.

[Eng00b] Engwall, O. (2000), “A 3D tongue model based on MRI data”, *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 901-904.

[Eng01] Engwall, O. (2001), “Synthesizing static vowels and dynamic sounds using a 3D vocal tract model”, *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire, Scotland.

[Eng03] Engwall, O. (2003), “Combining MRI, EMA and EPG measurements in a three-dimensional tongue model”, *Speech Communication*, 41(2-3), 303-329.

[Eng04] Engwall, O. (2004), “From real-time MRI to 3D tongue movements”, *Proceedings of the International Conference on Spoken Language Processing*, Jeju Island, Korea.

[Fan60] Fant, G. (1960), “Acoustic theory of speech production”, The Hague : Mouton.

[Fan73] Fant, G. (1973), “Speech Sounds and Features”, MIT, Cambridge, Mass.

[FB05] Fontecave, J. & Berthommier, F. (2005), “Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database”, *Proceedings of the European Conference on Speech Communication and Technology*, Lisboa, Portugal.

[FB06a] Fontecave, J. & Berthommier, F. (2006), “Extraction semi-automatique des mouvements du conduit vocal à partir de données cinéradiographiques”, *XXVIèmes Journées d'Etude sur la Parole*, Dinard, France.

[FB06b] Fontecave, J. & Berthommier, F. (2006), “Semi-automatic extraction of vocal tract movements from cineradiographic data”, *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, USA.

[FB06c] Fontecave, J. & Berthommier, F. (2006), “Articulatory synthesis driven by geometrical contours of the vocal tract extracted from cineradiographic data”, *Proceedings of the 7th Seminar on Speech Production*, Ubatuba, Brazil.

[Fla55] Flanagan, J. (1955), “A Difference Limen for Vowel Formant Frequency”, *Journal of the Acoustical Society of America*, 27(3), 613-617.

[Fla72] Flanagan, J. (1972), “Speech Analysis Synthesis and Perception”, Springer-Verlag, New-York.

[Fla84] Flament, B. (1984), “Recherche sur la mise en relief en français. Approche théorique et essai de caractérisation phonétique à partir de données de la mingographie et de la radiocinématographie”, *Doctorat d’Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg, France.

[FSN+95] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995), “Query by image and video content: the QBIC system”, *IEEE Computer*, 28(9), 23-32.

[GAB96] Guiard-Marigny, T., Adjoudani, A. & Benoît, C. (1996), “3D models of the lips and jaw for visual speech synthesis”, *Progress in speech synthesis*, J.P.H. van Stanten, R. W. Sproat, J.-P. Olive and J. Hirschberg (eds.), Springer-Verlag, New York.

[Gal97] Galvan-Rdz, A. (1997), “Etudes dans le cadre de l’inversion acoustico-articulatoire : Amélioration d’un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des plosives”, *Thèse de Doctorat*, Institut de la Communication Parlée, INP Grenoble, France.

[Gav00] Gavrila, D.M. (2000), “Pedestrian detection from a moving vehicle”, *Proceedings of the European Conference on Computer Vision*, Dublin, Ireland.

[GBP+91] Gay, T., Boé, L.-J., Perrier, P., Feng, G. & Swayne, E. (1991), “The acoustic sensitivity of vocal tract constrictions: a preliminary report”, *Journal of Phonetics*, 19, 445-452.

[GP03] Giese, M.A. & Poggio, T. (2003), “Neural mechanisms for the recognition of biological movements”, *Neuroscience*, 4(3),179-192.

[Gro05] Grosgeorges, A. (2005), “Etudes des caractéristiques perceptives de la parole réduite chez les sujets sains et les implantés cochléaires”, *Thèse de Doctorat*, INP Grenoble, France.

[Har72] Hardcastle, W.J. (1972). “The use of electropalatography in phonetic research”, *Phonetica*, 25, 197-215.

[HBSK00] Heckmann, M., Berthommier, F., Savariaux, C. & Kroschel, K. (2000), “Labeling audio-visual speech corpora and training an ANN/HMM audio-visual

speech recognition system", *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China.

[HBSK03] Heckmann, M., Berthommier, F., Savariaux, C. & Kroschel, K. (2003), "Effects of Image Distorsions on Audio-Visual Speech Recognition", *Proceedings of the Audio Visual Speech Processing*, St Jorioz, France, 163-168.

[HGW01] Hoole, P., Geng, C. & Winkler R. (2001), "Towards a speaker-independent representation of tongue-posturing for speech", *Proceedings of the 4th International Speech Motor Conference*, Nijmegen, The Netherlands.

[HI91] Henningsson, G. & Isberg, A. (1991), "A Cineradiographic Study of Velopharyngeal Movements for Deviant Versus Nondeviant Articulation", *The Cleft Palate-Craniofacial Journal*, 28(1), 115-118.

[HKSB02] Heckmann, M., Kroschel, K., Savariaux, C. & Berthommier, F. (2002), "DCT-based video features for Audio-Visual speech recognition", *Proceedings of the International Conference on Spoken Language Processing*, Denver, USA, 1925-1928.

[HLG77] Harshman, R., Ladefoged, P. & Goldstein, L. (1977), "Factor analysis of tongue shapes", *Journal of the Acoustical Society of America,* 62(3), 693-707.

[HS64] Heinz, J.M. & Stevens, K.N. (1964), "On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech", *Journal of the Acoustical Society of America,* 36(S4), 1037.

[HS65] Heinz, J.M. & Stevens, K.N. (1965), "On the relations between lateral cineradiographs area functions and acoustic spectra of speech", *Proceedings of the 5th International Congress of Acoustics,* Liège, Paper A44.

[HS82] Hashimoto, K. & Sasaki, K. (1982), "On the relationship between the shape and position of the tongue for vowels", *Journal of Phonetics*, 10, 291-299.

[KDLF97] Kobla, V., Doermann, D., Lin, K.-I. & Faloutsous, C. (1997), "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video", *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Database*, 200-210.

[KHM05] Kim, H., Honda, K. & Maeda, S. (2005), "Stroboscopic-cine MRI study of the phasing between the tongue and the larynx in the Korean three-way phonation contrast", *Journal of Phonetics*, 33, 1-26.

[KWT87] Kass, M., Witkin, A.P., & Terzopoulos, D. (1987), "Snakes: Active Contour Models", *International Journal of Computer Vision*, 4(1), 321-331.

[KO83] Keller, E. & Ostry, D. (1983), "Computerized measurement of tongue dorsum movement with pulsed echo ultrasound", *Journal of the Acoustical Society of America,* 73, 1309-1315.

[Isk05] Iskarous, K. (2005), "Patterns of tongue movement", *Journal of Phonetics*, 33(4), 363-381.

[JAB+02] Jiang, J., Alwan, A., Bernstein, L.E., Auer, E.T. & Keating, P.A. (2002), "Predicting face movements from speech acoustics using spectral dynamics",

Proceedings of the International Conference on Multimedia and Expo, Lausanne, Switzerland, 181-184.

[Joh73] Johansson, G. (1973), "Visual perception of biological motion and a model for its analysis", *Perception and Psychophysics*, 14, 201-211.

[Joo48] Joos, M. (1948), "Acoustic phonetics", *Baltimore, Linguistic Society of America*.

[KC77] Kozlowski, L.T. & Cutting, J. E. (1977), "Recognizing the gender of walkers from dynamic point-light displays", *Perception and Psychophysics*, 21, 575-580.

[KIF75] Kiritani, S., Itoh, K. & Fujimura, O. (1975), "Tongue-pellet tracking by a computer-controlled x-ray microbeam system", *Journal of the Acoustical Society of America*, 57, 1516-1520.

[KZ96] Keller, E. & Zellner, B. (1996), "A timing model for fast French", *York Papers in Linguistics*, University of York, 17, 53-75.

[LA84] Le Huche, F. & Allali, A. (1984), "La voix. Tome 1 : Anatomie et physiologie des organes de la voix et de la parole", Masson, Paris.

[Lal91] Lallouache, M.T. (1991), "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", *Thèse de Doctorat*, INP Grenoble, France.

[LB96] Laprie, Y. & Berger, M.-O. (1996), "Extraction of tongue contours in x-ray images with minimal user interaction", *Proceedings of the International Conference on Spoken Language Processing,* Philadelphia, USA, 268-271.

[LeG91] Le Gall, D. (1991), "MPEG: A video compression standard for multimedia applications", *Communications of ACM*, 34(4), 46-58.

[Lil71] Liljencrants, J. (1971), "Fourier series description of the tongue profile", *Speech Transmission Laboratory - Quarterly Progress Status Report* - Stockholm, (4), 9-18.

[LL72] Liljencrants, J. & Lindblom, B. (1972), "Numerical Simulations of vowel quality systems: the role of perceptual contrast" , *Language*, 48 : 839-862.

[LM85] Liberman, A.M. & Mattingly, I.G. (1985), "The motor theory of speech production revised", *Cognition*, 21, 1-36.

[LS71] Lindblom, B. & Sundberg, J. (1971), "Acoustical consequences of lip, tongue and jaw movements", *Journal of the Acoustical Society of America*, 50, 1166-1179.

[Mae78] Maeda, S. (1978), "Une analyse statistique sur les positions de la langue : Etude préliminaire sur les voyelles françaises", *IX^{èmes} Journées d'Etude sur la Parole*, Lannion, France, 191-199.

[Mae79] Maeda, S. (1979), "Un modèle articulatoire de la langue avec des composantes linéaires", *X^{èmes} Journées d'Etude sur la Parole*, Grenoble, France,152-164.

[Mae82] Maeda, S. (1982), “A digital simulation method of the vocal-tract system”, *Speech Communication*, 1, 199-229.

[Mae90] Maeda, S. (1990), “Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model”, *Speech Production and Speech Modeling*, Hardcastle, W.J. & Marchal, A. (eds), Kluwer Academic Publishers, 131-149.

[Mae91] Maeda, S. (1991), “On articulatory and acoustic variabilities”, *Journal of Phonetics*, 19, 321-331.

[Mak80] Makhoul, J. (1980), “A fast cosine transform in one and two dimensions”, *IEEE Transactions on Acoustic Speech and Signal Processing,* 28(1), 27-34.

[Mat99] Mathieu, B. (1999), “Modèles de production de parole et reconnaissance à partir d’automates”, *Thèse de Doctorat*, Université Heni-Poincaré, Nancy, France.

[MBVB96] Mawass, K., Badin, P., Vescovi, C. & Beautemps, D. (1996), “Evaluation d'un modèle de source de friction pour la synthèse articulatoire des consonnes fricatives”, *XXI^{èmes} Journées d'Etude sur la Parole*, Avignon, France, 367-370.

[McQ67] McQueen, J. (1967), “Some methods for classification and analysis of multivariate observations”, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

[Mer73] Mermelstein, P. (1973), “Articulatory model for the study of speech production”, *Journal of the Acoustical Society of America*, 53,1070-1082.

[Mer78] Mermelstein, P. (1978), “Difference limens for formant frequencies of steady-state and consonant-bound vowels”, *Journal of the Acoustical Society of America*, 63(2), 572-580.

[ML97] Mathieu, B. & Laprie, Y. (1997), “Adaptation of Maeda’s model for acoustic to articulatory inversion”, *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 2015-2018.

[MMC^{+}97] Mohammad, M., Moore, E., Carter, J.N., Shadle, C.H. & Gunn, S.J. (1997), “Using MRI to image the moving vocal tract during speech”, *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 2027-2030.

[MTH^{+}99] Masaki, S., Tiede, M.K., Honda, K., Shimada, Y., Fujimoto, I., Nakamura, Y. & Ninomiya, N. (1999), “MRI-based speech production study using a synchronized sampling method”, *Journal of the Acoustical Society of Japan*, 20(5), 375-379.

[MVT94] Munhall, K., Vatikiotis-Bateson, E., & Tohkura, Y. (1994), “X-ray film database for speech research”, *Technical report TR-H-116*, ATR Human Information Processing Laboratories, Kyoto.

[MVT95] Munhall, K.G., Vatikiotis-Bateson, E. & Tohkura, Y. (1995), “X-ray Film database for speech research”, *Journal of the Acoustical Society of America,* 98, 1222-1224.

[NA00] Narayanan, S. & Alwan, A. (2000), “Noise Source Models for Fricative Consonants”, *IEEE Transactions on Speech and Audio Processing*, 8(2), 328-344.

[NNL+04] Narayanan, S., Nayak, K., Lee, S., Sethy, A. & Byrd, S. (2004), “An approach to real-time magnetic resonance imaging for speech production”, *Journal of the Acoustical Society of America*, 115(4), 1771-1776.

[NT91] Nagasaka, A. & Tanaka, T. (1991), “Automatic Video Indexing and Full-Video Search for Object Appearances”, *Proceedings of the Working Conference on Visual Database Systems*, 119-133.

[Ohm66] Öhman, S. (1966), “Coarticulation in VCV utterances: Spectrographic measurements”, *Journal of the Acoustical Society of America,* 39(1), 151-168.

[OL00] Ouni, S. & Laprie, Y. (2000), “Improving acoustic-to-articulatory inversion by using hypercube codebooks”**,** *International Conference on Spoken Language Processing*, Beijing, Chine.

[OL05] Ouni, S. & Laprie, Y. (2005), “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”, *Journal of the Acoustical Society of America*, 118, 444-460.

[Osh84] O’Shaughnessy, D. (1984), “A multispeaker analysis of durations in read French paragraphs”, *Journal of the Acoustical Society of America,* 76, 1664-1672.

[PB52] Peterson, G. & Barney, H. (1952), “Control Methods used in a Study of the Vowels”, *Journal of the Acoustical Society of America,* 24(2).

[PBS92] Perrier, P., Boë, L.-J. & Sock, R. (1992), “Vocal tract area functions estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modelling the transition with two sets of coefficients”, *Journal of Speech and Hearing Research*, 35, 53-67.

[PCS+92] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. & Jackson M. (1992), “Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements”, *Journal of the Acoustical Society of America,* 92(6), 3078-3096.

[Per69] Perkell, J. (1969), “Physiology of speech production: results and implications of a quantitative cineradiographic study”, *M.I.T. Press*, Cambridge, MA.

[Per74] Perkell, J. (1974), “A Physiologically-Oriented Model of Tongue Activity during Speech Production”, *PhD Thesis*, M.I.T., Cambridge, MA.

[PGC98] Potamianos, G., Graf, H.P. & Cosatto, E. (1998), “An Image Transform Approach for HMM Based Automatic Lipreading”, *Proceedings of International Conference on Image Processing*, Chicago, USA, 3, 173-177.

[PL05] Potard, B. & Laprie, Y. (2005), “Using phonetic constraints in acoustic-to-articulatory inversion”, *Proceedings of the European Conference on Speech Communication and Technology,* Lisbon*,* Portugal*.*

[PO93] Parush, A. & Ostry, D.J. (1993), “Lower pharyngeal wall coarticulation in VCV syllables”, *Journal of the Acoustical Society of America*, 94, 715-722.

[PP93] Payan, Y. & Perrier, P. (1993), “Vowel normalization by articulatory normalization : first attempts for vowel transitions”, *Proceedings of the European Conference on Speech Communication and Technology,* Berlin, Germany, 417-420.

[PYS+06] De Paula, H., Yehia, H.C., Shiller, D., Jozan, G., Munhall, K.G. & Vatikiotis-Bateson, E. (2006), “Analysis of audiovisual speech intelligibility based on spatial and temporal filtering of visible speech information”, *Speech Production: Models, Phonetic Processes and Techniques*, Harrington & Tabain (eds), Psychology Press.

[RBF00] Rossato, S., Badin, P. & Feng, G. (2000), “Détermination de la position du voile du palais à partir du signal de parole pour les nasales du français“, *XXIII^{èmes} Journées d'Etude sur la Parole*, Aussois, France.

[RF81] Runeson, S. & Frykholm, G. (1981), “Visual perception of lifted weight”, *Journal of Experimental Psychology: Human Perception and Performance*, 7, 733-740.

[RFR03] Rein, S., Fitzek, F. & Reisslein, M. (2003), “Voice Quality Evaluation for Wireless Transmission with ROHC”, *Proceedings of Internet and Multimedia Systems and Applications*, Honolulu, USA.

[RGBV97] Revéret, L., Garcia, F., Benoit, C. & Vatikiotis-Bateson, E. (1997), “An hybrid approach to orientation free liptracking”, *Proceedings of Audio-Visual Speech Processing*, Rhodes, Grèce.

[RJ93] Rabiner, L. R., Juang, B. H. (1993), “Fundamentals of Speech Recognition”, Prentice-Hall, Englewood Cliffs, NJ.

[RJS96] Rosenblum, L.D. , Johnson, J.A. & Saldaña, H.M. (1996), “Visual kinematic information for embellishing speech in noise”, *Journal of Speech and Hearing Research*, 39(6), 1159-1170.

[Roy03] Roy, J.-P. (2003), “INTRIC, une interface de traitement d’images cinéradiographiques”, *Travaux de l’Institut de Phonétique de Strasbourg*, 163-177.

[RS98] Rosenblum, L.D. & Saldaña, H.M. (1998), “Time-varying information for visual speech perception”, *Hearing by Eye II*, R. Campbell, B. Dodd and D. Burnham (eds.), Psychology Press, 61-81.

[RSG+96] Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., Browman, C. (1996), “CASY and Extensions to the Task-Dynamic Model”, *Proceedings of the 1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics, 4th Speech Production Seminar: Models and Data,* Autrans, France.

[SB05] Serrurier, A. & Badin, P. (2005), “A Three-Dimensional Linear Articulatory Model of Velum Based on MRI Data”, *Proceedings of the European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2161-2164.

[SBVA97a] Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. (1997), “Major trends in vowel system inventories”, *Journal of Phonetics*, 25, 233-253.

[SBVA97b] Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. (1997), “The dispersion-focalization theory of vowel systems”, *Journal of Phonetics*, 25, 255-286.

[Sch67] Schroeder, M.R. (1967), “Determination of the geometry of the human vocal tract by acoustic measurements”, *Journal of the Acoustical Society of America,* 41(4B), 1002-1010.

[SDD+01] Stone, M., Davis, E.P., Douglas, A.S., Ness Aiver, M., Gullapalli, R., Levine, W.S. & Lundberg, A.J. (2001), “Modeling Tongue Surface Contours from Cine-MRI Images”, *Journal of Speech, Language and Hearing Research*, 44, 1026-1040.

[SFK80] Shaw, R.L., Folkins, J.W., & Kuehn, D.P. (1980), “Comparison of methods for measuring velar position from lateral-view cineradiography”, *The Cleft Palate-Craniofacial Journal*, 17, 326-329.

[Sha91] Shadle, C.H. (1991), “The effect of geometry on source mechanisms of fricative consonants”, *Journal of Phonetics*, 19, 409-424.

[SH55] Stevens, K.N. & House, A.S. (1955), “Development of a quantitative of vowel articulation”, *Journal of the Acoustical Society of America*, 27, 484-493.

[SLMD02] Soquet, A., Lecuit, V., Metens, T. & Demolin, D. (2002), “Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI”, *Speech Communication*, 36, 168-180.

[Son79] Sondhi, M.M. (1979), “Estimation of vocal-tract areas: the need for acoustical measurements” *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(3), 268-273.

[SS87] Sondhi, M.M. & Schroeter, J. (1987), “A hybrid time-frequency domain articulatory speech synthesizer”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7), 955-967.

[SSH+81] Sonies, B., Shawker, T., Hall, T., Gerber, L. & Leighton, S. (1981), “Ultrasonic visualization of tongue motion during speech”, *Journal of the Acoustical Society of America,* 70(3), 683-686.

[SSTR88] Stone, M., Shawker, T., Talbot, T. & Rich, A. (1988), “Cross-sectional tongue shape during the production of vowels”, *Journal of the Acoustical Society of America,* 83(4), 1586-1596.

[SSV+03] Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T. & Munhall, K.G. (2003), “Perceiving biological motion: Dissociating talking from walking”, *Journal of Cognitive Neuroscience*, 15(6), 800–809.

[Ste28] Stetson, R.H. (1928), “Motor Phonetics: a study of speech movements in action”, *Archives néerlandaises de phonétique expérimentale, 3*, 1-216. (edition, 1988, Kelso, J.A.S. & Munhall, K.G., Boston).

[Ste72] Stevens, K.N. (1972), “The Quantal Nature of Speech : Evidence from Articulatory-Acoustic Data” , *Human communication : a unified view (E.E. Davis, Jr & P.B. Denes, eds)*, 51-66, New-York : McGraw-Hill.

[Ste89] Stevens, K.N. (1989), “On the quantal nature of speech”, *Journal of Phonetics*, 17, 3-45.

[Ste98] Stevens, K.N. (1998), “Acoustic Phonetics”, *M.I.T. Press*, Cambridge, MA.

[Sto90] Stone, M. (1990), “A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data”, *Journal of the Acoustical Society of America*, 87, 2207-2217.

[Sum79] Summerfield, A.Q. (1979), “Use of visual information in phonetic perception”, *Phonetica*, 36, 314–331.

[Sun69] Sundberg, J. (1969), “Articulatory differences between spoken and sung vowels in singers”, *Speech Transmission Laboratory - Quarterly Progress Status Report* - Stockholm, 10(1), 33-46.

[SVSE76] Strange, W., Verbrugge, R., Shankweiler, D. & Edman, T. (1976), “Consonant environment specifies vowel identity”, *Journal of the Acoustical Society of America*, 60, 213-224.

[SZK+95] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonsky, J. & Ekelid, M. (1995), “Speech recognition with primarily temporal cues”, *Science*, 270, 303-304.

[TB01] Toyama, K. & Blake, A. (2001), “Probabilistic tracking in a metric space”, *Proceedings of the International Conference on Computer Vision*, Vancouver, Canada.

[Thi99] Thimm, G. (1999), “Segmentation of X-ray image sequences showing the vocal tract”, *IDIAP Research Report*, Martigny, Suisse.

[TL99] Thimm, G. & Luettin, J. (1999), “Extraction of articulators in x-ray image sequences”, *Proceedings of the European Conference on Speech Communication and Technology,* Budapest, Hungary, 157-160.

[TM03] Toutios, A. & Margaritis, K. (2003), “Acoustic-to-articulatory inversion of speech: a review”, *Proceedings of the International 12th Turkish Symposium on Artificial Intelligence and Neural Networks*, Canakkale, Turkey.

[Tod06] Toda, M. (2006), “Deux stratégies pour la réalisation du contraste acoustique des sibilantes /□/ et /□/ en français”, *XXVIèmes Journées d’Etude sur la Parole*, Dinard, France.

[Tri05] Triggs, W. (2005), “Reconstruction monoculaire du mouvement humain, et autres travaux 2000-2004”, *Habilitation à diriger des recherches*, INP Grenoble, France.

[TV94] Tiede, M.K. & Vatikiotis-Bateson, E. (1994), “Extracting articulator movement parameters from a videodisc-based cineradiographic database”, *Proceedings of the International Conference on Spoken Language Processing,* Yokohama, Japan, 45-48.

[Ull76] Ullman, S. (1976), “The interpretation of structure from motion”, MIT, Cambridge, USA.

[VAB98] Vilain, A., Abry, C. & Badin, P. (1998), “Coarticulation and degrees of freedom in the elaboration of a new articulatory plant: Gentiane”, *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 7, 3147-3150.

[Vil00] Vilain, A. (2000), “Apports de la modélisation des degrés de liberté articulatoires à l’étude de la coarticulation et du développement de la parole”, *Thèse de Doctorat*, Université Stendhal, Grenoble, France.

[VSR+03] Vaxelaire, B., Sock, R., Roy, J.-P., Asci, A. & Hecker, V. (2003), “Audible and Inaudible Anticipatory Gestures in French”, *Proceedings of the International Congress on Phonetics Sciences*, Barcelona, Spain, 447-450.

[VT00] Veltkamp, R.C. & Tanase, M. (2000), “Content-based image retrieval systems: a survey”, *Technical Report*, Utrecht University, Utrecht, The Netherlands.

[Wak73] Wakita, H. (1973), “Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms”, *IEEE Transactions on Audio and Electroacoustics*, 21(5), 417-427.

[Wal91] Wallace, G.K. (1991), “The JPEG still picture compression standard”, *Communications of ACM*, 34(4), 30-44.

[Wes91] Westbury, J. R. (1991), “The significance and measurement of head position during speech production experiments using the x-ray microbeam system”, *Journal of the Acoustical Society of America*, 89, 1782-1791.

[Wil95] Wilhelms-Tricarico, R. (1995), “Physiological modeling of speech production: Methods for modeling soft-tissue articulators”, *Journal of the Acoustical Society of America*, 97(5), 3085-3098.

[Wio85] Wioland, F. (1985), “Faits de jointure en français. Implications aux niveaux articulatoire et acoustique. Incidences sur le plan des fonctions linguistiques”, *Doctorat d’Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg, france.

[Woo79] Wood, S. (1979), “A radiographic examination of constriction location for vowels”, *Journal of Phonetics*, 7, 25-43.

[Woo91] Wood, S. (1991), “X-ray data on the temporal coordination of speech gestures”, *Journal of Phonetics*, 19(3-4), 281-292.

[WS03] Wang, J. & Singh, S. (2003), “Video analysis of human dynamics - a survey”, *Real-Time Imaging*, 9(5), 321-346.

[Yeh02] Yehia, H.C. (2002), “A study on the speech acoustic-to-articulatory mapping using morphological constraints”, *PhD Thesis*, Nagoya University, Graduate School of Engineering.

[YHC92] Yuille, A.L., Hallinan, P.W. & Cohen, D.S. (1992), “Feature Extraction from Faces using Deformable Templates”, *International. Journal of Computer Vision*, 8(2), 99-111.

[YRV98] Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. (1998), “Quantitative association of vocal-tract and facial behavior”, *Speech Communication*, 26, 24-43.

[YT97] Yehia, H. & Tiede, M. (1997), “A parametric three-dimensional model of the vocal tract based on MRI data”, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1619-1622.

[ZDB+87] Zimmermann, G., Dalston, R.M., Brown, C., Folkins, J.W., Linville, R.N. & Seaver, E.J. (1987), “Comparison of cineradiographic and photodetection techniques for assessing velopharyngeal function during speech”, *Journal of Speech and Hearing Research*, 30(4), 564-569.

[Zer84] Zerling, J.-P. (1984), “Phénomènes de nasalité et de nasalisation vocaliques : étude cinéradiographique pour deux locuteurs”, *Travaux de l'Institut de Phonétique de Strasbourg*, 16, 241-266.

[ZHP03] Zheng, Y., Hasegawa-Johnson, M. & Pizza, S. (2003), “PARAFAC analysis of the three dimensional tongue shape”, *Journal of the Acoustical Society of America*, 113(1), 478-486.

[ZLSW95] Zhang, H.J., Low, C.Y., Smoliar, S.W. & Wu, J.H. (1995), “Video parsing, retrieval and browsing: an integrated and content-based solution”, *Proceedings of the ACM Multimedia Conference*, 15-24.

ANNEXES

A1. QUELQUES ELEMENTS D'ANATOMIE DU CONDUIT VOCAL

Pour générer un message vocal, le locuteur procède à la production d'un ensemble de commandes qui, du système nerveux central jusqu'aux muscles via le système nerveux périphérique, vont permettre de piloter les évolutions de la forme du conduit vocal.

Les sons de la parole ont pour origine des phénomènes aérodynamiques et acoustiques. L'air emmagasiné dans les poumons fournit l'énergie nécessaire à la génération de la voix par le larynx qui excite le conduit vocal.

Le conduit vocal est un tube acoustique d'aire non uniforme limité à ses deux extrémités par la glotte et les lèvres. Les différences homme-femme se situent au niveau du larynx (les hommes ont le larynx plus bas et donc un pharynx plus long). La position et les mouvements des articulateurs vont modifier la forme du conduit vocal. Tout au long de celui-ci, la section peut être nulle (occlusion) et atteindre jusqu'à 20 cm² à l'extrémité labiale. Ces constrictions vont moduler les modes de résonance acoustiques et ainsi générer les différentes unités phonétiques (consonnes et voyelles).

Notons que les organes articulateurs mis en jeu dans la parole ne sont pas spécifiques à sa production. Pour parler, l'homme utilise les deux grandes fonctions physiologiques que sont la respiration et la digestion.

Nous décrivons ici en quelques mots l'organisation générale du conduit vocal ([Ste98], [LA84]). Les articulateurs sont regroupés selon leur localisation en deçà ou au-delà du larynx en deux sous-systèmes, subglottique ou supraglottique.

Le système subglottique englobe les poumons et la trachée. Cet ensemble se comporte comme un générateur de débit d'air qui alimente le larynx.

Le **larynx** contribue à la génération de la source vocale. Alimenté en pression par le système subglottique, il génère par vibration des cordes vocales l'onde qui véhiculera le signal sonore.

Le système supraglottique se compose en deux parties, la partie orale et la partie nasale. La première comprend le pharynx et un ensemble d'articulateurs.

On trouve plusieurs définitions pour les articulateurs. Certains considèrent comme articulateur toute partie mobile du conduit vocal sur laquelle on peut agir volontairement et qui est fonctionnelle dans la production des sons de parole. Dans cette définition, palais et dents ne sont pas compris dans les articulateurs. Pour d'autres, toute partie du conduit vocal est potentiellement un articulateur.

Nous ne détaillerons pas les muscles qui permettent les mouvements des articulateurs.

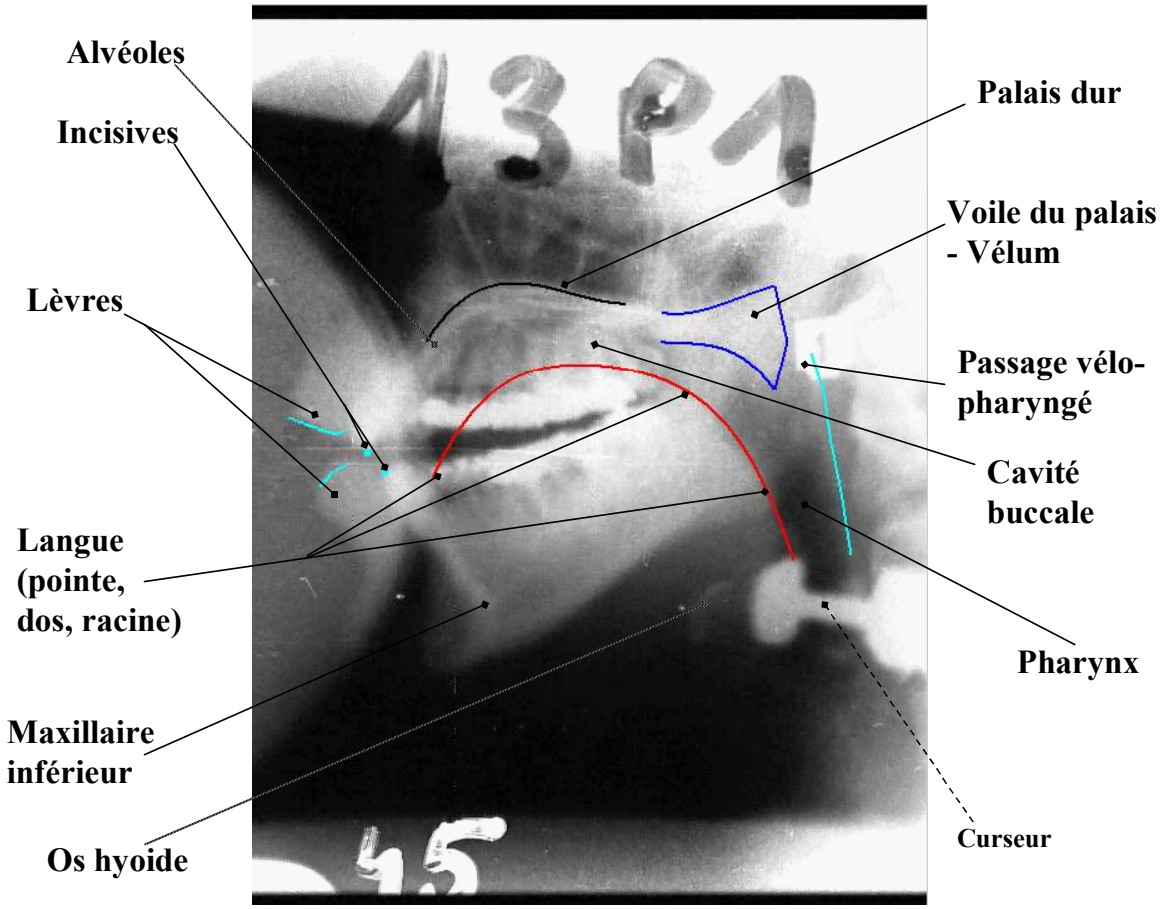


Figure 127 : Exemple de cliché radiologique de la séquence Wioland, avec annotations des articulateurs.

Le **pharynx** s'étend en avant des vertèbres cervicales, il est entouré de muscles, les constricteurs (inférieur, médian, supérieur), qui s'étendent du larynx (qui abritent les cordes vocales) jusqu'au vélum. La contraction de ces muscles réduit le passage de l'air dans le pharynx.

Le **voile du palais** est une cloison musculo-membraneuse mobile qui prolonge en bas et en arrière la voûte palatine. Sa face antéro-inférieure (buccale) est concave. Sa face postéro-supérieure est en continuité avec le plancher des fosses nasales. Il détermine deux configurations distinctes du flux d'air : passage par le nez ou pas. En réalité, une partie de l'air sort toujours par la bouche. En position haute, le vélum est en contact avec la paroi arrière du pharynx, fermant le passage de l'air entre les cavités orales et nasales. En position baissée, le voile du palais peut être en contact avec le dos de la langue (en fonction de la position de cette dernière) et une ouverture est créée entre le pharynx et les cavités nasales.

La voûte palatine ou le **palais dur** est de forme variable. Elle peut en effet être plate ou au contraire profonde, suivant les locuteurs.

Si on considère le pharynx, le vélum et le palais dur comme la partie supérieure du conduit vocal, alors la langue est considérée comme la partie inférieure.

La **langue** est une masse musculaire et muqueuse. Elle est portée par la mandibule et est fonctionnellement divisée en 3 : la racine, le dos (ou corps), et la pointe (ou apex). Son dos forme un arc de près de 90° à peu près circulaire. L'apex a une mobilité spécifique (indépendance vis à vis du reste de la langue). La langue a une action sur la configuration du conduit vocal, dans la production des voyelles comme dans celle des consonnes. On distingue deux grands types d'écoulement de l'air : central et latéral. La langue peut bloquer le flux sur toute la largeur ou laisser un passage par les côtés.

Sa face dorsale, ses bords, sa pointe et la partie antérieure de sa face inférieure sont revêtus par une muqueuse et sont libres dans la cavité buccale. La base ou la racine de la langue reçoit des vaisseaux et des nerfs et s'attache par de nombreux muscles à l'os hyoïde, au maxillaire inférieur, à la voûte palatine. Par ses muscles (17 au total), la langue est douée d'une grande mobilité, grâce à laquelle elle intervient dans la mastication, la déglutition et la phonation. Au repos, la langue occupe la majeure partie de la cavité buccale ; sa face dorsale reste à distance de la voûte palatine.

Le conduit vocal est terminé par l'ouverture formée par les **lèvres**, qui sont deux replis musculo-membraneux.

A2. CORPUS DES SEQUENCES TRAITEES

1. Séquence Wioland

1 – La vie limite
2 – Une belle horde
3 – Il lit posément
4 – Un bel ordre
5 – La vie imite
6 – Il l'y pose et ment
7 – J'étais à Sète, à Nîmes, à Sion
8 – La couple les comble
9 – Le couple est complet
10 – J'étais à cette animation
11 – Le couplet complet
12 – Le pape a dit
13 – Une étrange hernie
14 – Avec panache
15 – Le papa dit
16 – Avec une hache
17 – Une étrangère nie
18 – C'est un invalide
19 – Je vois trois petits trous
20 – L'abstinent s'affaiblit
21 – Il pense hardiment
22 – C'est un nain valide
23 – Je vois trois petites roues
24 – L'abstinence affaiblit
25 – Il pensa sa blessure
26 – Il pensa sa blessure
27 – Il pensa sa blessure
28 – C'est aérien
29 – Il t'a hérissé
30 – Il pense à sa blessure
31 – Il la jeta étonné
32 – Il l'acheta et toucha
33 – J'ai vu cette horde
34 – Il sait tordre
35 – Une robe havane
36 – Un héros bavard
37 – Une robe avantageuse
38 – C'est un faraud
39 – Une blague hasardeuse
40 – Tu blagues hardiment
41 – Mon frère hume
42 – Une rive hasardeuse
43 – Un vrai rhume
44 – Tu y vas à droite
45 – Il rase hardiment
46 – C'est un signalement
47 – Avec honte

48 – Il y deux fautes graves
49 – Ce sont neuf hautes maisons
50 – C'est une touche hantée
51 – Il faut vingt cordes
52 – Il a tout chanté
53 – Il y avait vingt-cinq hordes
54 – Pour ses comptes
55 – Un rang d'ormes
56 – Une grande horde
57 – C'est un signe hardi
58 – Pierre est à la porte et la montre
59 – C'est un signe allemand
60 – Pierrette alla porter la montre
61 – Cette accolade est sincère, j'le pense
62 – C'est ta colle à dessin, Serge le pense
63 – T'as vu le vieil hareng saur
64 – T'as vu le vieillard en sort
65 – Il pensa sa blessure

2. Séquence Flament

1 – C'est le pont qu'il prendra
2 – Il connaît le pont d'Avignon
3 – C'est ce taon qui t'a piqué
4 – Il marque le temps fort
5 – C'est ce coup qu'il encaisse
6 – Il donne le coup de grâce
7 – C'est cette baie qui est jolie
8 – La petite baie vitrée
9 – C'est le brun qui lui va
10 – Tu choisis le brun clair
11 – C'est ce vœu que tu as fait
12 – Elle prend le dé à coudre
13 – C'est ce gars qu'il connaît
14 – Les joyeux gars de la marine
15 – C'est ce fou qui arrive
16 – Tu te fous de lui
17 – C'est cette scie qu'il achète
18 – Il achète scies et marteaux
19 – C'est ce chat qui miaule
20 – Il voit le chat noir
22 – Le père le veut bien
23 – C'est ce zoo que tu connais
24 – Il voit le zoo de Vincennes
25 – C'est ce jus qui est amer
26 – Il aime le jus d oranges
27 – C'est ce lin qui est solide
28 – Un drap de lin blanc
29 – C'est ce riz que nous mangeons
30 – Il aime le riz au lait
31 – C'est ce mont qu'il a franchi
32 – Tu vois le Mont Blanc
33 – C'est le Nord qu'elle indique

34 – Il habite le Nord Ouest
35 – C'est cette gnole qui est forte
36 – Il n y a plus de gnole pour toi
37 – C'est un oui qu'il a dit
38 – Il a ouï faiblement
39 – C'est ce yacht qu'il a acheté
40 – Il a opté pour le yacht neuf
41 – C'est sur le huit qu'il a misé
42 – Il a vu huit ennemis
43 – Ce qu'il est paresseux
44 – Qu est-ce qu'il tâtonne
45 – C'est vraiment un cochon
46 – Ce sont de vrais bandits
47 – Il est assez paresseux
48 – Quelquefois il tâtonne
49 – Il a vu un cochon
50 – Ils ont pris les bandits
51 – C'est un vrai dégoûtant
52 – Le sol est assez dégoûtant
53 – C'est vraiment un gueulard
54 – On le tient pour un gueulard
55 – Ce qu il est furieux
56 – Il est assez furieux
57 – C'est un vrai cinglé
58 – Il était cinglé
59 – Ce qu'il est chouchouté
60 – La fille est chouchoutée
61 – C'est vraiment avec
63 – Ce qu'ils sont zélés
64 – Ils seront zélés
65 – C'est un bateau gigantesque
66 – Il voit un bateau gigantesque

3. Séquence Laval43

1 – Mes amis ont créé le théâtre
2 – Le léopard réintègre sa cage
3 – La trahison frappa les truands
4 – Le co-auteur est anéanti
5 – Le brouhaha incite au chahut
6 – Les enfants avancent cahin-cahan
7 – Les embruns invitent à la rêverie
8 – C'est une clarté extraordinaire
9 – C'est une situation engagée
10 – Il a obtenu un mets infecte
11 – La prohibition est une défense
12 – Les indo-européens importent
13 – Les documents indiens coïncident
14 – En coopération on est frères
15 – Louis est un enfant orgueilleux
16 – Le valet ignorait où j'étais
17 – Il est dans un état euphorique
18 – C'est une assemblée eucharistique

19 – Il a lu auprès de la fenêtre
20 – Le cadet emporta un ballon
21 – Il a un comportement hâbleur
22 – J'aimais obéir à mes parents
23 – A l'opéra on chante et on danse
24 – Martin identifie un objet
25 – L'hindou orphelin œuvre pour les pauvres
26 – Mon parrain et mon époux importent
27 – Les assistants oublient leur devoir
28 – Quelqu'un use mon crayon ou ma gomme
29 – On le lui reprocha âprement
30 – Le défunt est enfin amené
31 – Elle a formulé des vœux hâtifs
32 – Quelqu'un heureusement l'a emmené
33 – Mon mari avait ouvert la porte
34 – Il frappa au-dessus du heurtoir
35 – Gomel et Oufa sont des villes russes
36 – Quelqu'un autorisa une sortie
37 – Le divan couvert d'auréoles
38 – Il revint quand il en a eu un

A3. INTERFACE DE MARQUAGE GEOMETRIQUE

Le marquage manuel des images clefs choisies pour chaque articulateur est une étape essentielle de la méthode d'extraction quasi-automatique mise en place dans cette thèse. Comme mentionné dans le manuscrit, une interface de marquage a été élaborée pour permettre et faciliter cette étape. Il s'agit d'un programme Matlab.
D'abord élaborée pour le marquage des degrés de liberté de la langue de la séquence Wioland, elle a pu facilement être adaptée aux séquences Flament et Laval43 et à d'autres articulateurs.

1. Description

L'interface se présente de la façon suivante (ici avec Laval43, Fig. 128) :

- l'image à marquer : il peut s'agir de l'image complète (pour la langue en particulier) ou du cadre focalisé sur l'articulateur à marquer
- le numéro de l'image à marquer
- des boutons autour de l'image permettant à l'utilisateur de naviguer dans l'interface et surtout, de marquer les points décrivant l'articulateur considéré
    - ***precedente*** : pour aller à l'image à marquer qui précède
    - ***suivante*** : pour aller à l'image à marquer qui suit
    - ***image 1*** : pour revenir à la première image à marquer
    - ***fin du test*** : pour terminer et quitter l'interface
    - ***a marquer*** : pour visualiser sur l'image les points déjà marqués et ceux qui ne l'ont pas encore été. Ce bouton permet aussi de remettre le slider à sa valeur initiale.
    - ***slider*** : curseur pour faire défiler les images adjacentes à l'image à marquer et pour visualiser le mouvement de la langue autour de cette image. La barre devient noire lorsqu'il s'agit de l'image cible.
    - ***spline*** : pour visualiser sur l'image le contour marqué et lissé par une interpolation polynomiale entre les points.
    - ***1, 2, 3…*** : pour effectuer le marquage des points de la langue, hormis la pointe. Ces boutons permettent de marquer un à un chaque point : si le point possède un seul degré de liberté, une droite horizontale ou verticale apparaît (elle représente la coordonnée fixée). Ici, pour les points 1 à 5, une droite

verticale s'affiche pour indiquer l'abscisse fixée. Pour les points 6 à 9, une droite horizontale s'affiche pour indiquer l'ordonnée fixée.

- ***pointe*** : pour effectuer le marquage de la pointe de la langue, aucun degré de liberté n'est fixé, aucune droite ne s'affiche.
- ***pointe2*** : pour effectuer le marquage d'un point entre la pointe de la langue et le point 1. Il n'est possible de marquer ce point qu'une fois la pointe et le point 1 marqués. L'abscisse de pointe 2 est à mi-distance entre pointe et le point 1.
- ***epiglotte*** : pour effectuer le marquage de l'épiglotte.
- ***complet*** : pour effectuer le marquage complet de l'image. Il faut marquer les points les uns à la suite des autres sans repasser par le slider, dans l'ordre défini et indiqué soit par les droites horizontales ou verticales qui s'affichent au fur et à mesure, soit par une indication textuelle qui apparaît sur l'image.

Pour chacun des boutons 1, 2, 3, …, pointe, pointe2, epiglotte, complet, il est possible de cliquer à l'extérieur du cadre de l'image pour sortir de l'action sans marquer.
Chaque point marqué est enregistré au fur et à mesure.
Il est possible de revenir en arrière et de corriger le marquage réalisé. Lorsque pour une image, l'ensemble des points a été marqué, un contour de la langue (points reliés linéairement entre eux) s'affiche à l'écran. Il est alors encore possible de faire des modifications.

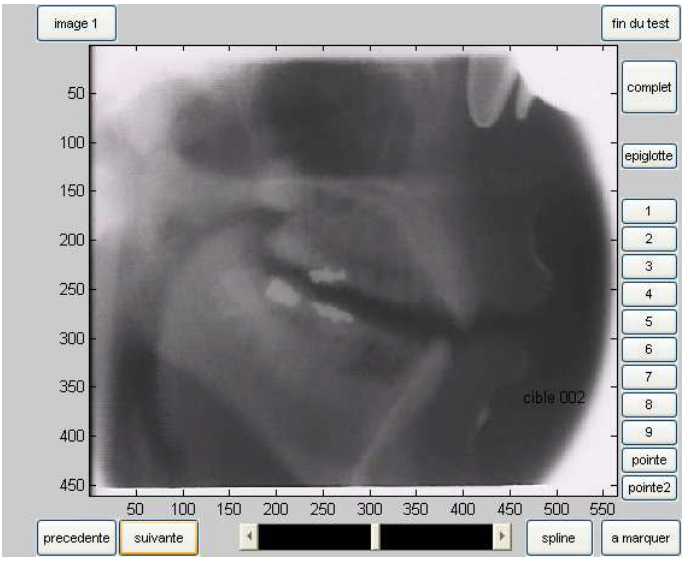


Figure 128 : Interface Matlab de marquage géométrique manuel d'images clefs. Exemple avec le marquage du contour de la langue de Laval43.

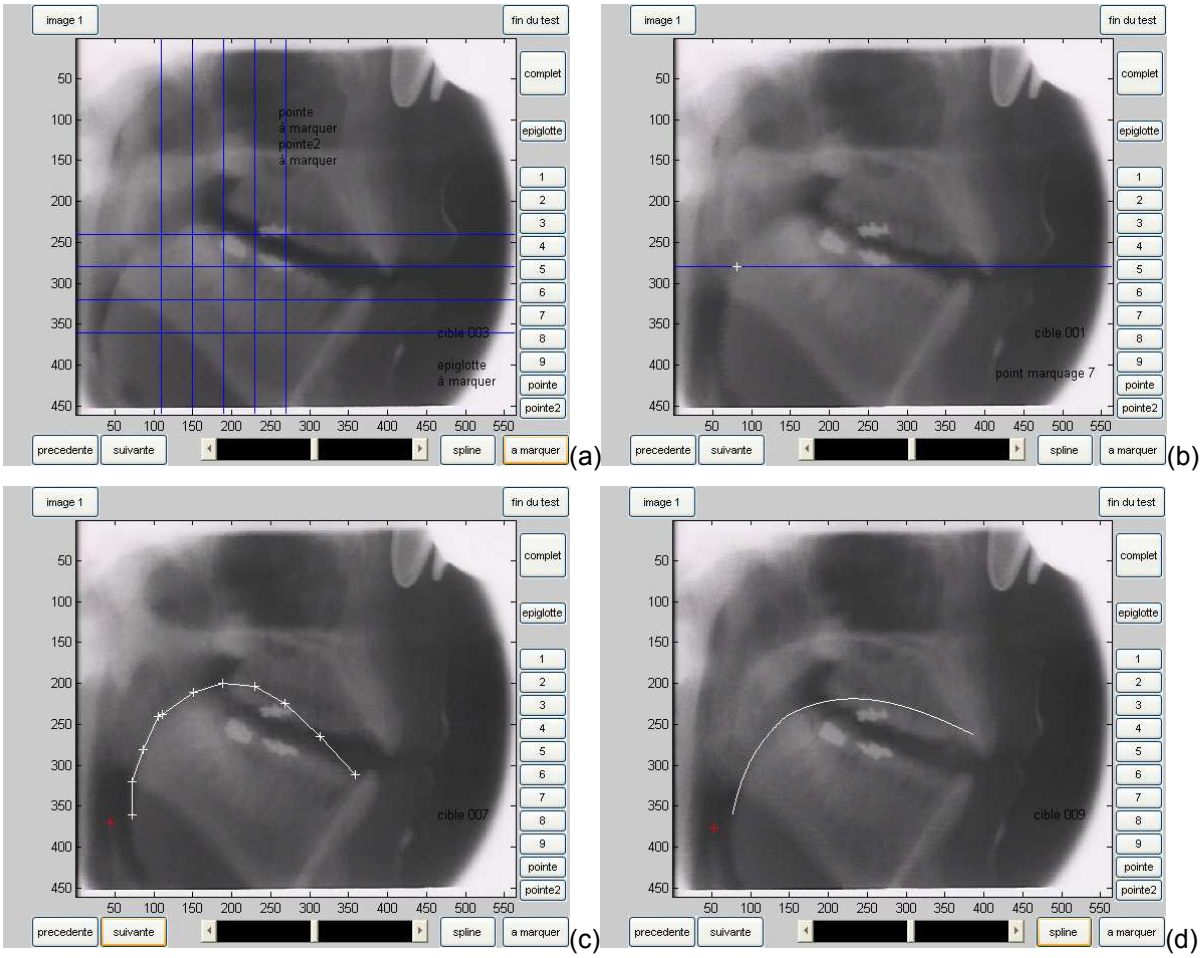


Figure 129 : (a) Aucun point de cette image n'a été marqué pour l'instant.
(b) Le point 7 a été marqué à l'aide du bouton 7. Une ligne bleue apparaît, elle fixe une des coordonnées (ici Y). Le point est à marquer à l'intersection entre la ligne et le contour de la langue. Une fois marqué, le point est sauvegardé. On peut utiliser le slider pour vérifier le marquage des points.
(c) Pour cette image, tous les points ont été marqués. Dès que les points sont tous marqués, un contour est tracé et relie les points entre eux.
(d) Le bouton spline sert à lisser le contour obtenu. On utilise des interpolations par des polynômes.

2. Remarques

Taille de l'image : On peut choisir la taille d'affichage de l'image à l'écran. Pour augmenter le contraste, il est parfois utile de réduire la taille de la fenêtre, mais on augmente la précision avec des fenêtres plus grandes. On peut envisager de marquer d'abord sur des petites images puis d'agrandir pour affiner ensuite le marquage.

Points : Pour la plupart des points à marquer, il y a un repère qui s'affiche, soit une droite verticale, soit une droite horizontale. Pour marquer précisément, il faut se placer à l'intersection entre cette droite et la langue, la droite de repère change alors de couleur.

Marquage : Une fois un des boutons pour marquer sélectionné, le programme attend un (ou plusieurs) clic(s) de la souris sur l'image. Il est possible de sortir de cette action en cliquant à l'extérieur de l'image. En particulier, il faut éviter d'utiliser le slider lorsque le programme attend un clic de la souris sur l'image.

Marquage complet : Il y a 2 techniques pour marquer complètement une image, soit marquer tous les points d'un coup avec le bouton « complet », soit marquer les points un par un avec les autres boutons.
Il est parfois plus facile de commencer à faire un premier marquage complet et ensuite de corriger les points un par un. En effet, lorsque le marquage est complet, les points sont reliés automatiquement et permettent de visualiser un contour superposé à l'articulateur.

Ordre des points : Suivant les images, il arrive que l'ordre des points varie un peu. Ceci sera pris en compte au moment de l'affichage du contour qui relie les points. Il faut continuer à marquer en cherchant l'intersection entre les droites de repère et l'articulateur.

« Voir » la langue : Il n'est pas évident de voir la langue sur les images statiques. Il ne faut surtout pas hésiter à utiliser le slider ou curseur dès le début et avant tout marquage pour faire défiler les images et observer la langue en mouvement de façon à bien la détecter.
Une fois le marquage effectué sur l'image (les points sont alors reliés), on peut à nouveau faire défiler les images avec le curseur pour vérifier le marquage.

Pointe : La pointe est difficile à marquer… Il n'y a pas d'aide ni pour l'ordonnée, ni pour l'abscisse. L'utilisation du curseur est très souvent nécessaire.
Le point pointe2 ne peut être marqué que lorsque les points 1 et pointe ont été marqués. Ensuite, chaque fois que l'on modifie l'un de ces 2 points, il est nécessaire de remarquer pointe2.

Autres articulateurs : Suivant les articulateurs, le choix des degrés de liberté est différent. Pour le vélum, on utilise une grille polaire pour marquer 12 des 13 points. Pour chaque ligne qui s'affiche, on marque 2 points (un à l'intersection entre la ligne et la partie supérieure du vélum et un à l'intersection entre la ligne et la partie inférieure du vélum). La pointe du vélum a 2 degrés de liberté.

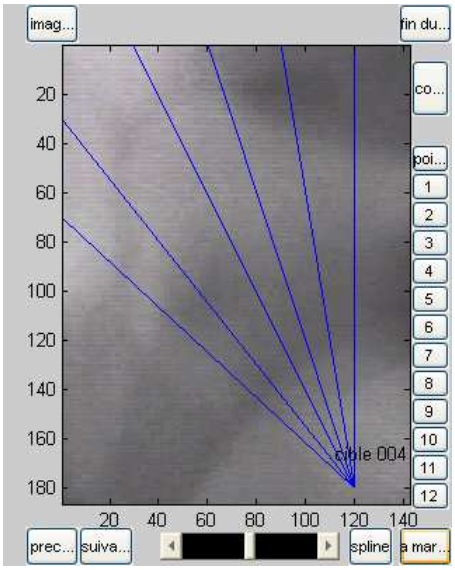


Figure 130 : Interface Matlab de marquage géométrique manuel d'images clefs. Exemple avec le marquage du contour du vélum de Laval43.

Un travail est actuellement en cours pour améliorer cette interface et faire en sorte de facilement marquer les autres séquences de la base d'ATR. Des étapes préliminaires sont mises en place dans l'interface de façon à rendre automatiques un certain nombre d'actions : le choix de la séquence, du cadre spécifique à l'articulateur à considérer, le choix des degrés de liberté spécifiques aussi à l'articulateur.

Une fois passées ces étapes, l'étape de marquage peut commencer.

A4. INTERFACE D'ETIQUETAGE AUDIO

Une partie importante des travaux que nous avons menés s'efforcent de mettre en parallèle les données articulatoires et acoustiques. Pour ces études, nous avons eu besoin de récupérer à la fois les instants de voyelles et ceux de consonnes.

Comme mentionné dans le manuscrit, une interface d'étiquetage audio a été élaborée pour permettre et faciliter la récupération de ces instants. Il s'agit d'une interface programmée sous Matlab.

Cette interface (Fig. 131) permet de sélectionner sur le signal audio (en bleu) un instant et de lui associer le phonème correspondant. Elle sauvegarde ainsi l'étiquetage du signal. On effectue de manière séparée l'étiquetage des voyelles et celui des consonnes (sauvegardes dans des fichiers différents), à partir de la même interface.

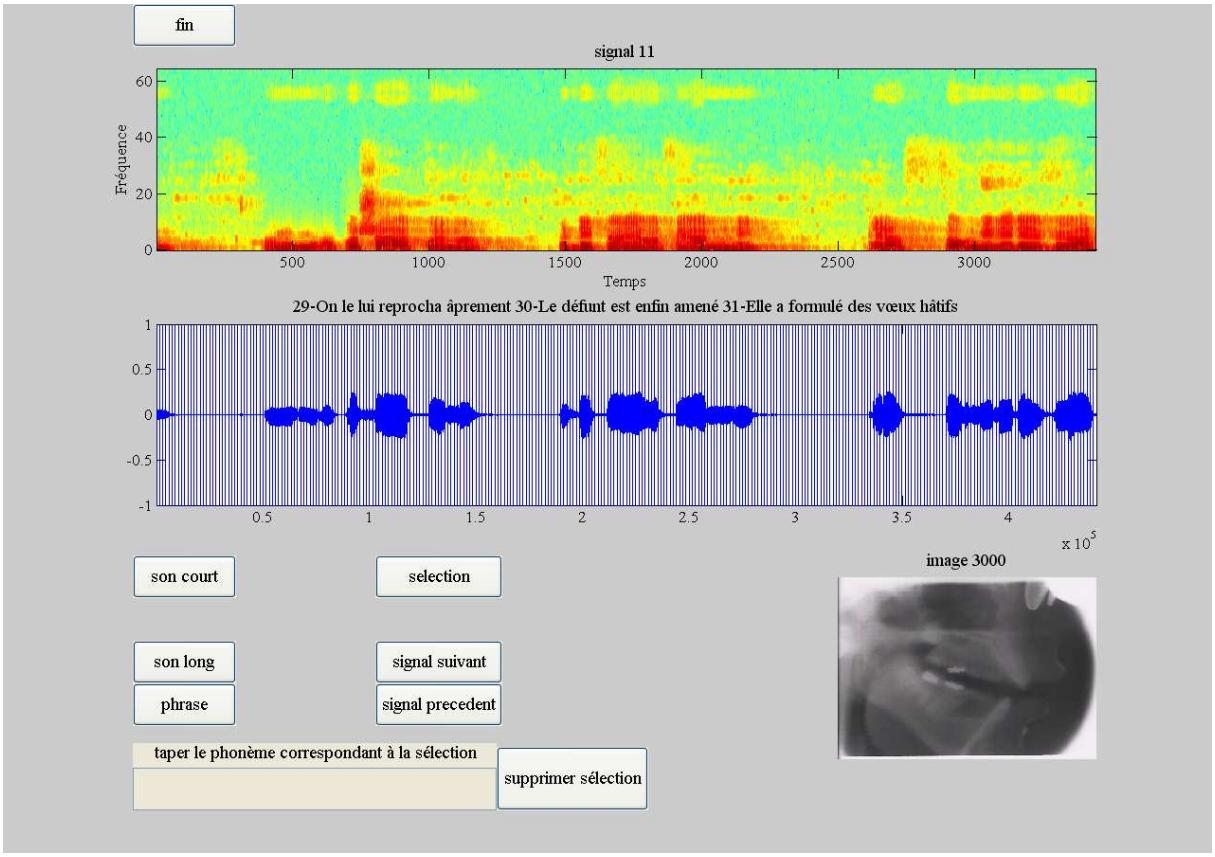


Figure 131 : Interface Matlab d'étiquetage audio.

L'interface se présente de la façon suivante :

- Le numéro de la portion de signal en cours d'étiquetage. Le signal complet a été découpé en portions de 10 secondes.
- Le ***spectrogramme*** de cette portion de signal.

- Au centre, le ***signal audio*** à étiqueter. Les droites verticales superposées au signal indiquent les images de la séquence Laval43 (une droite toutes les 33 ms).
- Les boutons son court, son long et phrase permettent d'écouter des bouts de signal.
    - ***Son court*** : pour écouter l'instant sélectionné, en réalité l'équivalent de 80ms correspondant à un peu plus de 2 trames du signal vidéo.
    - ***Son long*** : pour écouter un peu plus que l'instant sélectionné. On entend 1 seconde de signal à partir de l'instant sélectionné.
    - ***Phrase*** : pour écouter les 10 secondes de signal correspondant à la portion en cours d'étiquetage.
- Le bouton ***selection*** permet de prendre la main sur l'interface pour sélectionner un instant.
- On valide la sélection en écrivant dans le champ prévu à cet effet, le phonème correspondant.
- On peut corriger une sélection en écrivant un autre phonème ou en la supprimant avec le bouton ***supprimer selection***.
- L'***image radiographique*** affichée correspond à la trame sélectionnée, dont le numéro est noté au-dessus.

A5. RAPPORT PIXELS-CMS DANS WIOLAND

Pour établir le rapport entre les pixels des images de la séquence cinéradiographique Wioland et les centimètres, nous nous sommes plongés dans le manuscrit de Wioland et en particulier nous avons analysé l'établissement de sa grille de mesure.

Pour cela, une vue de profil avait été choisie comme référence. Sur cette vue (Fig. 132a), une circonférence a été tracée de telle sorte qu'elle passe par la pointe des incisives supérieures et que la voûte palatine à la limite palato-vélaire lui soit tangente. Sur l'image choisie, la paroi pharyngale est également tangente à la circonférence. Le centre O de la circonférence est à égale distance de la pointe des incisives inférieures (I) , de la limite palato-vélaire (J) et de la paroi pharyngale (K). Cette distance est de 5,5 cm.

L'image de référence choisie par Wioland est une image d'une séquence dont nous ne disposons pas. Nous ne pouvons pas directement réaliser nos mesures sur cette image.

Nous avons donc travaillé en moyenne sur l'ensemble des trames de la séquence. Pour chacune, nous avons tracé le cercle passant par la pointe des incisives supérieures et par le point à la limite palato-vélaire. Ce cercle est presque toujours quasiment tangent à la paroi pharyngale. Pour chaque image, nous trouvons une valeur en pixels pour le rayon de ce cercle (Fig. 132b). La valeur moyenne de ce rayon sur la séquence est mise en correspondance avec les 5,5 centimètres mesurés par Wioland.

Ceci nous permet ainsi d'évaluer qu'un centimètre correspond à 38 pixels. C'est ce rapport que nous utilisons dans l'évaluation de l'erreur (Chapitre 2)

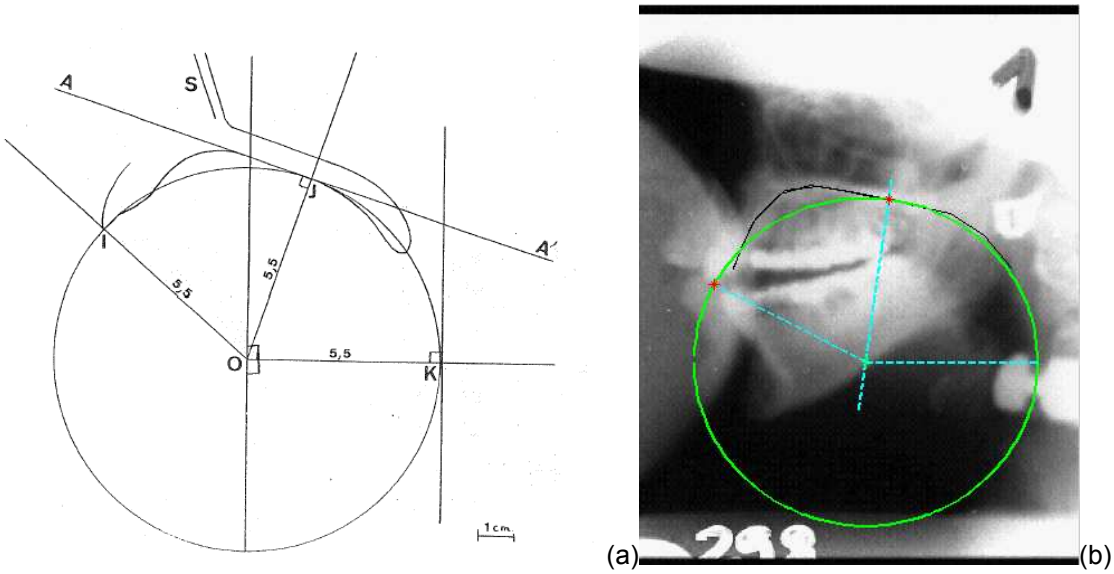


Figure 132 : (a) Croquis de la vue de profil de référence utilisée par Wioland pour l'établissement de sa grille de mesure.
(b) Circonférence superposée à une image de la séquence pour l'évaluation du rapport pixels/cms.