Inversion acoustique articulatoire à partir de coefficients cepstraux

Julie Busset

UNIVERSITÉ DE LORRAINE

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Inversion acoustique articulatoire à partir de coefficients cepstraux

THÈSE

présentée et soutenue publiquement le 25 mars 2013

pour l'obtention du

Doctorat de l'Université de Lorraine

(spécialité informatique)

par

Julie Busset

Composition du jury

Rapporteurs :	Olov ENGWALL, Professeur, *Centre for Speech Technology – TMH Stockholm (Suède)* Pierre BADIN, Directeur de Recherche, *CNRS, GIPSA-LAB*
Examinateurs :	Bernard GIRAU, Professeur, *Université de Lorraine* Rudolph SOCK, Professeur, *Université de Strasbourg – UNISTRA* Shinji MAEDA, Directeur de Recherche Emérite, *CNRS Telecom Paris*
Directeurs de thèse :	Yves LAPRIE, Directeur de Recherche, *CNRS LORIA* Martine CADOT, PRAG, *Université de Lorraine*

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Monsieur Yves Laprie, qui m'a dirigée durant toutes ces années au LORIA. Je lui suis reconnaissant pour ses encouragements, son enthousiasme et sa confiance. Je remercie également ma co-directrice de thèse Martine Cadot qui, tout au long de ces années m'a encouragée et conseillée.

J'adresse mes vifs remerciements à Monsieur Olov Engwall et Monsieur Pierre Badin pour avoir accepté d'être rapporteurs de ce travail. Je tiens également à remercier Monsieur Bernard Girau, Monsieur Rudolph Sock et Monsieur Shinji Maeda de m'avoir fait l'honneur de participer à ce jury de thèse.

Je souhaite également remercier mes collègue de l'équipe PAROLE notamment Utpala et Matthieu pour tous les moments agréables passés à discuter autour d'un café.

Je souhaite remercier spécialement Aurélien pour son soutien et sa patience tout au long de la thèse.

Enfin je remercie ma sœur Margaux et mes parents pour leur soutien au cours de ces longues années d'études et sans lesquels je n'en serais pas là aujourd'hui.

Table des matières

Chapitre 1
Introduction générale

Chapitre 2
Position du problème de l'inversion

2.1 Observation du conduit vocal 5
2.2 Position du problème 8
2.3 Inversion point à point pour un instant donné 8
2.3.1 Méthodes d'inversion directe 9
2.3.2 Méthodes d'inversion basées sur les codebooks 9
2.3.2.1 Présentation du problème 9
2.3.2.2 Construction et organisation des codebooks 10
2.3.2.3 Recherche de solutions dans le codebook 12
2.4 Inversion de segments acoustiques 13
2.5 Méthodes d'inversion utilisant un apprentissage statistique 13
2.5.1 Inversion basée sur les modèles de Markov cachés 14
2.5.2 Inversion basée sur les GMM 14
2.5.3 Inversion basée sur les réseaux de neurones artificiels 16
2.6 Conclusion 16

Partie I Modélisation du conduit vocal **19**

Chapitre 3
Modélisation du conduit vocal

3.1 Modèles à fonction d'aire 21
3.1.1 Modèles à trois paramètres 22
3.1.2 Modèles à concaténation de tubes 22
3.2 Modèles articulatoires 23
3.2.1 Modèles géométriques 23
3.2.2 Modèles statistiques 24
3.2.3 Modèles biomécaniques 25
3.3 Passage de la coupe sagittale à la fonction d'aire 26
3.4 Simulation acoustique 28
3.4.1 Equations de l'acoustique 28
3.4.1.1 Excitation du son dans le conduit vocal 31
3.5 Conclusion 31

Chapitre 4
Construction d'un nouveau modèle articulatoire

4.1 Traitement des données 33
4.1.1 Corpus 33
4.1.2 La mâchoire et les structures rigides 35
4.1.3 Les lèvres et l'épiglotte 38
4.1.3.1 Les lèvres 38
4.1.3.2 L'épiglotte et le larynx 40
4.1.4 La langue 40
4.1.5 Contours des articulateurs 41
4.2 Construction du modèle articulatoire 41
4.2.1 Soustraction du mouvement de la tête 42
4.2.2 La mâchoire 42
4.2.2.1 ACP sur les données 42
4.2.2.2 Soustraction de la contribution de la mâchoire 43
4.2.3 La langue 44
4.2.3.1 Différentes présentations des données de la langue 45
4.2.3.2 Différentes stratégies d'analyse 46
4.2.3.3 Choix des composantes de la langue 47
4.2.4 Les lèvres 51
4.2.4.1 Extraction des données 51
4.2.4.2 Analyse 51
4.2.4.3 Reconstruction des lèvres 52
4.2.5 L'épiglotte et le larynx 54

4.2.6 Assemblage du modèle articulatoire 56
4.3 Synthèse articulatoire 62
4.3.1 Passage à la fonction d'aire 62
4.3.2 Simulation acoustique 65
4.4 Conclusion 66

Partie II Inversion **69**

Chapitre 5
Construction d'un codebook hypercuboïdal

5.1 Espaces articulatoire et acoustique 71
5.1.1 Synthétiseur articulatoire 72
5.1.2 Représentations acoustiques 72
5.1.2.1 Lissage cepstral 72
5.1.2.2 Prédiction linéaire 75
5.1.2.3 Calcul du vecteur acoustique 77
5.2 Représentations mathématiques dans un codebook 78
5.2.1 Structure hypercuboïdale du codebook 78
5.2.2 Modélisation de la relation articulatoire-acoustique 79
5.3 Construction du codebook 79
5.3.1 Principe général 79
5.3.2 Test de linéarité à partir de la matrice jacobienne 79
5.3.2.1 Calcul de la matrice jacobienne 81
5.3.2.2 Test de linéarité 81
5.3.3 Choix des points de test 82
5.3.4 Subdivision d'un hypercuboïde 82
5.3.4.1 Conditions de subdivision 83
5.3.4.2 Méthode de subdivision 85
5.3.4.3 Terminaison de la récursivité 85
5.4 Évaluation expérimale du codebook 86
5.4.1 Valeur optimale du paramètre de la matrice jacobienne 86

5.4.2 Couverture de l'espace articulatoire 92
5.4.3 Seuils de subdivision et précision acoustique 94
5.5 Conclusion 98

Chapitre 6
Inversion par codebook hypercuboïdal

6.1 Principe général 101
6.2 Comparaison entre les vecteurs cepstraux naturels et synthétiques 103
6.2.1 Voyelles extraites du corpus 103
6.2.2 Adaptation cepstrale 106
6.2.3 Distorsion fréquentielle (ou *frequency warping*) 110
6.3 Optimisation de la recherche dans le codebook 113
6.3.1 Organisation du codebook 114
6.3.2 Appariement des pics des spectres naturels et synthétiques 114
6.4 Évaluation 117
6.4.1 Inversion de données synthétiques 117
6.4.1.1 Résultats de l'optimisation de la recherche dans le codebook . . . 117
6.4.1.2 Résultats de l'inversion 118
6.4.2 Inversion de données réelles 125
6.5 Modification du modèle 133
6.6 Conclusion 137

Conclusions et perspectives

Conclusion et perspectives

Annexes

Annexes

Annexe A
Composantes de la langue

Annexe B
Construction de codebook

Annexe C
Appariement des pics spectraux par programmation dynamique

C.1 Rappel de l'algorithme 161
C.2 Résolution 162

Bibliographie **165**

Table des figures

2.1 Vue sagittale du conduit vocal obtenue par cinéradiographie. 6
2.2 Image IRM représentant la coupe sagittale d'un /a/. Le locuteur n'a pas beaucoup ouvert la bouche ce qui donne un /a/ peu typique. 7

3.1 Modèle du conduit vocal à fonction d'aire proposé par Fant [Fan70]. 22
3.2 Les sept paramètres du modèle de Maeda : la position de la mâchoire (P1), la position du corps de la langue (P2), la forme du corps de la langue (P3), la position de l'apex de la langue (P4), l'ouverture des lèvres (P5), la protrusion des lèvres (P6) et la hauteur du larynx (P7). 25
3.3 Le modèle biomécanique de Perkell. Les éléments de tension sont représentés par les lignes continues et en pointillé, les points noirs sont les nœuds porteurs de masse. (D'après [Per74]). 26
3.4 Coupes transversales du conduit vocal réalisées à partir d'un moulage de conduit vocal réalisé sur un cadavre. a) coupe sagittale du conduit vocal, b) les sections transversales correspondantes (d'après [Cal89]). 27
3.5 La figure du haut montre la représentation du conduit vocal sous forme de concaténation de tubes. La figure du bas montre la fonction d'aire correspondant au conduit vocal du haut. 29
3.6 Représentation schématique du système vocal. 31

4.1 Vue sagittale du conduit vocal obtenue par cinéradiographie. Les principaux articulateurs visibles sont la mandibule inférieure, la langue, les lèvres, l'épiglotte et le voile du palais. 34
4.2 Suivi par corrélation du mouvement de la région de la mâchoire. L'image de gauche est l'image de référence où la région de la mâchoire a été tracée. L'image de droite présente le résultat du suivi ; la région s'est déplacée par rapport à l'image de référence. 36
4.3 Exemple de suivi par corrélation pour le mouvement de la mâchoire, le mouvement de la tête et le mouvement de l'os hyoïde. L'image de gauche est l'image de référence où les contours ont été tracés manuellement. L'image de droite est issue du suivi, les régions sont celles de l'image de référence auxquelles le mouvement calculé a été appliqué. 36

4.4 Exemple de suivi par corrélation pour le mouvement de la mâchoire, le mouvement de la tête et le mouvement de l'os hyoïde. Le contour du palais suit le mouvement de la région de la tête et le plancher de la bouche suit le mouvement de la mâchoire. Les cercles jaunes correspondent à la position des incisives supérieures et inférieures qui suivent respectivement le mouvement de la tête et le mouvement de la mâchoire. L'image de gauche est l'image de référence. L'image de droite est issue du suivi, les contours et les positions des incisives sont déplacés en fonction de la région auxquels ils sont reliés. 37
4.5 Exemple de contours obtenus par l'algorithme de suivi semi-automatique. Les nouveaux contours sont obtenus à partir de la moyenne des contours des trois images clés les plus proches. Les nouveaux contours sont replacés dans l'image originale. 39
4.6 Image cinéradiographique présentant les contours utilisés pour l'épiglotte et le larynx. 40
4.7 Images cinéradiographiques présentant les contours des différents articulateurs du conduit vocal. 41
4.8 Composantes obtenues par ACP sur les données de la mâchoire. Le mouvement est appliqué sur la région utilisée pour le suivi. Les composantes varient entre -3 écarts-types (ligne rouge) et +3 écarts-types (ligne bleue). La courbe noire étant la position moyenne. Les carrés représentent la position du bord de l'incisive inférieure. 43
4.9 Représentation de la grille polaire adaptative. La grille est définie à partir de son centre, de l'apex et de la racine de la langue. Les points rouges sont les points d'intersection entre la grille et le contour de la langue. 44
4.10 Données utilisées avec l'approche 2. Les points rouges sont les points d'intersection entre la grille et le contour de la langue. d_i est la distance entre le point d'intersection avec la $i^{ème}$ ligne de la grille et le centre de la grille. 45
4.11 Les quatre paramètres de la langue issus de l'approche 2 (distances et angles extrêmes) avec la stratégie d'analyse 2 (cascade). Chaque paramètre varie entre -3 et +3 écarts-types. La forme noire correspond à la forme neutre, la rouge à -3 écarts-types et la bleue à +3 écarts-types. 50
4.12 Les paramètres des lèvres. 51
4.13 Construction de la section représentant les lèvres. La section est définie par quatre points positionnés par rapport à la position de référence pour l'incisive, de la hauteur et de la protrusion. 53
4.14 Les deux composantes principales issues de l'analyse en composantes principales sur les données des lèvres. La première composante permet de contrôler la hauteur et la protrusion et représente 92,7% de la variance. Les courbes noires correspondent à la position moyenne, les rouges à -3 écarts-types et les bleues à +3 écarts-types. 54
4.15 Représentation des données de l'épiglotte et du larynx utilisées pour l'ACP. À gauche les trois contours obtenus par suivi automatique et à droite le contour de l'épiglotte et la position extérieure du larynx (caractérisée par la croix rouge). 55
4.16 Deux premières composantes issues de l'ACP pour l'épiglotte et le larynx. Les courbes noires correspondent à la position moyenne, les rouges à -3 écarts-types et les bleues à +3 écarts-types. 55

4.17 Le contour du plancher de la langue (en rouge) est utilisé pour compléter le contour intérieur lorsque la langue est reculée. 56
4.18 Intersections des contours intérieur et extérieur avec la grille semi-polaire. Les disques rouges correspondent à ces intersections. Les extrémités du larynx sont représentées. Deux points sont ajoutés entre les extrémités du larynx et les dernières intersections avec la grille. 57
4.19 Grille semi-polaire avec les différents paramètres. 58
4.20 Modèle articulatoire représenté avec la grille semi-polaire. Les contours rouges correspondent aux contours intérieur et extérieur du conduit vocal. La grille semi-polaire est formée de m_1 lignes pour la partie linéaire inférieure, m_2 lignes pour la partie polaire et m_3 lignes pour la partie linéaire antérieure. Les angles θ et Ω définissent la position des différentes parties de la grille. 60
4.21 Les sept paramètres du modèle articulatoire. Chaque paramètre varie dans l'intervalle $[-3; 3]$. 61
4.22 Détermination de la longueur l_k de la section k et de la distance sagittale. B_k est l'aire de la partie grise de la section k. 63
4.23 Détermination des huit régions du conduit sur le modèle articulatoire. 64
4.24 Exemple de passage de la coupe sagittale à la fonction. La figure de gauche représente une vue sagittale du conduit vocal et la figure de droite correspond à la fonction d'aire correspondante. 65
4.25 Synthèse articulatoire à partir de la forme sagittale du conduit vocal. La figure de gauche présente une forme de conduit vocal fournie par le modèle. La figure de droite représente les fonctions de transfert acoustique du conduit vocal (la source n'est pas prise en compte) obtenus par synthèse articulatoire : ligne noire (spectre obtenu par simulation acoustique) et ligne bleue (spectre issu d'un lissage cepstral). 66

5.1 Modèle simplifié de la production de la parole. 73
5.2 Exemple de lissage cepstral sur un signal de parole. 74
5.3 Schéma représentant le modèle de production de la parole utilisé pour la LPC. 75
5.4 Représentation d'un hypercuboïde dans un espace de dimension 3 caractérisé par son rayon $r = (r_1, r_2, r_3)$ et son centre P_0. 78
5.5 Subdivisions successives dans un hypercuboïde. Si la relation locale dans un sous hypercuboïde n'est pas linéaire, il y a subdivision en deux sous-hypercuboïdes dans une seule direction et ainsi de suite. La subdivision est effectuée jusqu'à ce que la relation soit linéaire ou que la taille minimale soit atteinte. 80
5.6 Représentation en deux dimensions d'un hypercuboïde avec les différents points de test. Les points rouges correspondent au centre, aux sommets, aux milieux des segments reliant deux sommets et les points utilisés pour le calcul du jacobien au centre à une distance proportionnelle au rayon. 83

5.7 Coupe l'espace articulatoire de Maeda selon deux composantes (mâchoire et position de la langue) variant entre -3 et $+3$ (d'après [Pot08a]). La zone en pointillés correspond à la zone synthétisable et la zone blanche à la zone non-synthétisable, la ligne rouge représente la limite entre les deux. Les rectangles correspondent aux coupes des hypercuboïdes. 84
5.8 Le graphique de gauche représente le volume total en fonction des différentes valeurs de ϵ_m et celui de droite le nombre d'hypercuboïdes en fonction des différentes valeurs de ϵ_m pour les trois zones. La zone 1 correspond aux courbes rouges, la zone 2 aux courbes bleues et la zone 3 aux courbes noires. 89
5.9 Graphique représentant le pourcentage de volume de l'espace articulatoire couvert en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3. 90
5.10 Graphique représentant l'erreur moyenne de resynthèse en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3. 91
5.11 Graphique représentant la valeur de la mesure homogène e en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3. 91
5.12 Exemple d'un hypercuboïde (en trois dimensions) situé sur la frontière. La partie grise correspond à l'espace articulatoire synthétisable et la partie blanche à la zone où les vecteurs articulatoires ne possèdent pas d'image acoustique. Cet hypercuboïde a la moitié de ses sommets sans image acoustique alors qu'il est presque entièrement défini. 92
5.13 Pourcentage de volume couvert par les trois zones étudiées en fonction du nombre de sommets utilisés. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3. 94
5.14 Nombre d'hypercuboïdes (graphique de gauche) et densité (graphique de droite) en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire). 96
5.15 Pourcentage de volume couvert en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire). 96
5.16 Erreur moyenne de resynthèse (graphique de droite) en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire). 97

6.1 Spectrogramme issu du corpus. Les lignes bleues en pointillés représentent les instants où une image cinéradiographique est disponible. Les flèches noires désignent les instants correspondants aux voyelles. 104
6.2 Erreur moyenne sur l'ensemble des voyelles entre les trois premiers pics des spectres lissés cepstralement issus des coefficients cepstraux synthétiques et réels en fonction du nombre de coefficients cepstraux utilisés pour l'adaptation. 107

6.3 Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle). 108
6.4 Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle). 109
6.5 Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle). 109
6.6 Coefficients cepstraux et spectres lissés cepstralement après adaptation avec 29 et 2 coefficients : ligne bleue (naturel), ligne noire (synthétique), ligne rouge (adapté avec 29 coefficients) et ligne rouge pointillée (adaptée avec 2 coefficients). 110
6.7 Exemples de transformations bilinéaires pour différentes valeurs de α. Dans le cas $\alpha = 0$ il n'y a pas de décalage. 112
6.8 Erreur moyenne sur l'ensemble des voyelles entre les trois premiers pics des spectres lissés cepstralement issus des coefficients cepstraux synthétiques et réels en fonction de la valeur de α. 112
6.9 Deux exemples de spectres lissés cepstralement : ligne bleue (naturel), ligne noire (synthétique), ligne rouge (après transformation bilinéaire). 113
6.10 Deux spectres pour illuster l'algorithme d'appariement des pics spectraux. 116
6.11 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /e/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 120
6.12 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /ø/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 121
6.13 Exemple d'inversion de données synthétiques pour un forme de conduit vocal correspondant à la voyelle /i/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 121
6.14 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /y/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 122
6.15 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /ε/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 122

6.16 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /a/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 123
6.17 Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /u/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 123
6.18 Conduits vocaux original et inversé à partir des paramètres synthétiques et paramètres articulatoires associés. La forme noire (resp. paramètres articulatoire) est la forme issue d'une image cinéradiographique d'un /ε/ et en rouge la solution inverse qui minimise l'erreur géométrique. 124
6.19 Conduits vocaux original et inversé à partir des paramètres synthétiques et paramètres articulatoires associés. La forme noire (resp. paramètres articulatoire) est la forme issue d'une image cinéradiographique d'un /i/ et en rouge la solution inverse qui minimise l'erreur géométrique. 124
6.20 La forme articulatoire noire est issue des images cinéradiographiques. Les courbes rouges correspondent aux 25 meilleures solutions en termes de distance géométrique. . 125
6.21 Différentes étapes pour l'inversion de données réelles 128
6.22 Exemple d'inversion à partir du signal réel pour la voyelle /e/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 129
6.23 Exemple d'inversion à partir du signal réel pour la voyelle /ø/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 129
6.24 Exemple d'inversion à partir du signal réel pour la voyelle /i/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 130
6.25 Exemple d'inversion à partir du signal réel pour la voyelle /y/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 130
6.26 Exemple d'inversion à partir du signal réel pour la voyelle /ε/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 131
6.27 Exemple d'inversion à partir du signal réel pour la voyelle /a/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 131
6.28 Exemple d'inversion à partir du signal réel pour la voyelle /u/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion). 132
6.29 Les deux graphiques présentent les conduits vocaux originaux (courbe noire) et inversés (courbe rouge) avec l'erreur acoustique minimale sans adaptation cepstrale (graphique de gauche) et avec adaptation cepstrale (graphique de droite). 132

6.30 Modification du modèle original (courbe en trait plein). La courbe de gauche (tirets) représente une augmentation de 10% de la taille et la courbe (tirets) de droite une diminution de 10%. 133
6.31 Exemple d'une solution d'inversion d'une voyelle (/y/) à partir du signal à un modèle où les cavités pharyngale et buccale sont raccourcies de 10%. Courbe rouge (solution inverse dans le modèle modifié) et courbe noire (forme de départ dans le modèle original). 137
6.32 Exemple d'inversion avec le modèle où les cavités pharyngale et buccale sont augmentées de 10%. Les conduits vocaux (à gauche) et les fonctions d'aire (à droite) sont représentés en noir pour la forme initiale et en rouge pour la solution inverse qui minimise l'erreur acoustique. 137

A.1 Cascade ; coordonnées euclidiennes 148
A.2 Cascade + corrélation ; coordonnées euclidiennes 149
A.3 Défaut ; coordonnées euclidiennes. 150
A.4 Cascade ; distances + angles extrêmes. 151
A.5 Cascade + corrélation ; distances + angles extrêmes. 152
A.6 Défaut ; distances + angles extrêmes. 153
A.7 Cascade ; coordonnées polaires. 154
A.8 Cascade + corrélation ; coordonnées polaires. 155
A.9 Défaut ; coordonnées polaires. 156

Chapitre 1

Introduction générale

La parole est un moyen de communication propre à l'être humain. Son apparente simplicité et la vitesse de transmission des informations qu'elle permet font d'elle un mode d'échange privilégié. La production d'un signal de parole résulte de l'action coordonnée d'un ensemble d'organes. La déformation des structures respiratoires, laryngées et du conduit vocal permet de créer des sources sonores, de les modifier, de les amplifier et de les filtrer. La production de la parole nécessite ainsi la contraction coordonnée de plus de 200 muscles impliquant la mâchoire, la langue, les lèvres, le voile du palais, le pharynx, le larynx et les muscles de la respiration. La phonétique cherche en particulier à découvrir les processus mis en place pour produire de la parole compréhensible et porteuse de sens. Les avancées dans le domaine de l'imagerie médicale et les progrès dans les techniques d'observation de l'activité des organes phonatoires ont permis des avancées très importantes dans la compréhension des mécanismes de production de la parole depuis les années cinquante.

La synthèse articulatoire se base sur des connaissances articulatoires, mécaniques et physiologiques des articulateurs de la parole modifiant la forme de la cavité buccale. À partir de la forme du conduit vocal, la synthèse articulatoire produit un signal acoustique via une simulation numérique.

Depuis plusieurs décennies, beaucoup d'études s'intéressent au cheminement inverse appelé « inversion acoustique-articulatoire », c'est-à-dire la détermination d'une trajectoire articulatoire du conduit vocal à partir d'un signal de parole. Un système capable d'approcher la position des articulateurs à partir d'un signal trouverait de nombreuses applications potentielles. Dans le domaine de la reconnaissance de la parole, l'utilisation d'une information articulatoire complémentaire pourrait améliorer les performances des systèmes de reconnaissance automatique, plus particulièrement dans le cas de parole bruitée, spontanée ou pathologique. L'inversion pourrait aussi compléter les techniques d'imagerie médicale qui peuvent être nocives pour la santé, trop chères ou présenter une durée d'acquisition trop longue. Cette technique serait également utile pour la synthèse de la parole. En effet, une information articulatoire peut être utilisée pour améliorer ou modifier les caractéristiques d'une parole synthétique. Une autre application possible serait une aide pour l'apprentissage des langues. La visualisation de la forme du conduit vocal à partir du signal prononcé permettrait de voir les erreurs et ainsi de proposer une meilleure stratégie articulatoire à l'apprenant. De la même

manière, un système d'inversion animant une tête parlante virtuelle permettrait aux malentendants de suivre une rééducation sans doute plus efficace.

L'inversion acoustique-articulatoire n'est pas un problème simple. La principale difficulté est liée au fait que la relation articulatoire-acoustique n'est pas biunivoque. En effet, plusieurs configurations du conduit vocal peuvent produire le même signal acoustique [ACMT78].

Les techniques d'inversion utilisant l'approche d'analyse par synthèse se basent sur l'utilisation d'un synthétiseur articulatoire associé à un modèle du conduit vocal contrôlé par un nombre réduit de paramètres. Le modèle est généralement construit à partir de données articulatoires extraites d'images cinéradiographiques présentant une vue sagittale du conduit vocal. Une table formée de couples de vecteurs articulatoire et acoustique permet de trouver un ensemble de solutions. Une méthode d'optimisation permet ensuite d'obtenir des trajectoires articulatoires réalistes.

Grâce au développement de techniques moins dangereuses pour la santé que les rayons X tel que les articulographes électromagnétiques (EMA) permettant l'acquisition de grands corpus associant données articulatoires et acoustiques, les méthodes d'apprentissage statistique se sont développées. Malheureusement, les données acquises correspondent aux positions d'un petit nombre de capteurs sur les principaux articulateurs et ne permettent pas de représenter l'ensemble du conduit vocal.

Une autre difficulté de l'inversion est l'évaluation des résultats. En effet, la quantité de données articulatoires disponibles est très faible. Les données ne sont pas toujours exploitables car les systèmes d'acquisition plus ou moins invasifs ne permettent pas de produire de la parole spontanée. Dans certains cas les enregistrements sonores ne sont pas disponibles, de mauvaise qualité ou non synchronisés avec les données articulatoires. La comparaison de l'inversion avec des données réelles est souvent difficile. Dans cette étude, nous nous plaçons dans un cas favorable car nous disposons de quatre films d'images cinéradiographiques extraits de la base DocVacim [SHL+11] enregistrés par le même sujet et pour lesquels l'enregistrement sonore correspondant est disponible et de bonne qualité.

Nous avons développé une méthode d'analyse par synthèse où l'inversion est réalisée en ajustant les paramètres d'un modèle articulatoire afin de faire correspondre les caractéristiques acoustiques à celles de la parole naturelle. Le modèle articulatoire utilisé est construit à partir d'images cinéradiographiques représentant une vue sagittale du conduit vocal. Sept paramètres contrôlent la forme sagittale du conduit vocal : un pour le mouvement de la mâchoire, quatre pour la position et la forme de la langue, un pour l'ouverture et la protrusion des lèvres et un pour la hauteur du larynx. Habituellement, le vecteur acoustique est composé des fréquences des trois premiers formants pour réaliser l'inversion. Cependant, l'extraction des formants à partir de la parole est souvent peu fiable, ce qui provoque des erreurs lors de l'inversion. L'objectif de la thèse est alors d'utiliser les coefficients cepstraux comme paramètres acoustiques d'entrée. Les vecteurs acoustiques comportent ainsi plus de coordonnées que le nombre de paramètres articulatoires, ce qui augmente le risque de ne pas trouver de solutions. De plus, l'effet de l'excitation de la source et les disparités entre le conduit vocal du locuteur et le modèle articulatoire peuvent être pris en compte explicitement en comparant les spectres naturels à ceux produits par le synthétiseur articulatoire car nous disposons des signaux réel et synthétique.

Ce mémoire se compose de deux parties. La première partie présente la construction d'un mo-

dèle articulatoire à partir d'images cinéradiographiques. Les techniques permettant de connaître la position des différents articulateurs seront détaillées. La seconde partie expose notre méthode d'inversion utilisant une analyse par synthèse. Une table représentant la relation acoustique-articulatoire est construite à partir de données synthétiques produites par le modèle articulatoire élaboré dans la première partie. Les vecteurs articulatoires sont composés des paramètres du modèle et les vecteurs acoustiques des coefficients cepstraux. La comparaison entre les signaux réels et synthétiques sera réalisée afin de permettre la recherche de solutions à partir de données réelles.

Chapitre 2

Position du problème de l'inversion

L'INVERSION acoustique-articulatoire vise à déterminer la géométrie du conduit vocal à partir d'un signal de parole. L'utilisation de l'imagerie médicale pour l'étude de la production de parole s'est largement imposée dans la communauté scientifique. Malgré les avancées technologiques, la mesure de la géométrie du conduit vocal reste difficile. Une technique d'imagerie médicale idéale permettrait de :

- couvrir l'ensemble du conduit vocal depuis la glotte jusqu'aux lèvres,
- avoir une fréquence d'acquisition suffisamment élevée pour capturer la dynamique du conduit vocal,
- ne présenter aucun risque pour la santé humaine,
- ne pas modifier l'articulation naturelle,
- ne pas perturber l'enregistrement sonore lors de l'acquisition.

Dans cette partie, nous présenterons brièvement les différentes approches d'inversion acoustique-articulatoire après une présentation des principales méthodes d'imagerie utilisées pour mesurer le conduit vocal (IRM, EMA).

2.1 Observation du conduit vocal

Il n'existe pas de technique répondant à tous les critères cités ci-dessus. Historiquement, la cinéradiographie fut la principale source d'informations des mécanismes de production de la parole. Le principal avantage de cette technique réside dans une acquisition à une fréquence compatible avec la dynamique des gestes articulatoires d'une vue sagittale de l'ensemble du conduit vocal. La principale difficulté est l'exploitation de ces images, car l'identification précise des différentes structures est très complexe à cause du recouvrement des organes. La figure 2.1 montre un exemple d'image cinéradiographique sur laquelle les principaux articulateurs sont visibles : la mâchoire, la langue, les lèvres, l'épiglotte, le voile du palais. L'utilisation des rayons X a été arrêtée presque complètement à la fin des années quatre-vingts à cause des dangers qu'elle faisait courir aux sujets.

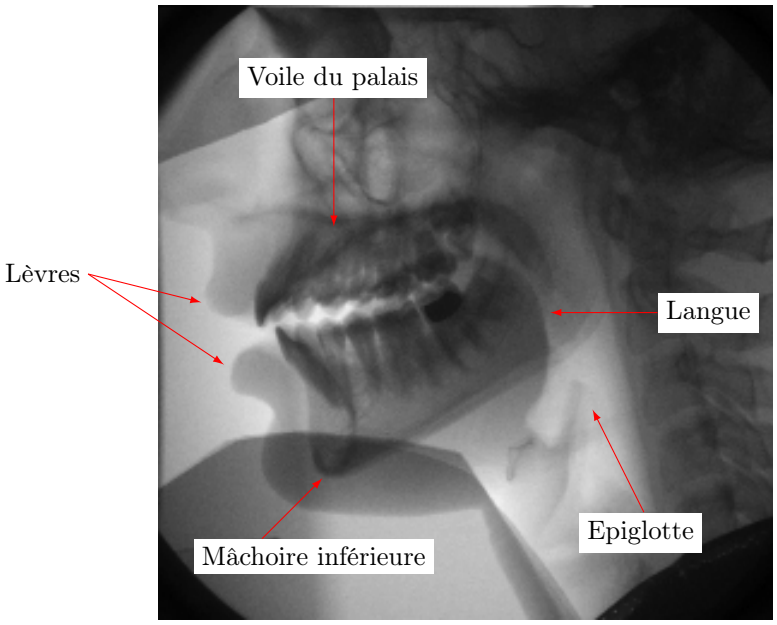


FIGURE 2.1 – *Vue sagittale du conduit vocal obtenue par cinéradiographie.*

Afin de réduire le temps d'exposition aux rayons X, Kiritani et al. [KIF75] utilisent les microfaisceaux de rayons X pour suivre les trajectoires de pastilles d'or fixées sur les principaux articulateurs. Le suivi est en général réalisé dans le plan médio-sagittal de la langue. Cette technique, qui s'avère très onéreuse, expose tout de même les sujets à des rayons X, malgré des doses beaucoup plus faibles que la cinéradiographie. Elle est désormais remplacée par une autre technique d'acquisition permettant de suivre la position de capteurs positionnés sur les principaux articulateurs : l'articulographie électromagnétique (EMA).

L'EMA est une méthode qui permet de suivre la position de capteurs en utilisant des champs magnétiques variables dans le temps. Le principe repose sur la mesure de la distance entre les bobines qui créent un champ magnétique et un capteur collé sur l'organe étudié. Les capteurs sont positionnés sur les lèvres inférieure et supérieure, sur la langue et sur l'incisive inférieure dans le plan sagittal. Les avantages de l'EMA sont une fréquence d'acquisition élevée et la possibilité d'étudier plusieurs articulateurs simultanément. Malheureusement, l'ensemble du conduit vocal ne peut être couvert, car le collage de capteurs à la base de la langue, ou au niveau du larynx est impossible. Le contour de la langue est donc incomplet. De plus, l'articulation est perturbée car il existe une compensation du sujet due aux fils reliant les capteurs au système.

L'imagerie par résonance magnétique (IRM) permet d'acquérir des informations en trois dimensions de l'ensemble du conduit vocal avec une bonne résolution [Eng00, BBR$^+$02]. Le principe de l'IRM consiste à utiliser les propriétés magnétiques des noyaux d'hydrogène présents dans les molécules d'eau. Une série d'images est alors obtenue en fonction d'une orientation : sagittale, coronale ou axiale (voir figure 2.2). Le conduit vocal et son volume peuvent être ainsi calculés directement en empilant successivement les images. L'IRM ne présente pas de risque connu pour la santé des sujets,

ce qui a permis l'acquisition de beaucoup de données. Cependant, cette technique présente quelques inconvénients notamment un temps d'acquisition élevé. En effet, le sujet doit rester immobile pendant toute la durée de l'acquisition puisqu'un léger mouvement peut provoquer des déformations de l'image. Environ 15 secondes sont nécessaires pour l'acquisition d'une vingtaine d'images espacées de 3 mm par exemple. Cette technique permet donc d'étudier seulement les positions statiques pouvant être maintenues dans le temps. De plus, les structures comme les os ou les dents n'apparaissent pas sur les images et se confondent donc avec l'air. Enfin, la position allongée du sujet et le bruit de la machine peuvent affecter l'articulation.

La technique d'IRM dynamique est une méthode permettant de visualiser une séquence d'images IRM montrant le mouvement des articulateurs lors de la production de la parole. La séquence d'images est reconstruite a posteriori après de nombreuses répétitions de la même séquence de parole. Shadle et al. [SMCJ99] ont utilisé une technique d'IRM dynamique sur plusieurs coupes sagittales afin de créer une séquence d'images du conduit vocal à partir d'un signal audio et des images acquises simultanément en répétant la même phrase à de nombreuses reprises.

Narayanan et al. [NNL+04] ont étudié une technique d'IRM en temps réel permettant d'acquérir 8 à 9 images par seconde couplée à une reconstruction a posteriori à une fréquence de 20 à 24 images par seconde. Cependant la qualité des images acquises en temps réel est très inférieure à celle obtenue avec des temps acquisition plus longs.

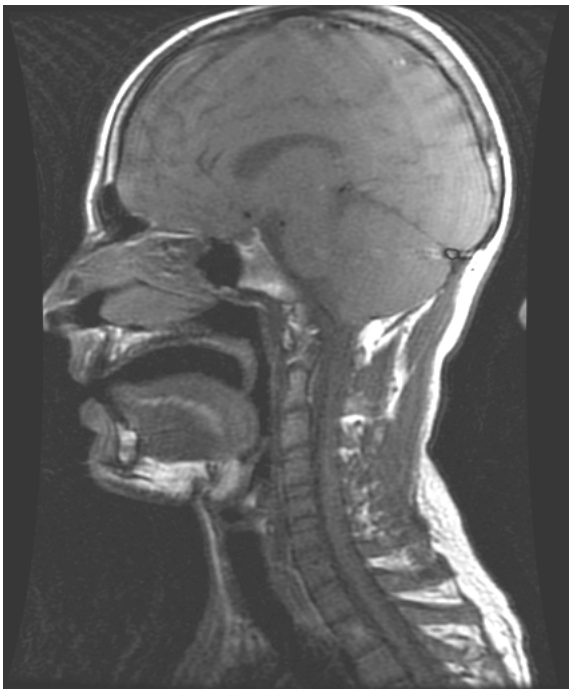


FIGURE 2.2 – *Image IRM représentant la coupe sagittale d'un /a/. Le locuteur n'a pas beaucoup ouvert la bouche ce qui donne un /a/ peu typique.*

2.2 Position du problème

L'inversion a pour objectif de retrouver la forme du conduit vocal à partir du son. En pratique, les espaces articulatoire et acoustique sont paramétrés afin de simplifier le problème. Le signal acoustique est réduit à un vecteur plus simple, comme les trois premiers formants ou les coefficients cepstraux. La forme du conduit vocal est réduite à un vecteur de petite taille, composé de paramètres contrôlant un modèle articulatoire ou les positions de capteurs (EMA).

L'inversion se heurte à plusieurs difficultés. Tout d'abord, l'existence d'une solution dépend du modèle articulatoire de production de la parole. En effet le modèle articulatoire ne peut pas approcher toutes les articulations que peut produire un locuteur humain : celle du /l/ par exemple ne peut être approchée dans le cas d'un modèle présentant une vue sagittale. Ceci n'est possible uniquement si le modèle utilisé représente une vue en trois dimensions (par exemple celui de Badin et al. [BBR$^+$02]).

Réciproquement, l'unicité de la solution n'est pas garantie. En effet, Atal [ACMT78] a montré qu'une infinité de fonctions d'aire peut produire le même triplet de formants. Enfin, la stabilité n'est pas toujours garantie en fonction de la méthode utilisée ; une petite perturbation du vecteur articulatoire peut entraîner une forte perturbation dans le domaine acoustique.

Les différentes méthodes d'inversion reposent sur plusieurs critères :
- les données articulatoires du conduit vocal utilisées (images cinéradiographiques, EMA, ...),
- le modèle de conduit vocal utilisé pour l'inversion,
- la représentation des données acoustiques,
- la prise en compte de la dimension temporelle,
- les contraintes réduisant la non-unicité de la relation articulatoire-acoustique,
- l'évaluation des performances de l'inversion.

Dans les sections suivantes, nous allons présenter les méthodes d'inversion point à point, puis les méthodes utilisant les codebooks. Enfin, comme souvent en traitement de la parole, il est possible d'utiliser une approche statistique.

2.3 Inversion point à point pour un instant donné

Les méthodes d'inversion point à point consistent à retrouver une solution à un instant précis, c'est-à-dire pour un vecteur acoustique donné, le vecteur articulatoire correspondant.

Nous présenterons dans cette partie les méthodes d'inversion directe et les méthodes basées sur les codebooks. Les méthodes d'inversion directe essaient de modéliser directement le lien entre les données acoustiques et la forme du conduit vocal. Elles n'utilisent généralement pas de modèle physique du conduit vocal mais des modèles de conduit vocal à fonctions d'aire destinés à approcher suffisamment finement la forme géométrique du conduit vocal. La fonction d'aire est souvent fortement simplifiée en utilisant un petit nombre de tubes, souvent moins de dix.

Les méthodes utilisant les tables articulatoires, ou *codebook* reposent généralement sur un modèle articulatoire du conduit vocal contrôlé par un petit nombre de paramètres.

2.3.1 Méthodes d'inversion directe

Les méthodes d'inversion directe se sont développées afin de palier les problèmes liés au manque de données articulatoires. Les premiers travaux ont porté sur des voyelles isolées. Le conduit vocal représenté par une fonction d'aire est estimé à partir d'un vecteur acoustique composé des premiers formants. Les tentatives pour résoudre ce problème analytiquement ont montré que la transformation acoustique-articulatoire n'a pas de solution unique. Mermelstein [Mer67] réalisa une étude sur le lien entre les fréquences propres et les fonctions d'aire d'un conduit vocal sans perte. La forme du conduit vocal est décomposée en termes de séries de Fourier. L'avantage de cette description mathématique est sa capacité à relier les paramètres du modèle (les coefficients de la série de Fourier) directement avec les formants. Cette description ne permet pas de modéliser une perturbation locale mais permet de modifier l'ensemble du conduit vocal.

La méthode de Schoentgen et Ciocea [SC95] utilise une formulation analytique du lien entre les formants et la fonction d'aire. Le modèle de conduit vocal est une concaténation de tubes variant dans le temps. Une solution unique est trouvée en imposant des contraintes sur les pseudo-énergies cinétiques et potentielles. La solution trouvée à t sert d'initialisation à $t+1$ et l'on part d'une solution neutre à $t = 0$.

2.3.2 Méthodes d'inversion basées sur les codebooks

L'approche par codebook consiste à construire une table formée de couples associant un vecteur articulatoire à son image acoustique représentée sous la forme d'un vecteur afin de fournir une version plus compacte de la relation articulatoire-acoustique. Les codebooks composés de couples de vecteurs sont organisés de façon à retrouver facilement les vecteurs articulatoires à partir d'un vecteur acoustique. Les codebooks peuvent être construits à partir des données synthétiques obtenues par un synthétiseur articulatoire ou alors à partir de données articulatoires et acoustiques réelles acquises simultanément. Un synthétiseur articulatoire se compose d'un modèle articulatoire gérant la forme du conduit vocal associé à une simulation acoustique. Il peut être plus ou moins réaliste du point de vue géométrique et du point de vue de la simulation acoustique. La représentation géométrique correspond à une approximation de la fonction d'aire ou de la forme du conduit vocal.

2.3.2.1 Présentation du problème

L'une des premières études marquantes sur l'inversion à partir de codebooks fut initiée par Atal [ACMT78]. En voici le principe. La relation articulatoire-acoustique peut être représentée comme une fonction multidimensionnelle f telle que $y = f(x)$, où x, y sont les vecteurs décrivant respectivement la forme du conduit vocal et les paramètres acoustiques correspondants. L'idée est de calculer pour un grand nombre de vecteurs articulatoires x, les vecteurs acoustiques $y = f(x)$ correspondants

et d'organiser les paires (x, y) en fonction des vecteurs acoustiques y. La recherche d'un vecteur x associé à un vecteur acoustique y consiste à chercher parmi toutes les valeurs de la table le vecteur donnant l'image acoustique la plus proche. En effet, si deux vecteurs articulatoires ou plus produisent un vecteur acoustique identique ou très proche, ces valeurs doivent être placées dans la même région du codebook.

L'espace articulatoire est échantillonné et la relation articulatoire-acoustique peut être approchée pour de petites régions du codebook. Le comportement de la relation articulatoire-acoustique peut être considéré comme linéaire dans un petit voisinage d'un vecteur articulatoire [ACMT78, Cha84, OL00, PLO04]. Soit M la dimension de l'espace articulatoire et N la dimension de l'espace acoustique, dans le voisinage d'un point x_0 de l'espace articulatoire, $y_0 = f(x_0)$ est approché par :

$$f(x_0) \approx y_0 + J(x - x_0) \tag{2.1}$$

avec J la matrice jacobienne de f. D'autres travaux proposent de caractériser la relation par des modélisations polynomiales [PL07] ou stochastiques [HH04, Ric01].

2.3.2.2 Construction et organisation des codebooks

La construction du codebook est un point important, le but étant d'avoir une couverture appropriée de l'espace articulatoire et une organisation efficace de la table. Différentes techniques d'échantillonnage ont été étudiées. La plupart des approches utilisent des synthétiseurs articulatoires, qui sont généralement des adaptations des modèles de Maeda [Mae79] ou de Mermelstein [Mer73], ou alors utilisent uniquement des données articulatoires mesurées sur des sujets [HLG$^+$96]. Les modèles articulatoires sont présentés dans la section 3.2.

Dans le cas des méthodes utilisant les modèles articulatoires, la quantité de données n'est en théorie pas limitée, mais la qualité des données dépend de la qualité du synthétiseur articulatoire utilisé. Pour les méthodes utilisant des données réelles, la quantité limitée de données est une source d'erreurs importante car seulement un sous-ensemble de l'espace articulatoire est exploré, ce qui induit une couverture partielle de l'espace acoustique. Au plus, l'inversion ne pourra explorer que l'espace couvert par le corpus, ce qui induit un biais important. De plus, la quantité de données n'est pas suffisante pour un apprentissage statistique. Nous décrivons maintenant quelques stratégies utilisées pour la construction des codebooks.

Codebooks à échantillonnage régulier : Pour explorer entièrement l'espace articulatoire, un échantillonnage régulier peut être réalisé. Atal et al. [ACMT78] ont utilisé cette approche pour construire leur codebook. L'espace articulatoire est celui des quatre paramètres : x_c la distance par rapport à la glotte de la constriction maximale, a_c l'aire transversale au niveau de la constriction x_c, a_m est l'aire au niveau de la bouche et l la longueur du conduit vocal. Les variables articulatoires sont échantillonnées linéairement pour la position de la constriction x_c et la longueur du conduit l et de façon logarithmique pour les aires a_c et a_m. Pour chaque vecteur articulatoire ainsi obtenu, l'image acoustique est calculée dans l'espace des trois premiers formants. L'espace acoustique à trois

dimensions est découpé en parallélépipèdes de même volume. Les données sont donc triées en fonction des valeurs des formants.

L'échantillonnage régulier permet d'explorer tout l'espace articulatoire mais un échantillonnage plus fin de l'espace articulatoire augmente le coût des calculs pour produire de très grands codebooks. De plus, les zones non-synthétisables de l'espace articulatoire sont explorées avec la même précision que les zones synthétisables.

Codebooks à partir de formes « racines » : Des codebooks construits à partir de formes de conduit vocal pertinentes ont été développés. Le but visé est de construire des codebooks contenant des formes du conduit vocal plus pertinentes et aussi d'obtenir un codebook plus petit mais pertinent.

Larar et al. [LSS88] ont échantillonné l'espace articulatoire à partir de formes dites « racines ». Ces formes ont été choisies afin de correspondre à l'ensemble des voyelles. L'espace articulatoire est l'espace des paramètres du modèle articulatoire de Mermetstein [Mer73]. L'échantillonnage est alors réalisé sur les lignes joignant chaque forme « racine » à une autre, produisant ainsi un codebook d'environ 10000 formes. Ces formes ont été rassemblées suivant une mesure de similarité acoustique basée sur les vecteurs LPC (voir section 5.1.2.2. Mais l'utilisation de formes racines connectées entre elles est trop simpliste. Il n'y a aucune garantie que les gestes réels suivent ces chemins de connexion. Et c'est même souvent faux.

Codebooks à échantillonnage aléatoire : Schroeter et al. [SMP90] ont proposé un échantillonnage aléatoire pour la construction du codebook. Des formes ont été choisies au hasard et leurs images acoustiques (représentées par les trois premiers formants) sont calculées. Les points trop similaires (c'est-à-dire les vecteurs articulatoires proches qui donnent des images acoustiques très proches) ont été supprimés du codebook. Deux modèles articulatoires ont été utilisés ; celui de Mermetstein [Mer73] et celui de Coker [Cok76].

Panchapagesan et Alwan [PA08] ont construit un codebook suivant la méthode de Schroeter et al. [SMP90] en utilisant le modèle de Maeda. Afin de réduire la taille de leur codebook, ils ont utilisé des données articulatoires issues d'acquisitions de micro-faisceaux de rayons X. Un contour partiel de la langue est approché à partir de la position des pastilles collées sur les articulateurs, qui est ensuite comparé aux contours contenus dans le codebook (obtenus par le modèle de Maeda). À partir des contours partiels obtenus sur des données réelles, ils ont éliminé tous les vecteurs articulatoires trop éloignés de ces formes. Bien que l'espace articulatoire soit caractérisé par les trois premiers formants, les coefficients cepstraux sont calculés pour chacune des configurations articulatoires du codebook épuré.

Boë et al. [BPB92] ont étudié la relation entre la géométrie du conduit vocal et les formants pour les voyelles. Un codebook a été généré à partir de formes aléatoires issues du modèle de Maeda. Seules les formes correspondant à des voyelles sont conservées. Pour chaque forme, des paramètres géométriques sont calculés : la position de la constriction (X_c), l'aire transversale de la section correspondant à la constriction (a_c) et l'aire de l'ouverture des lèvres (A_1). Ils ont montré que ces

variables géométriques peuvent être déduites du signal de parole si des contraintes articulatoires sont prises en compte dans la procédure d'inversion.

La génération de formes aléatoires permet théoriquement d'explorer pratiquement tout l'espace articulatoire ce qui permet de n'oublier aucune région. Cependant, un grand nombre de points est calculé dans les zones où il n'y a pas d'image acoustique. Les zones non-synthétisables sont donc explorées inutilement. Par contre les zones synthétisables (où les vecteurs articulatoires possèdent une image acoustique) sont explorées avec la même précision articulatoire. Or certaines zones devraient être échantillonnées plus finement, ce qui n'est pas possible.

Codebooks à échantillonnage adaptatif : Ouni et Laprie [OL00] ont utilisé un échantillonnage adaptatif dans l'espace articulatoire du modèle de Maeda. La méthode explore l'espace articulatoire de façon récursive sous forme d'hypercubes. La subdivision est guidée par la vérification de la linéarité de la relation dans un hypercube. Le critère de linéarité est évalué pour les sommets de chaque hypercube. Si la relation n'est pas linéaire, la subdivision a lieu dans les sept dimensions de l'espace articulatoire de Maeda ; ce qui conduit à l'exploration des 128 sous-hypercubes. L'espace acoustique est représenté par l'espace formé par les trois premiers formants. Potard [Pot08a] utilise également un échantillonnage adaptatif mais la subdivision est réalisée seulement dans une seule direction. Deux critères de subdivision ont été envisagés : soit diviser dans la direction qui maximise la « non-linéarité » , soit diviser dans la direction normale au demi-espace qui minimise la « non-linéarité » .

La construction de codebook à échantillonnage adaptatif permet de couvrir l'ensemble de l'espace articulatoire ; toutes les solutions sont donc présentes dans le codebook. De plus, l'échantillonnage adaptatif permet d'échantillonner plus finement certaines régions de l'espace articulatoire, notamment la frontière entre les zones synthétisables et non-synthétisables.

2.3.2.3 Recherche de solutions dans le codebook

Les codebooks sont des tables qui associent des données articulatoires à des données acoustiques. La recherche de solutions à partir du codebook dépend de sa construction et de son organisation.

Les codebooks construits à partir d'un échantillonnage aléatoire [SMP90, PA08] discrétisent l'espace acoustique en fonction généralement des trois premiers formants. Les images acoustiques de vecteurs articulatoires choisis au hasard sont rangées dans des cuboïdes de l'espace acoustique. La recherche de solutions consiste à choisir le cuboïde acoustique qui minimise la distance entre le vecteur acoustique à inverser et un vecteur acoustique représentant le cuboïde acoustique. Schroeter et Sondhi [SS92] minimisent la distance cepstrale entre les coefficients cepstraux dérivés du spectre naturel et les coefficients cepstraux dérivés du codebook.

Dans le cas des codebooks construits par échantillonnage adaptatif, une solution est calculée dans un cuboïde articulatoire dans lequel une approximation linéaire est faite. Ouni et Laprie [OL00], Potard [Pot08a] ont construit un codebook composé d'hypercubes. La recherche de solutions dans le

codebook consiste à explorer les hypercubes susceptibles de contenir un triplet de formants donnés s. Dans chaque hypercube H_c, la relation articulatoire-acoustique locale est approchée par P le polynôme d'interpolation. La recherche du vecteur articulatoire correspondant consiste à résoudre le système de M équations (non-linéaires) à N inconnues :

$$P(X) = s \tag{2.2}$$

avec $M = 7$ la dimension de l'espace articulatoire et $N = 3$ la dimension de l'espace acoustique. L'inversion consiste alors à résoudre un système sous-déterminé de 3 équations à 7 inconnues. Une solution est formée d'une solution particulière et d'un vecteur de l'espace nul. La résolution est réalisée par la méthode SVD (décomposition en valeurs singulières).

2.4 Inversion de segments acoustiques

Dans la section précédente, nous avons présenté des méthodes d'inversion statique dont le but est de trouver une ou plusieurs solutions avec une image acoustique proche des données à inverser. Le but est de réaliser une inversion dynamique de segments acoustiques, ce qui implique la recherche d'une trajectoire articulatoire.

Une trajectoire articulatoire représente l'évolution de la fonction d'aire ou de la forme du conduit vocal dans le temps. Une contrainte temporelle vient donc s'ajouter. Une trajectoire réaliste suppose que le mouvement des articulateurs soit continu. Par conséquent, la recherche de trajectoires implique donc une étape de lissage des mouvements articulatoires.

D'autres contraintes peuvent aussi être utilisées. Les formes s'éloignant trop de la position neutre des articulateurs peuvent être ainsi pénalisées [Per74].

La première étape consiste généralement à rechercher une trajectoire articulatoire initiale. La construction d'une trajectoire initiale s'effectue à partir d'une recherche dans le codebook, souvent à l'aide de la programmation dynamique (Mathieu [Mat99], Ouni [Oun01] et Potard et Laprie [PL09]).

2.5 Méthodes d'inversion utilisant un apprentissage statistique

Les principales méthodes basées sur un apprentissage statistique reposent sur des modèles de Markov cachés ou des réseaux de neurones. Ces modèles permettent d'associer des vecteurs articulatoires à des vecteurs acoustiques par un apprentissage. La phase d'apprentissage nécessite un grand volume de données. L'inversion basée sur des méthodes statistiques se réalise généralement à partir de données acquises à l'aide d'un articulographe électromagnétique qui permet de recueillir un volume de données important relativement facilement.

Les principales approches d'inversion basées sur un apprentissage statistique utilisées dans la littérature sont du type HMM (Modèles de Markov Cachés) [HH04, ZR08, BYBBH09, ZNT11] ou

GMM (Modèles de Mélanges de Gaussiennes) [TBT08] ou réseaux de neurones [PHT+92, SSJ91, Ric01].

2.5.1 Inversion basée sur les modèles de Markov cachés

Les approches utilisant les HMM cherchent à prendre en compte l'aspect temporel de la parole. Un des travaux importants d'inversion basée sur les HMM (Hidden Markov Model) a été développé par Hiroya et Honda [HH04]. Ils ont développé un modèle de production de la parole basé sur les modèles de Markov cachés (HMM). Chaque modèle de phonème se compose d'un HMM des paramètres articulatoires et d'une transformation qui permet de passer des paramètres articulatoires aux paramètres acoustiques pour chacun des états du modèle. Hiroya et Honda [HH04] ont proposé une régression linéaire pour modéliser la relation articulatoire-acoustique. La transformation articulatoire-acoustique d'un vecteur articulatoire x en un vecteur acoustique y à l'état j est alors approchée par la fonction affine :

$$y = A_j x + b_j \tag{2.3}$$

Chaque HMM λ relatif à un phonème est défini par :

$$\lambda = \{\bar{x_j}, \sigma_{x_j}, \sigma_{\omega_j}, A_j, b_j, \bar{y_j}, \sigma_{y_j}, a_{ij}\} \text{ pour tout état } j \text{ de l'HMM} \tag{2.4}$$

σ_{ω_j} est la covariance de l'erreur d'approximation linéaire de la transformation articulatoire-acoustique. Les valeurs moyennes des paramètres acoustiques et articulatoires sont $\bar{x_j}$ et $\bar{y_j}$ et σ_{x_j} et σ_{y_j} les covariances correspondantes. a_{ij} est la probabilité de transition de l'état i à l'état j.

Pour une séquence de vecteurs acoustiques donnée y et une séquence d'états q, on détermine une séquence de paramètres articulatoires x en maximisant la probabilité a posteriori $P(x|y, q, \lambda)$. La séquence d'états optimale pour une séquence de vecteurs acoustiques y est déterminée en cherchant le maximum de vraisemblance de la probabilité $P(y|\lambda)$ en utilisant l'algorithme de Viterbi.

L'approche par HMM permet de tenir compte de l'aspect temporel d'un signal de parole, cependant il est nécessaire de réaliser un étiquetage phonétique qui est très coûteux. L'approche GMM permet de réaliser un apprentissage non supervisé mais chaque instant est considéré indépendamment. Pour leur méthode d'inversion basée sur les HMM, Hiroya et Honda [HH04] obtiennent une erreur quadratique moyenne (RMSE) de 1,73 mm en utilisant uniquement une information acoustique et une erreur de 1,50 mm en ajoutant une information phonémique.

2.5.2 Inversion basée sur les GMM

Les méthodes basées sur les modèles de mélanges de gaussiennes (GMM) modélisent la distribution conjointe des vecteurs articulatoires et acoustiques par un modèle GMM.

Toda et al. [TBT08] ont décrit une approche statistique de l'inversion acoustique-articulatoire qui n'utilise pas d'information phonétique. La densité de probabilité conjointe d'un vecteur source

x_t et d'un vecteur cible y_t à l'instant t est modélisée par un GMM (Gaussian mixture model) par :

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \tag{2.5}$$

où z_t est le vecteur conjoint $z_t = [x_t^T, y_t^T]^T$. α_m est le poids de la $m^{ième}$ composante et M le nombre de composantes. La densité de probabilité d'une loi normale de moyenne μ et de covariance Σ est notée $\mathcal{N}(.; \mu, \Sigma)$. L'ensemble des paramètres du GMM est $\lambda^{(z)}$ comprend les poids, les vecteurs moyens et les matrices de covariance de chaque composante. $\mu_m^{(z)}$ et $\Sigma_m^{(z)}$ sont le vecteur moyen et la matrice de covariance de la $m^{ième}$ composante. Le GMM est entraîné avec l'algorithme de maximisation de l'espérance (EM).

La densité de probabilité conditionnelle de y_t sachant x_t, $P(y_t|x_t, \lambda^{(z)})$, est aussi représentée par un GMM. L'estimateur $\hat{y}_t$ du paramètre cible utilisant l'erreur quadratique moyenne (MMSE– Minimum mean-square error) est alors déterminé par :

$$\hat{y}_t = E[y_t|x_t] \tag{2.6}$$

où E est l'espérance mathématique d'une variable aléatoire.

L'utilisation de l'erreur quadratique moyenne (MMSE) donne d'assez bons résultats. Cependant ce n'est pas adapté pour les distributions de probabilités multiples car la covariance de chaque distribution est ignorée même quand elles sont différentes les unes les autres. De plus, les trajectoires peuvent présenter des mouvements aberrants dus à la résolution trame par trame. L'estimation du maximum de vraisemblance (MLE – Maximum likelihood estimation) est souvent utilisé à la place pour améliorer les performances ([TBT08]).

L'estimateur $\hat{y}_t$ du paramètre cible basé sur l'estimation du maximum de vraisemblance (MLE) est déterminé par :

$$\hat{y}_t = \arg\max_{y_t} P(y_t|x_t, \lambda^{(z)}) \tag{2.7}$$

Comme la densité de probabilité est aussi modélisée par un GMM, l'algorithme EM est utilisé pour maximiser la fonction de vraisemblance.

Toda et al. [TBT08] obtiennent pour la méthode basée sur l'erreur quadratique moyenne (MMSE) une erreur de 1,61 mm pour leur locutrice et de 1,53 mm pour leur locuteur. L'estimation du maximum de vraisemblance (MLE) permet d'améliorer les résultats : 1,45 mm pour la locutrice et 1,36 pour le locuteur.

L'approche basée sur les HMM utilise l'information phonétique pour relier l'acoustique à l'articulatoire alors que l'approche par GMM relie directement les paramètres acoustiques aux paramètres articulatoires.

2.5.3 Inversion basée sur les réseaux de neurones artificiels

Un réseau de neurones artificielsest constitué d'une ou de plusieurs couches de neurones connectés entre eux. Chaque connexion est associée à un poids. Le réseau de neurones reçoit en entrée les données et fournit les résultats du traitement en sortie. Une phase d'apprentissage est nécessaire afin de garantir la cohérence de la sortie par rapport à la sortie demandée en ajustant les différents poids de chaque connexion. Le réseau de neurones reçoit les paramètres acoustiques en entrée et fournit les paramètres articulatoires correspondants en sortie.

Papcun et al. [PHT+92] ont employé des réseaux de neurones appliqués à des données articulatoires acquises par micro-faisceaux de rayons X. Les données acoustiques et articulatoires sont utilisées pour entraîner le réseau : il n'y a donc pas de synthèse acoustique.

Soquet et al. [SSJ91] réalisent l'inversion à partir d'un réseau de neurones à trois couches. La fonction d'aire (composée de 30 sections) de onze voyelles du français a ainsi été estimée à partir des trois premiers formants.

Richmond [Ric01] utilise des réseaux à mélange de densités (mixture density networks – MDNs) afin d'obtenir des trajectoires articulatoires à partir du signal de parole. Il a aussi utilisé des réseaux de neurones artificiels pour réaliser l'inversion. Il a montré que les MDN sont plus performants pour aborder le problème de non-unicité que les réseaux de neurones artificiels.

Une fois la phase d'entraînement du réseau terminée, la phase d'inversion demande peu de ressources en termes de mémoire et de vitesse d'exécution.

La méthode basée sur les réseaux de neurones artificiels présentée par Uria et al. [UMRR12] permet d'obtenir une erreur de 0,885 mm.

2.6 Conclusion

Dans ce chapitre nous avons vu que les différentes méthodes d'inversion sont fortement liées à la nature et la quantité des données articulatoires disponibles.

Le développement de grands corpus de données articulatoires acquises par EMA a permis le développement de méthodes statistiques basées sur un apprentissage à partir de données. Malheureusement, les données acquises par cette méthode ne sont pas complètes car elles correspondent à la position d'un petit nombre de capteurs sur les articulateurs de la parole.

Lorsque les données articulatoires ne sont pas en quantité suffisante, il est possible de recourir à l'utilisation de modèles physiques du conduit vocal. Ces modèles généralement associés à une synthèse articulatoire créent des données articulatoires supplémentaires. Les modèles sont composés d'un nombre réduit de paramètres qui contrôlent la position et la forme des différents articulateurs conduisant à une représentation généralement sagittale du conduit vocal. Cependant, la relation entre le modèle physique et l'acoustique est complexe et nécessite une adaptation à tout nouveau locuteur.

Dans le cadre de cette thèse, nous disposons de données articulatoires acquises par cinéradiographie et du signal de parole correspondant. Nous souhaitons développer une méthode d'inversion basée sur une analyse par synthèse en utilisant une table associant des données articulatoires (paramètres du modèle) et acoustiques (coefficients cepstraux). Il est donc nécessaire de disposer d'un modèle articulatoire adapté à notre locuteur afin de produire un signal synthétique que l'on veut le plus proche possible du signal réel. À partir de nos données, il nous est possible d'associer une configuration articulatoire (contour du conduit vocal dans le plan sagittal) au signal correspondant. La construction du codebook se fera suivant la méthode proposée par Potard [Pot08a] mais les vecteurs acoustiques seront les coefficients cepstraux en lieu et place des formants. L'inversion recherchera dans le codebook les vecteurs articulatoires correspondant à un vecteur acoustique donné. Les données synthétiques permettent de construire des codebooks couvrant tout l'espace articulatoire possible et de compléter les données lorsqu'il est impossible de les acquérir avec une méthode d'imagerie. Il est donc nécessaire de comparer le signal réel et celui produit par la synthèse articulatoire car nous devons rechercher dans le codebook (construit à partir de données synthétiques) une solution correspondant à une donnée issue du signal réel. En effet, les différences entre les signaux réels et synthétiques doivent être compensées d'une façon ou d'une autre pour pouvoir réaliser une inversion acoustique-articulatoire de signaux réels à partir de données produites par la synthèse articulatoire.

Notre approche d'inversion présente des similitudes avec celle proposée par Panchapagesan et Alwan [PA08], [PA11]. En effet, ils proposent une méthode d'inversion d'analyse par synthèse.Les données articulatoires (associées à l'acoustique) utilisées sont issues de corpus acquis par microfaisceaux de rayons X et sont utilisées afin de réaliser une adaptation au locuteur du modèle de Maeda. L'ensemble du conduit vocal est modifié afin de s'adapter au locuteur et le contour extérieur (le palais et le pharynx) est également modifié. Leur codebook est construit à partir de vecteurs articulatoires choisis aléatoirement et regroupés suivant leur proximité acoustique. Ils supposent que les vecteurs articulatoires présents dans le codebook ne sont pas tous pertinents et ne conservent donc que les formes articulatoires proches des données articulatoires à leur disposition. Leur méthode d'inversion est basée sur l'optimisation d'une fonction de coût. La solution initiale est obtenue par une recherche dans le codebook. Un des points importants est le choix du vecteur acoustique : leur codebook est organisé suivant les trois premiers formants. Panchapagesan et Alwan utilisent les coefficients cepstraux pour identifier la région de l'espace acoustique contenant le vecteur acoustique à inverser. Dans les travaux [PA08], le terme acoustique de la fonction de coût mesure la distance entre les coefficients cepstraux.L'évaluation de leur méthode d'inversion est réalisée en calculant la distance entre la position de chacun des capteurs avec le contour obtenu par le modèle articulatoire adapté. Cette méthode ne permet pas d'apporter d'explications claires sur l'origine des erreurs observées. En particulier, il n'est pas possible de savoir si l'adaptation du modèle est incorrecte ou si l'approche elle-même est insuffisante.

Notre méthode d'inversion par analyse par synthèse utilise un modèle articulatoire construit à partir de données articulatoires issues d'images cinéradiographiques d'un locuteur. Le codebook sera construit suivant une méthode d'exploration récursive de l'espace articulatoire. Nous utiliserons comme vecteurs acoustiques les coefficients cepstraux au lieu des formants qui sont généralement utilisés.

Nous avons choisi de construire un nouveau modèle dans le but de réduire les erreurs liées à l'adaptation au locuteur à partir d'un modèle existant. Des erreurs liées au nouveau modèle seront présentes car il utilise uniquement un nombre réduit de modes de déformations. Nous espérons que l'élaboration d'un nouveau modèle sera bien supérieure en termes de fidélité à la géométrie du conduit vocal du sujet que la simple adaptation d'un modèle géométrique.

L'évaluation de notre méthode sera effectuée sur l'ensemble du conduui vocal. En effet, une mesure géométrique mesurera la distance entre la forme « réelle » et la forme inversée. La forme dite réelle est celle composée par les contours des articulateurs visibles sur les images cinéradiographiques. Pour chaque image, nous pourrons ainsi évaluer la qualité de l'inversion.

Première partie

Modélisation du conduit vocal

Chapitre 3

Modélisation du conduit vocal

L'INVERSION acoustique-articulatoire a pour but de retrouver la forme du conduit vocal à partir d'un son. La forme du conduit vocal humain est très complexe, d'où la nécessité d'utiliser une représentation simplifiée qui approche la forme réelle. Traditionnellement, la forme du conduit vocal est décrite par une fonction d'aire ou par un modèle articulatoire qui fournit une description géométrique.

Il existe de nombreuses modélisations du conduit vocal dans la littérature. Nous présenterons dans cette section les modèles les plus représentatifs ; les modèles à fonction d'aire et les modèles articulatoires.

Enfin, nous présenterons la synthèse articulatoire qui représente la relation entre le domaine articulatoire vers le domaine acoustique. En d'autres termes, la synthèse permet de passer de la forme articulatoire au signal sonore.

3.1 Modèles à fonction d'aire

La fonction d'aire d'un conduit vocal se définit par l'aire transversale le long de la ligne médiane du conduit de la glotte aux lèvres. Dans cette représentation, la forme géométrique de la section transversale n'est pas prise en compte. Seule l'aire de la section est prise en compte. De plus, l'angle formé par les cavités orale et pharyngale n'est pas représenté. Cependant, cette description simplifiée en une seule dimension est acoustiquement valide pour les fréquences inférieures à 4 kHz, où le principal mode de propagation du son s'effectue le long du conduit vocal [SH55].

La description du conduit vocal avec des fonctions d'aire permet une simplification des calculs des caractéristiques acoustiques associés. Le conduit vocal peut être représenté par une vingtaine de sections pour lesquelles l'aire transversale et la longueur varient au cours du temps. Au lieu de contrôler l'aire de chaque section, il est préférable de décrire la fonction d'aire avec un plus petit nombre de paramètres, qui peuvent avoir un lien avec des caractéristiques articulatoires, comme la position de la langue ou encore la position et l'aire de la constriction.

3.1.1 Modèles à trois paramètres

Stevens & House [SH55] et Fant [Fan70] ont proposé des modèles de fonction d'aire à trois paramètres : la position de la constriction, l'aire à la constriction et un paramètre pour l'ouverture des lèvres représenté par le rapport entre la longueur de la section des lèvres et l'aire de la section des lèvres. Le modèle proposé par Fant [Fan70] (voir figure 3.1) est formé de quatre tubes : un tube pour les lèvres (ouverture) (A_1), une cavité avant (A_2), un tube pour la constriction de la langue (A_3) et une cavité arrière (A_4). La longueur totale des tubes est de $16cm$ et l'aire transversale des cavités avant et arrière est fixée à $8cm^2$.

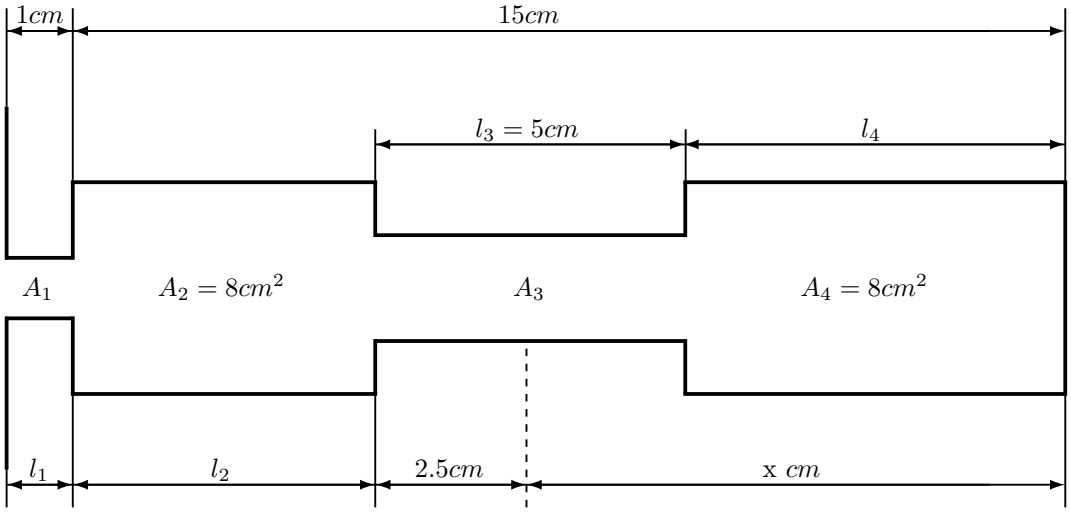


FIGURE 3.1 – *Modèle du conduit vocal à fonction d'aire proposé par Fant [Fan70].*

Bien que ces modèles soient très simples, ils capturent les configurations articulatoires pour les voyelles. Dans des versions plus sophistiquées, la section uniforme où se situe la constriction est remplacée par une fonction parabolique [SH55] ou une fonction hyperbolique [Fan70] afin de mieux représenter la forme arrondie du corps de la langue dans la fonction d'aire.

En fait, ces modèles ne sont pas vraiment des modèles de fonction d'aire, mais de cavités résonantes équivalentes d'un point de vue acoustique.

3.1.2 Modèles à concaténation de tubes

Le conduit vocal peut être représenté par une succession de tubes cylindriques mis bout à bout. Dans ce type de modèle, la fonction d'aire est décrite par les caractéristiques de chaque tube, c'est-à-dire la longueur et l'aire de la section. La précision peut être améliorée en ajoutant des tubes.

Cette représentation présente plusieurs inconvénients. Tout d'abord, un grand nombre de paramètres est nécessaire afin d'avoir une représentation assez précise. De plus, une fonction d'aire ne correspond pas forcément à une forme de conduit vocal humain, ce qui est gênant pour l'inversion acoustique-articulatoire.

Schoentgen et Ciocea [SC95] proposent un modèle de fonction d'aire un peu plus évolué. Les segments de conduit ont une forme conique, ce qui permet d'obtenir des fonctions d'aire continues.

3.2 Modèles articulatoires

Contrairement aux modèles à fonction d'aire, les modèles articulatoires représentent de façon géométrique la forme du conduit. La plupart des modèles offrent une représentation en deux dimensions dans le plan médio-sagittal parce qu'il s'agit de la forme la plus appropriée et que c'est une visualisation du conduit pour laquelle on dispose d'un grand nombre de données.

La coupe sagittale du conduit vocal est une représentation qui permet d'obtenir assez fidèlement l'acoustique à l'aide de modèle de passage de la coupe sagittale à la fonction d'aire [HS65].

Des modèles articulatoires contrôlés par un petit nombre de paramètres permettant de représenter les différentes coupes sagittales réalisables par un humain ont ainsi été développés.

Plusieurs solutions ont été envisagées pour la construction de modèles articulatoires du conduit vocal. Les modèles géométriques proposent une représentation des organes à partir de formes géométriques simples [Cok76], [Mer73]. D'autres modèles ont été construits à partir d'une analyse statistique de contours sagittaux extraits d'images aux rayons X [Mae90]. Enfin les modèles biomécaniques reposent sur l'intégration de propriétés physiologiques et des interactions entre les différentes structures [Per74]. Le développement des techniques d'imagerie médicale, telles que l'IRM, a permis l'élaboration de modèles en trois dimensions [BBR$^+$02].

Nous allons maintenant présenter les principaux modèles articulatoires existant dans la littérature.

3.2.1 Modèles géométriques

Les modèles géométriques sont construits à partir de primitives géométriques simples (des droites et cercles) correspondant à une représentation compacte des organes. Ils visent à décrire l'anatomie de chaque organe, mais cette simplification géométrique est à la fois très pauvre et sans contrainte assurant un certain réalisme.

Le modèle de Coker et Fujimura [Cok76] utilise cinq paramètres. La langue y est représentée par un cercle qui peut se déplacer dans deux directions dans le plan sagittal. Son mouvement est limité par le palais et la paroi pharyngale. La position de l'apex est contrôlée par une troisième variable. Deux paramètres pour les lèvres permettent de contrôler l'ouverture et la protrusion.

Le modèle de Mermelstein [Mer73] est plus évolué et propose une description des articulateurs de la parole à partir de la position de six structures principales : la mâchoire, le corps de la langue, l'apex de la langue, le voile du palais, les lèvres et l'os hyoïde. La position de ces articulateurs est contrainte dans l'espace.

Les modèles géométriques permettent de représenter les configurations du conduit vocal pour des voyelles et des consonnes à partir d'un petit nombre de paramètres. L'utilisation de formes géométriques simples ne permette pas de représenter certaines configurations articulatoires. Par exemple, la langue est représentée de façon circulaire, ce qui exclut de modéliser sa capacité d'aplatissement. De plus, ces modèles ne prennent pas en compte les influences croisées entre articulateurs. Les modèles géométriques représentent mal les gestes articulatoires. Ces modèles sont trop peu contraints et utilisent trop de paramètres.

3.2.2 Modèles statistiques

Les modèles statistiques sont plus réalistes car ils sont construits à partir de corpus de coupes sagittales. Ils sont fondés sur des formes réelles dérivées de données cinéradiographiques, ils fournissent donc une bonne capacité descriptive du conduit vocal. Une analyse statistique permet d'extraire des caractéristiques articulatoires pertinentes et interprétables.

L'un des principaux modèles articulatoires construit à partir d'images cinéradiographiques est le modèle de Maeda [Mae90]. Il résulte de l'observation de 400 contours de conduits vocaux extraits d'images cinéradiographiques enregistrés lors de la production de petites phrases. La forme du conduit est mesurée dans un système de coordonnées semi-polaires. La paroi intérieure se compose de la pointe, du corps et de la racine de la langue et de la partie supérieure du larynx. Le contour extérieur se compose lui des incisives supérieures, du palais dur, du voile du palais et des parois du pharynx et du larynx. On considère les intersections de ces contours avec la grille semi-polaire. Le système de coordonnées semi-polaires est inutile pour les tubes des lèvres et du larynx car ces articulateurs se déplacent dans une direction quasi perpendiculaire aux lignes de la grille. Le larynx est représenté par les coordonnées des bords intérieur et extérieur de l'extrémité du larynx. Le tube représentant le pavillon labial est modélisé par une ellipse contrôlée par trois variables : la hauteur, la largeur et la longueur (la protrusion). La recherche des composantes du modèle s'effectue à partir d'une analyse factorielle proposée par Overall [Ove62]. Maeda [Mae90] montrent que deux composantes suffisent à expliquer 90 % de la variance des données de contours de langue. Avec trois composantes retenues, ce taux s'élève à 98%. Le modèle (voir figure 3.2) ainsi obtenu se compose de sept paramètres :

- position de la mâchoire,
- position du corps de la langue,
- forme du corps de la langue,
- position de l'apex de la langue,
- ouverture des lèvres,
- protrusion des lèvres,
- hauteur du larynx.

Les paramètres articulatoires varient entre moins trois et plus trois écarts-types.

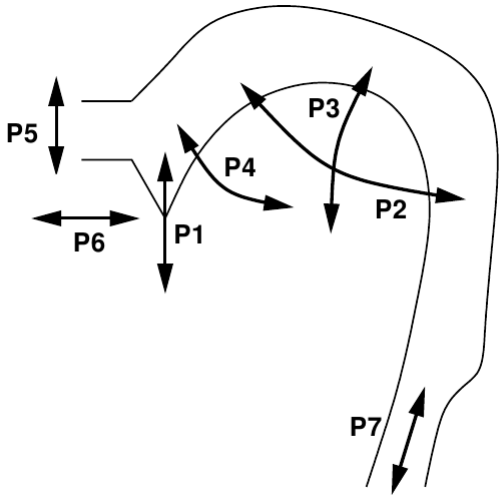


FIGURE 3.2 – *Les sept paramètres du modèle de Maeda : la position de la mâchoire (P1), la position du corps de la langue (P2), la forme du corps de la langue (P3), la position de l'apex de la langue (P4), l'ouverture des lèvres (P5), la protrusion des lèvres (P6) et la hauteur du larynx (P7).*

D'autres modèles dérivés du modèle de Maeda ont été développés. Galván-Rodrigez [GR97] a modifié l'apex de la langue pour pouvoir modéliser les fricatives. Mathieu et Laprie [ML97] ont proposé une adaptation du modèle de Maeda en adaptant le palais dur.

D'autres modèles basés sur une étude statistique des contours sagittaux existent. Beautemps et al. [BBB01] ont étudié les degrés de liberté pour la production de voyelles orales, de consonnes occlusives et fricatives du français à partir de contours sagittaux extraits de films cinéradiographiques. Le modèle articulatoire développé se compose de neuf paramètres ; un pour la mâchoire, un pour la hauteur du larynx, trois pour les lèvres, quatre pour la langue. L'étude utilise comme Maeda, l'intersection des contours de la langue avec une grille semi-polaire, mais Beautemps et al. [BBB01] ont utilisé une grille qui s'adapte en fonction du contour de la langue afin de garder un nombre constant d'intersections de la grille avec la langue.

La majorité des études effectuées concerne une vue sagittale du conduit vocal car les données articulatoires étaient principalement disponibles dans le plan sagittal. Les systèmes d'acquisition en trois dimensions ont permis l'étude de modèles statistiques en trois dimensions. Sur le même principe que Maeda, Badin et al. [BBR+02] proposent un modèle articulatoire basé sur une analyse factorielle sur les trois dimensions à partir d'images IRM et de labiofilms.

3.2.3 Modèles biomécaniques

Les modèles biomécaniques sont des modèles de conduits vocaux qui intègrent des propriétés physiologiques des articulateurs (os, muscles) et leurs interactions.

Perkell [Per74] a développé le premier modèle physiologique de la langue (voir figure 3.3). Ce modèle, de type masse-ressort, propose une représentation sagittale de la langue et du conduit

vocal. Le modèle est composé de seize nœuds porteurs de masses reliés entre eux ou à des éléments générateurs de tension. La génération de tension peut être active ou passive. Les éléments actifs correspondent aux tissus musculaires qui sont capables de développer des forces en réponse à une stimulation. Les éléments passifs représentent les tissus conjonctifs et les structures rigides ou molles du conduit vocal. L'organisation en éléments actifs ou passifs se base sur une étude anatomique de la langue.

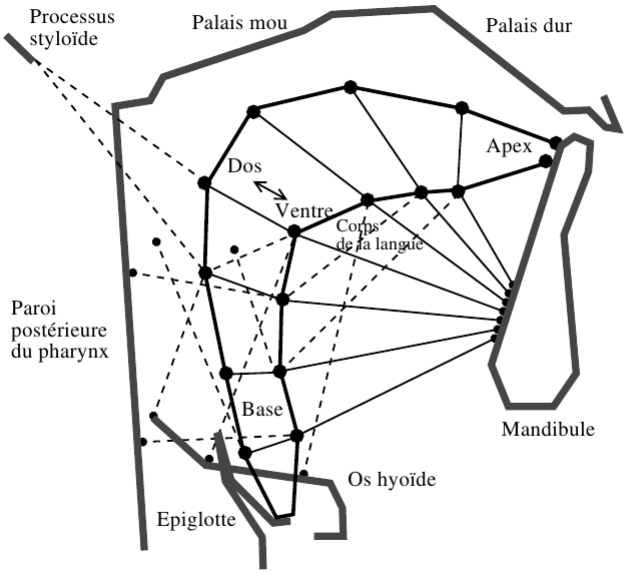


FIGURE 3.3 – *Le modèle biomécanique de Perkell. Les éléments de tension sont représentés par les lignes continues et en pointillé, les points noirs sont les nœuds porteurs de masse. (D'après [Per74]).*

Des travaux plus récents ont apporté des améliorations à l'aide de mesures des activités musculaires plus précises ; au développement des modèles tridimensionnels et à l'utilisation des méthodes par éléments finis [KM75], [PP97], [DH97], [WT95], [SLO98].

Plus récemment, Gérard et al. [GWTPP03] utilisent une méthode par éléments finis dans laquelle ils modélisent la langue en trois dimensions en se basant sur les lois physiques de l'élasticité non-linéaire.

Le nombre important de paramètres et la difficulté à les déterminer sont des inconvénients majeurs à l'utilisation de modèles biomécaniques dans le cadre de l'inversion malgré la modélisation réaliste du conduit vocal.

3.3 Passage de la coupe sagittale à la fonction d'aire

Dans de nombreuses études sur la production de la parole, la représentation du conduit vocal repose sur une coupe sagittale du conduit vocal. Cependant, afin d'étudier les relations entre la géométrie du conduit vocal et l'acoustique correspondante, il est nécessaire de transformer la coupe sagittale

en une fonction d'aire. La fonction d'aire du conduit vocal représente l'aire transversale pour chaque section de la coupe sagittale. La fonction d'aire représente un lien essentiel entre les configurations articulatoires et l'acoustique et nécessite suffisamment de précision pour son estimation.

Le développement de l'imagerie médicale, notamment l'IRM, permet une estimation plus précise de la forme en trois dimensions du conduit vocal. Le volume peut donc être calculé directement à partir des différentes coupes.

Disposant de données IRM de notre locuteur, des calculs de l'aire à partir des images IRM ont été effectués. Les essais réalisés n'étaient pas concluants car le locuteur utilisait une ouverture de la mâchoire trop faible lors de l'acquisition IRM et il était donc difficile d'extrapoler l'aire des coupes pour une ouverture de la mâchoire plus grande. Le maintien de l'articulation pendant plusieurs secondes et la position allongée du locuteur sont à l'origine de ces différences.

Notre étude porte sur des données cinéradiographiques qui nous fournissent une vue sagittale du conduit vocal, mais de telles données ne permettent pas de mesure directe de la fonction d'aire. Ceci explique la nécessité d'utiliser un modèle qui convertit la coupe sagittale en une fonction d'aire.

Le passage de la coupe sagittale du conduit vocal à la fonction d'aire correspondante n'est pas simple à cause de la forme irrégulière du conduit vocal. La figure 3.4 représente un exemple de l'irrégularité des sections transversales ; les dix coupes transversales ont été réalisées à partir d'un moulage du conduit vocal sur un cadavre.

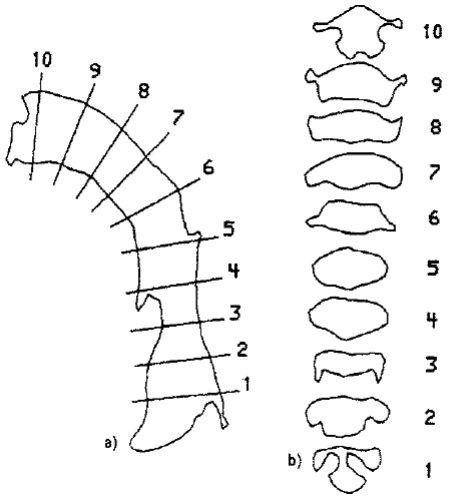


FIGURE 3.4 – *Coupes transversales du conduit vocal réalisées à partir d'un moulage de conduit vocal réalisé sur un cadavre. a) coupe sagittale du conduit vocal, b) les sections transversales correspondantes (d'après [Cal89]).*

Le passage de la coupe sagittale à la fonction d'aire a fait l'objet de nombreuses études. Le modèle proposé par Heinz et Stevens [HS65] est à la base de nombreuses études. À partir de mesures faites sur des cadavres, Heinz et Stevens ont créé un modèle basé sur une relation de puissance entre

la fonction d'aire et la coupe sagittale. La fonction d'aire A, c'est-à-dire l'aire transversale de chaque section, est calculée par :

$$A(x) = \alpha(x).d(x)^{\beta(x)} \tag{3.1}$$

où x est la distance à la glotte, A est l'aire transversale de la section et d est la distance sagittale dans une direction normale au flux d'air. Les coefficients α et β sont déterminés de façon ad hoc et varient suivant la région du conduit vocal.

Les modèles de passage de la coupe sagittale à la fonction d'aire sont spécifiques à un locuteur. Plusieurs travaux ont cherché à améliorer ce modèle et définir les paramètres α et β ([Mae90], [PBS92], [BBL95], [SLMD02]).

Perrier et al. [PBS92] ont proposé un modèle inspiré de Heinz et Stevens, basé sur l'analyse de coupes du conduit vocal (obtenues par tomographie) pour trois voyelles /i/, /a/, /u/ prononcées par un locuteur masculin. Le coefficient α est déterminé en fonction de la région du conduit vocal, qui est décomposé en sept régions. Le coefficient β est fixé à 1.5.

Beautemps et al. [BBL95] ont proposé une extension du modèle de [PBS92], où le coefficient α varie de façon continue le long de la ligne médiane du conduit vocal. Ce modèle optimisé pour un sujet permet de modéliser des consonnes fricatives et des voyelles.

Soquet et al. [SLMD02] ont comparé trois types de transformations : une transformation linéaire, une transformation polynomiale (d'ordre deux) et la transformation de Heinz et Stevens. Les aires transversales réelles ont été estimées à partir de données issues de l'IRM. Il en résulte que la transformation proposé par Heinz et Stevens est plus adaptée au problème.

De nombreuses études ont porté sur la modélisation du passage de la coupe sagittale à la fonction d'aire. La transformation de Heinz et Stevens semble la plus adaptée au problème. L'estimation des coefficients α et β en fonction du locuteur étudié permet d'améliorer les résultats.

3.4 Simulation acoustique

Dans cette section, nous présentons la simulation acoustique décrite à l'aide des équations de l'acoustique. Nous évoquerons uniquement le cas du conduit oral, ainsi que les sources d'excitation du conduit vocal.

3.4.1 Equations de l'acoustique

L'étude de la propagation des vibrations sonores dans le conduit vocal est très complexe [RS78]. Plusieurs simplifications doivent donc être faites. Le conduit vocal est modélisé par une concaténation de tubes cylindriques élémentaires. Les caractéristiques du conduit peuvent alors être représentées par la fonction d'aire qui spécifie l'aire transversale de chaque tube de la glotte jusqu'aux lèvres (voir la figure 3.5).

Les ondes sonores sont créées par une variation du débit acoustique et se propagent dans l'air par les vibrations des molécules de l'air. Ainsi, les lois de la physique servent de base pour décrire la génération et la propagation du son. Le mouvement de l'air dans le conduit vocal est décrit par un ensemble d'équations. La formulation et la recherche de solutions de ces équations est possible en utilisant des hypothèses très simples sur la forme du conduit vocal et sur les pertes d'énergie dans le conduit.

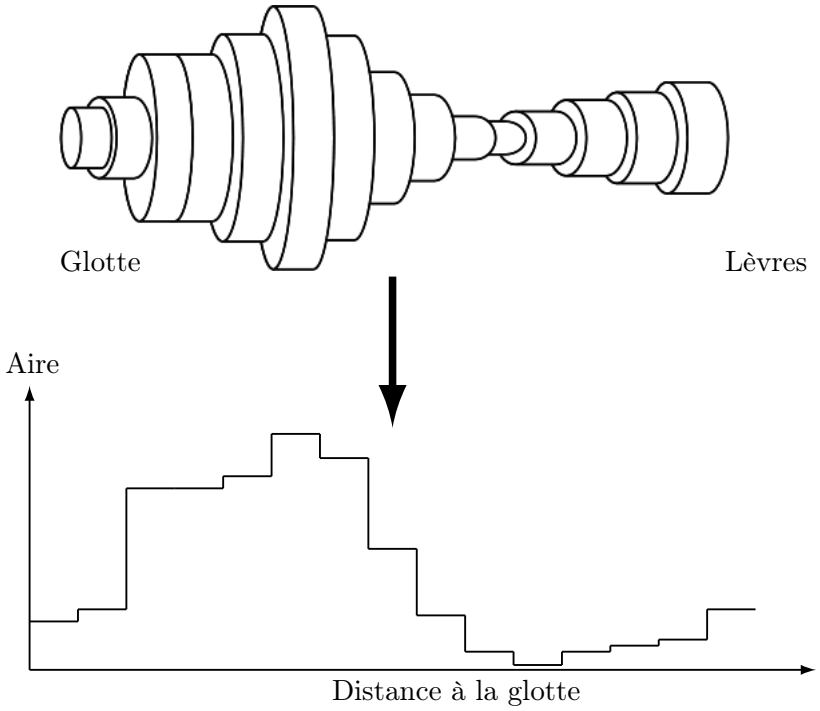


FIGURE 3.5 – *La figure du haut montre la représentation du conduit vocal sous forme de concaténation de tubes. La figure du bas montre la fonction d'aire correspondant au conduit vocal du haut.*

Prenons l'exemple d'un cas relativement simple où le conduit vocal est représenté par un tube cylindrique, uniforme et sans perte. Pour l'ensemble des équations suivantes, on utilisera les notations suivantes, en supposant la propagation réduite à des ondes planes acoustiques :

$p = p(x,t)$ la pression acoustique du son dans le tube à la position x et au temps t,

$u = u(x,t)$ le débit volumique à la position x et au temps t,

ρ la densité de l'air,

ρ_0 est la masse volumique de l'air,

c la célérité du son,

$A = A(x,t)$ l'aire transversale à la position x et au temps t.

La propagation du son dans un tube cylindrique, uniforme et sans perte est décrite par les équations de l'acoustique. La première équation utilisée est l'équation d'état. On suppose que la

propagation du son est adiabatique et que l'air est un gaz parfait. L'équation d'état se définit alors par :

$$\rho = \rho_0 + \frac{1}{c^2}p \tag{3.2}$$

L'équation d'Euler s'applique dans le cadre d'un fluide parfait, elle s'écrit :

$$-\frac{\partial p}{\partial x} = \rho_0 \frac{\partial u}{\partial t} \tag{3.3}$$

Enfin, l'équation de continuité s'écrit :

$$-\rho_0 \left(\frac{\partial A}{\partial x} u + \frac{\partial u}{\partial x} A \right) = \frac{A}{c^2} \frac{\partial p}{\partial t} \tag{3.4}$$

A partir des équations de l'acoustique (3.2), (3.3) et (3.4), on en déduit les équations (3.5) qui régissent la propagation d'une onde acoustique dans un tuyau rigide sans perte.

$$\begin{cases} -\dfrac{\partial p}{\partial x} = \dfrac{\rho}{A} \dfrac{\partial u}{\partial t} \\ -\dfrac{\partial u}{\partial x} = \dfrac{A}{\rho c^2} \dfrac{\partial p}{\partial t} \end{cases} \tag{3.5}$$

Les équations (3.5) peuvent être décrites par l'équation unidimensionnelle de Webster (3.6) en dérivant l'équation de continuité (3.4) par rapport au temps. L'équation de Webster s'écrit alors :

$$\frac{1}{A(x,t)} \frac{\partial}{\partial x} \left(A(x,t) \frac{\partial p(x,t)}{\partial x} \right) = \frac{1}{c^2} \frac{\partial^2 p(x,t)}{\partial t^2} \tag{3.6}$$

Un parallèle avec les lignes de transmission électriques uniformes sans perte peut être fait (voir le tableau 3.1).

Quantités acoustiques	Quantités électriques analogues
p - pression	v - tension
u - débit volumique	i - courant
$\frac{\rho}{A}$ - inductance acoustique	L - inductance
$\frac{A}{\rho c^2}$ - capacité acoustique	C - capacité

TABLEAU 3.1 – *Analogie entre les quantités acoustiques et électriques.*

La tension $v(x,t)$ et le courant $i(x,t)$ sur la ligne satisfont les équations :

$$\begin{cases} -\dfrac{\partial v}{\partial x} = L \dfrac{\partial i}{\partial t} \\ -\dfrac{\partial i}{\partial x} = C \dfrac{\partial v}{\partial t} \end{cases} \tag{3.7}$$

où L et C sont respectivement les inductance et capacité par unité de longueur. La résolution des équations est détaillée dans [Fla72].

3.4.1.1 Excitation du son dans le conduit vocal

La source sonore peut venir de plusieurs origines et les sons peuvent être classés en trois catégories en fonction de leur mode d'excitation.

1. Les *sons voisés* sont produits par le passage de l'air à travers la glotte dans laquelle les cordes vocales vibrent excitant ainsi le conduit vocal (voir figure 3.6).
2. Les *sons fricatifs* sont générés par la formation d'une constriction étroite dans le conduit vocal (généralement dans la cavité buccale du conduit) qui provoque un écoulement turbulent en aval de la constriction. Cela crée une source de bruit qui excite le conduit vocal.
3. Les *sons occlusifs* résultent de la fermeture complète du conduit vocal qui crée une pression derrière la constriction. Un bruit d'explosion est généré lors du relâchement de la occlusion.

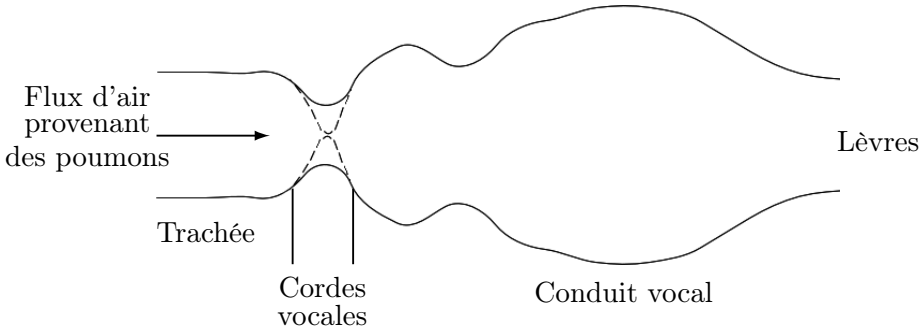


FIGURE 3.6 – *Représentation schématique du système vocal.*

3.5 Conclusion

Le conduit vocal humain est un organe complexe, donc difficile à représenter. Les premières représentations sont les fonctions d'aire qui expriment le volume des différentes cavités avec plus ou moins de précision. L'acquisition de corpus de plus en plus importants, a permis le développement de modèles plus réalistes qui offrent une description plus précise du conduit vocal. Les modèles articulatoires ont généralement été développés en deux dimensions, mais grâce au développement de nouvelles méthodes d'acquisition des modèles en trois dimensions ont été créés.

Dans cette étude, nous disposons de données en deux dimensions, une étape supplémentaire passant de la coupe sagittale à la fonction d'aire permet d'obtenir la troisième dimension. Le système

de production de la parole humain peut être ainsi simulé numériquement à partir d'une représentation simplifiée du conduit vocal par analogie avec les équations de l'électrique. Ainsi, il est possible d'associer à une représentation géométrique du conduit vocal la fonction de transfert correspondante.

Le modèle articulatoire est ici un point crucial pour la simulation. Notre méthode d'inversion nécessite l'utilisation d'un modèle articulatoire le plus fidèle possible à la géométrie du locuteur test. Il est ainsi possible d'utiliser des modèles existants et de les adapter pour approcher le conduit vocal du locuteur test.

Ici, nous sommes dans un cadre favorable car nous disposons de données articulatoires et le son correspondant de bonne qualité. Ainsi, une comparaison est donc réalisable entre les fonctions de transfert calculées sur le signal réel et celles obtenues via une simulation acoustique. Nous avons choisi de construire un modèle spécifique à notre locuteur test pour s'approcher de la géométrie du conduit vocal et réduire les erreurs liées à une mauvaise adaptation. Le chapitre suivant sera consacré à la construction du modèle articulatoire à partir des images cinéradiographiques présentant une vue sagittale du conduit vocal.

Chapitre 4

Construction d'un nouveau modèle articulatoire

Le codebook utilisé lors de l'inversion est construit à partir d'un synthétiseur articulatoire utilisant en particulier un modèle paramétrique du conduit vocal. Afin de se placer dans les meilleures conditions possibles, nous avons choisi de construire un modèle articulatoire adapté à partir des données d'un locuteur test. Cela nous permettra dans le futur d'étudier l'impact de l'adaptation du modèle à un nouveau locuteur sur la qualité des résultats. L'influence sur l'inversion des disparités entre le modèle et le conduit vocal pourra ainsi être étudiée en dégradant artificiellement la qualité de la représentation du modèle ce qui n'a jamais pu être étudié.

La construction d'un modèle est effectuée à partir de données articulatoires acquises par cinéradiographie du locuteur test. La première étape consiste à traiter les images afin d'obtenir une représentation sagittale du conduit vocal à partir de la position ou des contours des articulateurs. Les contours médio-sagittaux du conduit vocal constituent une bonne représentation de l'articulation de la parole. En effet, pour la plupart des phonèmes la forme du conduit vocal peut être déduite à partir du contour sagittal. De plus, les profils médio-sagittaux permettent de lier la forme géométrique du conduit à l'acoustique correspondante.

La seconde étape consiste à construire un modèle à partir de ces représentations sagittales à l'aide d'une analyse factorielle.

4.1 Traitement des données

4.1.1 Corpus

Le corpus a été enregistré dans les années 90 par un locuteur masculin français dans le but d'étudier la coarticulation en français [SHL$^+$11]. Il se compose de quatre films. Les deux premiers sont des séries de six courtes phrases allant de /se dø si yltɛR/ à /se dø sikst skyltɛR/ (pour chaque

phrase une consonne non-labiale est ajoutée entre /i/ et /y/ par rapport à la précédente) à des fréquences d'élocution normale et rapide. Les deux dernières séries sont des séries de /VCV/ (/aku iku uku atu itu utu/) à des fréquences d'élocution normale et rapide.

Malheureusement, ces quatre films ne sont pas phonétiquement équilibrés. En dépit de ces faiblesses, la taille, la couverture de l'ensemble du conduit vocal, les deux fréquences d'élocution et son caractère dynamique comparé aux IRM font de ce corpus une importante ressource articulatoire. Au total, ce corpus se compose de 946 images (256×256 pixels) à 25 images par seconde. Seules les images correspondant à de la parole, soit 672 images, sont conservées. Les images correspondant aux instants sans parole ou à la respiration ne sont pas prises en compte.

Nous avons comme objectif de construire un modèle articulatoire basé sur les données de notre locuteur test. Pour la construction de ce nouveau modèle, il est nécessaire de déterminer la position et la forme des principaux articulateurs à partir des images cinéradiographiques. La cinéradiographie fournit une image sur laquelle toutes les structures se projettent contrairement à une tomographie. Les principaux articulateurs de la parole sont plus ou moins visibles : la mâchoire, la langue, l'épiglotte, le larynx et les lèvres (voir la figure 4.1).

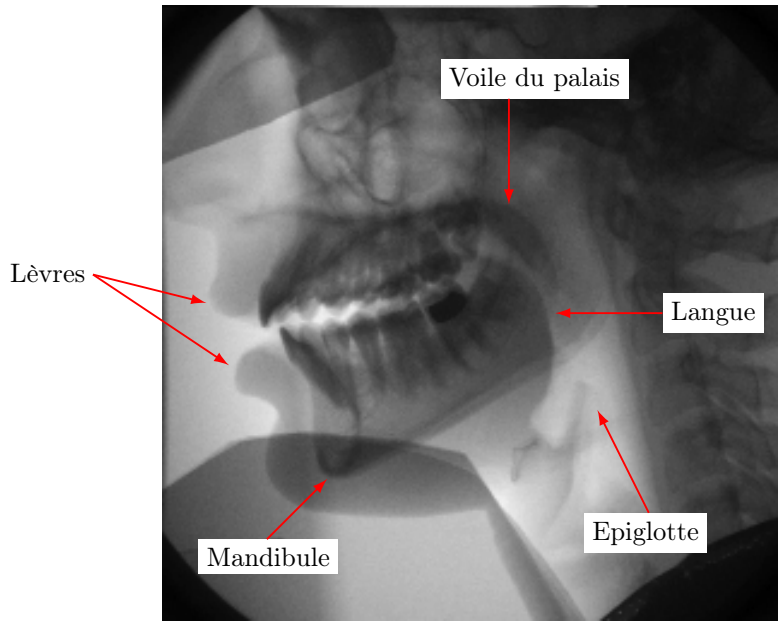


FIGURE 4.1 – *Vue sagittale du conduit vocal obtenue par cinéradiographie. Les principaux articulateurs visibles sont la mandibule inférieure, la langue, les lèvres, l'épiglotte et le voile du palais.*

Les contours des principaux articulateurs ont été extraits des images cinéradiographiques soit à la main soit automatiquement via le logiciel *Xarticul* qui propose des outils développés dans ce but [SHL+11]. *Xarticul* propose des outils automatiques afin de suivre les régions rigides, comme les os, des outils de suivi semi-automatique pour les lèvres, le larynx et l'épiglotte et des outils pour tracer le contour de la langue.

4.1.2 La mâchoire et les structures rigides

Le premier articulateur étudié est la mâchoire. En effet, la mâchoire est un des articulateurs les plus importants car elle influence directement les autres articulateurs, notamment la langue et les lèvres. Les différents mouvements de la mâchoire ont été étudiés à plusieurs reprises [Wes88, EH90, OVBG97]. La mâchoire est une structure rigide qui possède six degrés de liberté (trois rotations et trois translations). Cependant lors de la production de parole, le mouvement est effectué dans le plan médio-sagittal. La position d'un objet rigide dans le plan sagittal est uniquement déterminée par une rotation et une translation par rapport à un point de référence.

Pour son modèle géométrique, Mermelstein [Mer73] suppose une rotation de la mâchoire autour d'un point fixe et les configurations articulatoires correspondant à un mouvement latéral ne peuvent donc pas être prises en compte. Maeda [Mae79] calcule le mouvement de la mâchoire à partir la position de l'incisive inférieure. Le déplacement de l'incisive est approché par une ligne droite car l'amplitude de la rotation de la mâchoire est très petite.

Les images cinéradiographiques de notre locuteur montrent une projection dans le plan sagittal sur laquelle les structures osseuses de la tête sont visibles. L'os de la mâchoire inférieure est visible. La mâchoire inférieure constitue un ensemble rigide qui ne se déforme pas dans le temps. Son mouvement s'observe uniquement dans le plan sagittal et correspond à une rotation et une translation. Ainsi, un suivi automatique par corrélation est utilisé pour calculer le mouvement de cette structure. Le suivi par corrélation consiste à rechercher dans une image une région donnée correspondant à une région de référence. La région peut s'être déplacée suivant une rotation et une translation.

L'algorithme de suivi commence par le choix d'une image de référence. Sur cette image, une région associée à la mâchoire doit être déterminée de façon à ce qu'elle apparaisse sur toutes les autres images à un déplacement près. Cette région est choisie de telle façon qu'elle n'intersecte pas le filtre visible dans le coin en bas à droite et qu'elle limite son intersection avec la langue. Nous avons choisi une région qui a la forme d'un boomerang de façon à maximiser les régions de plus fort contraste qui se situent principalement au niveau des dents. La région utilisée pour le suivi est représentée dans la figure 4.2.

Pour chaque image de la séquence, on recherche la région de référence à une rotation et une translation près. Le centre de rotation est déterminé sur l'image de référence. Il s'agit du point en haut à gauche de la région. Les paramètres du mouvement ainsi obtenus constituent les paramètres du mouvement de la mâchoire. La figure 4.3 représente un exemple de résultat obtenu à partir du suivi par corrélation. L'image de gauche est l'image de référence utilisée sur laquelle une région pour la mâchoire a été définie. Sur l'image de droite, la région correspond à la région de l'image de référence à laquelle on a appliqué la rotation et la translation calculée.

De plus, il est possible de déduire du mouvement de la mâchoire, la position des incisives inférieures et le plancher de la bouche. La position des incisives inférieures peut être définie par un point dans le plan sagittal qui est placé sur la partie supérieure des incisives inférieures. Le plancher de la bouche est représenté par un contour tracé manuellement. La position des incisives inférieures et le plancher de la bouche (bien qu'il soit assez déformable) sont définis sur l'image de référence

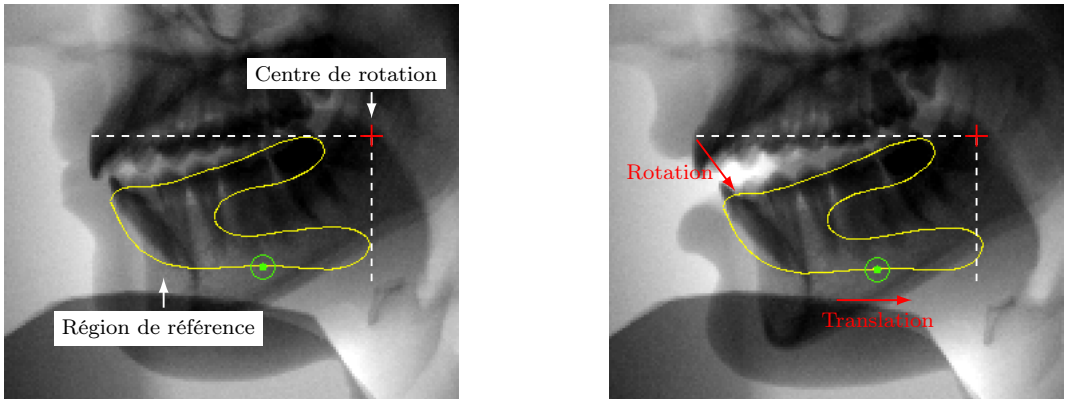


FIGURE 4.2 – *Suivi par corrélation du mouvement de la région de la mâchoire. L'image de gauche est l'image de référence où la région de la mâchoire a été tracée. L'image de droite présente le résultat du suivi ; la région s'est déplacée par rapport à l'image de référence.*

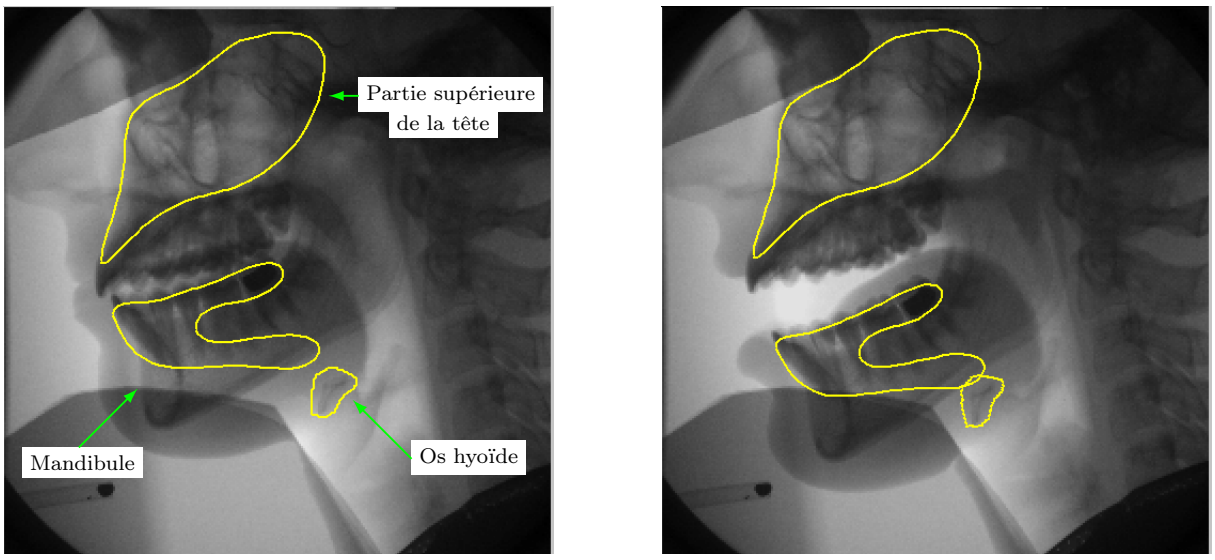


FIGURE 4.3 – *Exemple de suivi par corrélation pour le mouvement de la mâchoire, le mouvement de la tête et le mouvement de l'os hyoïde. L'image de gauche est l'image de référence où les contours ont été tracés manuellement. L'image de droite est issue du suivi, les régions sont celles de l'image de référence auxquelles le mouvement calculé a été appliqué.*

utilisée pour initialiser le suivi de la mâchoire. Leur position est obtenue en appliquant le mouvement obtenu pour la région de la mâchoire (voir figure 4.4).

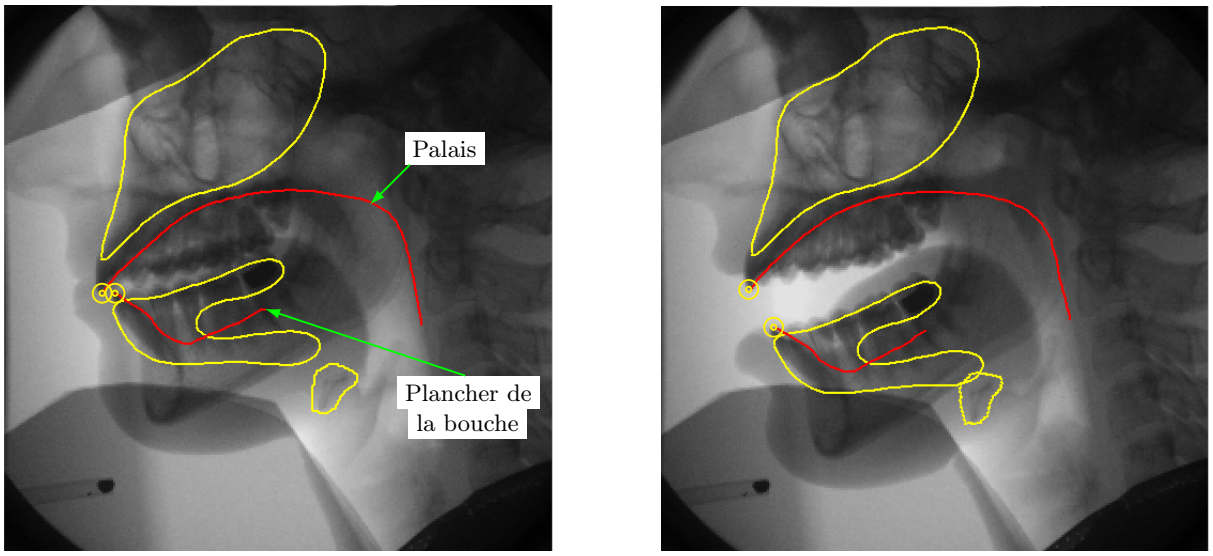


FIGURE 4.4 – *Exemple de suivi par corrélation pour le mouvement de la mâchoire, le mouvement de la tête et le mouvement de l'os hyoïde. Le contour du palais suit le mouvement de la région de la tête et le plancher de la bouche suit le mouvement de la mâchoire. Les cercles jaunes correspondent à la position des incisives supérieures et inférieures qui suivent respectivement le mouvement de la tête et le mouvement de la mâchoire. L'image de gauche est l'image de référence. L'image de droite est issue du suivi, les contours et les positions des incisives sont déplacés en fonction de la région auxquels ils sont reliés.*

L'algorithme de suivi par corrélation peut être utilisé pour toutes les structures rigides qui ne sont pas entièrement recouvertes par d'autres organes. Il est donc possible de suivre également le mouvement de l'os hyoïde à condition qu'il ne soit pas trop haut et donc partiellement ou complètement caché par la mandibule (voir figure 4.3).

Le suivi est aussi utilisé pour connaître le mouvement de la tête. En effet, bien que la tête du sujet soit maintenue, de légers mouvements de la tête sont visibles. Le principe est le même que pour la mâchoire : on définit une région dans la partie supérieure de la tête et son mouvement (une rotation et une translation) est calculé pour les autres images (voir figure 4.3).

Le mouvement de la tête permet d'obtenir le contour du palais dur. En effet, le palais dur suit le même mouvement que la tête.Le contour du palais dur est alors tracé sur l'image de référence auquel on applique le mouvement de la tête pour les autres images (voir figure 4.4). On fait l'hypothèse que le sujet ne change pas l'orientation de la tête par rapport au pharynx, ce qui n'est pas tout à fait vrai.

4.1.3 Les lèvres et l'épiglotte

4.1.3.1 Les lèvres

Contrairement à Maeda [Mae90], nous ne disposons pas de labiofilm pour les séquences cinéradiographiques. L'étude des lèvres n'est donc réalisée que dans le plan sagittal. Les lèvres se déplacent et se déforment légèrement dans le temps : donc il n'est pas possible de suivre leur mouvement automatiquement par corrélation avec une image de référence. Les contours des lèvres sont visibles et ne sont pas recouverts par d'autres organes. De plus, l'observation visuelle des images montre qu'il existe beaucoup de formes qui sont très proches. Le tracé sur toutes les images serait inutile car beaucoup de formes sont très proches. Le suivi semi-automatique proposé par Fontecave et Berthommier [FB06] est utilisé. L'idée du suivi est de tracer les contours sur un nombre réduit d'images clés et de calculer les contours sur les autres images à partir des images clés.

Nous allons maintenant présenter l'algorithme de suivi semi-automatique plus en détail [FB06]. Tout d'abord, les images sont découpées de façon à ne conserver que la partie étudiée c'est-à-dire les lèvres. Une image toutes les dix images est choisie comme image clé. Sur chacune des images clés, les contours sont tracés avec le logiciel *Xarticul*. Chaque image clé possède alors le même nombre de contours et chaque contour a le même nombre de points. Pour chaque image, les coefficients DCT (*Discrete Cosine Transform – Transformée en cosinus discrète*) sont calculés. L'application de la DCT permet de passer dans le domaine fréquentiel. La formule (4.1) donne la définition de la DCT en deux dimensions d'une image dont la matrice A représente le niveau de gris de chaque pixel, le résultat est dans l'image B.

$$B(i,j) = \alpha_i \alpha_j \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A(m,n) \cos \frac{\pi(2m+1)i}{2M} \cos \frac{\pi(2n+1)j}{2N} \tag{4.1}$$

pour

$$0 \leq i \leq M-1$$
$$0 \leq j \leq N-1$$

où α_i et α_j sont définis par :

$$\alpha_i = \begin{cases} \frac{1}{\sqrt{M}} & \text{, si } i = 0 \\ \sqrt{\frac{2}{M}} & \text{, si } 1 \leq i \leq M-1 \end{cases}$$

et

$$\alpha_j = \begin{cases} \frac{1}{\sqrt{N}} & \text{, si } j = 0 \\ \sqrt{\frac{2}{N}} & \text{, si } 1 \leq j \leq N-1 \end{cases}$$

avec M et N les dimensions de l'image A.

La DCT possède une capacité de « regroupement » de l'énergie ; les premiers coefficients portent en effet l'essentiel de l'information. Un triangle dans le coin supérieur gauche de la matrice B est utilisé pour le calcul d'une mesure de similarité. Cette mesure correspond à la distance euclidienne entre les coefficients DCT des deux images. Pour chacune des images (non-clés), on calcule donc la distance entre les coefficients DCT de l'image et ceux de toutes les images clés. Puis, pour chaque image, les trois images clés les plus proches en terme de distance entre les coefficients DCT sont conservées. Les nouveaux contours sont obtenus en calculant les points moyens des trois images clés les plus proches. La dernière étape consiste à replacer les contours dans l'image originale (c'est-à-dire non découpée). La figure 4.5 montre un exemple de calcul de contour à partir des trois images clés les plus proches.

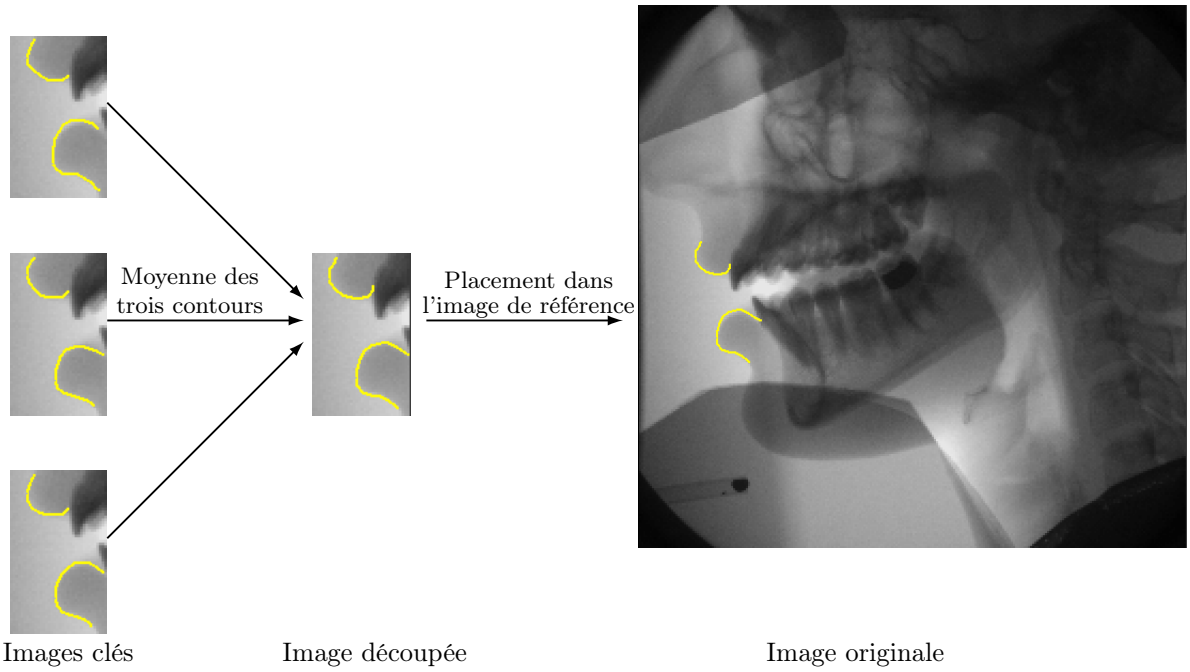


FIGURE 4.5 – *Exemple de contours obtenus par l'algorithme de suivi semi-automatique. Les nouveaux contours sont obtenus à partir de la moyenne des contours des trois images clés les plus proches. Les nouveaux contours sont replacés dans l'image originale.*

Dans notre cas, les lèvres sont représentées par deux contours : un contour pour la lèvre supérieure et un pour la lèvre inférieure. Les contours sont tracés à l'aide du logiciel *Xarticul* et nous nous avons assuré que la direction du tracé soit la même pour toutes les images clés. Le point de départ de la courbe est situé à la jonction entre la lèvre et l'incisive et le point final est fixée de façon arbitraire. Une *spline* est appliquée sur les contours tracés de façon manuelle afin qu'ils aient tous le même nombre de points. Nous avons ensuite appliqué le suivi sur l'ensemble des images.

Le nombre d'images étant relativement peu élevé, nous avons examiné visuellement les résultats afin de garantir que les contours soient corrects. Quand un contour est mal suivi cela signifie que l'image concernée est trop éloignée des images clés et elle est donc ajoutée comme image clé.

4.1.3.2 L'épiglotte et le larynx

L'épiglotte et le larynx possèdent les mêmes caractéristiques que les lèvres, c'est-à-dire qu'ils se déforment légèrement, possèdent un contour visible et ne sont pas recouverts par d'autres organes. Le même algorithme de suivi que pour les lèvres (voir paragraphe précédent) a donc été utilisé.

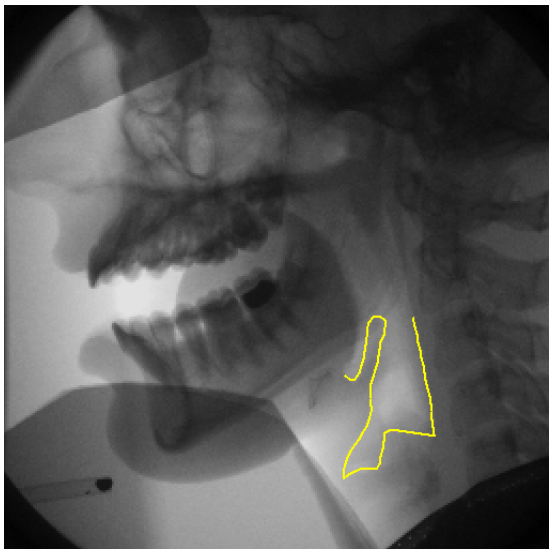


FIGURE 4.6 – *Image cinéradiographique présentant les contours utilisés pour l'épiglotte et le larynx.*

Le suivi semi-automatique permet d'obtenir les contours de certains articulateurs en contrepartie du traçage d'environ 10% des images. La précision des contours ne peut être mesurée. L'examen des contours obtenus montre que le résultat du suivi semi-automatique est cohérent.

4.1.4 La langue

Le dernier articulateur étudié est la langue. La langue est l'articulateur le plus difficile à suivre automatiquement. Plusieurs études sur le suivi automatique de la langue ont montré que cette tâche n'est pas aisée [BML95, TL99, FB06].

Tout d'abord, la forme de la langue peut conduire à plusieurs contours dans le plan sagittal. En effet, la langue présente un sillon central, l'image cinéradiographique étant une projection dans le plan sagittal deux contours (voire trois si la langue n'est pas symétrique du tout) peuvent apparaître. De plus, le contour peut être caché par d'autres organes comme les dents ou les os de la mâchoire. Les contours de la langue ont donc été tracés manuellement avec le logiciel *Xarticul*. Le contour de la langue tracé est celui correspondant au sillon central.

Afin de faciliter le traçage des contours, le logiciel *Xarticul* permet d'afficher rapidement les images précédentes ou suivantes de l'image en cours d'analyse. En effet, la visualisation du mouvement aide souvent à situer le contour de la langue qui peut être confondu avec d'autres organes.

4.1.5 Contours des articulateurs

La figure 4.7 représente une image cinéradiographique avec l'ensemble des positions et des contours des articulateurs. Les régions représentées en jaune sont les régions suivies automatiquement par corrélation ; la mâchoire inférieure, le crâne et l'os hyoïde. Les cercles correspondent aux positions des incisives dépendant du mouvement des régions rigides et les contours rouges sont les contours du palais et du plancher de la bouche liés à la région du crâne et de la mâchoire. Les contours bleus sont les contours utilisant le suivi semi-automatique ; les lèvres, l'épiglotte et le larynx. Le contour orange est celui de la langue tracé à la main.

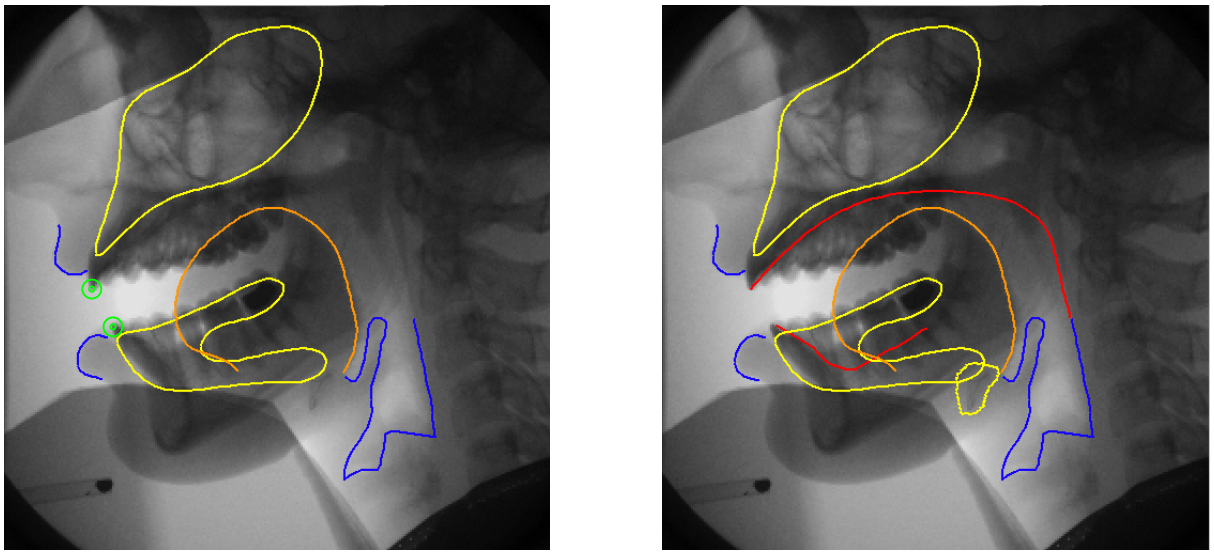


FIGURE 4.7 – *Images cinéradiographiques présentant les contours des différents articulateurs du conduit vocal.*

L'étape suivante est de construire un modèle articulatoire contrôlé par un petit nombre de paramètres à partir de toutes ces données articulatoires obtenues à partir des images cinéradiographiques.

4.2 Construction du modèle articulatoire

L'objectif maintenant est de construire un modèle articulatoire contrôlé par un petit nombre de paramètres et adapté à notre locuteur. L'analyse factorielle permet de fournir une description plus compacte d'un point de vue statistique.

Pour chaque articulateur, nous allons réaliser une analyse en composantes principales (ACP) dans le but d'obtenir un petit nombre de paramètres afin d'expliquer le plus de variance possible. Ce traitement statistique est purement mathématique : il n'est donc pas évident d'obtenir des compo-

santes physiquement interprétables. Dans cette section, nous décrivons les analyses statistiques pour la mâchoire, la langue, l'épiglotte et le larynx.

4.2.1 Soustraction du mouvement de la tête

Avant d'effectuer des analyses sur les données articulatoires calculées dans la section 4.1, un premier traitement doit être effectué. Malgré la contention de la tête du locuteur lors de l'enregistrement des données, il subsiste un léger mouvement de la tête. Ce mouvement doit donc être retiré des contours et des positions calculés précédemment.

Le mouvement de la tête a été calculé à l'aide d'un suivi automatique (voir § 4.1.2). Il est composé d'une rotation et d'une translation. Il est donc possible de retirer ce mouvement des contours obtenus en appliquant le mouvement inverse.

4.2.2 La mâchoire

La première analyse factorielle est effectuée pour la mâchoire. En effet, la mâchoire est l'un des articulateurs les plus importants car elle influence fortement les autres articulateurs notamment la langue et les lèvres.

4.2.2.1 ACP sur les données

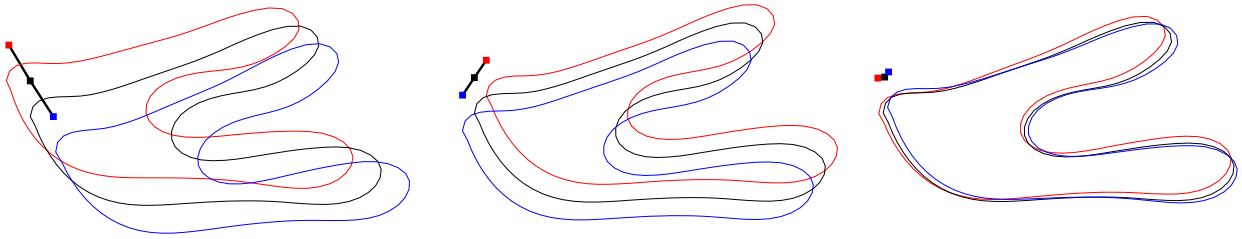
Le mouvement de la mâchoire correspond au mouvement de la région définie dans le paragraphe 4.1.2. Il se compose d'une rotation, définie par son angle (le centre étant fixe défini en fonction de la position de la région sur l'image de référence ; il n'est pas pris en compte dans l'analyse), et une translation, définie par deux coordonnées x et y. La mâchoire est ainsi contrôlée par trois paramètres. Nous utiliserons une analyse en composantes principales afin de réduire le nombre de paramètres pour le mouvement de la mâchoire. La variance expliquée par la première composante est de 64% et de 26% pour la seconde (voir le tableau 4.1).

Pourcentage de variance	Pourcentage cumulé
64.42 %	64.42 %
26.23 %	90.65 %
9.35 %	100 %

TABLEAU 4.1 – *Tableau des pourcentages de variance expliquée résultant de l'ACP sur les données de la mâchoire.*

La figure 4.8 montre le mouvement obtenu à partir des différentes composantes, il est appliqué sur la région utilisée pour le suivi automatique du déplacement de la mâchoire. La première composante (voir figure 4.8(a)) est un mouvement d'ouverture et de fermeture de la mâchoire accompagné

d'une légère translation. Ce mouvement se rapproche d'une rotation. La deuxième composante (voir figure 4.8(b)) est totalement différente, le mouvement décrit est « à peu près orthogonal » à celui de la première composante. La rotation n'est pas très importante, la translation décrit un mouvement d'avancement/descente à un mouvement de retrait vers le haut. La dernière composante (voir figure 4.8(c)) décrit un faible déplacement de la mâchoire.



(a) Première composante (64,42%) (b) Deuxième composante (26,23%) (c) Troisième composante (9,35%)

FIGURE 4.8 – *Composantes obtenues par ACP sur les données de la mâchoire. Le mouvement est appliqué sur la région utilisée pour le suivi. Les composantes varient entre -3 écarts-types (ligne rouge) et +3 écarts-types (ligne bleue). La courbe noire étant la position moyenne. Les carrés représentent la position du bord de l'incisive inférieure.*

L'objectif étant de converser un petit nombre de composantes, nous avons choisi de conserver seulement la première composante comme paramètre de la mâchoire. Contrairement à d'autres approches, cette composante contrôle la rotation et la translation. Les trois paramètres du mouvement, l'angle de rotation θ_M, les coordonnées de la translation x_M et y_M, sont donc contrôlés par le paramètre de mâchoire M.

4.2.2.2 Soustraction de la contribution de la mâchoire

La mâchoire étant liée à la langue, la contribution de son mouvement doit d'abord être soustraite aux mouvements de la langue. Il est très difficile de séparer les mouvements de la langue provoqués par celui de la mâchoire et ceux issus de l'action des muscles de la langue à cause de la complexité des liens entre les muscles de la langue et la mâchoire [SLO98]. Bailly et al. [BBV98] ont montré que les mouvements propres à la langue sont plus importants que ceux associés à la mâchoire. Le mouvement passif de la langue dû au mouvement de la mâchoire doit donc être déterminé.

Lors de la construction de son modèle, Maeda [Mae79] utilise comme unique paramètre pour la mâchoire : la position de l'incisive inférieure.

Beautemps et al. [BBB01] se sont inspirés de la méthode utilisée par Maeda mais la mâchoire est contrôlée par deux paramètres : la hauteur et l'avancement.

La composante de la mâchoire est connue, elle doit maintenant être soustraite des données de la langue et des autres articulateurs avant d'appliquer l'analyse factorielle. Deux possibilités sont envisagées.

La première, qui est généralement adoptée, consiste à soustraire la corrélation entre les données de la mâchoire et les données de la langue (Maeda [Mae79]). La seconde consiste à soustraire géométriquement le mouvement de la mâchoire du contour de la langue. Ainsi, il n'y a plus d'influence du mouvement de la mâchoire sur le contour de la langue. D'un autre côté, d'autres interactions plus complexes entre la mâchoire et la langue subsistent.

La première stratégie est la meilleure pour réduire la variance dans le corpus analysé. Cependant, il faudrait que le contenu articulatoire du corpus d'images soit phonétiquement équilibré, ce qui est rarement vrai. Nous explorerons donc les deux stratégies.

4.2.3 La langue

Pour la langue nous disposons des contours sagittaux tracés à la main. L'utilisation des coordonnées euclidiennes des points des contours n'est pas judicieuse pour une analyse factorielle. En effet les points ne se correspondent pas dans le temps d'une image à une autre car la langue se déforme et se déplace.

Nous avons choisi d'utiliser une grille polaire adaptative qui s'adapte à chaque contour de langue. Les extrémités de la grille sont la racine et l'apex de la langue. Le centre de la grille est un point défini sur la mâchoire et qui suit son mouvement. La grille polaire est définie à partir de ces trois points et les lignes de la grille sont espacées régulièrement (voir la figure 4.9). La grille utilisée comporte cent lignes.

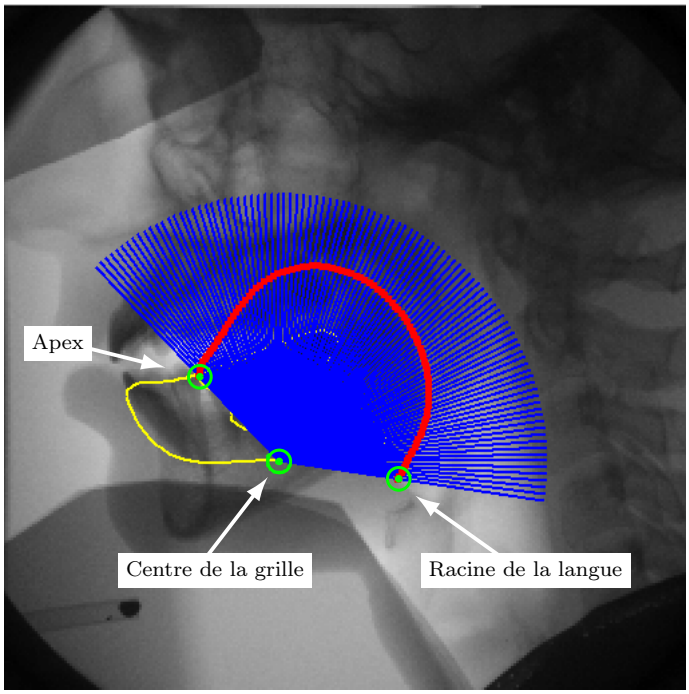


FIGURE 4.9 – *Représentation de la grille polaire adaptative. La grille est définie à partir de son centre, de l'apex et de la racine de la langue. Les points rouges sont les points d'intersection entre la grille et le contour de la langue.*

4.2.3.1 Différentes présentations des données de la langue

Les données utilisées dans l'analyse factorielle sont des vecteurs unidimensionnels dans le cas d'une grille semi-polaire : les intersections du contour de la langue avec les lignes de la grille. Il n'est pas possible d'utiliser des vecteurs unidimensionnels dans le cas d'une grille polaire adaptative puisque les lignes de la grille ne sont pas fixes. Les données sont donc dépendantes de la position de la grille polaire. Trois approches ont donc été envisagées afin de palier cet inconvénient.

Approche 1 (Coordonnées euclidiennes) : La première approche consiste à prendre les coordonnées (x, y) des points d'intersection de la grille avec le contour de la langue. Les coordonnées du centre de la grille sont soustraites des coordonnées de chaque point d'intersection.

Approche 2 (Distances et angles extrêmes) : La deuxième approche n'utilise pas les coordonnées des points d'intersection mais la distance entre le centre de la grille et chaque point d'intersection. À ces distances deux autres paramètres viennent s'ajouter : deux angles correspondants à la racine de la langue θ_R et à l'apex θ_A. Puisque les angles entre les lignes de la grille sont régulièrement espacés entre le racine de la langue et l'apex, la même information que pour l'approche 1 est contenue dans ces données mais en divisant la quantité de données par deux. La figure 4.10 illustre l'approche 2.

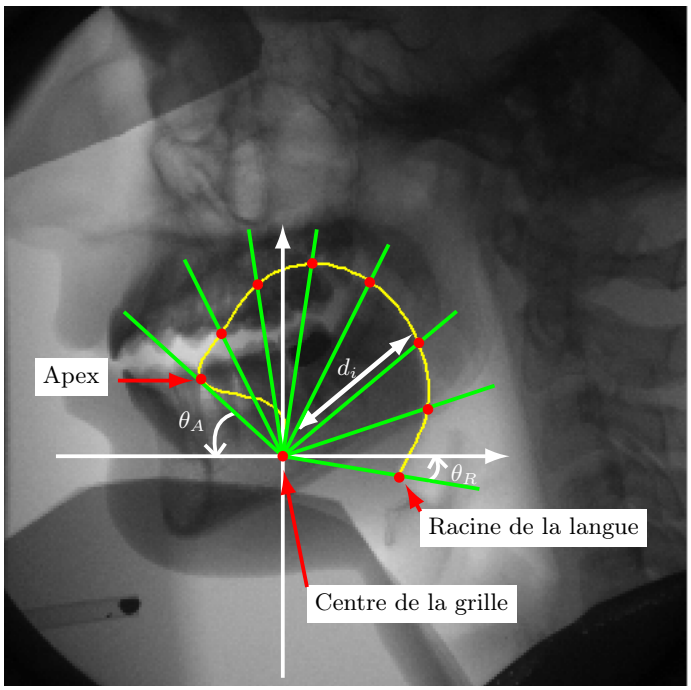


FIGURE 4.10 – *Données utilisées avec l'approche 2. Les points rouges sont les points d'intersection entre la grille et le contour de la langue. d_i est la distance entre le point d'intersection avec la $i^{ème}$ ligne de la grille et le centre de la grille.*

Approche 3 (Coordonnées polaires) : La dernière approche calcule les distances entre le centre de la grille et chaque point d'intersection comme pour l'approche 2 mais les angles correspon-

dants à chaque ligne de la grille sont calculés (les coordonnées polaires par rapport au centre de la grille). La quantité de données est comparable à l'approche 1.

Nous avons testé ces trois approches parce qu'elles offrent des points de vue légèrement différents. La première solution est la plus naturelle, mais elle sépare les coordonnées x et y dans l'analyse. La seconde solution semble plus adaptée à la nature de la langue mais elle combine deux sortes de données (des distances et des angles). La dernière solution utilise des données redondantes puisque les angles peuvent être obtenus à partir des deux angles extrêmes mais garde une homogénéité entre les données (chaque point est défini par un angle et une distance). Les deux dernières solutions offrent l'avantage d'expliquer la nature radiale de la langue ce qui peut être observé lorsque la langue se contracte lors de l'articulation d'un /u/ par exemple. Ceci devrait favoriser l'émergence de cette déformation dans l'analyse.

4.2.3.2 Différentes stratégies d'analyse

Une analyse statistique des données doit être faite dans l'objectif de représenter les données de manière plus compacte. Parmi les différentes méthodes d'analyse factorielle, nous pouvons citer l'analyse en Composantes Principales (ACP) ou encore l'analyse en Composantes Indépendantes (ACI) qui cherche à extraire des composantes linéaires indépendantes. L'ACP fournit une description hiérarchique des données et extrait des composantes (décorrélées et orthogonales) par ordre décroissant d'énergie. De plus, l'ACP est toujours possible alors l'ACI n'est possible que si les composantes existent car son objectif est de trouver des composantes « physiquement significatives ». Nous avons choisi une ACP car elle simple à mettre en œuvre contrairement à l'ACI pour laquelle il n'existe pas d'algorithme universel.

L'ACP est donc appliquée sur les trois approches utilisées pour décrire les données de la langue afin d'en fournir une description plus compacte.

Comme nous l'avons vu dans le paragraphe 4.2.2.2, le mouvement de la mâchoire doit être soustrait des données de la langue. La première possibilité consiste à soustraire la corrélation entre les données de la mâchoire et celle de la langue. La seconde possibilité soustrait géométriquement le mouvement de la mâchoire du contour de la langue en appliquant le mouvement inverse au données de la langue.

L'analyse factorielle de la mâchoire utilisée est celle présentée dans le paragraphe 4.2.2.1. Trois stratégies ont été envisagées pour réaliser l'analyse factorielle de la langue.

Stratégie 1 (Indépendance) : La première stratégie consiste à traiter indépendamment la mâchoire et la langue. L'ACP est réalisée sur les données de la langue auxquelles nous avons retiré géométriquement le mouvement de la mâchoire (en utilisant les paramètres de mouvement). Ainsi, les données utilisent le centre réel (c'est-à-dire celui calculé à partir du mouvement réel de la mâchoire). Cette stratégie suppose qu'il n'existe plus d'influence de la mâchoire sur la langue. Les composantes de la mâchoire déplacent le contour de la langue sans en changer la forme.

Stratégie 2 (Cascade) : La deuxième stratégie utilise l'analyse effectuée sur la mâchoire. Le mouvement de la mâchoire est estimé à partir de la première composante extraite de l'ACP sur les données de la mâchoire. Le centre de la grille est ainsi estimé à partir de ce mouvement. Ensuite, les données de la langue (suivant les trois approches) sont calculées avec une grille polaire adaptative utilisant le nouveau centre. Finalement, une ACP est appliquée à ces données.

Stratégie 3 (Cascade et corrélation) : La dernière stratégie est similaire à la précédente sauf que la corrélation entre la mâchoire (représentée par la première composante extraite de l'ACP) et les données de la langue est soustraite des données de la langue. Cette stratégie est très similaire à celle adoptée par Maeda [Mae90] pour supprimer l'influence de la mâchoire sur les données de la langue.

4.2.3.3 Choix des composantes de la langue

Les composantes obtenues en fonction des différentes approches de présentation des données et des différentes stratégies d'analyse sont présentées en annexe A. Les nomogrammes résultants des analysessont présentées dans les figures A.1, A.2, A.3, A.4, A.5, A.6, A.7, A.8 et A.9. Les paramètres varient entre -3 et 3 écarts-types. Ce choix d'intervalle provient de l'inégalité de Bienaymé-Tchebychev :

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{pour tout } k \text{ réel strictement positif} \tag{4.2}$$

pour X une variable aléatoire d'espérance μ et de variance finie σ^2. Cela signifie que pour $k = 3$ par exemple la proportion de valeurs de la distribution se situant dans l'intervalle $[-3\sigma, 3\sigma]$ est d'au moins 88,8%. Nous avons choisi de faire varier les paramètres entre -3 et 3 écarts-types ce qui nous garantit la couverture d'au moins 88,8% des données.

Dans tous les cas, la première composante correspond à un mouvement d'aplatissement/arrondissement de la langue. On observe dans les approches 1 et 3 (Coordonnées euclidiennes et polaires) que ce mouvement s'effectue verticalement ; la langue s'arrondit vers le haut et s'aplatit vers le bas (voir figures A.1(b), A.2(b), A.3(b), A.7(b), A.8(b) et A.9(b)). Alors que pour l'approche 2 (Distances et angles extrêmes), l'aplatissement s'accompagne d'un recul de la langue vers l'arrière (voir figures A.4(b), A.5(b) et A.6(b)).

La deuxième composante se comporte différemment en fonction des différentes approches. Les approches 1 et 3 (Coordonnées euclidiennes et polaires) associées à la stratégie d'analyse séparant la langue et la mâchoire fournissent un mouvement haut/bas (voir figures A.3(c) et A.9(c)). Alors que les stratégies d'analyse soustrayant la corrélation conduisent à un mouvement avant/arrière (voir figures A.1(c), A.2(c), A.7(c) et A.8(c)). Dans le cas de l'approche 2 utilisant les distances et les angles extrêmes, le mouvement correspond à un aplatissement vers l'avant et un arrondissement vers l'arrière de la langue (voir figures A.4(c), A.5(c) et A.6(c)).

Comme pour la deuxième composante, on peut faire un rapprochement pour la troisième composante entre les approches 1 et 3 associées à la stratégie d'indépendance. Elle correspond à un mouvement latéral avant/arrière (voir figures A.3(d) et A.9(d)). Pour les autres stratégies associées

aux approches 1 et 3 le mouvement s'oppose à la deuxième composante. Le mouvement de la troisième composante peut être qualifié d'arrondissement arrière vers le haut jusqu'à un léger aplatissement vers l'avant (voir figures A.1(d), A.2(d), A.7(d) et A.8(d)). Pour l'approche 2, les déformations de la troisième composante concernent principalement l'apex quelle que soit la stratégie d'analyse utilisée (voir figures A.4(d), A.5(d) et A.6(d)).

Les composantes quatre, cinq et six représentent des déformations plus mineures avec dans certains cas une influence sur l'apex de la langue.

L'observation des composantes principales extraites des différentes analyses nous montre qu'il existe quelques disparités et aussi des similitudes. Les approches 1 et 3 (Coordonnées euclidiennes et polaires) présentent des similitudes alors que l'approche 2 (distances associées aux angles extrêmes) a un comportement complètement différent. Les deux stratégies d'analyse utilisant l'estimation du mouvement de la mâchoire à partir du paramètre associé produisent des résultats similaires pour chaque approche.

Le tableau 4.2 présente les pourcentages de chaque composante et les pourcentages cumulés en fonction des différentes approches et stratégies utilisées. Dans presque toutes les situations, six composantes permettent d'expliquer environ 99% de la variance. Le maximum de variance expliquée avec six composantes est obtenu par l'approche 3 (Coordonnées polaires) avec la stratégie 2 (Cascade). Avec quatre et cinq composantes, c'est l'approche 2 (Distances et angles extrêmes) avec la stratégie 2 (Cascade) qui explique le plus de variance. 98,53% de la variance est expliquée avec cinq composantes et 97,57% avec quatre composantes.

On constate que les performances sont proches (voir tableau 4.2) et le choix des composantes n'est pas simple. En effet, notre objectif est de construire un modèle articulatoire avec un petit nombre de paramètres. Nous avons décidé de limiter le nombre de composantes pour la langue à quatre. Pour quatre composantes, la variance expliquée est toujours supérieure à 95%, c'est l'approche 2, utilisant les distances et les angles extrêmes, qui permet d'expliquer le plus de variance quel que soit l'analyse effectuée. La combinaison qui explique le plus de variance avec quatre composantes est l'approche 2 utilisant les distances et les angles extrêmes associée à la stratégie 2 (Cascade) (97.57% de variance expliquée voir le tableau 4.1(b)). Il n'y a pas de différence significative entre quatre, cinq et six composantes car les cinquième et sixièmes composantes expliquent peu de variance.

Nous avons choisi de réaliser une ACP afin d'obtenir un petit nombre de paramètres, mais nous n'avons pas la garantie d'obtenir des composantes physiquement interprétables. Notre choix va donc être orienté par la possibilité ou non d'interpréter les composantes obtenues. Pour chaque approche de présentation des données, on remarque que les composantes présentent des similarités quel que soit l'analyse effectuée. Dans tous les cas les premières composantes ont un comportement interprétable comme étant un mouvement haut/bas, arrondissement/aplatissement ou encore avant/arrière. On remarque que seule l'approche 2 (distances et angles extrêmes) possède une composante qui contrôle le mouvement de l'apex seul. Si on le souhaitait, l'apex aurait pu être séparé dans l'analyse pour mieux faire apparaître sa contribution. Notre choix porte donc sur l'approche 2. Les trois analyses factorielles conduisent à des composantes très similaires, nous avons décidé de prendre celle qui explique le plus de variance.

(a) Approche 1 (Coordonnées euclidiennes)

Composantes	Stratégie 1		Stratégie 2		Stratégie 3	
1	47,67 %	47,67 %	46,12 %	46,12 %	50,08 %	50,08 %
2	26,10 %	73,77 %	28,84 %	74,96 %	24,38 %	74,46 %
3	16,58 %	90,35 %	16,59 %	91,55 %	16,03 %	90,49 %
4	5,28 %	95,63 %	4,02 %	95,57 %	4,52 %	95,01 %
5	1,70 %	97,33 %	1,92 %	97,49 %	2,18 %	97,19 %
6	1,64 %	98,97 %	1,58 %	99,07 %	1,75 %	98,94 %

(b) Approche 2 (Distances et angles extrêmes)

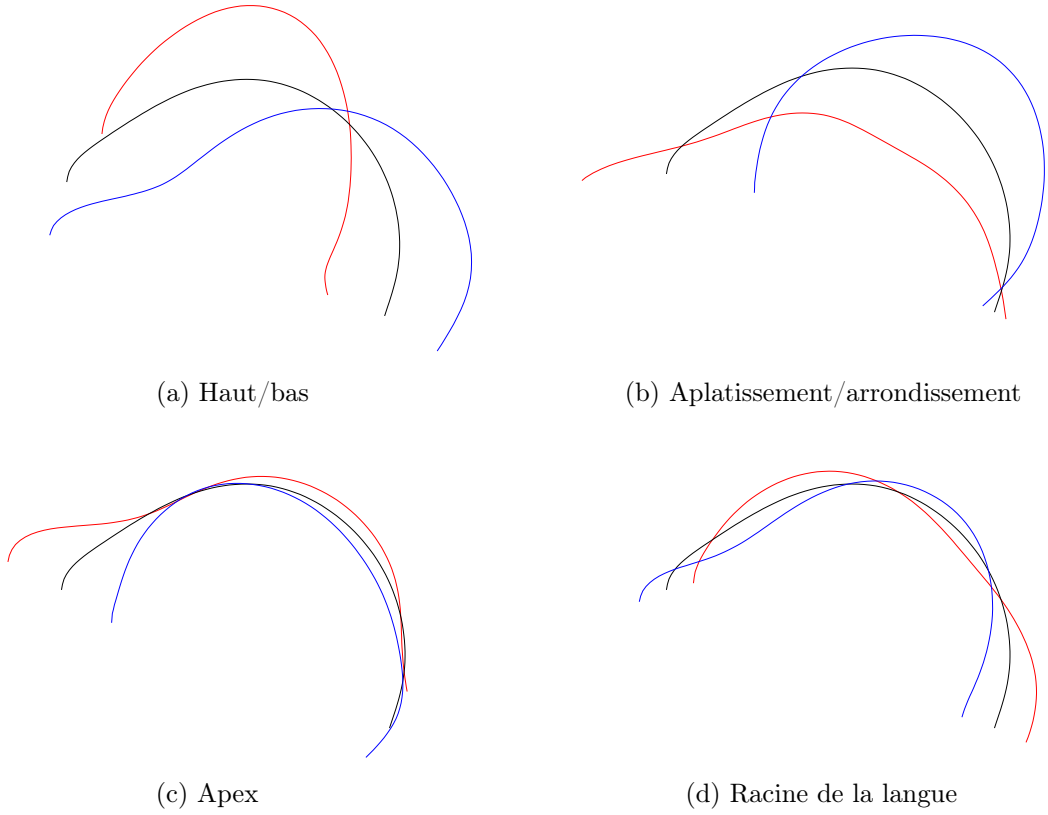
Composantes	Stratégie 1		Stratégie 2		Stratégie 3	
1	48,02 %	48,02 %	53,26 %	53,26 %	52,67 %	52,67 %
2	31,63 %	79,65 %	32,16 %	85,42 %	29,86 %	82,53 %
3	12,86 %	92,51 %	7,53 %	92,95 %	9,70 %	92,23 %
4	5,03 %	97,54 %	4,62 %	**97,57 %**	5,03 %	97,26 %
5	0,96 %	98,50 %	0,96 %	**98,53 %**	1,10 %	98,36 %
6	0,51 %	99,01 %	0,56 %	99,09 %	0,51 %	99,00%

(c) Approche 3 (Coordonnées polaires)

Composantes	Stratégie 1		Stratégie 2		Stratégie 3	
1	47,67 %	47,67 %	47,99 %	47,99 %	47,67 %	47,67 %
2	25,74 %	73,41 %	29,05 %	77,04 %	25,74 %	73,41 %
3	16,54 %	89,95 %	15,68 %	92,72 %	16,54 %	89,95 %
4	5,84 %	95,79 %	3,30 %	96,02 %	5,84 %	95,79 %
5	2,11 %	97,90 %	1,98 %	98,00 %	2,11 %	97,90 %
6	1,19 %	99,09 %	1,19 %	**99,19 %**	1,19 %	99,09 %

TABLEAU 4.2 – *Pourcentages de variance expliquée pour les données de la langue en fonction des différentes approches et stratégies d'analyse factorielle.*

Les composantes retenues ont l'avantage de représenter chacune un mouvement différent. En effet, la première composante correspond à un mouvement haut/bas, la deuxième un mouvement d'aplatissement/arrondissement, la troisième correspond au mouvement de l'apex et la dernière est un mouvement de la racine de la langue (voir figure 4.11).



(a) Haut/bas

(b) Aplatissement/arrondissement

(c) Apex

(d) Racine de la langue

FIGURE 4.11 – *Les quatre paramètres de la langue issus de l'approche 2 (distances et angles extrêmes) avec la stratégie d'analyse 2 (cascade). Chaque paramètre varie entre -3 et +3 écarts-types. La forme noire correspond à la forme neutre, la rouge à -3 écarts-types et la bleue à +3 écarts-types.*

4.2.4 Les lèvres

4.2.4.1 Extraction des données

Les lèvres sont représentées uniquement dans le plan médio-sagittal par les contours des lèvres supérieure et inférieure. Aucune information sur l'ouverture des lèvres n'est disponible car il n'y a pas eu de labiofilms enregistrés en même temps que les images aux rayons X. Dans le plan médio-sagittal, deux paramètres sont calculés à partir des deux contours : la hauteur et la protrusion (voir la figure 4.12).

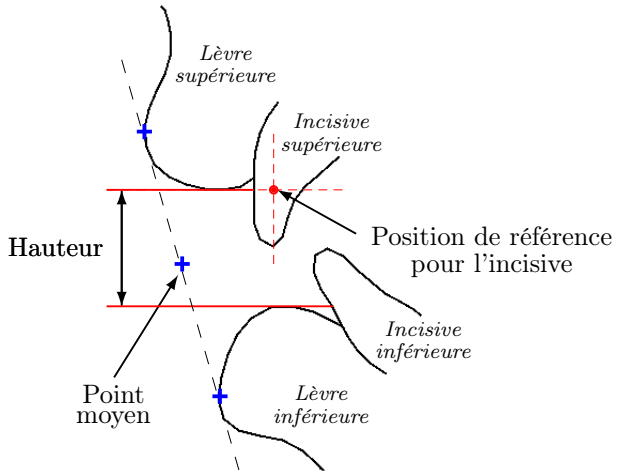


FIGURE 4.12 – *Les paramètres des lèvres.*

La hauteur des lèvres se définit comme la différence entre les ordonnées du point le plus bas de la lèvre supérieure et du point le plus haut de la lèvre inférieure. La protrusion correspond à l'avancement des lèvres. À partir des points les plus en avant des lèvres supérieure et inférieure, on définit un point moyen qui est le centre du segment formé par les deux points cités précédemment. La figure 4.12 montre un exemple de calcul des paramètres pour les lèvres.

4.2.4.2 Analyse

Pour l'analyse, nous avons choisi d'utiliser comme paramètres la hauteur h et l'abscisse x_p du milieu du segment formé par les deux points extrêmes des lèvres.

La mâchoire ayant une influence sur les lèvres, la corrélation avec le premier facteur de la

mâchoire est soustraite de ces données. Deux nouvelles variables sont créées h_n et x_{pn} définies par :

$$\begin{cases} h_n = h - \rho_{(h,M)}.\dfrac{\rho_h}{\rho_M}.M \\ x_{pn} = x_p - \rho_{(x_p,M)}.\dfrac{\rho_{x_p}}{\rho_M}.M \end{cases} \tag{4.3}$$

avec ρ_h, ρ_{x_p} et ρ_M les écarts-types des variables h, x_p et M. $\rho_{(h,M)}$ (resp. $\rho_{(x_p,M)}$) est la corrélation entre le paramètre de hauteur h (resp. le paramètre de protrusion x_p) et le premier facteur de la mâchoire M.

La première idée est d'utiliser les variables indépendamment et de créer deux paramètres pour contrôler la hauteur et la protrusion des lèvres. La hauteur et la protrusion sont alors contrôlées par les paramètres LH et LP par :

$$\begin{cases} h_n = \rho_{h_n}.LH + \overline{h_n} \\ x_{pn} = \rho_{x_{pn}}.LP + \overline{x_{pn}} \end{cases} \tag{4.4}$$

avec $\overline{h_n}$ (resp. $\overline{x_{pn}}$) la moyenne de la variable h_n (resp. x_{pn}) et $\rho^2_{h_n}$ (resp. $\rho^2_{x_{pn}}$) la variance de variable h_n (resp. x_{pn}). Cette approche permet bien sûr d'expliquer 100% de la variance. Cependant nous voulons réduire le nombre de paramètres.

Bien qu'il n'y ait que deux paramètres, nous avons tout de même effectué une ACP sur ces données afin de savoir s'il est possible de supprimer une variable. Le résultat de l'ACP sur ces données montre que la première composante explique 92.7% de la variance. Les lèvres peuvent donc être contrôlées par une seule composante. Contrairement à d'autres approches, celle de Maeda en particulier, une seule composante permet de contrôler la hauteur et la protrusion.

Les lèvres étant représentées par un seul paramètre, on peut donc utiliser un paramètre supplémentaire pour mieux représenter la langue.

4.2.4.3 Reconstruction des lèvres

L'ACP pour les lèvres utilise deux paramètres, la hauteur et l'abscisse d'un point moyen pour la protrusion. Nous allons voir comment nous allons modéliser les lèvres en utilisant ces données. Les lèvres ne sont pas reconstruites comme la langue mais à l'aide de deux segments pour reconstruire la section correspondante (voir figure 4.13). La section est caractérisée par une hauteur et une longueur (paramètre de protrusion). Il faut donc déterminer la position de cette section qui servira de référence pour toutes les images. Ce point de référence est déterminé en fonction d'un point fixe : l'incisive supérieure et de la position moyenne de la lèvre supérieure.

Tout d'abord, nous avons mesuré une position de référence pour l'incisive supérieure (voir figure 4.12). Ce point correspond à un point moyen calculé à partir de toutes les images utilisées pour l'analyse. Son abscisse correspond à la moyenne des abscisses de la position de l'incisive supérieure (voir figure 4.4 où la position est représentée par un cercle jaune) et son ordonnée à la moyenne des ordonnées du point le plus bas de la lèvre supérieure.

La construction de la section correspondant aux lèvres dépend de la position de référence pour l'incisive supérieure. La figure 4.13 montre un exemple de reconstruction des lèvres. Les points bleus sont les points délimitant la section des lèvres. La position de référence pour l'incisive est fixée, les points correspondant à l'extrémité du conduit ont pour abscisse la valeur du paramètre de la protrusion (x_p) ; le point supérieur a la même ordonnée que la position de référence et le point inférieur correspond au point supérieur auquel on a soustrait la hauteur. Le dernier point a la même abscisse que la référence et a pour ordonnée l'ordonnée de la référence moins la hauteur.

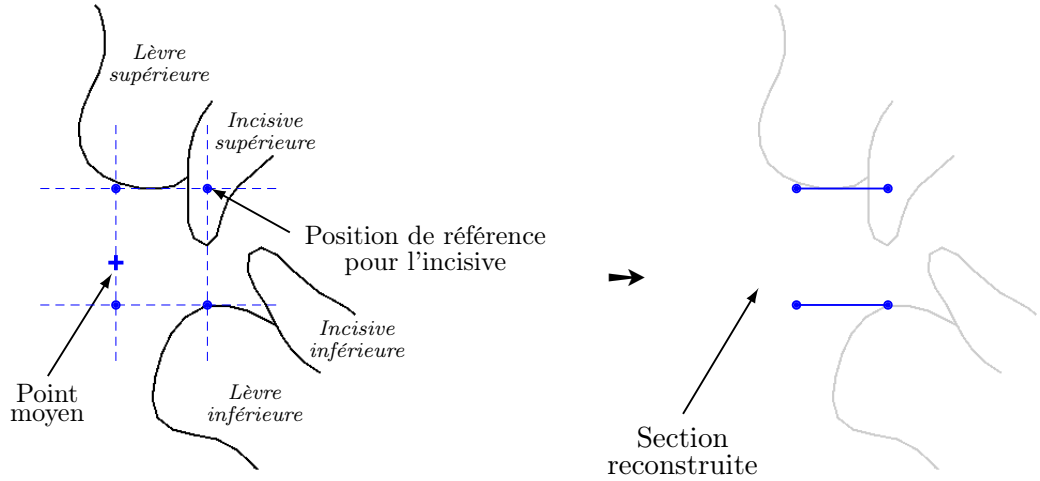
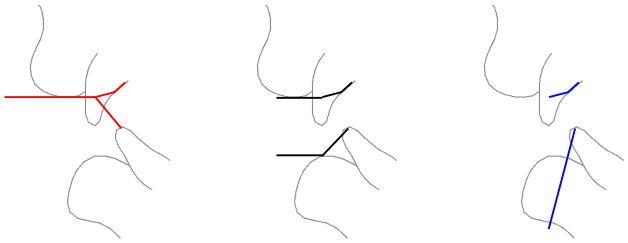
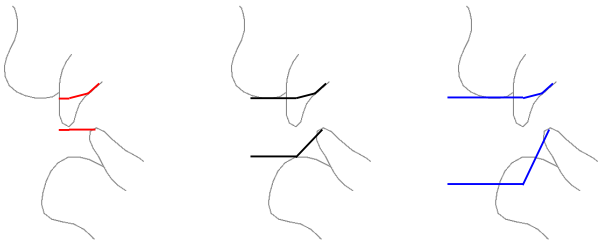


FIGURE 4.13 – *Construction de la section représentant les lèvres. La section est définie par quatre points positionnés par rapport à la position de référence pour l'incisive, de la hauteur et de la protrusion.*

La figure 4.14 montre les deux composantes obtenues par ACP. La première composante représente 92,7% de la variance et la deuxième 7,3%. Pour la première composante, lorsque la hauteur diminue la protrusion augmente et inversement lorsque l'ouverture augmente les lèvres sont moins protruses. La deuxième qui a une contribution beaucoup plus faible que la première, correspond à un mouvement opposé à celui de la première composante.



a. Première composante (92,7%)



b. Seconde composante (7,3%)

FIGURE 4.14 – *Les deux composantes principales issues de l'analyse en composantes principales sur les données des lèvres. La première composante permet de contrôler la hauteur et la protrusion et représente 92,7% de la variance. Les courbes noires correspondent à la position moyenne, les rouges à -3 écarts-types et les bleues à +3 écarts-types.*

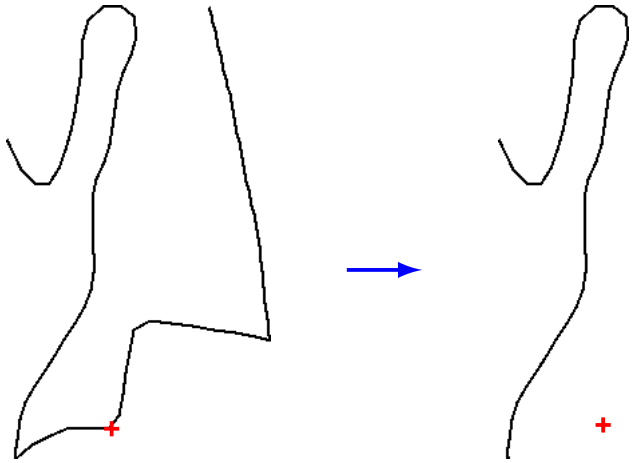
4.2.5 L'épiglotte et le larynx

L'épiglotte et le larynx sont définis par trois contours obtenus par un suivi semi-automatique (voir figure 4.6). Chaque contour est défini avec le même nombre de points de contrôle. Pour notre modèle, nous devons déterminer la position de l'épiglotte ainsi que la position des extrémités de la glotte. Nous avons ainsi uniquement conservé le contour de l'épiglotte et la position extérieure du larynx (voir figure 4.15).

Afin de réduire l'influence de la mâchoire sur l'épiglotte et le larynx, nous avons appliqué le même traitement que pour les lèvres ; la corrélation avec la première composante de la mâchoire a été soustraite. Pour chacune des données Lrx_i, on soustrait la corrélation. Pour chaque variable i, la nouvelle variable $Lrxn_i$ est définie par :

$$Lrxn_i = Lrx_i - \rho_{(Lrx_i,M)}.\frac{\rho_{Lrx_i}}{\rho_M}.M \tag{4.5}$$

La figure 4.16 présente les deux premières composantes principales issues de l'ACP. Bien que la première composante ne représente uniquement que 47,41% de la variance, nous avons choisi de conserver uniquement cette composante car ce n'est pas un des articulateurs les plus importants.



Trois contours orginaux

Données utilisées pour l'analyse

FIGURE 4.15 – *Représentation des données de l'épiglotte et du larynx utilisées pour l'ACP. À gauche les trois contours obtenus par suivi automatique et à droite le contour de l'épiglotte et la position extérieure du larynx (caractérisée par la croix rouge).*

a. Première composante (47,41%)

b. Seconde composante (36,85%)

FIGURE 4.16 – *Deux premières composantes issues de l'ACP pour l'épiglotte et le larynx. Les courbes noires correspondent à la position moyenne, les rouges à -3 écarts-types et les bleues à +3 écarts-types.*

4.2.6 Assemblage du modèle articulatoire

A cette étape, tous les paramètres de contrôle du modèle ont été calculés. Il faut maintenant assembler les différents modes de déformation afin de constituer un modèle articulatoire.

Le modèle doit produire un contour sagittal continu : or lorsque la langue s'arrondit il n'existe pas de contour entre les dents et la langue. Pour compléter le contour, nous avons utilisé le contour du plancher de la bouche. La figure 4.17 montre un exemple où le plancher de la bouche doit être utilisé. Pour cela, on calcule l'intersection entre le contour de la langue et le plancher si elle existe sinon les deux contours sont joints. Pour le palais, nous avons tracé un contour sur l'image de référence (voir figure 4.17).

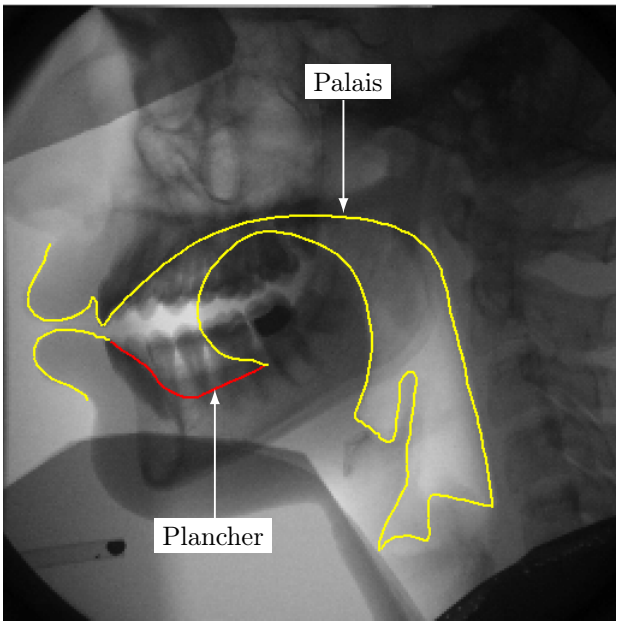


FIGURE 4.17 – *Le contour du plancher de la langue (en rouge) est utilisé pour compléter le contour intérieur lorsque la langue est reculée.*

A ce stade, nous avons un modèle articulatoire possédant sept paramètres : un pour l'ouverture de la mâchoire, quatre pour la forme et la position de la langue, un pour l'ouverture et la protrusion des lèvres et un pour l'épiglotte et le larynx. La prochaine étape est de calculer une fonction d'aire (voir §4.3) associée à la forme du conduit vocal. Le conduit vocal doit être découpé en sections. L'utilisation d'une grille semi-polaire comme Maeda [Mae82] permet de conserver un nombre constant de sections (voir figure 4.18).

Le système de coordonnées semi-polaires est formé d'une partie polaire et de deux parties linéaires. La grille doit être placée correctement. Bien que certains travaux proposent de déterminer la position du centre par rapport à des repères anatomiques visibles sur les images [JM08], la positionnement du centre est quelque peu arbitraire. Nous avons placé notre grille de façon à ce que les contours intersectent tous la grille, c'est-à-dire que tous les contours soient recouverts par la grille. Un autre critère est le fait que les traits de la grille soient parallèles au front d'onde. Le centre de la

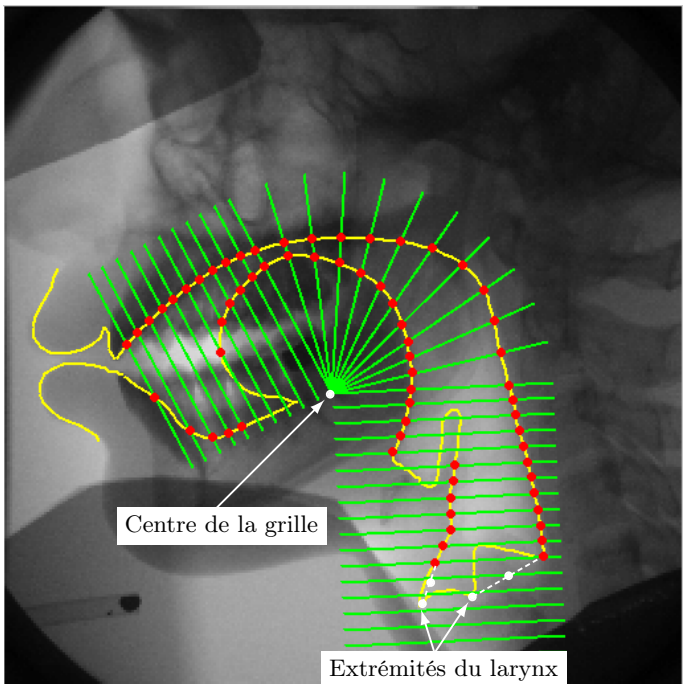


FIGURE 4.18 – *Intersections des contours intérieur et extérieur avec la grille semi-polaire. Les disques rouges correspondent à ces intersections. Les extrémités du larynx sont représentées. Deux points sont ajoutés entre les extrémités du larynx et les dernières intersections avec la grille.*

grille est ainsi défini de façon ad hoc sur l'image de référence utilisée pour les suivis automatiques. La figure 4.18 montre la grille semi-polaire et ces intersections avec les contours du conduit vocal.

Les coordonnées de la grille sont calculées en fonction de la région du conduit vocal considérée. La région linéaire du pharynx est composée de m_1 lignes, la région vélaire de m_2 lignes et la région palato-dentale de m_3 lignes. Les coordonnées de la grille sont données par les deux vecteurs de coordonnées *igd* pour la partie intérieure et *egd* pour la partie extérieure. Deux angles Ω et θ permettent de contrôler l'orientation des deux cavités (voir figure 4.19).

Les coordonnées linéaires de la grille dans la région du pharynx sont définies pour $1 \leq i \leq m_1$ par :

$$\begin{cases} x_{igd(i)} = s_P \times d_1 x_i \times (m_1 - i) \\ y_{igd(i)} = s_P \times d_1 y_i \times (m_1 - i) \\ x_{egd(i)} = s_P \times d_1 x_e + x_{igd(i)} \\ y_{egd(i)} = s_P \times d_1 y_e + y_{igd(i)} \end{cases} \quad \text{avec} \quad \begin{cases} d_1 x_i = d_l \times \cos\left(\Omega - \dfrac{\pi}{2}\right) \\ d_1 y_i = d_l \times \sin\left(\Omega - \dfrac{\pi}{2}\right) \\ d_1 x_e = r \times \cos(\Omega) \\ d_1 y_e = r \times \sin(\Omega) \end{cases} \tag{4.6}$$

où r est la longueur des traits de la grille et s_P le coefficient d'élongation pour la cavité pharyngale.

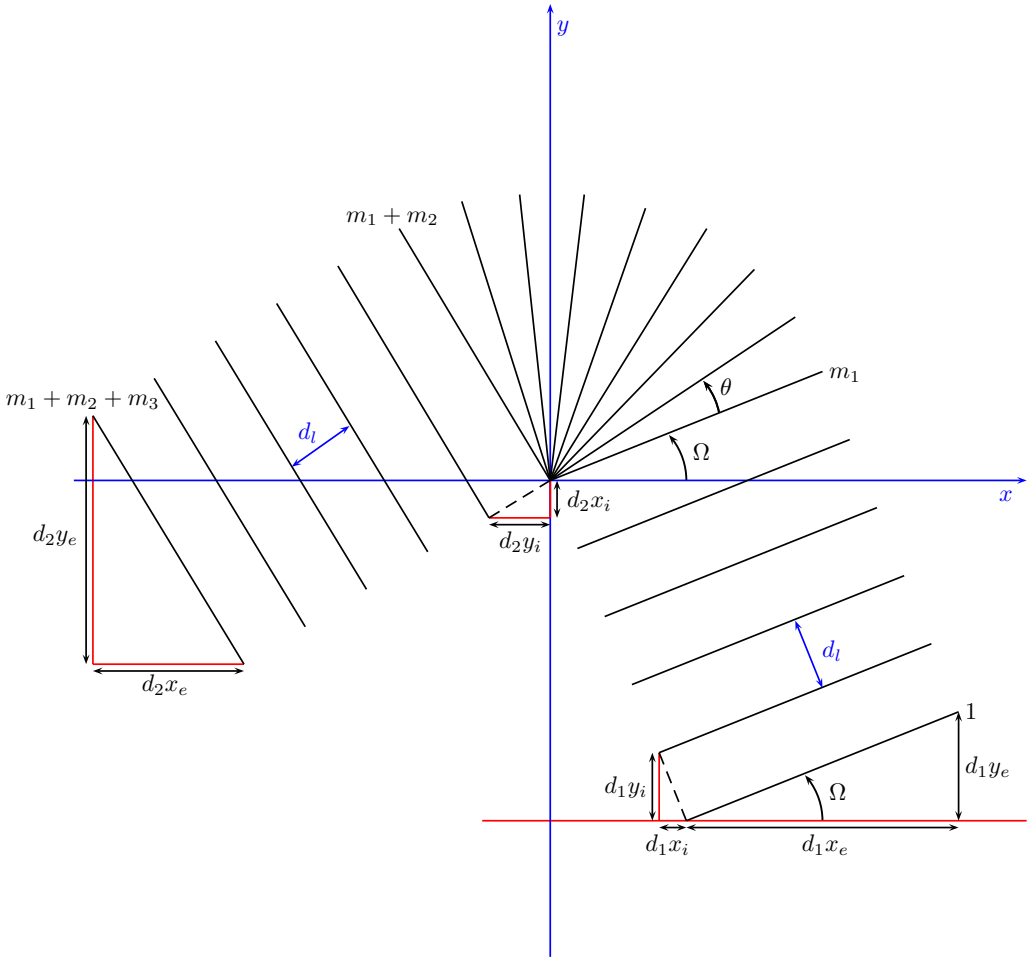


FIGURE 4.19 – *Grille semi-polaire avec les différents paramètres.*

Les coordonnées polaires de la grille dans la région vélaire sont définies pour $m_1 < i \leq m_1 + m_2$ par :

$$\begin{cases} x_{igd(i)} = 0 \\ y_{igd(i)} = 0 \\ x_{egd(i)} = s(i) \times r \times \cos(\Gamma_i) \\ y_{egd(i)} = s(i) \times r \times \sin(\Gamma_i) \end{cases} \text{ avec } \begin{cases} s(i) = (s_B - s_P) \times \dfrac{i - m_1}{m_2} + s_P \\ \Gamma_i = \theta \times (i - m_1) + \Omega \end{cases} \tag{4.7}$$

où d_l est l'espacement entre les lignes, r la longueur des traits de la grille, s_P le coefficient d'élongation pour la cavité pharyngale et s_B le coefficient d'échelle pour la cavité buccale.

Les coordonnées linéaires de la grille dans la région palato-dentale sont définies pour $m_1 + m_2 < i \leq m_1 + m_2 + m_3$ par :

$$\begin{cases} x_{igd(i)} = d_2x_i \times (i - m_1 - m_2) \\ y_{igd(i)} = d_2y_i \times (i - m_1 - m_2) \\ x_{egd(i)} = d_2x_e + x_{igd(i)} \\ y_{egd(i)} = d_2y_e + y_{igd(i)} \end{cases} \text{ avec } \begin{cases} \Gamma = \theta \times m_2 + \Omega \\ d_2x_i = s_B \times d_l \times \cos\left(\Gamma - \dfrac{\pi}{2}\right) \\ d_2y_i = s_B \times d_l \times \sin\left(\Gamma - \dfrac{\pi}{2}\right) \\ d_2x_e = s_B \times r \times \cos(\Gamma) \\ d_2y_e = s_B \times r \times \sin(\Gamma) \end{cases} \tag{4.8}$$

où d_l est l'espacement entre les lignes, r la longueur des traits de la grille et s_B le coefficient d'échelle pour la cavité buccale.

La partie linéaire de la grille correspondant au pharynx-larynx est composée de 19 lignes parallèles espacées de 0,36 cm et la partie linéaire avant de la cavité buccale de 10 lignes parallèles. La partie polaire de la grille est composée de 11 lignes. Les parties linéaires de la grille sont positionnées par rapport aux angles θ pour la partie avant et Ω pour la partie du larynx. θ et Ω sont définis sur la figure 4.20. La partie polaire se calcule à partir de ces deux angles.

Les contours intérieur et extérieur ne sont pas réguliers, de plus plusieurs intersections du contour avec la même ligne sont possibles. La figure 4.18 montre un exemple d'intersections avec la grille semi-polaire. Lorsque la langue est arrondie et reculée, une ligne de la grille peut intersecter plusieurs fois le contour intérieur. Dans ce cas, l'intersection la plus proche du contour extérieur est choisie. De plus, lorsque plusieurs intersections sont possibles au niveau de l'épiglotte, l'intersection avec la langue est choisie. Sinon, l'intersection la plus proche du contour extérieur est choisie dans le cas de l'épiglotte. Lorsque la langue entre en contact avec le palais, le point pris en compte pour représenter la langue est l'intersection de la ligne de la grille avec le palais. C'est une solution simple pour réaliser le « clipping » de la langue.

Les lèvres sont représentées par une section définie par quatre points (voir section 4.2.4.3). L'extrémité du larynx est définie par un point intérieur et un extérieur. Deux sections sont définies à partir des extrémités et de la fin de l'épiglotte (voir figure 4.18). Le tube laryngé étroit n'étant pas toujours visible sur les images, il est peut être pas assez long.

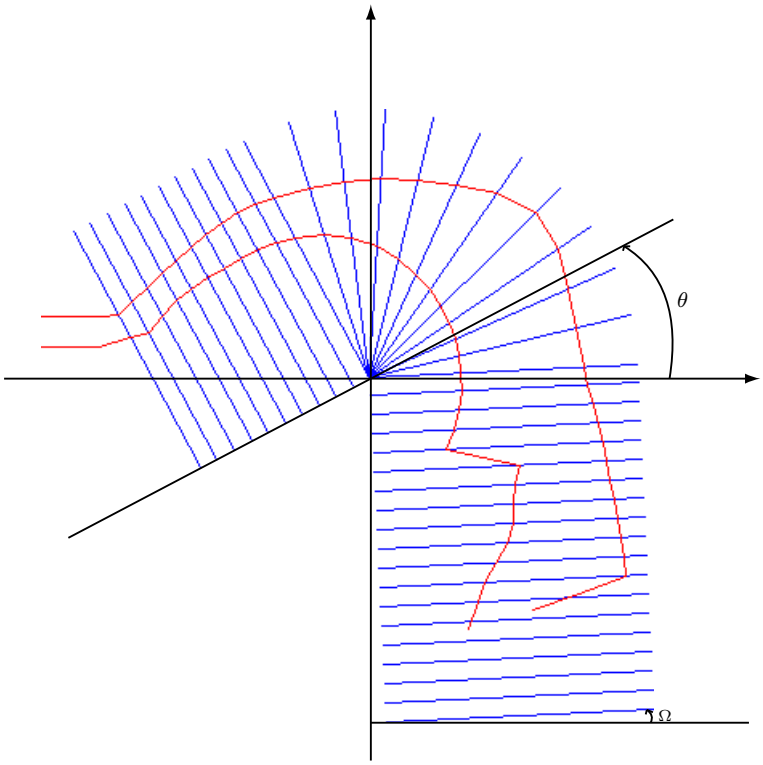
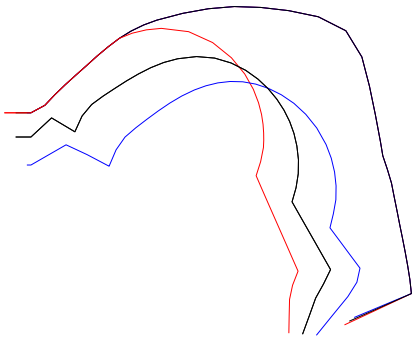
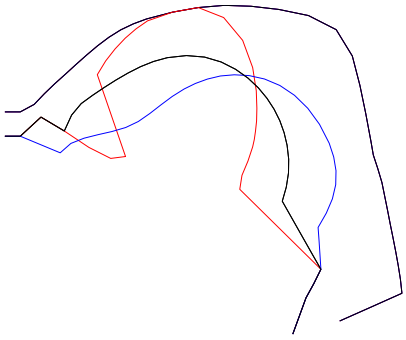


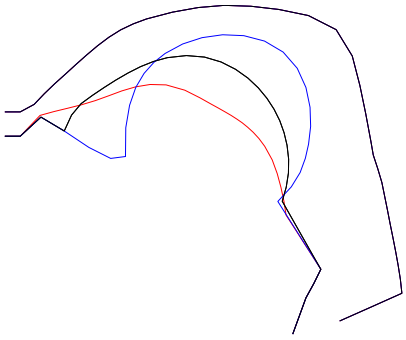
FIGURE 4.20 – *Modèle articulatoire représenté avec la grille semi-polaire. Les contours rouges correspondent aux contours intérieur et extérieur du conduit vocal. La grille semi-polaire est formée de m_1 lignes pour la partie linéaire inférieure, m_2 lignes pour la partie polaire et m_3 lignes pour la partie linéaire antérieure. Les angles θ et Ω définissent la position des différentes parties de la grille.*



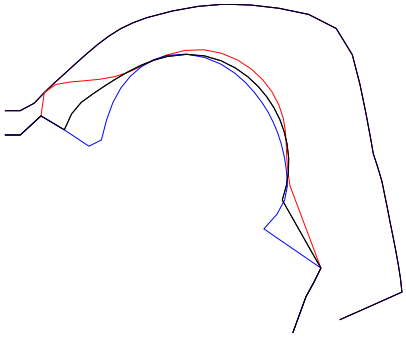
(a) Mâchoire



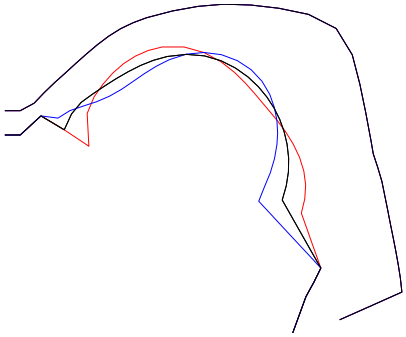
(b) Langue 1



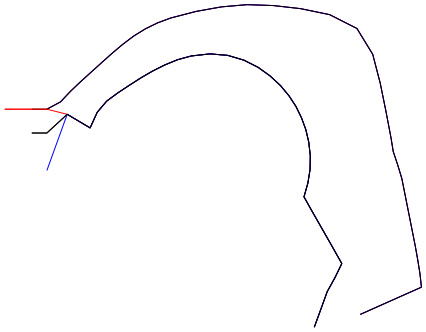
(c) Langue 2



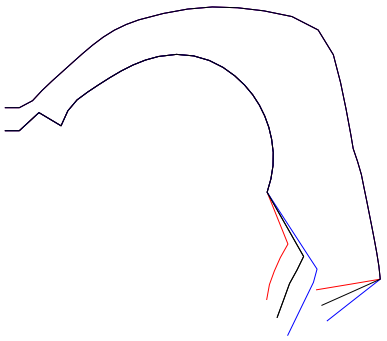
(d) Langue 3



(e) Langue 4



(f) Lèvres



(g) Epiglotte

FIGURE 4.21 – *Les sept paramètres du modèle articulatoire. Chaque paramètre varie dans l'intervalle* $[-3; 3]$.

Le modèle articulatoire est maintenant défini par rapport à la grille semi-polaire (voir figure 4.20). Chaque contour intersecte chaque ligne de la grille au maximum une fois. Les lèvres et l'extrémité du larynx n'intersectent pas la grille. Les différents paramètres du modèle sont présentés par la figure 4.21.

4.3 Synthèse articulatoire

Les contours médio-sagittaux seuls ne permettent pas de calculer les paramètres acoustiques correspondants. Les modèles acoustiques se basent généralement sur un conduit vocal représenté par une concaténation de tubes acoustiques qui permet de réaliser une simulation acoustique dans le domaine temporel ou fréquentiel afin d'obtenir un signal de parole ([Mae82], [BF84]).

La première étape consiste alors à passer du contour sagittal à la fonction d'aire correspondante. Puis la seconde étape est la synthèse acoustique.

4.3.1 Passage à la fonction d'aire

Le conduit vocal doit être représenté par une fonction d'aire. En pratique il est nécessaire de discrétiser la fonction d'aire. Elle est donnée par l'aire de tuyaux élémentaires issus du découpage du conduit vocal en différentes sections caractérisées par la longueur de la section et l'aire transversale.

Tout d'abord, il faut découper le conduit vocal en sections. Le conduit vocal est représenté dans le plan sagittal par deux contours : le contour « intérieur » (composé de la langue, de l'épiglotte, de la mâchoire et de la lèvre inférieure) et le contour « extérieur » (composé du palais dur, du voile du palais et de la partie postérieure du pharynx). Le découpage du conduit vocal en sections se fait à l'aide d'une grille semi-polaire (voir figure 4.18).

Le conduit est maintenant découpé en sections ; la longueur et l'aire transversale de chaque section peuvent donc être calculées. La fonction d'aire est estimée à partir de la transformation proposée par Heinz et Stevens [HS65], où l'aire transversale de chaque section, est calculée par la formule :

$$A(x) = \alpha(x).d(x)^{\beta(x)} \tag{4.9}$$

avec d la distance sagittale.

La première étape est de déterminer la longueur des sections et la distance sagittale. Pour chaque ligne de la grille, on calcule le milieu du segment formé par les intersections de la grille avec les contours intérieur et extérieur. La longueur de la section correspond à la distance entre les milieux des segments pour les lignes de la grille définissant cette section. (voir figure 4.22). La distance sagittale d_k de la section k est obtenue par :

$$d_k = \frac{B_k}{l_k} \tag{4.10}$$

avec l_k la longueur de la section et B_k l'aire du quadrilatère défini par les quatre intersections du contour avec les lignes délimitant la section (voir figure 4.22).

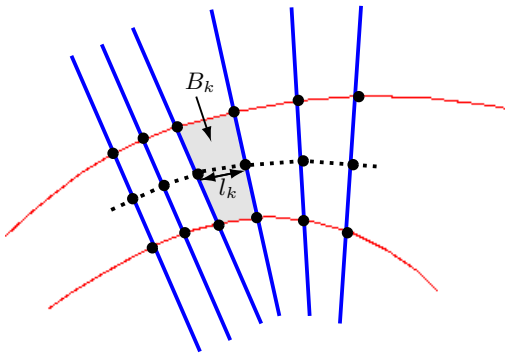


FIGURE 4.22 – *Détermination de la longueur l_k de la section k et de la distance sagittale. B_k est l'aire de la partie grise de la section k.*

L'utilisation d'un modèle de passage de la coupe sagittale à la fonction basée sur une transformation du type $A(x) = \alpha(x).d(x)^{\beta(x)}$ implique la détermination des paramètres α et β. De nombreux travaux ont été consacrés à la détermination de valeurs de α et β optimales, le dernier en date est celui de McGowan, Jackson et Berger [MTTB12]. Même si le calcul de α et β dépend de la forme globale, le bénéfice obtenu n'est pas considérable. Qui plus est une étude de Ericsdotter [Eri07] montre que la troisième dimension joue un rôle finalement assez limité. Nous avons donc retenu le choix de Soquet et al. [SLMD02]. Les paramètres α et β ont été estimés à partir de coupes obtenues par IRM pour les voyelles du français. La valeur de ces paramètres varient en fonction de la position dans le conduit vocal. Huit régions, représentées dans la figure 4.23, ont été définies dans le conduit vocal :

- le larynx,
- le pharynx bas,
- le pharynx moyen,
- l'oropharynx,
- le vélum,
- le palais dur,
- la zone alvéolaire,
- la zone labiale.

Le tableau 4.3 présente les différentes valeurs de α et β obtenus par Soquet et al. [SLMD02] pour un locuteur masculin en fonction des huit zones qui partagent le conduit vocal.

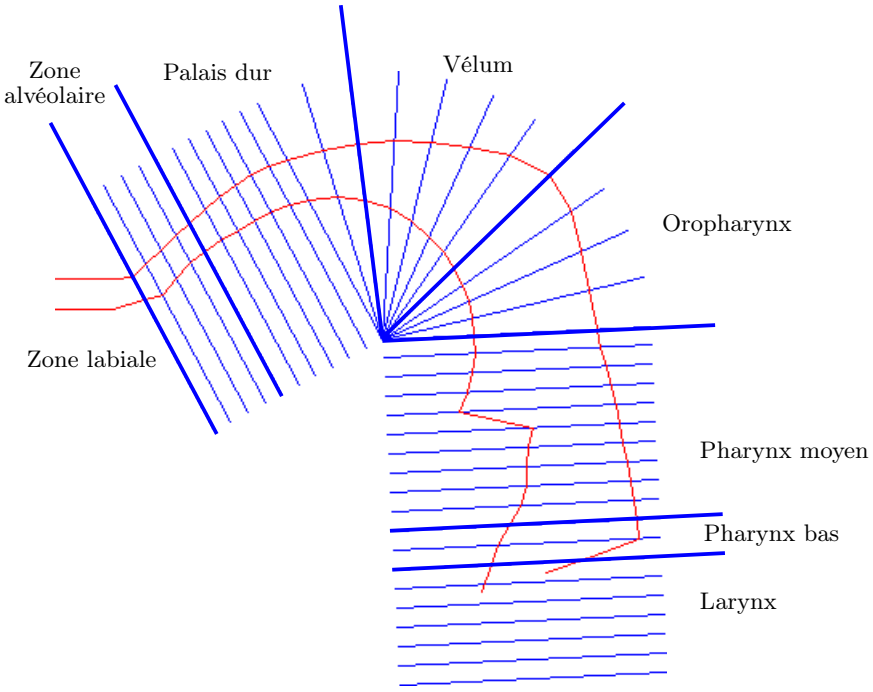


FIGURE 4.23 – *Détermination des huit régions du conduit sur le modèle articulatoire.*

Zones	α	β
Larynx	1,11	2,35
Pharynx bas	1,79	1,38
Pharynx moyen	1,34	1,62
Oropharynx	0,73	1,81
Vélum	1,39	1,08
Palais dur	1,34	1,51
Zone alvéolaire	1,92	1,20
Zone labiale	4,72	2,48

TABLEAU 4.3 – *Paramètres alpha et β obtenus par [SLMD02] pour un locuteur masculin en fonction des huit régions du conduit vocal.*

La figure 4.24 montre un exemple de passage de la coupe sagittale à la fonction d'aire.

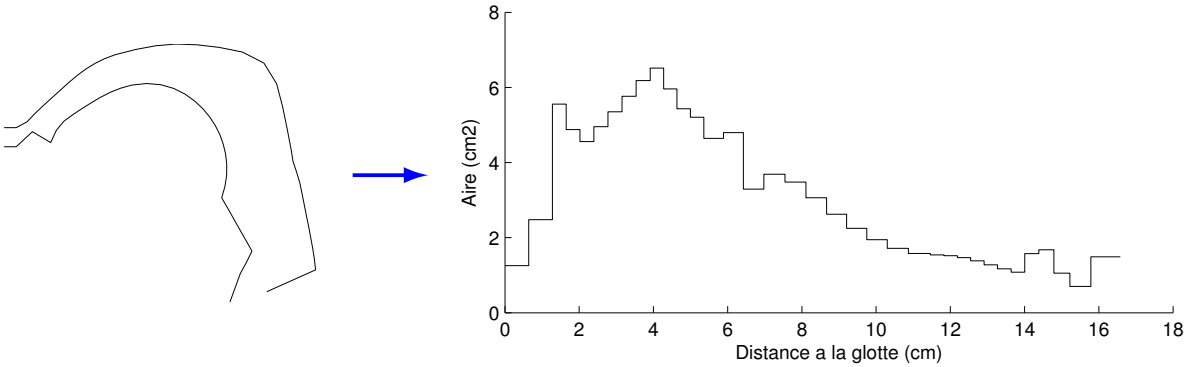


FIGURE 4.24 – *Exemple de passage de la coupe sagittale à la fonction. La figure de gauche représente une vue sagittale du conduit vocal et la figure de droite correspond à la fonction d'aire correspondante.*

4.3.2 Simulation acoustique

Nous ainsi construit un modèle articulatoire à partir d'images cinéradiographiques d'un même locuteur. Ce modèle comporte un petit nombre de paramètres de contrôle permettant d'obtenir une vue sagittale du conduit vocal. Chaque paramètre varie dans l'intervalle $[-3; 3]$. Les différents paramètres sont représentés sur la figure 4.21 :

- un paramètre pour le déplacement de la mâchoire,
- quatre paramètres pour la position et la forme de la langue,
- un paramètre pour la hauteur et la protrusion des lèvres,
- un pour le larynx et l'épiglotte.

Nous avons vu dans la section précédente le passage de la coupe sagittale à la fonction d'aire à partir d'un modèle basée sur une transformation du type $A(x) = \alpha(x).d(x)^{\beta(x)}$. Il est possible de réaliser une simulation acoustique (cf §3.4) en résolvant les équations de l'acoustique qui régisssent la propagation de l'air dans un conduit.

Ce nouveau modèle est associé à une synthèse articulatoire développée par Maeda ([Mae82]). La figure 4.25 montre un exemple de synthèse articulatoire. Un signal de parole synthétique est créé à partir d'une forme sagittale donnée. Nous avons donc construit un synthétiseur articulatoire qui permet de passer d'un vecteur articulatoire composé de sept paramètres au spectre synthétique correspondant.

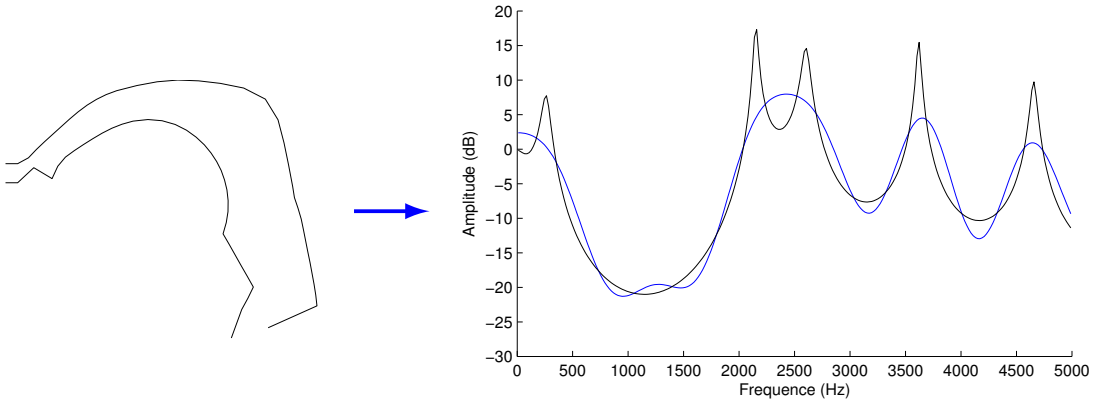


FIGURE 4.25 – *Synthèse articulatoire à partir de la forme sagittale du conduit vocal. La figure de gauche présente une forme de conduit vocal fournie par le modèle. La figure de droite représente les fonctions de transfert acoustique du conduit vocal (la source n'est pas prise en compte) obtenus par synthèse articulatoire : ligne noire (spectre obtenu par simulation acoustique) et ligne bleue (spectre issu d'un lissage cepstral).*

4.4 Conclusion

Dans ce chapitre, nous avons expliqué les différentes stratégies utilisées pour obtenir les positions des articulateurs de la parole. Des suivis automatique par corrélation et semi-automatique ont servi à déterminer le mouvement et le contour d'articulateurs. En revanche le contour de la langue a été tracé à la main pour éviter les erreurs.

Ces informations sur les articulateurs de la parole ont servi à construire un modèle articulatoire. Ce modèle permet à partir d'un petit nombre de paramètres de contrôler la position et la forme des positions des articulateurs de la parole. Les différents paramètres sont obtenus par ACP (Analyse en Composantes Principales) successives. L'utilisation de l'ICA (*Independent Component Analysis*) aurait sans doute été similaire et qu'à résultats équivalents nous avons préféré l'analyse qui fait le moins d'hypothèse. Le modèle obtenu comporte sept paramètres : un pour contrôler l'ouverture de la mâchoire, quatre pour la position et la forme de la langue, un pour la hauteur et la protrusion des lèvres et un pour la hauteur du larynx. Ce modèle articulatoire a été associé à un synthétiseur articulatoire qui permet de générer de la parole synthétique.

Le synthétiseur articulatoire nous permet de passer d'un vecteur de paramètres contrôlant la forme du conduit vocal au son associé. Il représente donc la relation articulatoire-acoustique.

Les chapitres suivants seront consacrés au cheminement inverse, l'inversion acoustique-articulatoire. L'objectif est d'utiliser le synthétiseur articulatoire ainsi construit afin de construire un codebook qui représentera la relation articulatoire-acoustique adaptée à notre locuteur. L'étape de construction de modèle articulatoire est nécessaire afin de se placer dans les meilleures conditions possibles. En effet, le but est que le signal synthétique soit aussi proche que possible du signal réel afin de réaliser l'inversion sur des données. Bien que notre modèle soit spécifiquement construit pour

notre locuteur, il subsiste des différences entre les deux signaux dues aux disparités entre le conduit vocal du locuteur et le modèle et à l'absence de source pour le signal synthétique. Ces différences seront étudiées afin de les compenser dans la partie suivante.

Deuxième partie

Inversion

Chapitre 5

Construction d'un codebook hypercuboïdal

Notre méthode d'inversion utilise une approche d'analyse par synthèse qui nécessite une table composée de couples de vecteurs articulatoires-acoustiques souvent appelée *codebook*. La création du codebook utilise le synthétiseur articulatoire construit dans les chapitres précédents en faisant varier les paramètres du modèle.

Le codebook représente donc la relation articulatoire-acoustique de façon compacte. Sa construction repose sur une exploration récursive de l'espace articulatoire afin de le diviser en petites régions où la relation articulatoire-acoustique est considérée comme linéaire.

La construction du codebook a été étudiée en détail par Potard [Pot08a] en utilisant les formants comme paramètres acoustiques. Notre contribution porte sur l'utilisation de coefficients cepstraux à la place des formants, ce qui donne lieu à des modifications assez importantes.

Dans ce chapitre, nous définirons la structure élémentaire, appelée *hypercuboïde*, qui compose le codebook. La méthode de construction exploite la structure utilisée et permet une exploration exhaustive de l'espace articulatoire. Enfin, nous évaluerons la fidélité du codebook par rapport à la réalité.

5.1 Espaces articulatoire et acoustique

La construction d'un codebook permet d'associer des paramètres articulatoires à des paramètres acoustiques. Le passage des paramètres articulatoires au paramètres acoustiques s'effectue à l'aide d'un synthétiseur articulatoire.

La méthode de construction du codebook ne dépend ni du modèle ni du synthétiseur articulatoires utilisés.

Nous revenons dans cette section, sur le synthétiseur articulatoire utilisé et sur notre paramétrisation acoustique.

5.1.1 Synthétiseur articulatoire

Nous utilisons le modèle articulatoire construit dans la partie précédente à partir de contours extraits d'images cinéradiographiques. Les différents paramètres du modèle ont été obtenus à partir d'analyses factorielles. Nous utiliserons comme espace articulatoire, l'espace des paramètres contrôlant notre modèle. Il se compose de sept paramètres : un pour la mâchoire (ouverture), quatre pour la langue (position, forme), un pour les lèvres (ouverture et protrusion) et un pour la hauteur du larynx. Chaque paramètre exprimé en écart-type varie dans l'intervalle $[-3; 3]$.

Le modèle articulatoire fournit uniquement la forme géométrique du conduit vocal dans le plan médiosagittal. Une étape supplémentaire permet d'estimer la fonction d'aire associée. La fonction de transfert est obtenue à partir de l'estimation de la fonction d'aire. Le synthétiseur utilisé a été développé par Maeda [Mae82] et utilise une méthode de simulation dans le domaine temporel.

5.1.2 Représentations acoustiques

Généralement, les formants sont utilisés pour réaliser l'inversion car ils permettent de bien caractériser les voyelles [Sor92, OL05, Pot08b]. Cependant, l'extraction des formants peut être difficile pour les locuteurs dont la fréquence fondamentale est aiguë et pour les consonnes, ce qui réduit fortement la pertinence de l'inversion.

Il est très difficile de comparer directement un spectre synthétique (obtenu via un synthétiseur articulatoire) avec un spectre calculé sur le signal réel. Le spectre réel doit donc être lissé et l'effet de la source doit être compensé. Nous réaliserons ces opérations dans le domaine cepstral.

5.1.2.1 Lissage cepstral

L'inversion acoustique-articulatoire se focalise sur la contribution du conduit vocal, il est donc important d'annuler les effets de la source dans la représentation spectrale.

Le lissage cepstral décrit dans Rabiner et al. [RS78] a pour principe de séparer les contributions de la source et du conduit vocal du signal de parole. D'après le modèle source-conduit, la parole résulte de la convolution d'une source par le filtre constitué par le conduit vocal et de la radiation aux lèvres (voir la figure 5.1). Le signal de parole $s(n)$ est obtenu par l'équation :

$$s(n) = e(n) * h(n) \tag{5.1}$$

avec $e(n)$ et $h(n)$ les contributions de la source et du conduit vocal respectivement.

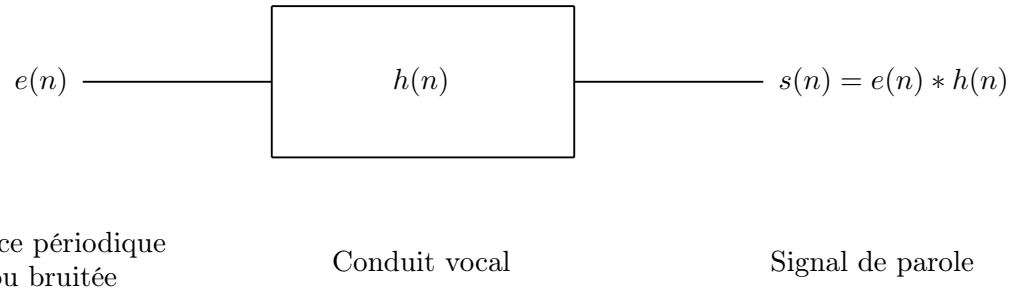


FIGURE 5.1 – *Modèle simplifié de la production de la parole.*

Le principe de l'analyse cepstrale est d'appliquer un traitement homomorphique qui transforme l'opération de convolution en un opérateur plus simple qui sépare clairement les deux contributions. Pour cela, nous appliquons tout d'abord la transformée de Fourier à l'opérateur de convolution, ce qui permet de passer à un produit car la transformée de Fourier d'une convolution est égale au produit de la transformée de Fourier de chaque facteur. D'où :

$$S(\omega) = E(\omega).H(\omega) \tag{5.2}$$

où $S(\omega)$, $E(\omega)$ et $H(\omega)$ les transformées de Fourier de $s(n)$, $e(n)$ et $h(n)$. Ensuite, nous utilisons une propriété de la fonction logarithme : le logarithme d'un produit est égal à la somme des logarithmes de chaque facteur. D'où :

$$\hat{S}(\omega) = \ln\big(S(\omega)\big) = \ln\big(E(\omega)\big) + \ln\big(H(\omega)\big) = \hat{E}(\omega) + \hat{H}(\omega) \tag{5.3}$$

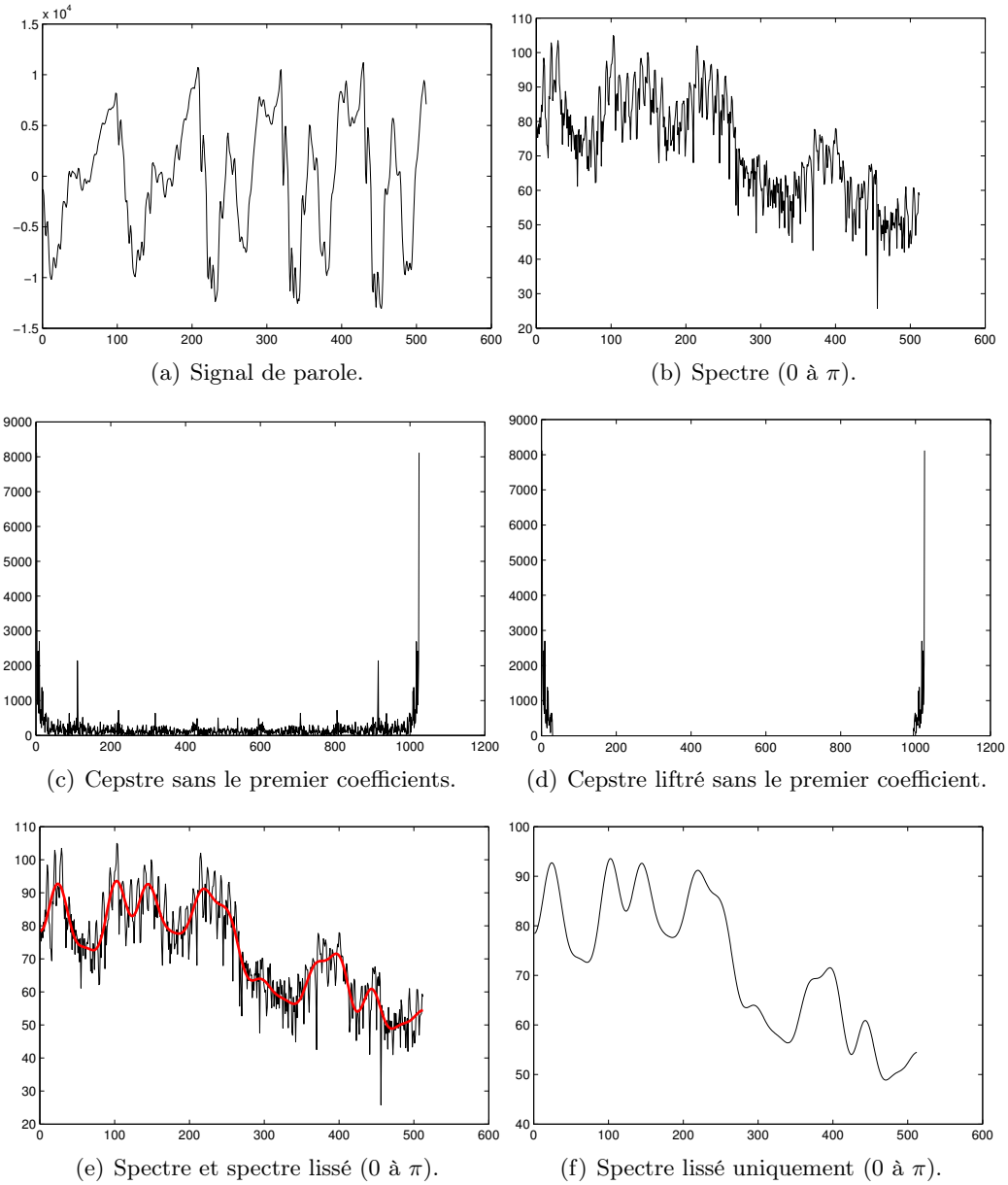
Puis, nous appliquons la transformée de Fourier inverse afin de revenir dans le domaine temporel. Le signal reste additif.

$$\hat{s}(n) = \hat{e}(n) + \hat{h}(n) \tag{5.4}$$

Les éléments du vecteur $\hat{s}(n)$ sont appelés les *coefficients cepstraux*. Les premiers coefficients représentent les variations lentes de la forme du spectre (c'est-à-dire la contribution du conduit vocal) et les derniers coefficients représentent les variations rapides (c'est-à-dire les harmoniques). Un simple filtrage, appelé *liftrage* car il s'effectue sur les coefficients cepstraux, consistant à conserver uniquement les premiers coefficients permet d'isoler la contribution du conduit vocal. Une nouvelle application de la transformée de Fourier fournit un spectre lissé.

La figure 5.2 montre un exemple de lissage cepstral.

En reconnaissance de la parole, les coefficients « mel-cepstraux » (ou MFCC – Mel Frequency Cepstral Coefficient) sont extraits à partir d'une échelle fréquentielle non-linéaire basée sur la perception humaine. Ce calcul n'est pas applicable dans le cas de la synthèse car il est impossible de retrouver l'enveloppe spectrale originale à partir des coefficients mel-cepstraux. Puisque nous souhaitons préserver les pics spectraux correspondant aux formants, nous n'utiliserons pas les MFCC.



(a) Signal de parole.

(b) Spectre (0 à π).

(c) Cepstre sans le premier coefficients.

(d) Cepstre liftré sans le premier coefficient.

(e) Spectre et spectre lissé (0 à π).

(f) Spectre lissé uniquement (0 à π).

FIGURE 5.2 – *Exemple de lissage cepstral sur un signal de parole.*

5.1.2.2 Prédiction linéaire

Le signal de parole n'est pas aléatoire ; les échantillons successifs sont donc corrélés. L'analyse par prédiction linéaire (*Linear Predictive Coding, LPC*) exploite cette corrélation afin de réduire la quantité de données ([MG76], [RS78]). Un échantillon de signal de parole à l'instant n, $s(n)$, est approché par une combinaison linéaire des p échantillons précédents :

$$s(n) \approx \hat{s}(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \tag{5.5}$$

$$= \sum_{k=1}^{p} a_k s(n-k) \tag{5.6}$$

Le modèle de production de la parole utilisée pour la modélisation LPC est représenté par la figure 5.3. Le filtre linéaire de prédiction correspondant est défini par :

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5.7}$$

où $H(z)$, $S(z)$ et $U(z)$ sont respectivement les transformées en z du filtre composé par le conduit vocal, du signal et de l'excitation et G le gain.

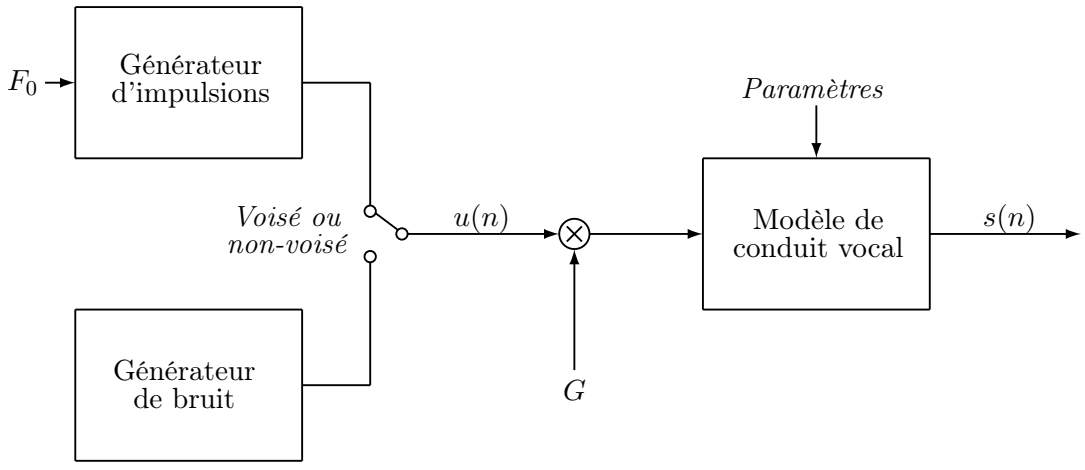


FIGURE 5.3 – *Schéma représentant le modèle de production de la parole utilisé pour la LPC.*

Les échantillons $s(n)$ sont alors liés à l'excitation $u(n)$ par l'équation :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{5.8}$$

Lerreur de prédiction est définie commme :

$$e(n) = s(n) - \hat{s}(n) \tag{5.9}$$

L'énergie de l'erreur E_n est obtenue par :

$$\begin{aligned} E_n &= \sum_m e_n^2(m) \\ &= \sum_m \big(s(m) - \hat{s}(m)\big)^2 \\ &= \sum_m \left(s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k)\right)^2 \end{aligned} \tag{5.10}$$

où $s_n(m)$ est un segment de parole choisi dans le voisinage de l'échantillon n tel que :

$$s_n(m) = s(m+n) \tag{5.11}$$

Les bornes de la sommation sur m dans l'équation (5.10) sont définies en fonction de la méthode de résolution.

Les valeurs des a_k qui minimisent E_n dans l'équation (5.10) sont obtenues en résolvant $\dfrac{\partial E_n}{\partial a_i} = 0$ pour $i = 1, \ldots, p$. On obtient alors les équations :

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} a_k \sum_m s_m(m-i)s_n(m_k) \quad 1 \le i \le p \tag{5.12}$$

Si on pose :

$$\Phi_n(i,k) = \sum_m s_m(m-i)s_n(m-k) \tag{5.13}$$

l'équation (5.12) peut être écrite :

$$\sum_{k=1}^{p} a_k \Phi_n(i,k) = \Phi_n(i,0) \quad 1 \le i \le p \tag{5.14}$$

ce qui correspond à un système de p équations à p inconnues. En utilisant les équations (5.10) et (5.12), l'erreur E_n peut s'écrire :

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} a_k \sum_m s_n(m)s_n(m_k) \tag{5.15}$$

et en utilisant l'équation (5.14), on en déduit :

$$E_n = \Phi_n(0,0) - \sum_{k=1}^{p} a_k \Phi_n(0,k) \tag{5.16}$$

La recherche des coefficients de prédiction optimaux passe par le calcul des quantités $\phi_n(i,k)$ pour $1 \leq i \leq p$ et $0 \leq k \leq p$. Après nous devons résoudre l'équation (5.14) pour obtenir les a_k. Dans le principe, la résolution des équations est assez simple mais le calcul des $\phi_n(i,k)$ est un peu plus compliqué.

Diverses méthodes de résolution existent, elles se décomposent en deux familles :
- les méthodes d'autocorrélation dans lesquelles la plage d'existence de s est infinie,
- les méthodes de covariance dans lesquelles le signal est connu uniquement sur une durée limitée.

Il est possible à partir des coefficients LPC de calculer les coefficients cepstraux c_i correspondants [RJ93].

$$c_0 = \ln G \tag{5.17a}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad \text{pour } 1 \leq m \leq p \tag{5.17b}$$

$$c_m = \sum_{k=m-p}^{m-1} \frac{k}{m} c_k a_{m-k} \quad \text{pour } m > p \tag{5.17c}$$

avec G le terme de gain du modèle LPC et p le nombre de coefficients LPC utilisés.

5.1.2.3 Calcul du vecteur acoustique

Le calcul du vecteur acoustique est une étape importante dans notre méthode d'inversion. Habituellement, les formants sont utilisés pour former le vecteur acoustique. Ici nous avons choisi de prendre les coefficients cepstraux. Le vecteur acoustique est composé d'un nombre de paramètres plus important que lors de l'utilisation des formants.

Le synthétiseur fréquentiel associé au modèle articulatoire contruit dans le chapitre 4 ne possède pas de modèle de source glottique et calcule uniquement la fonction de transfert du conduit vocal. Les coefficients cepstraux synthétiques sont déterminés à partir de cette fonction de transfert (voir §5.1.2.1).

Les coefficients cepstraux calculés sur le signal réel sont obtenus après une analyse LPC sur le spectre discrétisé (voir §5.1.2.2). Nous avons choisi de conserver les trente premiers coefficients cepstraux ; le premier coefficient cepstral relatif à l'énergie est ignoré. Panchapagesan [Pan08] utilise vingt coefficients, ici nous utilisons plus de coefficients mais le nombre n'est pas fixé de façon définitive.

L'inversion sera effectuée uniquement à partir des coefficients cepstraux mais les formants seront

calculés pour le centre des hypercuboïdes. Cette information sur les formants sera utilisée uniquement pour sélectionner les hypercuboïdes mais pas pour réaliser l'inversion. Les formants sont estimés à l'aide d'une analyse LPC (Potard [Pot08a]).

5.2 Représentations mathématiques dans un codebook

Le codebook doit représenter l'espace articulatoire de façon exhaustive, c'est-à-dire l'espace des paramètres contrôlant le modèle articulatoire. Nous utilisons le modèle articulatoire décrit dans la partie précédente. Ce modèle est contrôlé par sept paramètres variant dans l'intervalle $[-3\sigma; 3\sigma]$. L'espace articulatoire est donc contenu dans un hypercube de dimension sept de coté six. Nous avons choisi d'utiliser la représentation proposée par Potard [Pot08a] à l'aide de parallélépipèdes, où *hypercuboïdes*, qui est une généralisation du rectangle en dimension supérieure à deux.

5.2.1 Structure hypercuboïdale du codebook

La structure élémentaire utilisée pour le codebook est l'hypercuboïde défini par Potard [Pot08a]. Soit M est la dimension de l'espace articulatoire, un hypercuboïde H_c est défini par :

$$H_c = \{P_0 + x, x \in \mathbb{R}^M | \forall i \in \{1, \ldots, M\} | x_i | \leq r_i\} \tag{5.18}$$

où P_0 est le centre géométrique de l'hypercuboïde et $r \in \mathbb{R}^M$ le rayon de l'hypercuboïde. Un hypercuboïde est alors caractérisé par son centre et son rayon. La figure 5.4 montre une représentation d'un hypercuboïde dans un espace de dimension 3.

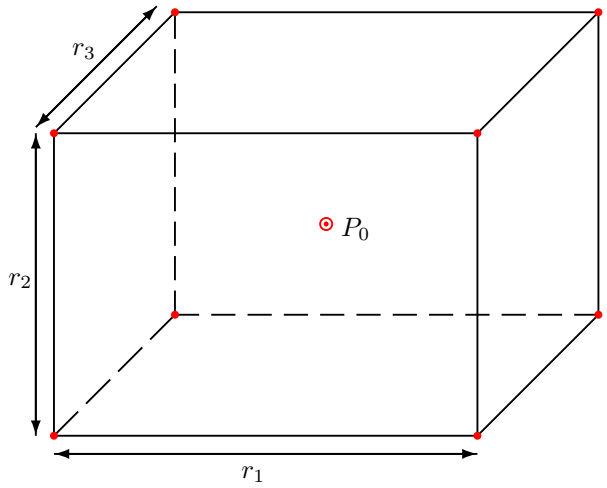


FIGURE 5.4 – *Représentation d'un hypercuboïde dans un espace de dimension 3 caractérisé par son rayon* $r = (r_1, r_2, r_3)$ *et son centre* P_0.

Le codebook est alors composé d'une collection d'hypercuboïdes de différentes tailles.

5.2.2 Modélisation de la relation articulatoire-acoustique

Dans chaque hypercuboïde du codebook, la relation articulatoire-acoustique locale est approchée par des fonctions simples. Les fonctions utilisées pour le comportement local de la relation articulatoire-acoustique peuvent être du type :

- fonctions constantes par morceaux [ACMT78, SS92, BPB92],
- fonctions linéaires par morceaux [Cha84, ST96, OL00],
- ou plus généralement des polynômes multivariés [Pot08a].

Dans le cadre de cette thèse, l'approximation locale dans un hypercuboïde est réalisée à l'aide d'une fonction affine.

5.3 Construction du codebook

5.3.1 Principe général

La méthode de construction du codebook repose sur une exploration exhaustive de l'espace articulatoire proposée par Ouni [Oun01] et Potard [Pot08a]. L'espace articulatoire est contenu dans un hypercube de dimension sept et dont chacun des côtés est égal à six (chaque paramètre varie entre -3σ et 3σ). Cet hypercube est divisé de façon récursive en petits hypercuboïdes. Tant que la relation articulatoire-acoustique locale n'est pas considérée comme suffisamment linéaire dans un hypercuboïde, il y a subdivision. L'hypercuboïde est divisé en deux sous-hypercuboïdes suivant une seule direction choisie de façon optimale ([Pot08a]). L'opération est répétée jusqu'à ce que la relation locale ait le comportement souhaité ou que la taille minimale de la structure soit atteinte. L'hypercuboïde est soit rejeté si la relation articulatoire-acoustique n'est pas linéaire, soit conservé dans le codebook si la relation peut être considérée linéaire.

La figure 5.5 montre la méthode de division en dimension deux pour la représentation du codebook. Les hypercuboïdes verts représentent les hypercuboïdes où la relation est considérée linéaire, il n'y a plus de subdivision. Sinon, les subdivisions continuent jusqu'à ce que la taille minimale ou la quasi-linéarité soit atteinte.

5.3.2 Test de linéarité à partir de la matrice jacobienne

Un des points importants est de déterminer si la relation articulatoire-acoustique se comporte comme l'approximation attendue. Le test est réalisé à partir de la matrice jacobienne [Pot08a] en un point précis de l'hypercuboïde (généralement le centre). À partir de ce point et de la matrice jacobienne autour de ce point, il est possible de calculer l'image acoustique de n'importe quel point de l'hypercuboïde en faisant l'hypothèse de linéarité. Pour vérifier l'hypothèse de linéarité, les images estimées sur les points de test (cf §5.3.3) sont comparées aux images réelles calculées avec le synthétiseur articulatoire.

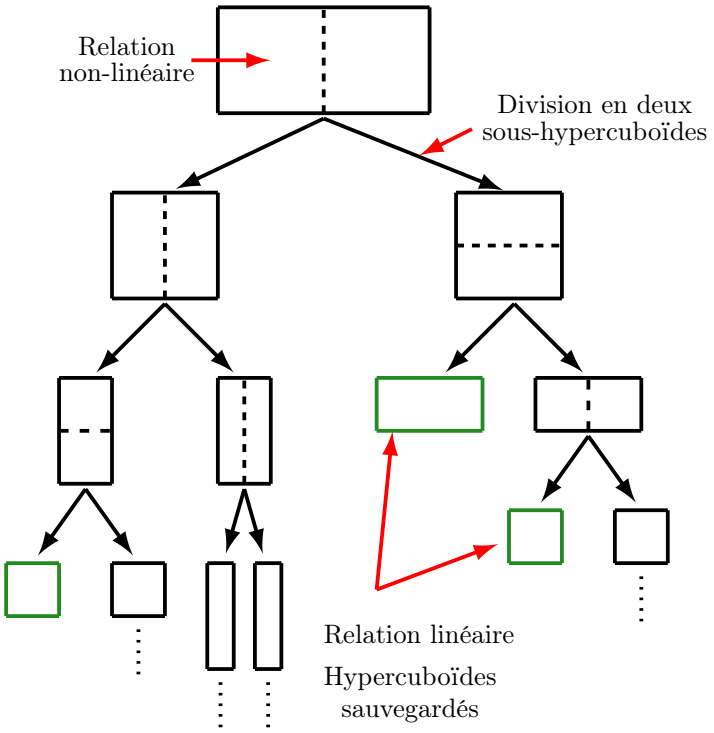


FIGURE 5.5 – *Subdivisions successives dans un hypercuboïde. Si la relation locale dans un sous hypercuboïde n'est pas linéaire, il y a subdivision en deux sous-hypercuboïdes dans une seule direction et ainsi de suite. La subdivision est effectuée jusqu'à ce que la relation soit linéaire ou que la taille minimale soit atteinte.*

5.3.2.1 Calcul de la matrice jacobienne

La matrice jacobienne est la matrice des dérivées partielles d'une fonction de plusieurs variables en un point donné.

Soient M et N les dimensions des espaces articulatoire et acoustique. On définit $A_r \in \mathbb{R}^M$ et $A_c \in \mathbb{R}^N$ les espaces articulatoire et acoustique. La fonction du synthétiseur articulatoire $f : A_r \longrightarrow A_c$ est définie par :

$$f : \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix} \longmapsto \begin{pmatrix} f_1(\alpha_1, \alpha_2, \ldots, \alpha_M) \\ f_2(\alpha_1, \alpha_2, \ldots, \alpha_M) \\ \vdots \\ f_N(\alpha_1, \alpha_2, \ldots, \alpha_M) \end{pmatrix} \tag{5.19}$$

La matrice jacobienne en un point $x_0 \in A_r$ est définie par :

$$\Delta f(x_0) = \left[\frac{\partial f_j}{\partial p_i}(x_0)\right]_{1 \leq i \leq M, 1 \leq j \leq N} \tag{5.20}$$

Dans un petit voisinage de x_0, on a la relation suivante :

$$f(x) \approx f(x_0) + \Delta f(x_0).(x - x_0) \tag{5.21}$$

Dans notre cas, le centre de l'hypercuboïde est choisi pour calculer la matrice jacobienne car il a la propriété d'être à la même distance de tous les sommets.

La matrice jacobienne est évaluée numériquement dans un petit voisinage du centre de l'hypercuboïde. L'évaluation s'effectue à une distance ϵ_m proportionnelle au rayon de l'hypercuboïde. On note δ_i le vecteur de $\mathbb{R}^M$ dont toutes les composantes sont nulles sauf la $i^{ième}$ qui est égale à 1. La matrice jacobienne J pour le centre P_0 d'un hypercuboïde est obtenue par :

$$J_{j,i}(P_0) = \left.\frac{f_j(P_0 + \epsilon_m.r_i.\delta_i) - f_j(P_0 - \epsilon_m.r_i.\delta_i)}{2\epsilon_m.r_i}\right|_{1 \leq i \leq M, 1 \leq j \leq N} \tag{5.22}$$

5.3.2.2 Test de linéarité

Le test de linéarité consiste à comparer les vecteurs acoustiques (cf §5.1.2.3) obtenues par le synthétiseur articulatoire et ceux obtenus par interpolation afin de vérifier que la relation articulatoire-acoustique locale peut être approchée par une fonction linéaire.

L'approximation linéaire dans un hypercuboïde est réalisée à partir de la matrice jacobienne

calculée autour du centre de l'hypercuboïde. En d'autres termes : soit H_c un hypercuboïde et $P_0 \in H_c$ le centre de l'hypercuboïde H_c, la matrice jacobienne en ce point $\Delta f(P_0)$ est définie par :

$$\Delta f(P_0) = \begin{bmatrix} \frac{\partial f_1}{\partial \alpha_1}(P_0) & \frac{\partial f_1}{\partial \alpha_2}(P_0) & \dots & \frac{\partial f_1}{\partial \alpha_M}(P_0) \\ \frac{\partial f_2}{\partial \alpha_1}(P_0) & \frac{\partial f_2}{\partial \alpha_2}(P_0) & \dots & \frac{\partial f_2}{\partial \alpha_M}(P_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial \alpha_1}(P_0) & \frac{\partial f_N}{\partial \alpha_2}(P_0) & \dots & \frac{\partial f_N}{\partial \alpha_M}(P_0) \end{bmatrix} \tag{5.23}$$

Si on note $F_0 = f(P_0)$ l'image acoustique du centre P_0 de l'hypercuboïde, l'hypothèse de linéarité nous permet d'écrire l'approximation suivante :

$$f(P_x) \approx F_0 + \Delta f(P_0).(P_x - P_0) \quad \text{pour tout } P_x \in H_c \tag{5.24}$$

Le test de linéarité est réalisé à partir de l'approximation locale donnée par l'équation (5.24). Un ensemble de points de test défini dans le paragraphe 5.3.3 est utilisé pour déterminer si la relation est considérée comme linéaire.

Pour chaque point de test, on vérifie que la distance entre l'image acoustique interpolée et l'image acoustique calculée par le synthétiseur n'est pas trop grande. Le test de linéarité s'écrit alors de la manière suivante :

$$d\big(F_0 + \Delta f(P_0).(P_x - P_0), f(P_x)\big) \leq \epsilon \tag{5.25}$$

où d est la distance euclidienne et ϵ un seuil donné.

5.3.3 Choix des points de test

Les points de test sont des points qui permettent de tester si l'hypercuboïde doit être divisé ou non. Les points de test utilisés sont le centre de l'hypercuboïde, les sommets, les milieux des segments reliant deux sommets, les points utilisés pour le calcul du jacobien au centre à une distance proportionnelle au rayon ou encore des points choisis de façon aléatoire. Ces points sont utilisés pour tester la linéarité. En effet, pour chacun des points nous vérifions que l'image acoustique obtenue par le synthétiseur articulatoire est proche de l'image obtenue par interpolation linéaire. La figure 5.6 montre la représentation en deux dimensions d'un hypercuboïde ; les points de test sont les points dessinés en rouge.

5.3.4 Subdivision d'un hypercuboïde

Dans cette section, nous allons présenter la façon dont la subdivision s'effectue. La construction du codebook est récursive ce qui signifie qu'à chaque subdivision d'un hypercuboïde le nombre total

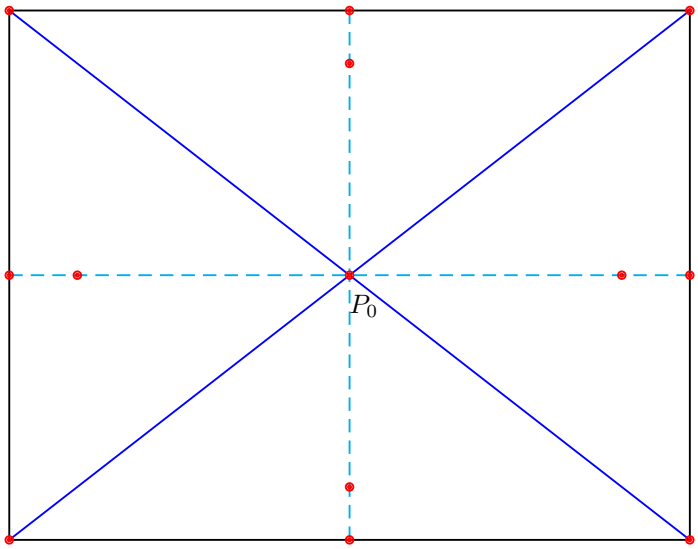


FIGURE 5.6 – *Représentation en deux dimensions d'un hypercuboïde avec les différents points de test. Les points rouges correspondent au centre, aux sommets, aux milieux des segments reliant deux sommets et les points utilisés pour le calcul du jacobien au centre à une distance proportionnelle au rayon.*

d'hypercuboïdes croît rapidement. Il est important de limiter le nombre de subdivisions (et ainsi le nombre d'hypercuboïdes). En effet le codebook doit être le plus concis possible mais également conserver une bonne précision. Une bonne méthode de subdivision doit permettre de supprimer les zones non-synthétisables (dans le cas où le conduit vocal possède une occlusion, il n'est pas possible de calculer une image acoustique) et de conserver le plus grand nombre de zones synthétisables (celles pour lesquelles il existe une image acoustique).

5.3.4.1 Conditions de subdivision

La frontière de l'espace articulatoire concentre la majorité des subdivisions car il y existe un problème de non-linéarité. Si l'on souhaite approcher finement la frontière, les hypercuboïdes seront très petits afin d'épouser la frontière ce qui engendre un grand nombre de subdivisions. En effet, la frontière de l'espace articulatoire définit la limite entre les régions synthétisables et non-synthétisables. Un hypercuboïde situé sur la frontière possède donc des points synthétisables et non-synthétisables, il est divisé en deux sous-hypercuboïdes afin d'affiner les contours de la région non-synthétisable. À chaque étape de division, on obtient donc des hypercuboïdes de plus en plus petits et les limites de l'espace articulatoire sont ainsi approchées. Cependant, si l'on souhaite approcher plus finement les limites, un grand nombre d'hypercuboïdes de très petite taille sont créés. Ces régions seront alors définies par un grand nombre d'hypercuboïdes alors qu'elles seront rarement atteintes par le locuteur car elles correspondent à un effort articulatoire important. Les limites ne seront pas approchées finement, nous utiliserons un échantillonnage qui permet de délimiter ces régions sans perdre trop de points.

La figure 5.7 illustre le problème de délimitation des frontières de l'espace articulatoire pour

le modèle de Maeda. La limite entre les zones synthétisables et non-synthétisables (où les vecteurs articulatoire n'ont pas d'image acoustique car il existe une occlusion dans le conduit vocal) est représentée par la ligne rouge. La zone en gris foncé est la zone synthétisable obtenue à partir d'un codebook et celle en gris clair est une zone synthétisable mais elle n'est pas prise en compte. Cette figure permet de voir qu'une partie de l'espace articulatoire du codebook n'est pas prise en compte dans le codebook.

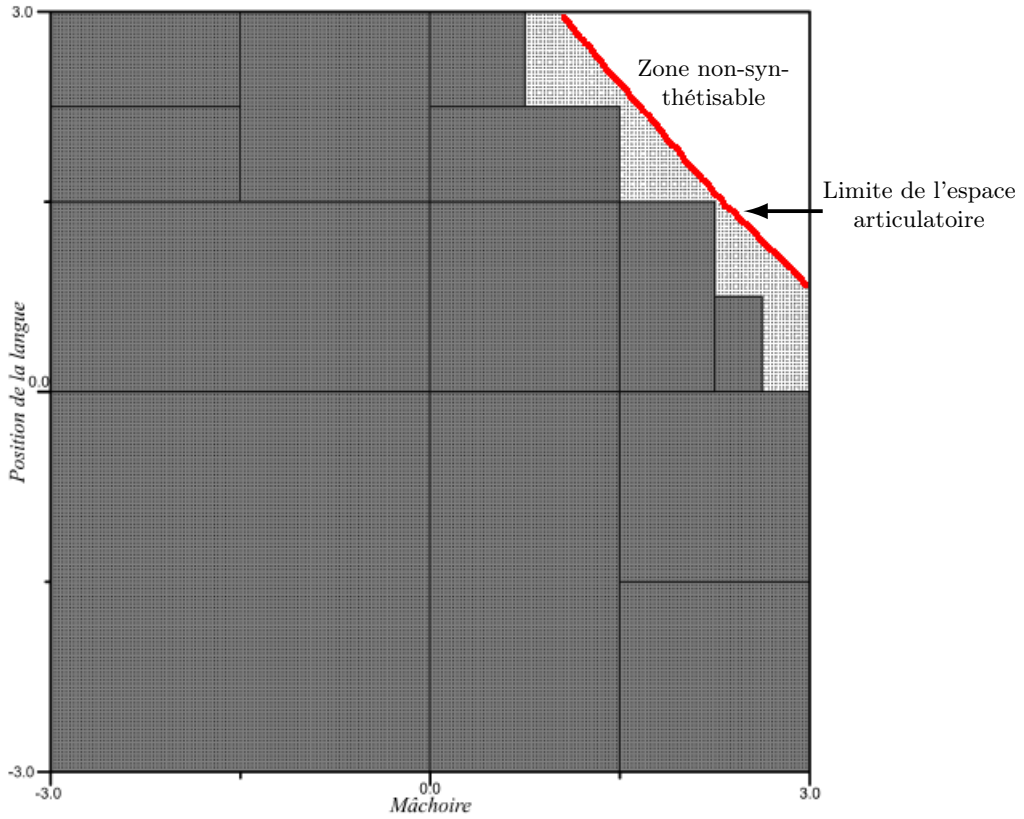


FIGURE 5.7 – *Coupe l'espace articulatoire de Maeda selon deux composantes (mâchoire et position de la langue) variant entre* -3 *et* $+3$ *(d'après [Pot08a]). La zone en pointillés correspond à la zone synthétisable et la zone blanche à la zone non-synthétisable, la ligne rouge représente la limite entre les deux. Les rectangles correspondent aux coupes des hypercuboïdes.*

Pour une grande partie des hypercuboïdes le long de la frontière, seul un petit nombre de sommets n'est pas défini. Dans ce cas, on conserve ces hypercuboïdes en définissant un seuil de sommets calculables et en imposant que la matrice jacobienne soit entièrement calculable.

La subdivision intervient dans deux cas de figure, la délimitation de la frontière et la non-linéarité dans un hypercuboïde. Lors de l'exploration d'un hypercuboïde, les premiers tests vérifient s'il se situe sur la frontière de l'espace articulatoire. Pour cela, on vérifie que le centre de l'hypercuboïde possède une image acoustique, que la matrice jacobienne est calculable et que les sommets

possèdent une image acoustique. Si le nombre de sommets calculables est insuffisant par rapport à un seuil fixé, cela signifie que l'on se situe au niveau de la frontière ;l'hypercuboïde sera alors subdivisé.

Si les tests précédents ont réussi, c'est-à-dire que l'on se situe dans l'espace articulatoire synthétisable, on réalise le test de linéarité en utilisant la matrice jacobienne. Pour chaque point de l'ensemble de test, nous comparons l'image acoustique obtenue par le synthétiseur articulatoire avec l'image acoustique obtenue par interpolation linéaire. Si la relation n'est pas linéaire, l'hypercuboïde est divisé.

5.3.4.2 Méthode de subdivision

L'espace articulatoire est approché par une exploration récursive de l'espace des paramètres articulatoires du modèle. À chaque étape, si la subdivision est nécessaire l'hypercuboïde doit être divisé en plusieurs hypercuboïdes et l'analyse est faite dans chaque hypercuboïde. Pour cela, il est nécessaire de déterminer comment les hypercuboïdes doivent être divisés.

L'approche utilisée par Ouni [Oun01] consiste à subdiviser dans chacune des sept dimensions de l'espace articulatoire de Maeda. Chaque subdivision conduisait alors à l'exploration de $2^7 = 128$ hypercubes. Cette méthode de construction était très coûteuse en temps et en espace. En effet, chaque niveau de subdivision multipliait la taille du codebook par 128. Le temps de calcul était donc très long à cause du grand nombre d'hypercubes. De plus, de nombreux hypercubes avaient des comportements acoustiques très proches et auraient donc pu être regroupés.

Afin de réduire le nombre de subdivisions et donc le nombre d'hypercuboïdes, Potard [Pot08a] proposa de ne plus diviser dans toutes les directions de l'espace articulatoire mais uniquement dans une seule direction. La structure hypercuboïdale permet la division en deux sous-hypercuboïdes. Le choix de la direction de subdivision doit permettre de limiter le nombre de divisions dans les directions où l'approximation locale est fidèle à la relation articulatoire-acoustique. Deux possibilités ont été envisagées afin de calculer la direction de subdivision :

1. déterminer la direction qui maximise la « non-linéarité » et diviser dans cette direction,
2. déterminer le demi-espace qui minimise la « non-linéarité » et diviser dans la direction normale à ce demi-espace.

Dans les cas où le centre n'aurait pas d'image acoustique et la matrice jacobienne ne serait pas calculable, la direction de subdivision correspond au rayon le plus grand de l'hypercuboïde. Ces situations se produisent lorsque que l'on se situe dans la zone non-synthétisable ou alors à la frontière.

5.3.4.3 Terminaison de la récursivité

La subdivision intervient lors du suivi de frontière et lors de la non-linéarité de la relation articulatoire-acoustique. L'arrêt de la récursivité correspond au rejet ou à la conservation de l'hypercuboïde.

Un hypercuboïde est conservé dans le cas où le test de linéarité réussit. Si la linéarité n'est

pas vérifiée et que la taille minimale (ou volume de l'hypercuboïde) est atteinte, l'hypercuboïde est conservé uniquement si le nombre de sommets calculables est suffisant. Le nombre de sommets est déterminé dans la section 5.4.2.

Le rejet d'un hypercuboïde intervient lorsque l'on se situe dans la zone non-synthétisable ou au niveau de la frontière de l'espace articulatoire avec cette zone. En effet, si le centre d'un hypercuboïde n'a pas d'image acoustique ou si la matrice jacobienne n'est pas calculable, ou encore si le nombre de sommets sans image acoustique est important (cf §5.4.2), cela signifie que l'on se situe dans ou à proximité d'une zone non-synthétisable. Si la subdivision n'est pas possible car l'hypercuboïde a atteint la taille minimale, il est rejeté.

5.4 Évaluation expérimale du codebook

Dans les sections précédentes, nous avons vu que la construction du codebook dépend de plusieurs paramètres qui doivent être choisis judicieusement.

Tout d'abord, nous déterminerons la distance ϵ_m optimale pour l'évaluation de la matrice jacobienne au centre de l'hypercuboïde. Puis, nous optimiserons la couverture de l'espace articulatoire notamment au niveau de la frontière. Enfin, nous évaluerons la taille minimale d'un hypercuboïde et le seuil acoustique minimal.

Afin d'évaluer la précision acoustique du codebook, les vecteurs acoustiques obtenus par interpolation sont comparés à ceux issus du synthétiseur. La resynthèse d'un vecteur articulatoire choisi au hasard consiste à identifier à quel hypercuboïde il appartient et à partir de l'interpolation linéaire locale de calculer son image acoustique.

La couverture de l'espace articulatoire par le codebook est évaluée en calculant son volume en dimension M (sept paramètres du modèle) et en le comparant au volume réel estimé.

5.4.1 Valeur optimale du paramètre de la matrice jacobienne

Dans le paragraphe 5.3.2.1, la matrice jacobienne est évaluée à une distance proportionnelle ϵ_m à son rayon. La valeur optimale de ϵ_m sera déterminée expérimentalement en construisant différents codebooks en faisant varier la valeur de ϵ_m.

Le choix de la valeur optimale pour ϵ_m n'est pas simple car plusieurs critères doivent être pris en compte dans le but d'avoir un codebook le plus concis et précis possible.

Le premier critère de choix est la concision du codebook caractérisée par le nombre d'hypercuboïdes le constituant. Plus le nombre d'hypercuboïdes est petit, plus le codebook sera concis. La concision du codebook ne doit pas être basée uniquement sur le nombre d'hypercuboïdes mais également sur le volume de l'espace articulatoire contenu dans le codebook. Le « volume » du code-

book (c'est-à-dire la somme des volumes de chacun des hypercuboïdes) doit donc s'approcher le plus possible du « volume » de l'espace articulatoire en conservant un nombre d'hypercuboïdes limité.

Le second critère porte sur la précision acoustique du codebook. La resynthèse d'un vecteur articulatoire doit être la plus fidèle possible. En effet les vecteurs acoustiques obtenus par synthèse sont comparés à ceux calculés par interpolation linéaire. L'évaluation de la précision acoustique est liée à la concision du codebook. Effectivement si les zones difficilement linéarisables ne sont pas présentes, l'évaluation portera uniquement sur les hypercuboïdes où la relation est considérée comme linéaire donnant alors le résultat attendu. La comparaison de l'erreur entre deux codebooks n'ayant pas le même volume n'a donc pas de sens. Un codebook avec un petit volume contenant uniquement des hypercuboïdes où la linéarité est « plus facilement » modélisée fournira de meilleurs résultats pour la resynthèse qu'un codebook couvrant plus d'espace contenant des zones situées sur la frontière qui sont « plus difficilement » synthétisables et donc fournissent de moins bonnes erreurs de resynthèse. La comparaison de l'erreur de resynthèse moyenne doit donc se faire sur des codebooks de taille similaire. L'erreur moyenne est calculée en resynthétisant 100000 vecteurs articulatoires choisis aléatoirement.

La recherche de la valeur optimale de ϵ_m est effectuée en construisant différents codebooks en faisant varier ϵ_m. Pour nos tests, nous avons choisi d'étudier trois zones différentes du codebook possédant des caractéristiques différentes : une zone possédant beaucoup de contacts avec la frontière, une zone presque entièrement synthétisable et une dernière zone choisie au hasard dont le volume synthétisable est inférieur à la deuxième. Chaque zone explore 1/128 de l'espace de départ dont le volume correspond au volume d'un hypercube de dimension sept et de côté six, soit $6^7 = 279936$. Chacune des trois zones explore donc un volume égal à $\frac{279936}{128} = 3^7 = 2187$.

Les trois zones n'occupent pas toutes le même volume. Nous avons déterminé empiriquement le volume de l'espace articulatoire contenu dans ces zones. Les volumes sont respectivement de 840,95 pour la première zone, 2163,51 pour la deuxième et 1338,04 pour la dernière. L'espace articulatoire occupe pratiquement toute la deuxième zone, le nombre de subdivisions liées au suivi de la frontière ne sera donc pas très important contrairement à la première zone dont le volume est plus faible.

Pour chaque zone, nous avons calculé les paramètres suivants : le nombre d'hypercuboïdes, le volume (somme des volumes de chaque hypercuboïde en dimension $M = 7$), le pourcentage de volume obtenu par rapport au volume réel et l'erreur de resynthèse moyenne. Les résultats sont présentés dans le tableau 5.1. On remarque que le nombre d'hypercuboïdes et le volume augmentent lorsque ϵ_m diminue (voir figure 5.8). Pour les valeurs élevées de ϵ_m, le volume représenté est très faible, ce qui signifie qu'une grande partie de l'espace articulatoire a été perdue car la relation n'était pas linéarisable.

	Zone 1				Zone 2				Zone 3			
ϵ_m	n_{Hc}	v	$\%v$	Δc	n_{Hc}	v	$\%v$	Δc	n_{Hc}	v	$\%v$	Δc
1,5	180	96,11	11,43%	1,85	3221	1740,63	80,45%	0,74	383	316,09	23,62%	0,48
1,2	284	151,64	18,03%	1,83	3435	1862,37	86,08%	0,72	498	403,65	30,17%	0,45
1,0	283	204,50	24,33%	1,73	3518	1904,01	88,01%	0,70	574	444,23	33,20%	0,42
0,9	402	214,64	25,52%	1,70	3570	1929,64	89,19%	0,69	612	460,25	34,40%	0,42
0,8	471	251,64	29,92%	1,70	3654	1972,89	91,19%	0,69	773	546,75	40,86%	0,45
0,7	569	303,81	36,13%	1,63	3706	2000,12	92,45%	0,69	853	591,60	44,21%	0,42
0,6	615	328,37	39,05%	1,62	3737	2011,87	92,99%	0,68	928	634,32	47,41%	0,43
0,55	656	350,26	41,61%	1,62	3772	2029,48	93,80%	0,69	1057	704,79	52,61%	0,45
0,5	687	366,81	43,62%	1,61	3781	2032,16	93,93%	0,69	1080	714,01	53,36%	0,45
0,45	707	377,49	44,89%	1,66	3792	2038,57	94,23%	0,70	1099	721,88	53,95%	0,45
0,4	732	390,84	46,48%	1,71	3869	2073,27	95,83%	0,72	1254	800,90	59,86%	0,48
0,35	776	414,33	49,27%	1,77	3897	2086,62	96,45%	0,74	1287	813,72	60,76%	0,49
0,3	823	439,43	52,25%	1,81	3920	2096,23	96,89%	0,75	1419	880,99	65,84%	0,53
0,25	858	458,12	54,48%	1,91	3943	2107,98	97,43%	0,78	1462	901,28	67,36%	0,55
0,2	921	491,75	58,48%	2,00	3951	2111,18	97,58%	0,80	1482	909,83	68,00%	0,58
0,1	1023	546,22	64,95%	2,12	3974	2122,39	98,10%	0,91	1699	1026,22	76,70%	0,66
0,05	1052	561,70	66,79%	2,27	3979	2126,13	98,27%	0,99	1835	1099,91	82,20%	0,74
0,02	1098	586,26	69,71%	2,43	3983	2128,27	98,37%	1,01	1903	1136,21	84,92%	0,75
0,01	1130	603,35	71,75%	2,42	3983	2128,27	98,37%	0,97	1922	1146,89	85,71%	0,73
0,005	1132	604,42	71,87%	2,40	3983	2128,27	98,37%	0,96	1931	1146,89	85,71%	0,75
0,002	1155	616,69	73,33%	2,54	3983	2128,27	98,37%	0,83	1927	1150,10	85,95%	1,04
0,001	1155	616,69	73,33%	1,99	3983	2128,27	98,37%	0,82	1923	1150,10	85,95%	0,73

TABLEAU 5.1 – *Principales caractéristiques de trois zones du codebook en fonction de différentes valeurs de ϵ_m. n_{Hc} le nombre d'hypercuboïdes, v le volume, $\%v$ le pourcentage de volume par rapport au volume estimé et Δc l'erreur de resynthèse moyenne. La précision acoustique est estimée en resynthétisant 10000 points dans chaque zone et en comparant l'image obtenue à partir du synthétiseur et celle calculée à l'aide du jacobien.*

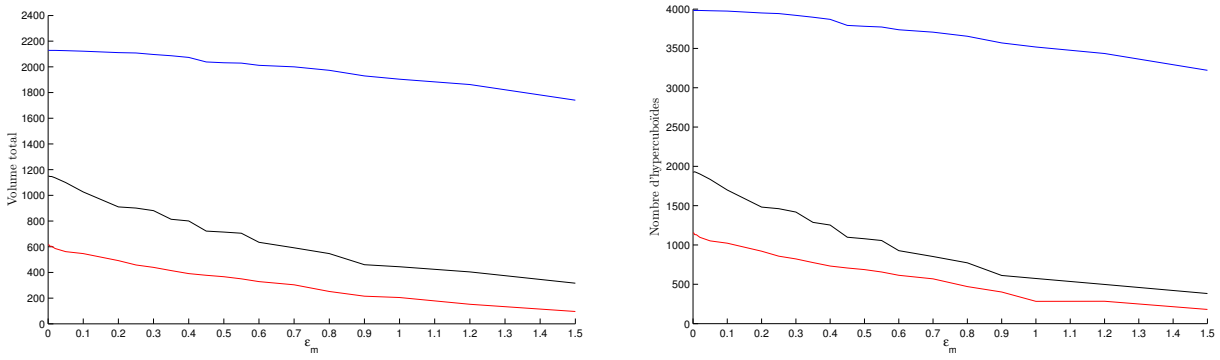


FIGURE 5.8 – *Le graphique de gauche représente le volume total en fonction des différentes valeurs de ϵ_m et celui de droite le nombre d'hypercuboïdes en fonction des différentes valeurs de ϵ_m pour les trois zones. La zone 1 correspond aux courbes rouges, la zone 2 aux courbes bleues et la zone 3 aux courbes noires.*

On observe également que le pourcentage de volume de l'espace articulatoire contenu dans les codebooks n'est pas le même pour les trois zones. Pour les première et dernière zones le volume varie beaucoup, alors que pour la deuxième zone la variation est moins importante. Cela confirme notre hypothèse que la relation articulatoire-acoustique au niveau de la frontière n'est pas linéaire contrairement au reste de l'espace articulatoire. Si l'on se situe au niveau de la frontière, le jacobien ne doit pas être calculé très loin du centre de l'hypercuboïde car le volume s'effrondre. Par exemple, si le jacobien est calculé avec $\epsilon_m = 1,5$ le volume obtenu est seulement de $11,43\%$ pour la première zone alors que pour la deuxième zone le volume utile est de $80,45\%$ (voir tableau 5.1 et figure 5.9). Les valeurs de ϵ_m élevées ne sont pas retenues car il est impossible de linéariser la relation articulatoire-acoustique. Pour les valeurs de ϵ_m supérieures à 0,5 la matrice est en fait calculée à l'extérieur de l'hypercuboïde. Le calcul en dehors de l'hypercuboïde est possible car la relation peut être linéaire même en dehors de l'hypercuboïde. Ce calcul est juste informatif.

La précision acoustique atteinte n'est pas identique dans tout le codebook. L'erreur moyenne de resynthèse est différente pour les trois zones du codebook étudiées ici alors que le seuil choisi pour la construction est le même, ce qui nous indique que le comportement du codebook varie fortement suivant la zone où l'on se situe (voir figure 5.10). Pour la première zone étudiée, le volume obtenu ne dépasse pas 74% du volume de l'espace articulatoire, l'espace manquant correspond donc aux hypercuboïdes situés sur la frontière qui sont plus difficilement linéarisables. Le volume de la deuxième zone atteint 98% du volume attendu, ce qui nous indique que dans ce cas le suivi de frontière est moins important. Le choix de la valeur optimale de ϵ_m ne peut donc pas se baser uniquement sur la valeur de l'erreur de resynthèse. Dans les exemples utilisés ici, on voit que lorsque l'erreur de resynthèse est minimale le volume utile est faible. Ainsi l'erreur de resynthèse moyenne n'est pas trop mauvaise car les « mauvais » hypercuboïdes ne sont pas inclus et donc seuls les « bons » hypercuboïdes avec une relation linéaire bien définie sont utilisés pour le calcul de l'erreur. Pour la première zone, l'erreur Δ_c est minimale pour $\epsilon_m = 0,5$ or dans ce cas le volume est seulement de $43,62\%$ du volume utile, de même pour la troisième zone où Δ_c est minimale pour $\epsilon_m = 1$ pour un volume très faible de

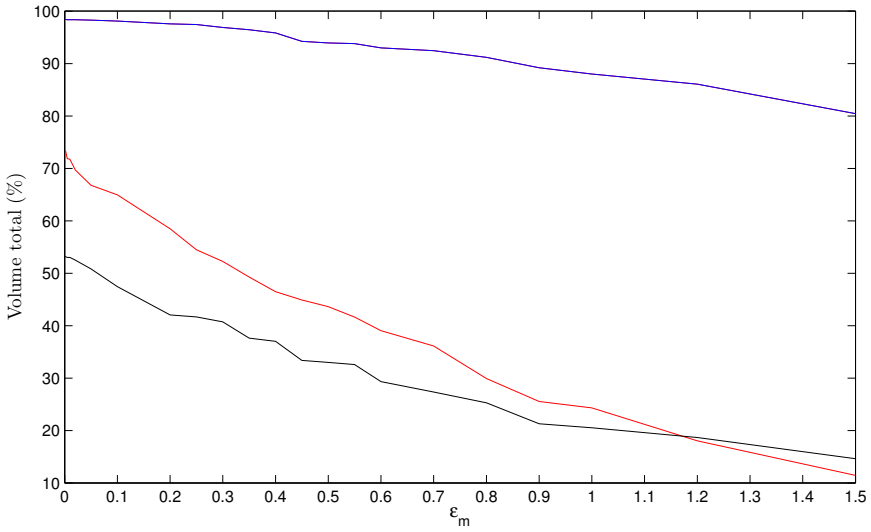


FIGURE 5.9 – *Graphique représentant le pourcentage de volume de l'espace articulatoire couvert en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3.*

$33,20\%$ (voir le tableau 5.1). Les valeurs de $\epsilon_m = 1$ pour lesquelles le volume est trop faible seront retirées.

On a vu que la précision acoustique entre les différents codebooks construits n'est pas toujours comparable. En effet, si les hypercuboïdes difficilement linéarisables sont absents du codebook, la précision acoustique est meilleure. La précision acoustique de différents codebooks est donc comparable uniquement si les codebooks représentent le même espace articulatoire. La recherche du ϵ_m optimal ne peut pas dépendre uniquement de l'erreur de resynthèse car elle ne permet pas de rendre compte de la concision du codebook. Une mesure plus significative consiste à utiliser l'erreur de resynthèse pondérée par une mesure de la densité du codebook. En d'autres termes, une mesure homogène de l'erreur tenant compte de la densité du codebook est donnée par :

$$e = \Delta c. \sqrt[M]{\frac{n_{Hc}}{v}} \tag{5.26}$$

où M la dimension de l'espace articulatoire, v le volume du codebook, n_{Hc} le nombre d'hypercuboïdes du codebook et Δc erreur moyenne de resynthèse. La mesure de l'erreur permet ainsi de rendre compte de la précision due au paramétrage de ϵ_m (voir figure 5.11).

On constate que pour les valeurs de ϵ_m supérieure à $0,1$ l'erreur e varie très peu pour les zones 2 et 3. L'erreur atteint toujours son minimum pour $\epsilon_m = 0,5$. Afin de ne pas utiliser les limites de l'hypercuboïde, nous avons choisi $\epsilon_m = 0,45$.

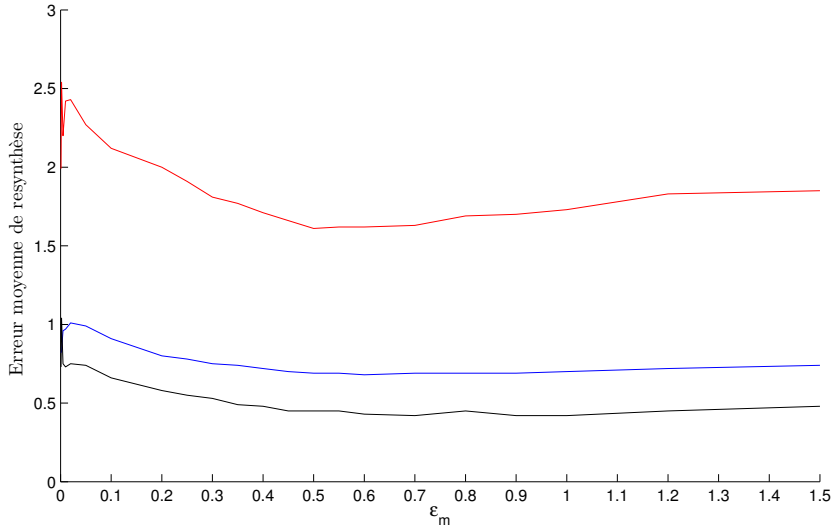


FIGURE 5.10 – *Graphique représentant l'erreur moyenne de resynthèse en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3.*

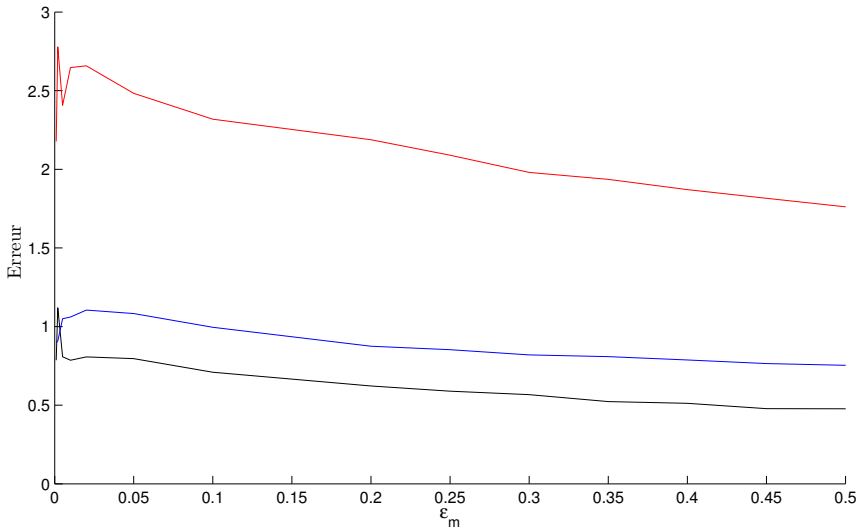


FIGURE 5.11 – *Graphique représentant la valeur de la mesure homogène e en fonction des différentes valeurs de ϵ_m. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3.*

5.4.2 Couverture de l'espace articulatoire

Notre codebook doit offrir une bonne couverture de l'espace articulatoire, c'est-à-dire que les zones synthétisables (là où il existe une image acoustique) doivent donc être présentes. Nous avons vu dans les sections précédentes que les frontières de l'espace concentrent un grande nombre d'hypercuboïdes mal définis. En d'autres termes, les sommets d'un hypercuboïde situés sur la frontière ne sont pas tous définis. La figure 5.12 montre un exemple d'hypercuboïde situé sur la frontière en dimension trois.

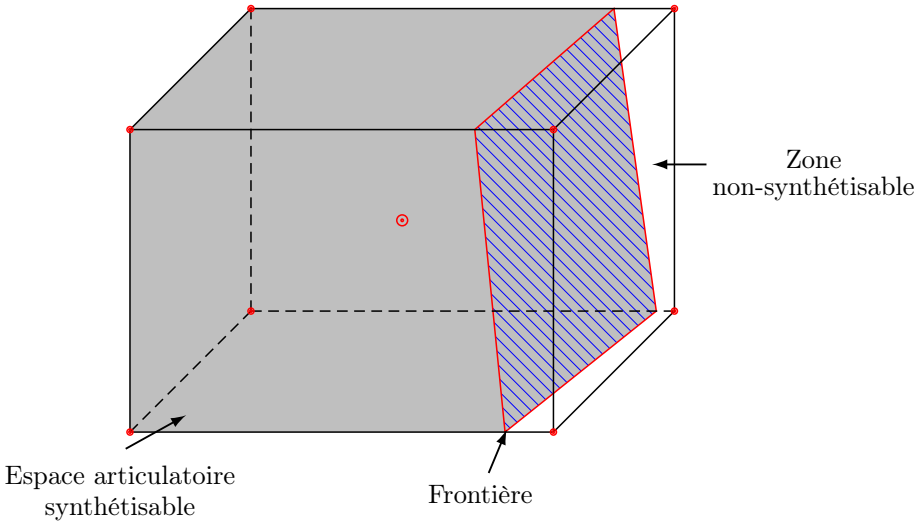


FIGURE 5.12 – *Exemple d'un hypercuboïde (en trois dimensions) situé sur la frontière. La partie grise correspond à l'espace articulatoire synthétisable et la partie blanche à la zone où les vecteurs articulatoires ne possèdent pas d'image acoustique. Cet hypercuboïde a la moitié de ses sommets sans image acoustique alors qu'il est presque entièrement défini.*

En dimension sept, chaque hypercuboïde possède 128 sommets. Afin de déterminer le nombre de sommets définis minimal acceptable pour le conserver, nous avons construit plusieurs codebooks en faisant varier le nombre de sommets définis minimal. Pour réaliser ce test nous avons utilisé les trois zones utilisées dans la section précédente car elles ont l'avantage de représenter des zones différentes du codebook. Les résultats sont présentés dans le tableau 5.2. Nous observons que lorsque le nombre de sommets définis diminue le volume occupé augmente.

La figure 5.13 montre l'évolution du volume couvert pour les trois régions en fonction du nombre de sommets. La zone 1 (courbe rouge) est située sur la frontière de l'espace articulatoire ; la plupart des hypercuboïdes la constituant n'ont pas tous leurs sommets définis. En effet, lorsque l'on impose que tous les sommets soient définis (128 sommets), seulement 14,01% du volume défini dans cette zone est atteint. Cela signifie qu'il existe très peu d'hypercuboïdes ayant tous leurs sommets définis. En diminuant ce seuil, le volume augmente pour atteindre 73% en imposant que la moitié des sommets soit définie. Au contraire, la zone 2 ne possédant qu'une petite région de frontière a un

	Zone 1				Zone 2				Zone 3			
nb	n_{Hc}	v	$\%v$	Δc	n_{Hc}	v	$\%v$	Δc	n_{Hc}	v	$\%v$	Δc
128	333	117,80	14,01%	2,05	3434	1847,42	85,39%	0,69	841	585,19	43,73%	0,44
125	410	218,91	26,03%	2,01	3489	1876,80	86,75%	0,70	857	593,74	44,37%	0,45
122	480	256,29	30,48%	2,00	3544	1906,15	88,10%	0,69	1117	732,56	54,75%	0,47
119	528	281,92	33,52%	1,99	3642	1958,48	90,52%	0,69	1275	816,92	61,05%	0,51
116	571	304,88	36,25%	1,99	3688	1983,04	91,66%	0,69	1389	877,79	65,60%	0,53
113	613	327,30	38,92%	1,98	3719	1999,59	92,42%	0,69	1405	886,33	66,24%	0,53
110	647	345,46	41,08%	1,97	3753	2017,74	93,26%	0,70	1463	917,30	68,56%	0,54
107	681	363,61	43,24%	1,96	3825	2056,19	95,04%	0,70	1537	956,81	71,51%	0,55
104	720	384,43	45,71%	1,95	3854	2071,67	95,76%	0,70	1585	982,44	73,42%	0,57
101	737	393,51	46,79%	1,95	3869	2079,68	96,13%	0,70	1634	1008,60	75,38%	0,58
98	787	420,21	49,97%	1,94	3899	2095,70	96,87%	0,71	1670	1027,83	76,82%	0,59
95	885	472,53	56,19%	2,00	3908	2100,50	97,09%	0,71	1676	1031,03	77,06%	0,59
92	956	510,44	60,70%	2,02	3910	2101,57	97,14%	0,71	1678	1032,10	77,14%	0,59
89	1004	536,07	63,75%	2,01	3912	2102,64	97,18%	0,71	1680	1034,23	77,29%	0,59
86	1045	557,96	66,35%	2,01	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
83	1077	575,05	68,38%	2,02	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
80	1104	589,46	70,09%	2,02	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
77	1126	601,21	71,49%	2,02	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
74	1144	610,82	72,63%	2,02	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
71	1150	614,02	73,02%	2,01	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
68	1153	615,63	73,21%	2,01	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59
65	1154	616,16	73,27%	2,01	3914	2103,71	97,24%	0,71	1682	1034,23	77,29%	0,59

TABLEAU 5.2 – *Caractéristiques des trois zones utilisées en fonction de différentes valeurs pour le nombre de sommets définis.* n_{Hc} *le nombre d'hypercuboïdes,* v *le volume,* $\%v$ *le pourcentage de volume par rapport au volume réel et* Δ_c *l'erreur moyenne de resynthèse.*

volume important ; le volume occupé est de 85,39% lorsque l'on impose que tous les sommets soient définis. Pour les zones 2 et 3, le volume occupé tend à se stabiliser au dessous de 98 sommets définis, pour la zone 1 vers 77 sommets.

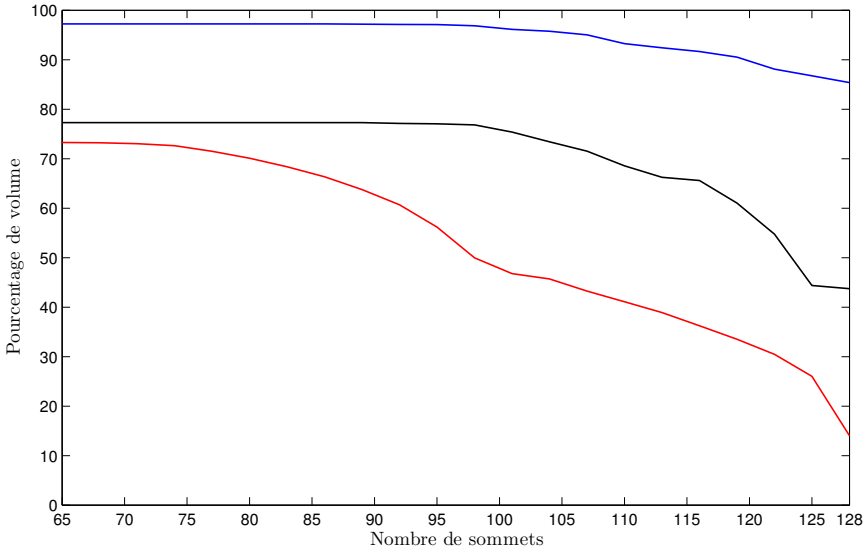


FIGURE 5.13 – *Pourcentage de volume couvert par les trois zones étudiées en fonction du nombre de sommets utilisés. La courbe rouge correspond à la zone 1, la bleue à la zone 2 et la noire à la zone 3.*

Le choix du nombre de sommets synthétisables doit être déterminé de façon à obtenir un volume occupé assez important sans trop dégrader l'erreur de resynthèse. Pour les zones bien couvertes (là où la zone de frontière n'est pas importante), le nombre de sommets synthétisables n'influe plus le volume lorsqu'il est inférieur à 75% des sommets de l'hypercuboïde (soit environ 96 sommets synthétisables). Lors du suivi de la frontière, on constate que le nombre d'hypercuboïdes avec un petit nombre de sommets synthétisables est important. Bien sûr, si le volume minimal d'un hypercuboïde diminue le nombre d'hypercuboïdes avec tous les sommets synthétisables devraient augmenter. Afin de conserver un volume suffisant et une erreur de resynthèse pas trop importante, nous avons choisi 77 sommets synthétisables comme nombre minimal de sommets pour conserver un hypercuboïde.

5.4.3 Seuils de subdivision et précision acoustique

Le codebook doit offrir une représentation compacte et fidèle de la relation articulatoire-acoustique afin de permettre une inversion rapide et conforme à la réalité. Le choix des différents seuils doit donc être un compromis entre le niveau de subdivision et la précision acoustique. La taille du codebook résultant dépend également des choix des paramètres de construction. Le temps de construction peut aussi être pris en compte, mais ce n'est pas le plus important car le codebook n'est construit qu'une seule fois.

Le choix du seuil de subdivision maximal doit permettre de conserver une bonne résolution de l'espace articulatoire en éliminant les régions non-définies. La frontière de l'espace articulatoire est la limite entre les zones synthétisables et non-synthétisables. Dans cette zone, les subdivisions pourraient

se poursuivre à l'infini, alors que ces zones ne sont pas les plus intéressantes pour l'inversion et sont plus difficilement linéarisables.

Les seuils permettant de conserver un codebook suffisamment compact et une bonne résolution de l'espace articulatoire sont déterminés expérimentalement en construisant des codebooks avec différents seuils acoustiques et résolutions articulatoires. On mesure le volume du codebook, le nombre d'hypercuboïdes et la précision acoustique moyenne de la resynthèse.

Les premiers tests sont réalisés sur les trois zones utilisées dans le paragraphe 5.4.1. Elles ont un comportement différent les unes des autres ; la zone 1 est située sur la frontière et la zone 2 ne touche que très peu la frontière. Les détails des résultats en fonction des différents volumes minimaux et seuils acoustiques sont présentés dans les tableaux B.1, B.2 et B.3 dans l'annexe B. Pour nos tests, nous avons fait varier le volume minimal d'un hypercuboïde et le seuil acoustique. Le volume minimal pour un hypercuboïde est de 4,27 $\left(\frac{6^7}{2^{16}}\right)$ soit seize subdivisions maximum jusqu'à 0,033 $\left(\frac{6^7}{2^{23}}\right)$ soit vingt-trois subdivisions maximum. Les graphiques 5.14, 5.15 et 5.16 présentent le nombre d'hypercuboïdes, la densité, le pourcentage de volume occupé et l'erreur de resynthèse moyenne pour un seuil acoustique de 2 et le volume minimal variant de 0,033 à 4,27. L'échelle logarithmique est utilisée pour représenter le volume minimal. Chaque graduation correspond au doublement du volume minimal quand on se déplace vers la droite de cet axe.

La figure 5.14 présente l'évolution du nombre d'hypercuboïdes dans le codebook et la densité en fonction du volume minimal d'un hypercuboïde pour un seuil acoustique de 2. Ce seuil est choisi arbitrairement ; il correspondent à la distance euclidienne entre le vecteur cepstral obtenu par la synthèse et celui par interpolation linéaire. La densité est calculée en divisant le nombre d'hypercuboïdes par le volume. Lorsque le volume minimal devient de plus en plus petit, le nombre d'hypercuboïdes augmente. Cette augmentation est plus importante pour la zone 2 (courbe bleue) qui comporte toujours un nombre d'hypercuboïdes supérieur aux deux autres zones. En effet, la zone 2 a un volume nettement supérieur aux autres. Pour pouvoir comparer les trois zones, il faut utiliser une mesure de densité. La figure 5.14 droite représente l'évolution de la densité en fonction du volume minimal. Nous avions vu que la zone 2 comportait le plus grand nombre d'hypercuboïdes mais c'est la zone 1 (courbe rouge) qui a une densité élevée. En d'autres termes, il faut plus d'hypercuboïdes par unité de volume pour décrire la zone 1. Ce qui est tout à fait réaliste car la zone 1 est une zone possédant une grande région de frontières alors pour la zone 2 la frontière est très réduite. On voit bien que pour approcher plus finement la frontière le nombre d'hypercuboïdes est important pour décrire un petit volume car dans ces régions le comportement de la relation articulatoire-acoustique n'est pas linéaire.

Le deuxième paramètre calculé est le volume occupé par le codebook afin d'évaluer la couverture de l'espace articulatoire. Dans le paragraphe 5.3.2.1, nous avons évalué le volume des trois zones. C'est à partir du volume théorique que l'on calcule le pourcentage de volume expliqué (voir figure 5.15). Tout d'abord, nous constatons que les trois zones ont un comportement différent. La zone 1 (courbe rouge) qui a pourtant le volume le plus petit est la moins bien couverte. Ceci s'explique car elle comporte une grande région de frontière. Même en réduisant le volume minimal la

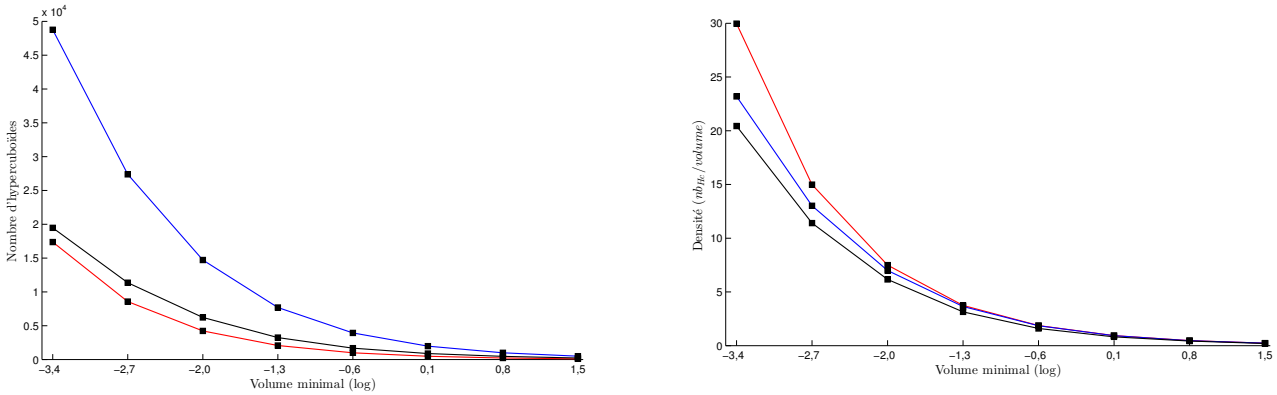


FIGURE 5.14 – *Nombre d'hypercuboïdes (graphique de gauche) et densité (graphique de droite) en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire).*

couverture atteinte ne dépasse pas 70%. Ce qui signifie que si l'on souhaite approcher finement cette zone le volume minimal doit être extrêmement petit. Or les zones situées sur la frontière ne sont pas les plus importantes pour le codebook, car elles sont rarement atteintes. Par contre, il ne faut pas que le volume manquant pour les autres zones soit trop importante. Prenons la zone 2 (courbe bleue), qui possède peu de frontière. Dans ce cas le volume occupé est d'environ 98% quel que soit le volume minimal. Cela nous indique que la couverture de l'espace articulatoire dans les zones hors frontière est bonne. Par contre la diminution du volume minimal devrait nous permettre d'obtenir une meilleure resynthèse.

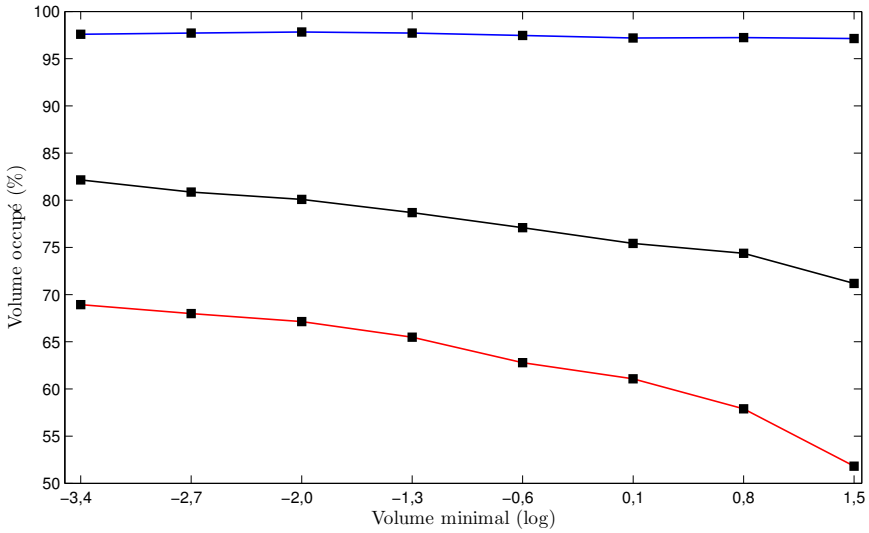


FIGURE 5.15 – *Pourcentage de volume couvert en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire).*

L'erreur de resynthèse moyenne est calculée pour 100000 vecteurs articulatoires choisis aléatoirement. Pour chaque vecteur articulatoire, on calcule la distance entre le vecteur acoustique obtenu par interpolation linéaire et celui obtenu par le synthétiseur articulatoire. La figure 5.16 présente l'évolution de l'erreur de synthèse en fonction du volume minimal pour un hypercuboïde. La diminution du volume minimal permet de réduire l'erreur de resynthèse. Cette diminution est plus grande pour la zone 1 (courbe rouge). De plus, la zone 1 est celle qui présente l'erreur la plus importante. Cela confirme que dans les zones de frontière, le comportement de la relation articulatoire-acoustique n'est pas linéaire.

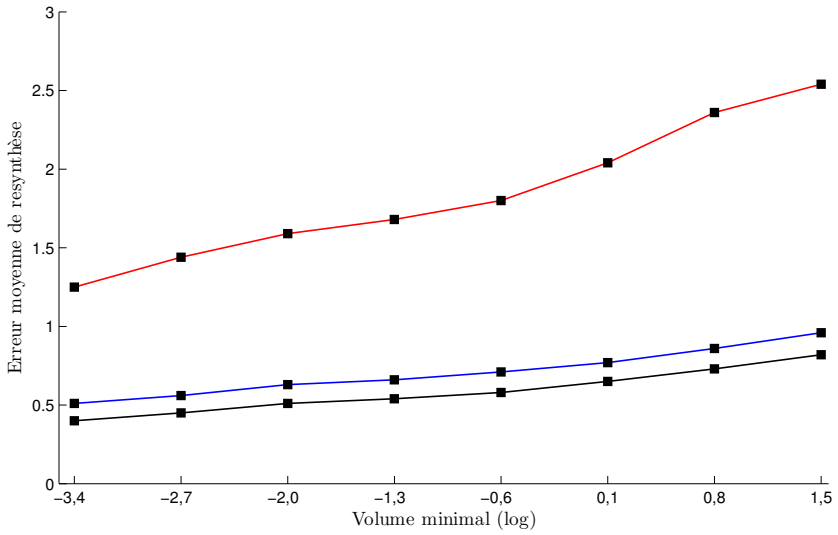


FIGURE 5.16 – *Erreur moyenne de resynthèse (graphique de droite) en fonction du volume minimal pour un hypercuboïde pour les trois zones : zone 1 (courbe rouge), zone 2 (courbe bleue) et zone 3 (courbe noire).*

Nous avons vu qu'en fonction de la zone du codebook, les résultats ne sont pas les mêmes. Un volume minimal très petit permet une meilleure précision mais fait croître très rapidement le nombre d'hypercuboïdes et ainsi la taille du codebook. La majorité des erreurs se concentre dans les zones de frontières.

Maintenant, nous construisons des codebooks entiers pour vérifier les observations précédentes. Le tableau 5.3 présente les résultats sur l'ensemble de l'espace articulatoire en fonction de différentes valeurs pour le seuil acoustique et le volume minimal d'un hypercuboïde. On détermine le nombre d'hypercuboïdes, le volume total du codebook et les erreurs moyennes calculées sur 100000 points en fonction de différentes valeurs pour le volume minimal d'un hypercuboïde et pour le seuil acoustique.

Tout d'abord, on constate que lorsque le seuil acoustique et le volume minimal diminuent le volume du codebook augmente. Le volume stagne lorsque le volume minimal est inférieur ou égal à 2

pour les trois seuils acoustiques utilisés et le nombre d'hypercuboïdes croît alors légèrement lorsque le volume diminue.

On constate également que lorsque l'on augmente la précision l'erreur acoustique moyenne de resynthèse moyenne diminue également.

Volume minimal (subdivisions)	Seuil acoustique	Nombre d'hyper-cuboïdes	Volume total	Erreur moyenne de resynthèse
4,27 (16)	5	35523	188953,03	1,48
	2	44290	189179,05	1,42
	1	44297	189179,05	1,42
1,07 (18)	5	105218	185250,73	1,27
	2	177981	192009,44	1,15
	1	179985	192272,82	1,14
0,53 (19)	5	176980	188209,64	1,20
	2	352905	193635,91	1,06
	1	363062	193936,84	1,05

TABLEAU 5.3 – *Caractéristiques des différents codebooks construits avec différentes valeurs pour la taille minimale et le seuil acoustique (nombre d'hypercuboïdes, volume total, erreur moyenne de resynthèse).*

5.5 Conclusion

Nous avons présenté dans ce chapitre la méthode utilisée pour la construction de codebook. Cette méthode repose sur l'exploration récursive de l'espace articulatoire à l'aide d'hypercuboïdes proposée par Potard [Pot08a]. À la différence de Potard les vecteurs acoustiques ne sont pas les formants mais les coefficients cepstraux. L'utilisation des coefficients cepstraux impliquera des changements importants lors de l'inversion.

Nous avons dans ce chapitre que la construction d'un codebook impose le choix de plusieurs paramètres. La méthode étant récursive, le premier critère d'arrêt est la taille minimale d'un hypercuboïde. D'autres paramètres permettent de modifier la précision du codebook. En effet, dans chaque hypercuboïde la relation articulatoire-acoustique est modélisée par une fonction linéaire calculée à partir de la matrice jacobienne au centre à une distance proportionnelle au rayon. Nous avons ainsi déterminé la valeur optimale.

L'étude des différentes zones de l'espace articulatoire a permis d'identifier une partie des erreurs, ou plutôt le comportement de la relation en fonction de la proximité avec la frontière, qui concentre la majorité des erreurs.

Nous avons observé que l'espace articulatoire contenu dans le codebook n'est pas complet, ce qui est dû au suivi de la frontière de l'espace synthétisable. En effet, l'espace articulatoire manquant se situe au niveau de la frontière. La réduction du volume manquant est possible mais engendrerait

un grand nombre de calculs et un grand nombre d'hypercuboïdes dans des zones où la relation articulatoire-acoustique n'est pas linéaire et rarement atteinte par le locuteur. Ces zones manquantes ne poseront pas de problème pour l'inversion de voyelles mais pourraient le devenir si nous inversions les consonnes et les voyelles.

À ce stade nous avons construit le codebook et l'étape suivante est l'inversion acoustique-articulatoire. Nous recherchons dans le codebook les configurations articulatoires permettant de produire un vecteur acoustique fourni en entrée. L'utilisation des coefficients cepstraux entraîne l'utilisation de la programmation quadratique. De plus, une étape d'adaptation entre le signal réel et synthétique est nécessaire afin de limiter les différences entre les deux types de signaux et de permettre l'inversion de données réelles.

Chapitre 6

Inversion par codebook hypercuboïdal

La méthode d'inversion utilisée cherche à déterminer, à l'aide du codebook, un ensemble de vecteurs articulatoires dont l'image acoustique est proche d'un vecteur acoustique donné. Les coefficients cepstraux sont utilisés comme paramètres acoustiques.

L'inversion d'un signal de parole passe par une phase d'adaptation entre les paramètres acoustiques réels et synthétiques. Contrairement à d'autres études, nous pouvons dans notre cas comparer les signaux naturels aux signaux synthétiques car nous disposons des images cinéradiographiques à l'origine du signal de parole naturelle qui a été enregistré en même temps que les images.

Enfin, différentes stratégies nous permettent d'optimiser la recherche de solutions dans le codebook en trouvant rapidement les hypercuboïdes candidats.

6.1 Principe général

Dans cette partie, nous allons décrire la méthode d'inversion d'un vecteur acoustique donné, c'est-à-dire l'obtention d'un ensemble de vecteurs articulatoires dont l'image acoustique est proche du vecteur acoustique à inverser. L'inversion est réalisée uniquement à un instant précis, l'exploration des différents hypercuboïdes nous permet d'obtenir plusieurs solutions.

Soit c un vecteur cepstral donné et f la fonction du synthétiseur articulatoire. L'inversion acoustique-articulatoire consiste alors à chercher les vecteurs articulatoires $x \in A_r$ tels que leur image acoustique $f(x) \in A_c$ soit proche du vecteur cepstral c donné. En d'autres termes, on cherche $x \in A_r$ tel que :

$$f(x) = c \tag{6.1}$$

L'équation (6.1) correspond à un système de N équations à M inconnues où M et N les dimensions des espaces acoustique et articulatoire respectivement.

Ouni [Oun01] et Potard [Pot08a] utilisent comme paramètres acoustiques les trois premiers

formants et comme paramètres articulatoires les sept paramètres contrôlant le modèle de Maeda. L'équation (6.1) est donc un système sous-déterminé de trois équations à sept inconnues. Une solution est formée d'une solution particulière et d'un vecteur de l'espace nul. La résolution est réalisée par la méthode SVD (décomposition en valeurs singulières).

Dans notre cas, les formants ne sont plus utilisés comme paramètres acoustiques. À la place, nous avons utilisé les trente premiers coefficients cepstraux (sans le premier coefficient qui est relatif à l'énergie). Le système est alors sur-déterminé, car $N > M$. En général, un tel système ne possède pas de solution. Une solution approchée x peut être trouvée au sens des moindres carrés. On définit le résidu z tel que :

$$z = f(x) - c \tag{6.2}$$

La solution du système (6.1) au sens des moindres carrés est obtenue en minimisant $||z||^2$. La recherche d'une solution dans un hypercuboïde H_c consiste alors à résoudre :

$$\min_{x \in H_c} ||f(x) - c||^2 \tag{6.3}$$

De plus, les solutions doivent appartenir à l'hypercuboïde considéré, ce qui conduit aux deux contraintes :

$$\begin{cases} x_i < P_{0i} + \dfrac{r_i}{2} \\ x_i > P_{0i} - \dfrac{r_i}{2} \end{cases} \tag{6.4}$$

où $r \in \mathbb{R}^M$ la taille et P_0 le centre de l'hypercuboïde H_c. Dans chaque hypercuboïde f est linéarisée et ainsi $f(x)$ est approchée par Ax (avec A une matrice). La distance acoustique à minimiser est donc :

$$\begin{aligned} D(x) &= (c - Ax)^\intercal (c - Ax) \\ &= c^\intercal c + x^\intercal \underbrace{A^\intercal A}_{H} x - (\underbrace{c^\intercal A}_{q} x + x^\intercal \underbrace{A^\intercal c}_{q^\intercal}) \\ &= x^\intercal H x - 2q^\intercal x + s \end{aligned} \tag{6.5}$$

L'équation (6.3) sous les contraintes (6.4) est résolue par une méthode de programmation quadratique proposée par Goldfarb et Idnani [GI83] qui minimise une fonction définie par :

$$D(x) = q^\intercal x + \frac{1}{2} x^\intercal H x \tag{6.6}$$

avec les contraintes :

$$v(x) = C^\intercal x - B \geq 0 \tag{6.7}$$

où C et B sont obtenus à partir de P_0 et r.

6.2 Comparaison entre les vecteurs cepstraux naturels et synthétiques

L'objectif est d'inverser de la parole naturelle, ce qui induit une recherche dans le codebook consistant à comparer les vecteurs cepstraux naturels et synthétiques. Cette comparaison ne peut pas être effectuée directement. En effet, la source n'est pas prise en compte lors de la synthèse acoustique. En outre, il existe des disparités entre le conduit vocal du locuteur et la géométrie du modèle articulatoire. Les vecteurs cepstraux calculés sur la parole naturelle et ceux calculés sur la parole synthétique sont donc très différents et il n'est pas possible de réaliser directement une recherche d'un vecteur acoustique réel dans le codebook construit à partir de données synthétiques.

Traditionnellement, un liftre [MSS91] est appliqué afin d'atténuer la contribution des premiers coefficients liés à la pente spectrale (*spectral tilt*) et des derniers coefficients liés aux harmoniques de la source d'excitation.

Cependant, une observation plus précise des spectres naturels et de ceux produits par le synthétiseur articulatoire (spectres lissés cepstralement) montre qu'ils sont aussi légèrement décalés en fréquence. Ceci est sans doute lié aux erreurs du modèle (la longueur du conduit vocal) et de la synthèse acoustique. Comme nous disposons des films cinéradiographiques et du signal associé, il est possible de comparer les spectres naturels et synthétiques correspondant à la même forme de conduit vocal.

Deux types de transformations ont été étudiés :
- une transformation affine des coefficients cepstraux,
- un déplacement en fréquence, ou « *frequency warping* », via une transformation bilinéaire.

Les différentes transformations étudiées seront effectuées à partir des formes correspondantes aux voyelles présentes dans le codebook. Nous présenterons, tout d'abord, les différentes voyelles extraites du corpus.

6.2.1 Voyelles extraites du corpus

Notre corpus se compose d'images cinéradiographiques ainsi que du signal de parole correspondant. Après une synchronisation entre les images et le son, il nous est possible d'associer une forme articulatoire du conduit vocal avec le spectre naturel correspondant. La figure 6.1 montre un exemple de synchronisation entre le signal et les images. Les lignes bleues en pointillés sont les instants où il existe une image.

Les différentes analyses seront effectuées à partir de formes correspondant aux voyelles. Nous avons donc extrait les différentes formes articulatoires des voyelles issues de notre corpus. La fréquence d'acquisition des images cinéradiographiques n'est pas très élevée, le nombre de formes de voyelles n'est alors pas très important. Sur la figure 6.1, les instants associés à une voyelle sont marqués par

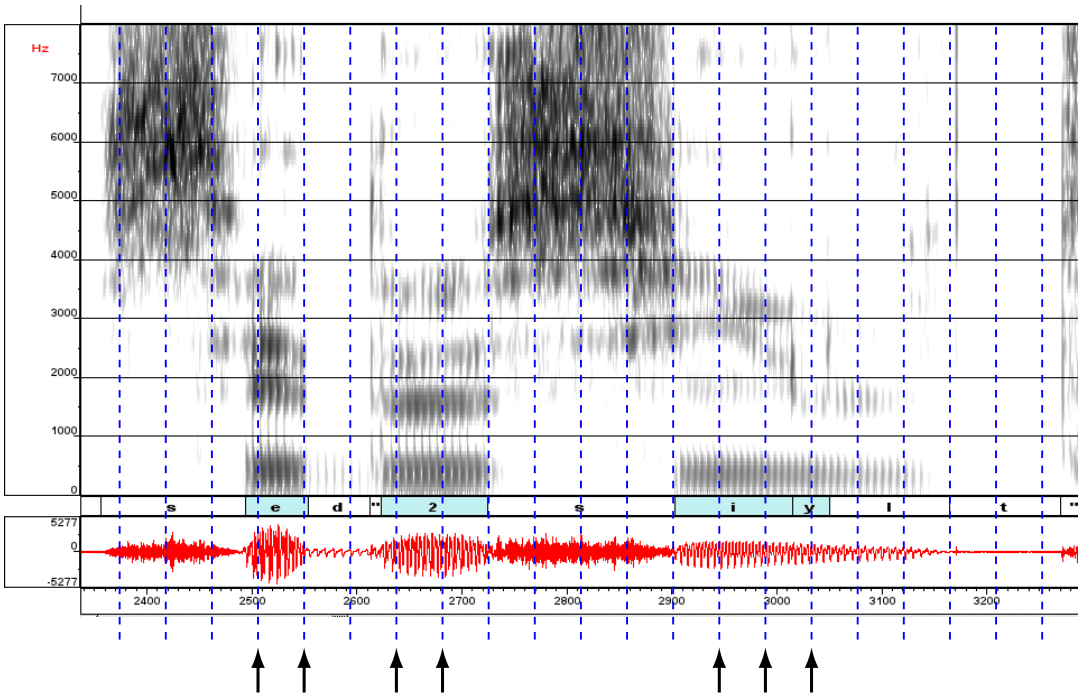


FIGURE 6.1 – *Spectrogramme issu du corpus. Les lignes bleues en pointillés représentent les instants où une image cinéradiographique est disponible. Les flèches noires désignent les instants correspondants aux voyelles.*

une flèche noire. Nous constatons que l'image peut se situer au début ou à la fin d'une voyelle. Dans ce cas la forme du conduit vocal sera très proche de celle de la consonne qui la précède ou la suit. Ces formes n'ont pas été prise en compte.

De plus, le corpus n'étant pas phonétiquement équilibré, les voyelles ne sont pas représentées dans les mêmes proportions (voir tableau 6.1). Ces différences s'expliquent également par la durée des voyelles, le /y/ est court et n'est pas forcément capturé alors que le /ɛ/ est plus long et peut être capturé par plusieurs images. Au total, nous disposons de 133 formes articulatoires de voyelles associées au signal original.

Phonème	Nombre d'images
/e/	9
/ø/	19
/i/	22
/y/	7
/ɛ/	30
/a/	8
/u/	38
Moyenne	133

TABLEAU 6.1 – *Répartition des 133 voyelles capturées dans le corpus.*

6.2.2 Adaptation cepstrale

Mokhtari et al. [MKTH04] ont réalisé une étude portant sur la comparaison entre des données réelles et synthétiques dans une situation similaire puisque des images IRM et le signal de parole correspondant étaient disponibles pour le même locuteur. Mokhtari et ses collègues ont utilisé une inversion par prédiction linéaire et ont compensé les fréquences des formants et les largeurs de bande via une transformation affine afin de garantir une meilleure correspondance entre les fonctions d'aire réelles et inversées. Les coefficients de la transformation affine, un pour chaque fréquence de formant et pour chaque largeur de bande, ont été obtenus à partir d'un ensemble de cinq voyelles.

De façon similaire, nous considérons une transformation affine afin de rapprocher les coefficients cepstraux extraits des signaux naturels et synthétiques. La régression linéaire est effectuée sur chaque coefficient cepstral séparément. Ainsi chaque coefficient synthétique est approché par une transformation affine du coefficient calculé sur le signal réel.

En d'autres termes, le $n^{ième}$ coefficient cepstral synthétique, c'_n, est approché par la transformation affine suivante :

$$c'_n \approx a_n.c_n + b_n \tag{6.8}$$

où c_n est le coefficient calculé sur le signal de parole naturelle. Les coefficients de la régression linéaire a_n et b_n sont déterminés en minimisant l'erreur E_n sur l'ensemble des coefficients cepstraux obtenus pour les voyelles. L'erreur est définie par :

$$E_n = \sum_k ||c'_{nk} - (a_n.c_{nk} + b_n)||^2 \tag{6.9}$$

où n est le numéro du coefficient, k le numéro de la forme utilisée et $||.||$ la norme euclidienne.

Pour chaque forme articulatoire sélectionnée, les coefficients cepstraux synthétiques sont obtenus après synthèse articulatoire à partir de la fonction d'aire issue des contours du conduit vocal. Les coefficients cepstraux naturels issus du fichier sonore sont calculés à partir d'une analyse LPC d'ordre 18. Pour notre analyse, nous utilisons 29 coefficients ($1 \leq n \leq 29$) ; le premier coefficient c_0 est exclu de l'analyse car il est relatif à l'énergie.

L'erreur E_n est minimisée par la méthode des moindres carrés sur l'ensemble des 133 voyelles du corpus. Le tableau 6.2 présente les différentes valeurs obtenues pour les coefficients a_n et b_n pour $1 \leq n \leq 29$.

Afin d'évaluer l'intérêt de cette transformation, nous calculons l'erreur moyenne sur l'ensemble des 133 voyelles entre les pics spectraux calculés via les coefficients cepstraux réels et synthétiques (voir figure 6.2). On constate que l'erreur est minimale avec 29 coefficients adaptés. De plus, lorsque 13 coefficients sont utilisés, la valeur de l'erreur est proche de l'erreur minimale.

n	a_n	b_n
1	0,138	-2,615
2	0,292	0,011
3	0,391	0,886
4	0,050	-0,197
5	0,047	1,043
6	0,585	0,405
7	0,061	0,613
8	0,075	0,337
9	0,184	0,631
10	0,009	0,804

n	a_n	b_n
11	0,390	0,288
12	0,190	0,201
13	0,393	-0,262
14	-0,063	0,235
15	-0,048	0,240
16	0,291	0,226
17	0,614	0,444
18	-0,987	-0,741
19	-0,139	-0,297
20	-0,675	-0,274

n	a_n	b_n
21	0,254	0,445
22	0,019	-0,170
23	0,252	0,182
24	0,142	-0,023
25	0,065	-0,087
26	-0,311	-0,236
27	-0,204	-0,277
28	-0,334	-0,092
29	-0,015	-0,253

TABLEAU 6.2 – *Coefficients a_n et b_n issus de l'adaptation cepstrale.*

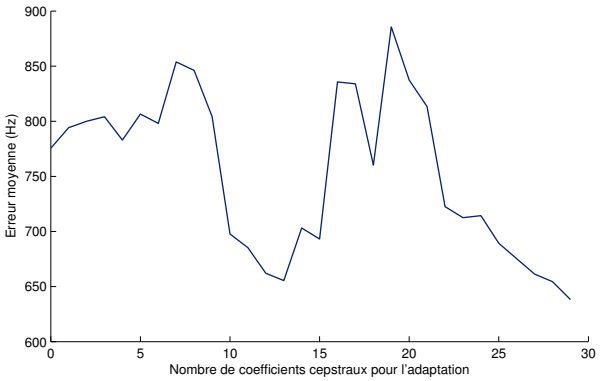
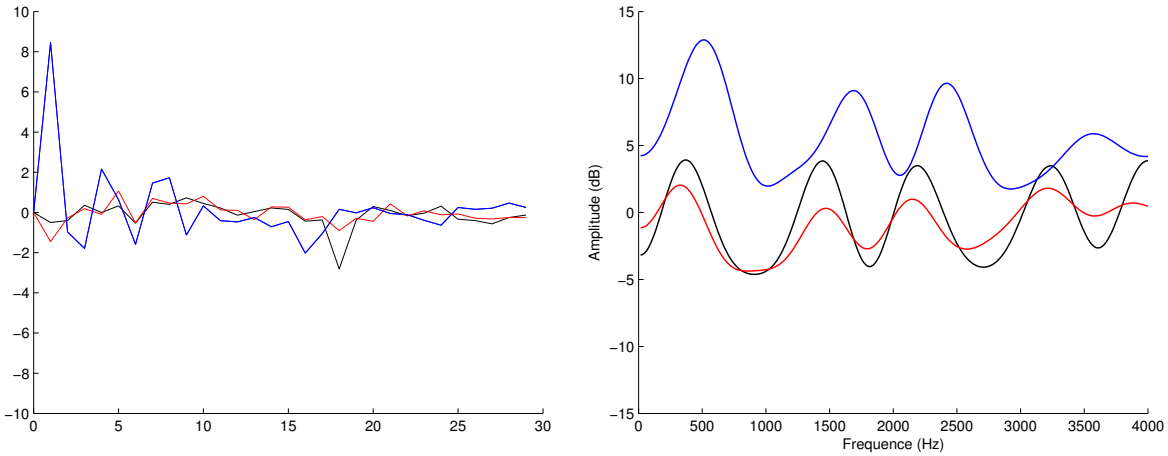


FIGURE 6.2 – *Erreur moyenne sur l'ensemble des voyelles entre les trois premiers pics des spectres lissés cepstralement issus des coefficients cepstraux synthétiques et réels en fonction du nombre de coefficients cepstraux utilisés pour l'adaptation.*

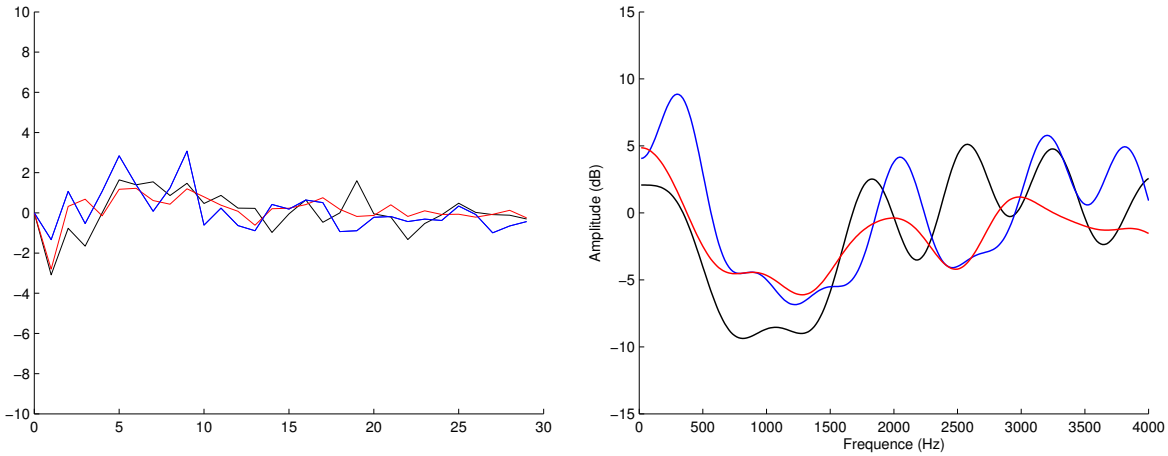
Les figures 6.3, 6.4 et 6.5 montrent les différents cas que l'on obtient en réalisant une transformation affine des 29 coefficients cepstraux. Le premier cas (voir figure 6.3) est un exemple où l'adaptation a permis de rapprocher les deux signaux. En effet, les pics spectraux naturels (courbe bleue) après transformation (courbe rouge) se sont rapprochés des pics spectraux synthétiques (courbe noire). Nous remarquons tout de même que les pics obtenus après transformation sont moins marqués que les pics synthétiques mais sont placés au bon endroit. Le deuxième cas (voir figure 6.4) concerne une transformation qui ne fournit pas le résultat escompté. En d'autres termes, les pics spectraux après transformation ne sont pas du tout placés au bon endroit. Le dernier cas de figure (voir figure 6.5) est aussi un des points faibles de cette technique, le spectre obtenu par transformation est trop plat. Les pics sont alors très mal définis, l'adaptation est donc un échec.



L'adaptation permet de déplacer les pics spectraux.

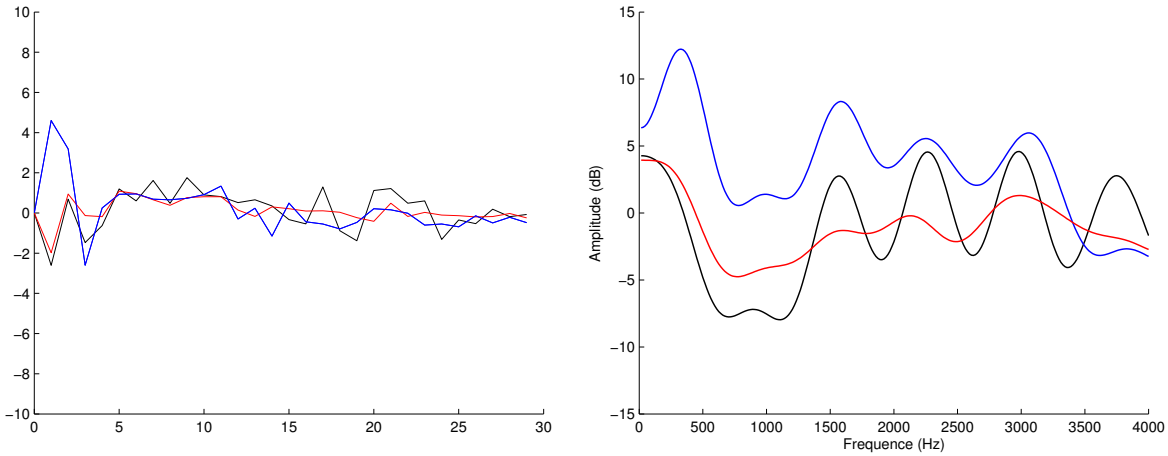
FIGURE 6.3 – *Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle).*

Nous avons vu que la transformation affine des coefficients permet effectivement de rapprocher les coefficients cepstraux naturels et synthétiques mais elle ne garantit pas que le spectre résultant soit exploitable. Dans certains cas, l'adaptation se comporte correctement puisque les pics du spectre naturel après transformation affine sont proches des pics synthétiques. Mais dans d'autres cas, l'adaptation échoue dans le rapprochement des deux spectres, les pics correspondant aux formants ont presque disparu car le spectre obtenu est trop lisse. Malheureusement même si le modèle articulatoire a été construit spécifiquement pour un locuteur une parfaite correspondance ne peut être atteinte et la minimisation tend à trop lisser le spectre naturel pour le rapprocher du spectre synthétique en moyenne.



L'adaptation ne fournit pas de résultat exploitable.

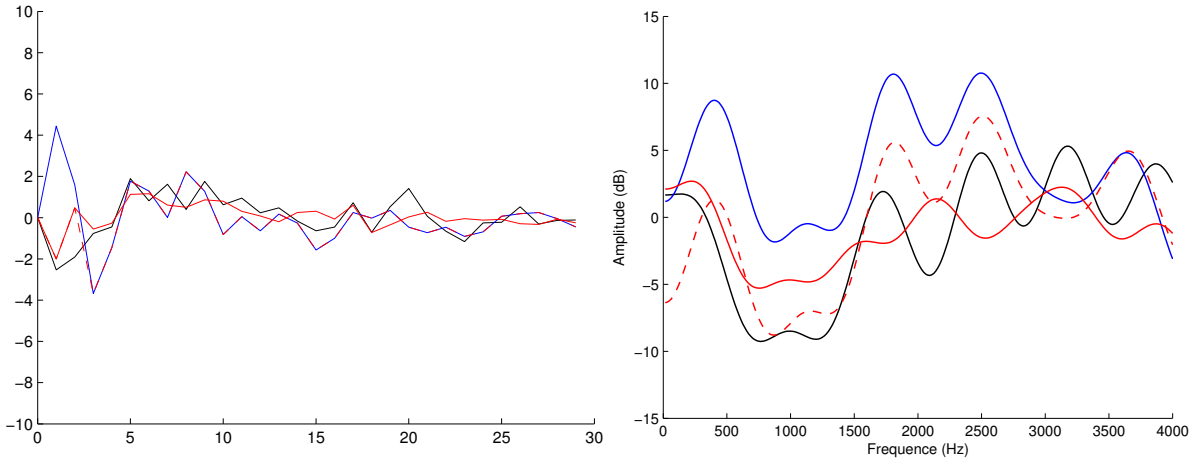
FIGURE 6.4 – *Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle).*



L'adaptation donne un spectre trop plat. Les pics formantiques ne sont pas bien définis.

FIGURE 6.5 – *Adaptation cepstrale par transformation affine des coefficients cepstraux (graphique gauche) et spectres correspondant lissés cepstralement (graphique droit). Courbes bleues (parole naturelle), courbes noires (parole synthétique) et courbes rouges (transformation affine des coefficients cepstraux issus de la parole naturelle).*

Nous avons utilisé 29 coefficients pour réaliser la transformation. En fait, le lissage est d'autant plus important lorsque le nombre de coefficients cepstraux adaptés est élevé. La figure 6.6 montre un exemple de transformation avec 2 et 29 coefficients cepstraux. Nous constatons que l'utilisation de deux coefficients permet de conserver la structure formantique mais les pics ne se sont pas assez déplacés.



(a) Coefficients cepstraux (b) Spectres lissés

FIGURE 6.6 – *Coefficients cepstraux et spectres lissés cepstralement après adaptation avec 29 et 2 coefficients : ligne bleue (naturel), ligne noire (synthétique), ligne rouge (adapté avec 29 coefficients) et ligne rouge pointillée (adaptée avec 2 coefficients).*

Il apparaît clairement que même si les pics spectraux sont déplacés correctement en utilisant vingt-neuf coefficients la structure formantique est dégradée ce qui compromet l'exploration du codebook pendant l'inversion.

L'adaptation est donc pertinente quand elle s'applique sur les premiers coefficients pour capturer la pente spectrale (*spectral tilt*) de la parole mais est contre-productive si elle s'applique sur tous les coefficients cepstraux. La source de la phonation a pour effet de diminuer la pente spectrale de 12dB par octave et le rayonnement acoustique aux lèvres d'augmenter de 6 dB par octave ainsi l'effet global sur la pente spectrale est de -6dB par octave [Man13]. Cependant la pente spectrale n'est pas constante et la transformation affine des coefficients cepstraux est parfois incapable de la compenser.

Dans la section suivante nous allons étudier la distorsion fréquentielle ou *frequency warping* pour déplacer les pics de façon à obtenir une meilleure correspondance entre les spectres naturel et synthétique sans aplanir le spectre.

6.2.3 Distorsion fréquentielle (ou *frequency warping*)

L'observation des spectres naturels et synthétiques montre que les deux spectres sont souvent décalés en fréquence. L'idée est de décaler l'échelle de fréquence afin de faire correspondre les pics spectraux sans les aplanir comme pour l'adaptation cepstrale par transformation affine.

Traditionnellement, la distorsion fréquentielle (ou *frequency warping*) est utilisée en reconnaissance automatique de la parole pour effectuer l'adaptation au locuteur [EG96]. Dans notre cas le modèle articulatoire a été construit à partir d'images du locuteur pour lequel la parole est inversée mais il subsiste encore des différences entre les signaux synthétiques et naturels. La distorsion fréquentielle est donc destinée à compenser ces différences de fréquences dues aux erreurs du modèle ou au calcul de la ligne médiane utilisée pour découper le conduit vocal en tubes élémentaires.

La normalisation de la longueur du conduit vocal s'effectue en réalisant un décalage de l'échelle de fréquence du spectre [PMSN01]. La fonction de décalage g_α est supposée inversible (strictement monotone et continue) :

$$\begin{array}{rcl} g_\alpha : [0,\pi] & \rightarrow & [0,\pi] \\ \omega & \rightarrow & \tilde{\omega} = g_\alpha(\omega) \end{array} \tag{6.10}$$

Un outil classique pour la normalisation de la longueur du conduit vocal est la transformation bilinéaire (Oppenheim et Johnson [OJ72]). La nouvelle fréquence $\tilde{z}$ est donnée par l'expression suivante :

$$\tilde{z} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad -1 < \alpha < 1 \tag{6.11}$$

où α est le paramètre de décalage. Le coefficient α modifie toute l'échelle des fréquences [OJ72]. La transformation de la fréquence est obtenue en faisant les substitutions $z = e^{j\omega}$ et $\tilde{z} = e^{j\tilde{\omega}}$ dans l'équation (6.11). On obtient :

$$\tilde{\omega} = \omega + 2\tan^{-1}\frac{\alpha \sin\omega}{1 - \alpha\cos\omega} \tag{6.12}$$

La figure 6.7 montre plusieurs représentations de la formule (6.12) pour plusieurs valeurs de α. Le cas $\alpha = 0$ correspond à l'application de la fonction identité c'est-à-dire pas de transformation de l'échelle des fréquences.

Une autre solution pour réaliser le *frequency warping* est d'utiliser une fonction linéaire par morceaux. En comparaison avec la transformation bilinéaire qui affecte l'ensemble de l'échelle de fréquence, la transformation linéaire par morceaux peut être facilement focalisée sur le domaine spectral correspondant aux formants F1 à F3, en d'autres termes les formants les plus importants du point de vue de l'inversion. Cependant ceci requiert l'ajustement d'au moins deux paramètres (un pour décomposer le domaine fréquentiel et un pour fixer le niveau de décalage). Ainsi, nous avons choisi la transformation bilinéaire qui nécessite seulement un paramètre.

La distorsion fréquentielle (ou *frequency warping*) correspond à une transformation linéaire dans le domaine cepstral [Ace90, McD98]. La relation entre les coefficients cepstraux c avant transformation et c_ω après transformation peut s'exprimer par la multiplication par une matrice :

$$c_\omega = L(\alpha)c \tag{6.13}$$

où L la matrice de décalage est fonction de α (voir Acero [Ace90] pour les détails).

Le coefficient α doit être ajusté afin de minimiser les différences entre les pics des spectres

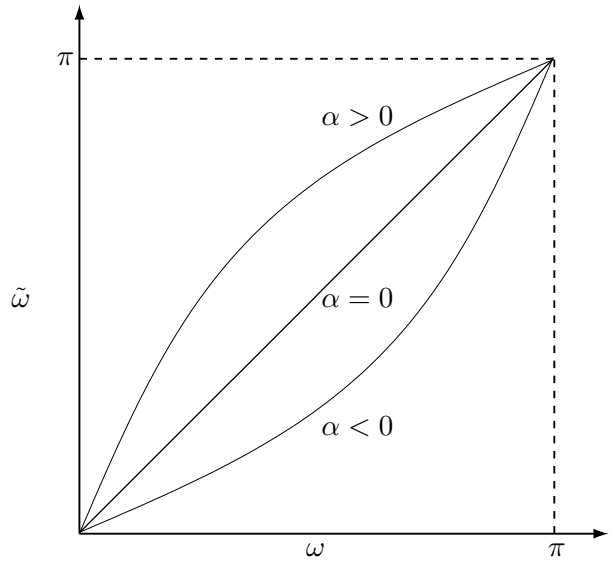


FIGURE 6.7 – *Exemples de transformations bilinéaires pour différentes valeurs de α. Dans le cas $\alpha = 0$ il n'y a pas de décalage.*

naturels et synthétiques. La figure 6.8 présente l'évolution de l'erreur moyenne en fonction de la valeur de α pour l'ensemble des 133 voyelles. Le minimum est atteint pour $\alpha = -0,08$.

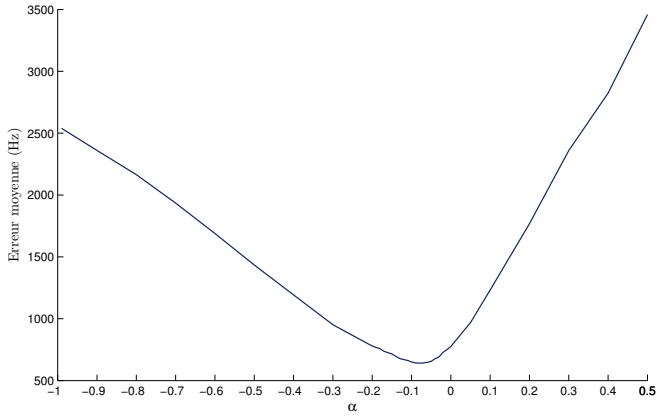


FIGURE 6.8 – *Erreur moyenne sur l'ensemble des voyelles entre les trois premiers pics des spectres lissés cepstralement issus des coefficients cepstraux synthétiques et réels en fonction de la valeur de α.*

La figure 6.9 montre l'effet de la distorsion fréquentielle. On peut voir que la structure des formants est bien préservée ce qui garantit que les régions pertinentes de l'espace articulatoire seront explorées lors de l'inversion.

Lors de l'inversion nous utiliserons donc cette technique car elle permet de converser la structure

formantique et permet de rapprocher le signal naturel du signal synthétique, ce qui permettra une recherche dans le codebook des hypercuboïdes susceptibles de contenir une solution. L'inversion sera effectuée à partir d'un vecteur acoustique composé des coefficients cepstraux calculés à partir du spectre modifié. Sans cette modification du spectre naturel, l'inversion est presque impossible car on ne peut pas garantir que les coefficients cepstraux naturels et synthétiques soient comparables.

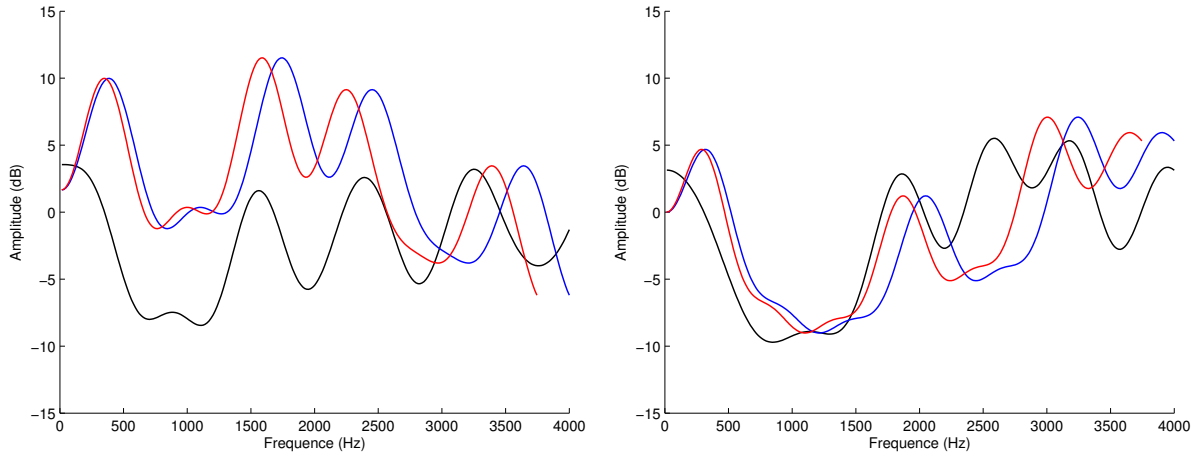


FIGURE 6.9 – *Deux exemples de spectres lissés cepstralement : ligne bleue (naturel), ligne noire (synthétique), ligne rouge (après transformation bilinéaire).*

6.3 Optimisation de la recherche dans le codebook

Notre méthode d'inversion repose sur la recherche de solutions dans le codebook. La rapidité de l'inversion dépend donc fortement du nombre d'hypercuboïdes explorés. En effet, pour chaque hypercuboïde exploré la recherche de solutions par la programmation quadratique est une opération relativement coûteuse en calculs. Il est donc important de déterminer le plus rapidement possible si un hypercuboïde contient ou non une solution. Les hypercuboïdes qui ne contiendraient a priori pas de solutions ne méritent bien sûr pas d'être explorés.

Ouni [Oun01] explorait tous les hypercubes présents dans le codebook. Potard [Pot08a] proposa une organisation de son codebook en fonction des formants afin de réduire le nombre d'hypercuboïdes explorés.

Dans cette partie, nous allons montrer comment le nombre d'hypercuboïdes à explorer a été réduit. Tout d'abord, le codebook sera organisé de la même manière que Potard [Pot08a]. Puis, nous proposerons un algorithme qui permet d'apparier des pics spectraux afin de les comparer à ceux d'un hypercuboïde et de décider s'il est exploré.

6.3.1 Organisation du codebook

Le but est d'organiser le codebook de façon à pourvoir trouver les hypercuboïdes candidats très rapidement. L'idée est d'utiliser les pics spectraux (ou *formants*) pour sélectionner les hypercuboïdes pertinents. Une indexation en fonction des formants permet une organisation efficace des différents hypercuboïdes du codebook. Cependant, il faut souligner que les pics spectraux sont uniquement utilisés pour réaliser l'indexation et non pas comme paramètres acoustiques pour réaliser l'inversion.

Lors du calcul d'un hypercuboïde, on détermine les valeurs minimales et maximales pour les trois premiers formants pour tout vecteur articulatoire présent dans l'hypercuboïde. Ces valeurs sont obtenues à partir des points de test utilisés notamment les sommets. Il est donc possible de déterminer à partir des pics extraits des spectres lissés cepstralement si un vecteur acoustique (par l'intermédiaire des pics spectraux) appartient ou non à un hypercuboïde particulier.

Puisque le domaine de fréquence du deuxième formant est le plus vaste ce formant est utilisé pour réaliser l'indexation. Un hypercuboïde sera conservé uniquement si le pic correspondant au deuxième formant est compris entre le minimum et le maximum de la valeur pour cet hypercuboïde.

La méthode utilisée pour réaliser l'indexation est très simple et très efficace en temps lors de la recherche dans le codebook. Un tableau est créé en indexant les différentes valeurs formantiques possibles. Chaque entrée du tableau correspond à une liste des hypercuboïdes candidats. Bien que la construction du tableau soit longue et que le tableau prenne une place relativement importante, cela permet de réduire le nombre d'hypercuboïdes à explorer pour un surcoût en mémoire raisonnable et un temps de traitement réduit. Le surcoût lié au parcours de la table est négligeable.

6.3.2 Appariement des pics des spectres naturels et synthétiques

Après la sélection des hypercuboïdes basée sur le deuxième formant, l'objectif est de trouver parmi les hypercuboïdes sélectionnés, ceux qui pourraient correspondre à un vecteur cepstral donné. La recherche est effectuée en appariant les pics du spectre naturel avec les formants du spectre synthétique. Cet appariement doit être robuste au décalage en fréquence, à la présence de faux pics, et inversement à l'absence de certains pics. Nous avons conçu un algorithme de programmation dynamique pour réaliser cet appariement. Ici encore, il faut que cette étape d'appariement soit rapide pour ne pas pénaliser la rapidité de l'inversion. La complexité de l'algorithme est faible puisqu'il y a peu de pics à considérer en général.

Voici l'algorithme réalisant l'appariement entre deux ensembles de pics spectraux.

Soient $P = \big[p(m)\big] = p(1) \dots p(m) \dots p(M)$ et $Q = \big[q(n)\big] = q(1) \dots q(n) \dots q(N)$ deux ensembles correspondant aux pics de deux spectres, où $p(m)$ (resp. $q(n)$) correspond au $m^{ième}$ (resp. $n^{ième}$) pic où M (resp. N) le nombre de pics de P (resp. Q).

Pour P et Q, on souhaite extraire un sous-ensemble de pics représenté par les séquences d'indices I et J.

$$I = \big[i(k)\big] = i(1) \dots i(k) \dots i(K) \text{ avec } K \leq M$$

$$J = \big[j(k)\big] = j(1)\dots j(k)\dots j(K) \text{ avec } K \leq N$$

I et J doivent préserver la monotonie de i et de j : $i(k) < i(k+1)$ et $j(k) < j(k+1)$.

Les séquences de pics $\bar{P}$ et $\bar{Q}$ correspondent aux séquences d'indices I et J.

$$\bar{P} = \big[p(i(k))\big] = p\big(i(1)\big)\dots p\big(i(k)\big)\dots p\big(i(K)\big)$$

$$\bar{Q} = \big[q(j(k))\big] = q\big(j(1)\big)\dots q\big(j(k)\big)\dots q\big(j(K)\big)$$

La détermination de i et j nécessite un critère qui minimise la distance entre deux pics de P et Q.

$$d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) = \Big|p\big(i(k)\big) - q\big(j(k)\big)\Big|; \quad \forall k \in K$$

Le critère global est défini par :

$$\min_{K,I,J} \sum_{k=1}^{K} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \tag{6.14}$$

$B\Big(p\big(i(k)\big)\Big) = b$ correspond au terme bonus avec b constant déterminé de façon ad hoc.

Ce problème est résolu à l'aide de la programmation dynamique.

On définit la mesure partielle :

$$D(m,n) = \min_{i,j} \sum_{k=1}^{k^*} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \tag{6.15}$$

avec $i(k^*) = m$ et $j(k^*) = n$.

On décompose la somme en deux parties :

$$\begin{aligned} D(m,n) = \min_{i,j} \Bigg\{ & d\Big(p\big(i(k^*)\big), q\big(j(k^*)\big)\Big) - B\Big(p\big(i(k^*)\big)\Big) \\ & + \sum_{k=1}^{k^*-1} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \Bigg\} \end{aligned} \tag{6.16}$$

$$\tag{6.17}$$

On pose $i(k^*-1) = l_1$ et $j(k^*-1) = l_2$, la formule de récursivité est donnée par :

$$D(m,n) = \min_{l_1<m, l_2<n} \Big\{ d\big(p(m), q(n)\big) - B\big(p(m)\big) + D(l_1, l_2) \Big\} \tag{6.18}$$

La recherche des séquences optimales de pics est détaillée dans l'annexe C.

L'appariement entre les pics s'effectue entre les pics extraits de l'image acoustique du centre des hypercuboïdes potentiels (sélectionnés grâce à l'indexation du deuxième formant) et ceux extraits de l'image acoustique du vecteur acoustique à inverser. La figure 6.10 représente deux spectres dont on souhaite apparier les pics. Dans cet exemple, les valeurs des trois premiers pics de la courbe noire sont 375Hz, 1688Hz, 2334Hz et celles de la courbe rouge sont 342Hz, 1016Hz, 1670Hz et 2451Hz. On remarque que la courbe rouge possède un pic supplémentaire à 1016Hz qui n'est pas présent sur la courbe noire. L'algorithme permet de vérifier que l'appariement entre les pics est possible en ne tenant pas compte du deuxième pics de la courbe rouge. Le résultat de l'algorithme est présenté dans le tableau 6.3. Le pic supplémentaire de la courbe n'a pas été pris en compte, l'algorithme a permis de rassembler les pics avec des valeurs proches ; les premiers pics des courbes noire et rouge (375Hz et 342Hz), le deuxième pic de la courbe noire (1688Hz) avec le troisième pic de la courbe rouge (1670Hz) et le troisième pic de la courbe noire (2334Hz) avec le quatrième pic de la courbe rouge (2451Hz).

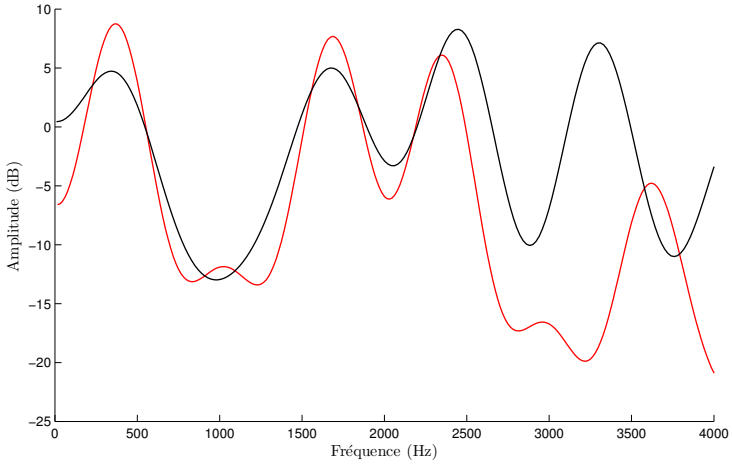


FIGURE 6.10 – *Deux spectres pour illuster l'algorithme d'appariement des pics spectraux.*

Spectre 1 (noir)		Spectre 2 (rouge)	
N° du pic	Fréquence	N° du pic	Fréquence
1	375Hz	1	342Hz
2	1688Hz	3	1670Hz
3	2334Hz	4	2451Hz

TABLEAU 6.3 – *Résultats de l'algorithme d'appariement sur les deux spectres de la figure 6.10.*

Si l'appariement entre deux spectres est impossible cela signifie que les spectres n'ont pas la même forme ou alors que le nombre de pics est insuffisant. Cet algorithme permet d'exclure les hypercuboïdes dont les pics spectraux du centre ne pourrait pas s'apparier avec les pics spectraux issus du vecteur acoustique à inverser.

6.4 Évaluation

Dans cette section, des expériences d'inversion statique ont été réalisées, c'est-à-dire que l'on réalise l'inversion à partir de formes sans rechercher de trajectoire. À cette étape, nous effectuons uniquement une recherche dans le codebook afin de vérifier qu'il est possible de trouver des solutions à partir du codebook.

Deux codebooks seront utilisés : le codebook n°1 avec un volume minimal de 4,27 et un seuil acoustique à 5 et le codebook n°2 avec un volume minimal de 1,07 et un seuil acoustique à 2 (voir tableau 5.3). Le codebook n°2 est plus précis que le n°1 ; le volume expliqué est plus important et l'erreur de resynthèse moyenne est plus faible.

Nous avons testé l'inversion sur des voyelles synthétiques produites par le synthétiseur articulatoire, puis sur les voyelles réelles.

6.4.1 Inversion de données synthétiques

La première série de tests réalisée a pour but de vérifier que la méthode d'inversion statique proposée précédemment fonctionne et que la construction du codebook est correcte.

Pour cela, nous avons réalisé une inversion statique à partir de données synthétiques correspondant aux voyelles prononcées. Pour toutes les formes correspondant aux voyelles, nous avons estimé les paramètres articulatoires et les paramètres acoustiques ont été calculés via le synthétiseur articulatoire. Ces paramètres acoustiques servent de vecteurs à inverser. Les 133 voyelles présentes dans le corpus ont servi à tester l'inversion.

6.4.1.1 Résultats de l'optimisation de la recherche dans le codebook

Les tableaux 6.4 présentent les résultats lors de l'optimisation de la recherche dans deux codebooks différents. La première colonne correspond aux pourcentages d'hypercuboïdes explorés suivant le formant F2 et la seconde colonne aux pourcentages après appariement des pics. Le codebook n°2 est plus précis que le codebook n°1.

L'organisation du codebook permet effectivement de réduire le nombre d'hypercuboïdes à explorer. En effet, moins de la moitié des hypercuboïdes sont conservés en réalisant une sélection suivant le deuxième formant, 40,43% pour le codebook n°1 et 32,49% pour le codebook n°2. Des différences existent entre les différents types de voyelles, pour /i/ nous conservons 16,79% des hypercuboïdes alors que pour le /a/ 55,72% pour le codebook n°1 simplement car /i/ est la voyelle qui a le deuxième formant le plus élevé.

La seconde étape d'optimisation est l'appariement des pics spectraux issus du vecteur acoustique à inverser et le centre de l'hypercuboïde étudié. Au final, nous obtenons 10,52% d'hypercuboïdes conservés pour le codebook n°1 et 5,71% pour le codebook n°2.

Codebook n°1

	Pourcentages d'hypercuboïdes conservés	
Phonème	Filtre F2	Appariement
/e/	38,28%	10,08%
/ø/	47,72%	13,96%
/i/	16,79%	5,97%
/y/	33,62%	10,32%
/ɛ/	46,90%	15,34%
/a/	55,72%	9,06%
/u/	44,01%	8,82%
Moyenne	40,43%	10,52%

Codebook n°2

	Pourcentages d'hypercuboïdes conservés	
Phonème	Filtre F2	Appariement
/e/	31,30%	5,23%
/ø/	39,97%	7,33%
/i/	12,42%	3,12%
/y/	26,89%	5,35%
/ɛ/	39,23%	9,24%
/a/	45,16%	5,54%
/u/	32,44%	4,13%
Moyenne	32,49%	5,71%

TABLEAU 6.4 – *Résultats de l'optimisation de la recherche dans le codebook suivant la recherche suivant F2 et l'appariement des pics spectraux.*

Nous constatons que ces deux étapes permettent de réduire considérablement le nombre d'hypercuboïdes dans lesquels nous recherchons une solution. L'organisation permet de ne conserver qu'un tiers des hypercuboïdes et l'appariement des pics spectraux conserve uniquement 5% des hypercuboïdes ce qui permet de rendre l'inversion plus rapide notamment pour des codebooks précis contenant de nombreux hypercuboïdes.

6.4.1.2 Résultats de l'inversion

Maintenant nous allons évaluer les résultats d'inversion calculés avec les hypercuboïdes conservés. Pour les 133 formes géométriques des voyelles du corpus, nous avons réalisé l'inversion à partir des vecteurs acoustiques synthétiques. L'évaluation des résultats va être réalisée en comparant les formes articulatoires utilisées pour calculer les vecteurs acoustiques synthétiques et celles obtenues par inversion. Nous disposons de beaucoup de solutions fournies par le codebook.

Le tableau 6.5 présente les erreurs géométriques et acoustiques minimales moyennes pour chaque type de voyelles. L'erreur géométrique mesure la distance moyenne entre chaque point de la forme synthétique à inverser et la forme inversée. Chaque forme synthétique est composée de 37 points : deux pour le larynx, trente trois pour l'épiglotte, la langue et le plancher de la langue et deux pour les lèvres. L'erreur acoustique est la distance entre le vecteur acoustique synthétique à inverser et le vecteur acoustique obtenu par inversion. Les vecteurs acoustiques sont composés de 29 coefficients cepstraux.

Les figures 6.11, 6.12, 6.13, 6.14, 6.15, 6.16 et 6.17 montrent des exemples de résultats pour chaque type de voyelles. La solution affichée est celle qui minimise la distance géométrique. On constate que la recherche dans le codebook permet de retrouver une forme articulatoire proche de la forme originale.

Codebook n°1

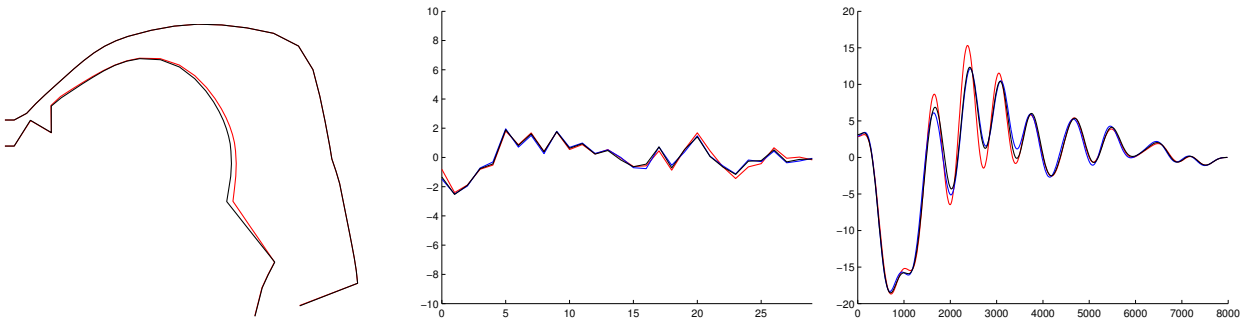
	Erreur minimale moyenne pour chaque phonème	
Phonème	Géométrique (mm)	Acoustique
/e/	0,4	0,71
/ø/	0,4	0,68
/i/	0,4	0,56
/y/	0,5	0,96
/ɛ/	0,2	0,29
/a/	0,4	0,54
/u/	0,5	0,70
Moyenne	0,4	0,59

Codebook n°2

	Erreur minimale moyenne pour chaque phonème	
Phonème	Géométrique (mm)	Acoustique
/e/	0,3	0,53
/ø/	0,3	0,59
/i/	0,3	0,38
/y/	0,4	0,79
/ɛ/	0,2	0,19
/a/	0,3	0,46
/u/	0,3	0,48
Moyenne	0,3	0,43

TABLEAU 6.5 – *Erreurs géométriques et acoustiques minimales moyennes pour chaque voyelle dans le corpus.*

La construction de notre codebook associant des vecteurs articulatoires formés par les paramètres du modèle articulatoire à des vecteurs acoustiques composés des premiers coefficients cepstraux nous permet de réaliser l'inversion acoustique-articulatoire. L'inversion de données synthétiques nous a permis de vérifier qu'il nous est possible de retrouver une forme articulatoire à partir des coefficients cepstraux. Mais il faut encore vérifier que l'utilisation de données acoustiques issues d'un signal réel nous fournisse des solutions.



/e/

FIGURE 6.11 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /e/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

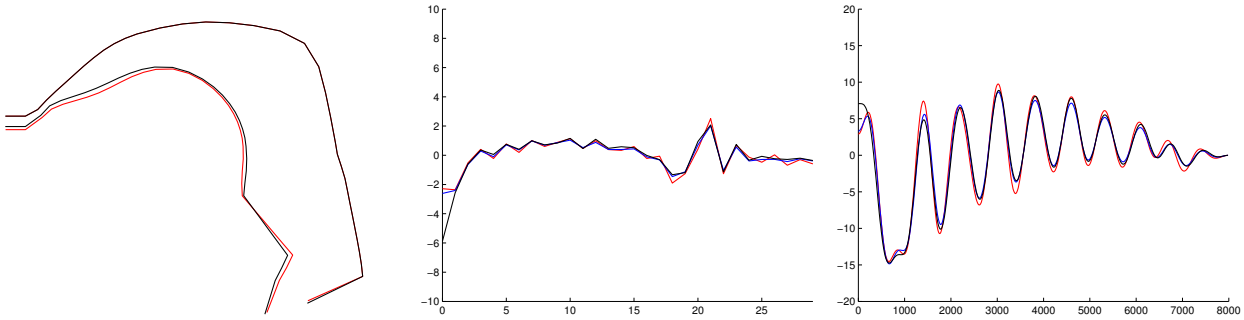


FIGURE 6.12 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /ø/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

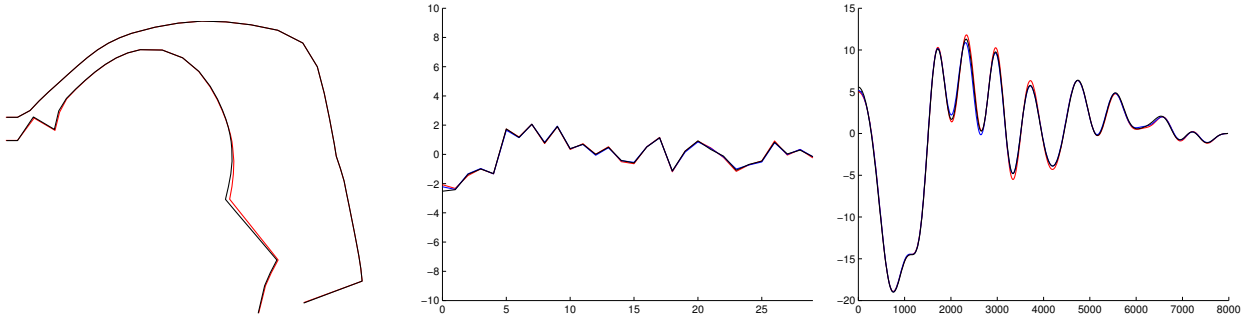
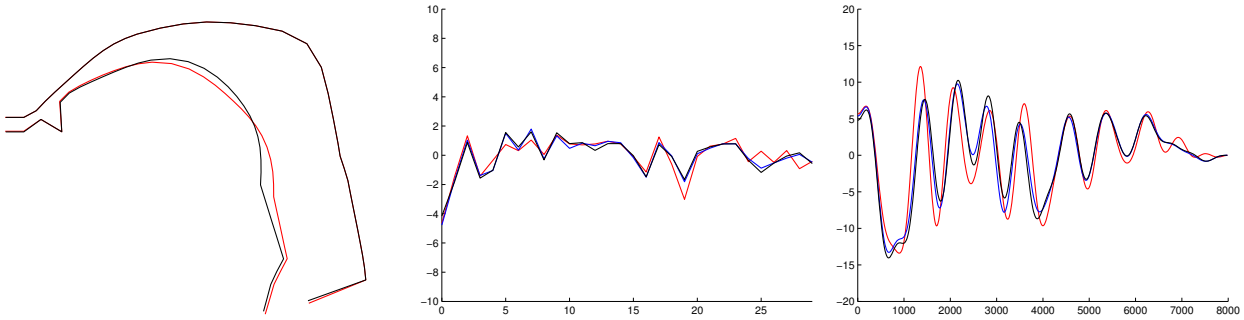
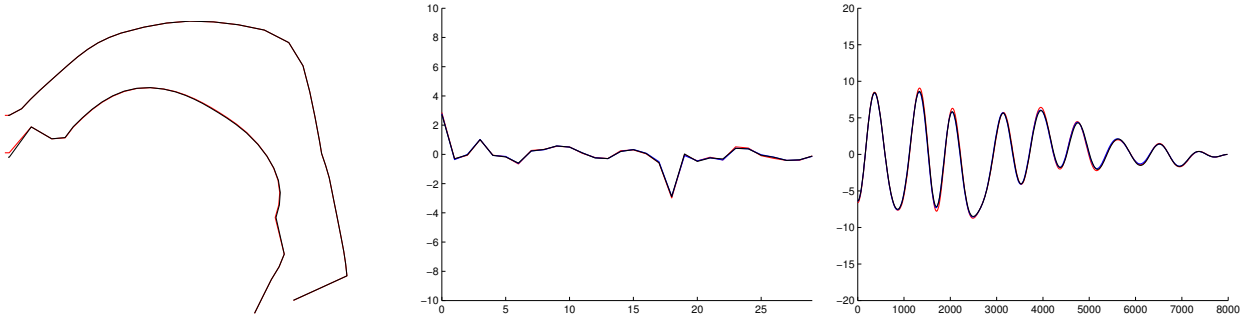


FIGURE 6.13 – *Exemple d'inversion de données synthétiques pour un forme de conduit vocal correspondant à la voyelle /i/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



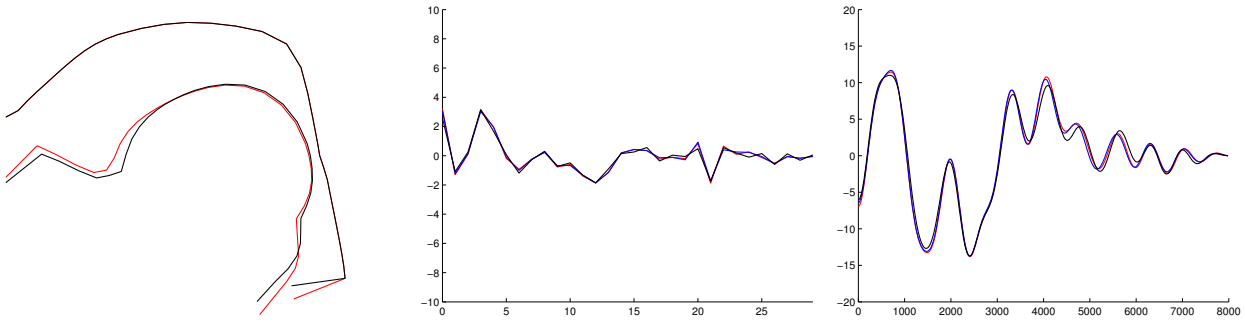
/y/

FIGURE 6.14 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /y/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



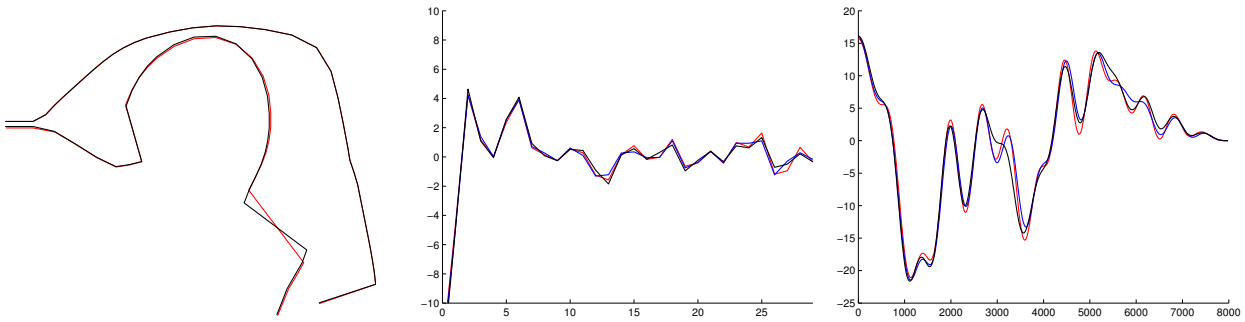
/ε/

FIGURE 6.15 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /ε/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



/A/

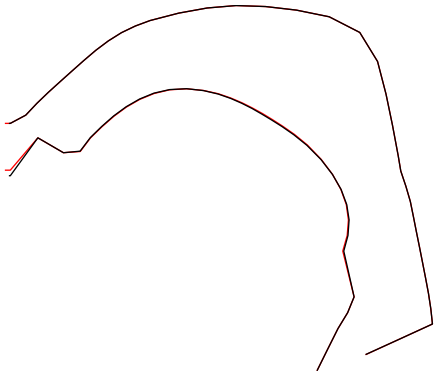
FIGURE 6.16 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /a/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



/u/

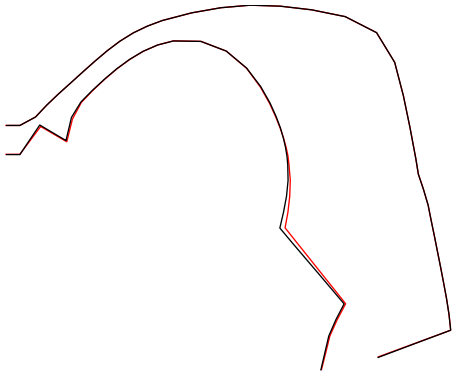
FIGURE 6.17 – *Exemple d'inversion de données synthétiques pour une forme de conduit vocal correspondant à la voyelle /u/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur minimale entre les coefficients cepstraux) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

Les résultats de l'inversion nous montrent que la solution avec l'erreur minimale est très proche de la forme de départ. Nous avons vu que les formes articulatoires sont très proches mais il n'est pas évident que les paramètres articulatoires obtenus par inversion soient proches des originaux. La figure 6.18 est un exemple où les paramètres de la solution avec une erreur géométrique minimale sont très proches des originaux. La figure 6.19 montre un autre cas où les paramètres sont un peu plus éloignés mais les formes géométriques obtenues sont très similaires.



	Paramètres articulatoires	
	Originaux	**Inversés**
Mâchoire	0,88	0,87
Langue 1	1,00	1,00
Langue 2	-0,78	-0,75
Langue 3	0,81	0,82
Langue 4	-2,44	-2,34
Lèvres	1,16	0,77
Epiglotte et larynx	0,18	0,19

FIGURE 6.18 – *Conduits vocaux original et inversé à partir des paramètres synthétiques et paramètres articulatoires associés. La forme noire (resp. paramètres articulatoire) est la forme issue d'une image cinéradiographique d'un /ɛ/ et en rouge la solution inverse qui minimise l'erreur géométrique.*



	Paramètres articulatoires	
	Originaux	**Inversés**
Mâchoire	-0,67	-0,52
Langue 1	-1,20	-1,26
Langue 2	-0,69	-0,76
Langue 3	1,13	0,99
Langue 4	0,37	0,00
Lèvres	0,61	0,48
Epiglotte et larynx	-0,79	-0,83

FIGURE 6.19 – *Conduits vocaux original et inversé à partir des paramètres synthétiques et paramètres articulatoires associés. La forme noire (resp. paramètres articulatoire) est la forme issue d'une image cinéradiographique d'un /i/ et en rouge la solution inverse qui minimise l'erreur géométrique.*

Nous avons montré que notre méthode d'inversion permet de trouver une forme géométrique proche de la forme originale. Mais si l'on regarde les autres solutions, nous constatons qu'il existe un grand nombre de solutions proches. La figure 6.20 montre des exemples d'inversion où les vingt cinq meilleures solutions sont dessinées pour trois formes de conduit vocal. Nous observons que notre

méthode d'inversion nous fournit des solutions très proches de la forme de départ. Ce qui signifie que l'utilisation des coefficients cepstraux est possible pour réaliser une inversion acoustique-articulatoire. Ces expériences ont permis de montrer que le codebook que nous avons construit permet de récupérer des solutions et que sa construction et son organisation sont correctes. L'étape suivante est de réaliser l'inversion à partir de données acoustiques mesurées sur le signal de parole réel.

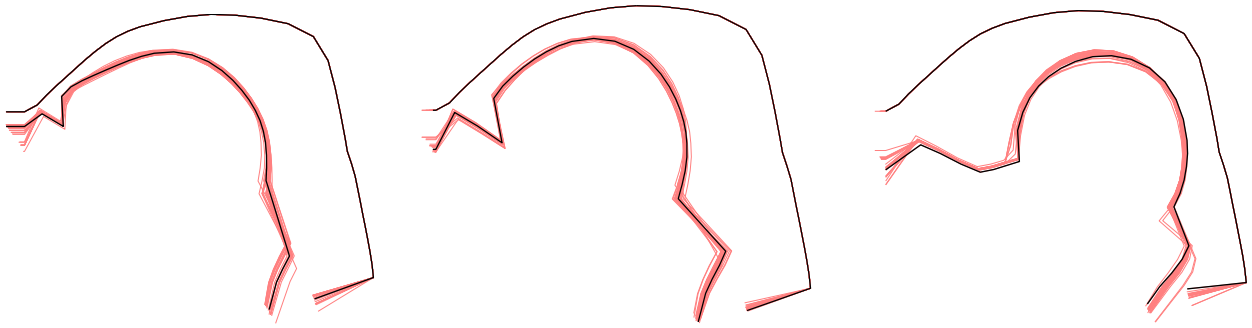


FIGURE 6.20 – *La forme articulatoire noire est issue des images cinéradiographiques. Les courbes rouges correspondent aux 25 meilleures solutions en termes de distance géométrique.*

6.4.2 Inversion de données réelles

Nous avons vu que l'inversion d'un vecteur acoustique synthétique est réalisable. L'utilisation des coefficients cepstraux comme paramètres acoustiques permet de s'approcher des paramètres articulatoires correspondants. Le but est de réaliser une inversion acoustique-articulatoire à partir de données réelles. Dans notre cas, nous disposons de l'enregistrement associé aux formes articulatoires, il nous est donc facile d'obtenir le spectre correspondant à une image (et donc à la forme géométrique associée) et ainsi évaluer les résultats obtenus avec la forme du conduit vocal. Nous allons donc réaliser l'inversion à partir des coefficients cepstraux calculés sur le signal et vérifier la proximité des formes géométriques obtenues avec celles issues des images radiographiques. Nous avons repris les voyelles mais cette fois l'inversion se fait à partir des coefficients cepstraux calculés sur le signal réel.

L'inversion à partir des coefficients cepstraux calculés sur le signal réel ne donne aucune solution, ce qui était prévisible car les disparités entre les signaux synthétiques et réels n'ont pas été prises en compte. La modification de ces données réelles est donc nécessaire afin de permettre l'utilisation du codebook construit à partir de données synthétiques.

Tout d'abord, nous calculons les spectres LPC à partir du signal original. Puis, nous réalisons une distorsion fréquentielle avec une transformation bilinéaire avec la valeur optimale de α obtenue dans la section 6.2.3. Les coefficients cepstraux sont calculés sur le spectre obtenu après la distorsion fréquentielle.

À ce stade, nous avons réalisé la distorsion fréquentielle sur les spectres LPC issus du signal réel. Cette transformation permet de décaler les pics spectraux afin de les rapprocher. Mais la pente spectrale (ou « spectral tilt ») n'est pas prise en compte. Nous avons vu dans la section 6.2.2 qu'une

adaptation cepstrale par transformation affine des coefficients cepstraux permettait de réduire les différences entre les coefficients cepstraux. De plus, lorsque l'on utilise un petit nombre de coefficients adaptés (deux ou trois), les pics spectraux sont mieux conservés et les différences entre les pentes spectrales sont réduites. Pour apprécier l'effet de l'adaptation cepstrale réalisée après le décalage en fréquence, nous avons fait une inversion sur les voyelles à partir du signal réel avec et sans adaptation cepstrale.

Le tableau 6.6 montre les résultats obtenus pour l'inversion de l'ensemble des voyelles du corpus à partir du signal réel après un décalage en fréquence du spectre LPC avec ou sans adaptation cepstrale effectuée après le décalage. Dans le tableau 6.6, l'erreur géométrique calculée sur la solution qui a l'erreur acoustique minimale apparaît entre paranthèse. Les solutions dont l'erreur acoustique est minimale ne sont donc pas celle qui minimisent l'erreur géométrique.

Codebook n°1

	Erreur minimale moyenne pour chaque phonème			
	Sans adaptation		Avec adaptation	
Phonème	Géométrique (mm)	Acoustique (géo)	Géométrique (mm)	Acoustique (géo)
/e/	1,2	5,66 (7,0)	0,9	3,59 (3,7)
/ø/	1,3	6,91 (6,8)	1,0	3,83 (4,2)
/i/	1,1	3,64 (4,6)	0,9	2,77 (3,2)
/y/	0,9	4,80 (5,7)	0,8	3,31 (3,9)
/ɛ/	1,2	8,20 (4,6)	1,1	3,97 (4,2)
/a/	2,4	6,82 (7,4)	2,1	4,54 (6,2)
/u/	1,4	5,83 (7,1)	1,1	4,03 (7,1)
Moyenne	1,3	6,18 (5,9)	1,0	3,69 (4,8)

Codebook n°2

	Erreur minimale moyenne pour chaque phonème			
	Sans adaptation		Avec adaptation	
Phonème	Géométrique (mm)	Acoustique (géo)	Géométrique (mm)	Acoustique (géo)
/e/	1,0	5,75 (7,0)	0,8	3,63 (3,5)
/ø/	1,1	7,08 (5,6)	0,8	3,91 (4,8)
/i/	1,0	3,67 (4,2)	0,9	2,87 (3,8)
/y/	0,8	4,85 (5,0)	0,7	3,25 (3,6)
/ɛ/	1,3	8,42 (4,4)	1,2	4,01 (3,8)
/a/	2,6	6,85 (7,5)	2,5	4,30 (6,8)
/u/	1,7	5,88 (6,5)	1,4	4,33 (6,7)
Moyenne	1,2	6,20 (5,5)	1,0	3,82 (4,8)

TABLEAU 6.6 – *Erreurs géométriques et acoustiques minimales moyennes pour chaque voyelle dans le corpus. L'erreur acoustique est accompagnée entre parenthèses de l'erreur géométrique pour la solution correspondant à l'erreur acoustique minimale.*

La première constatation que l'on peut faire est que le décalage en fréquence permet de trouver des solutions, ce qui n'était pas possible sans. Pour les voyelles /e/, /ø/, /i/ et /y/, on constate que l'erreur géométrique minimale moyenne est meilleure dans le codebook n°2. Inversement pour /ɛ/, /a/ et /u/, les résultats ne sont pas meilleurs lorsque l'on utilise un codebook n°2 qui est plus précis. /a/ possède une grande variabilité articulatoire, il est donc normal de la retrouver ici. Au contraire /u/ demande une articulation précise, la section de la bouche a souvent une ouverture très limitée correspondant à la forte protrusion et la petite ouverture des lèvres. Ces formes sont donc proches de la fermeture complète. Elles se situent près de la frontière de l'espace articulatoire. Or, nous avons vu que ces régions sont plus difficilement linéarisables, d'où une erreur plus importante pour ces configurations articulatoires.

On remarque également que l'erreur minimale moyenne n'est pas forcément meilleure avec un codebook plus précis. C'est le cas du /e/ pour laquelle l'erreur acoustique minimale moyenne est de 5,66 pour le codebook n°1 alors qu'elle est de 5,75 pour le codebook n°2. Cela ne signifie pas que le codebook n°2 est moins bon que le n°1 mais que les hypercuboïdes que l'on a sélectionnés donnent des erreurs acoustiques plus importantes. Les solutions inverses sont calculées uniquement sur les hypercuboïdes conservés après une sélection selon le deuxième formant et l'appariement des pics. Certains hypercuboïdes sont donc absents de la sélection.

Les deux dernières colonnes du tableau 6.6 présentent les résultats de l'adaptation cepstrale réalisée après le décalage en fréquence. Pour cela, nous avons utilisé les coefficients cepstraux calculés sur le spectre LPC décalé et les coefficients cepstraux synthétiques correspondants. Puis nous avons réalisé la même opération de transformation affine que la section 6.2.2. Nous avons adapté uniquement les deux premiers coefficients. La distance acoustique est calculée entre les coefficients cepstraux réels adaptés et les cepstres resynthétisés avec les paramètres articulatoires obtenus par inversion.

Pour les deux codebooks utilisés, on constate que l'adaptation cepstrale effectuée après le décalage en fréquence permet de réduire les erreurs. L'erreur géométrique minimale moyenne est meilleure pour chacune des voyelles lorsque l'adaptation est réalisée.

L'adaptation permet de rapprocher le vecteur acoustique à inverser des vecteurs acoustiques présents dans le codebook, ce qui se vérifie car l'erreur acoustique minimale moyenne est plus faible après adaptation.

L'erreur acoustique diminue de 51% pour le /ɛ/ ou encore de 45% pour le /ø/. C'est pour le /i/ que la réduction de l'erreur est la moins importante mais cette voyelle donne lieu à l'erreur la plus faible.

Les différentes étapes de notre inversion de données réelles sont présentées dans la figure 6.21.

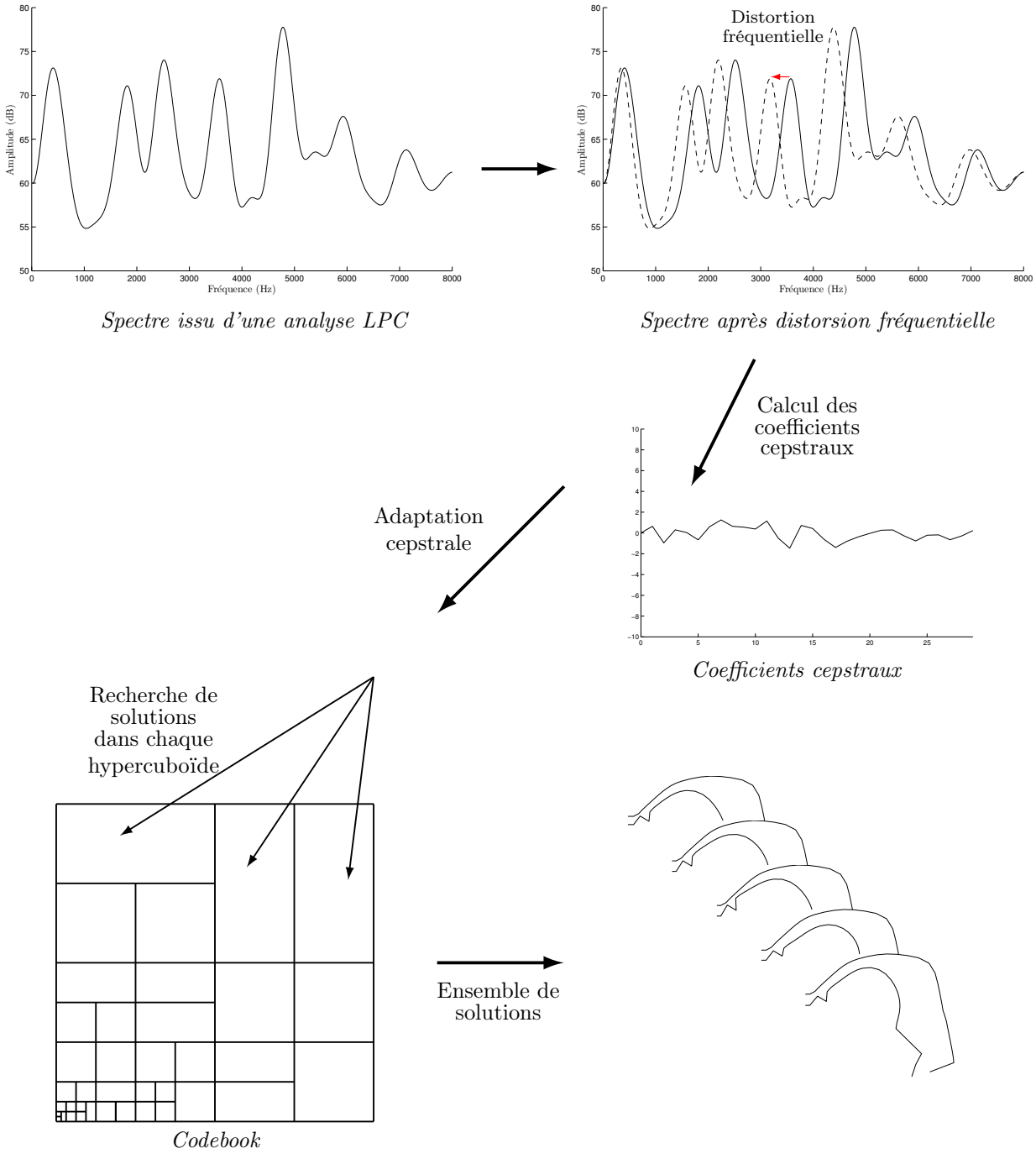
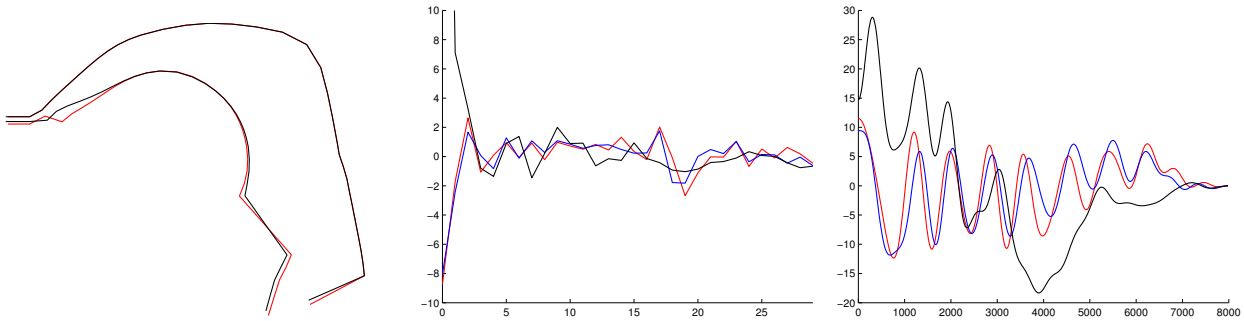


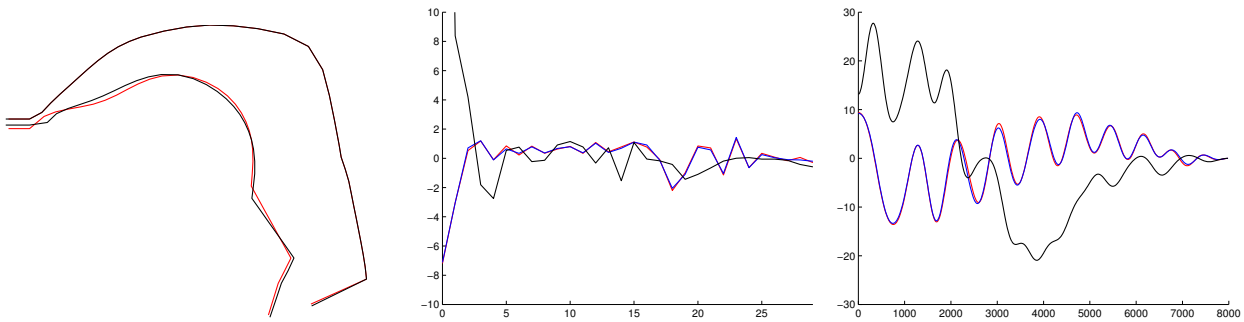
FIGURE 6.21 – *Différentes étapes pour l'inversion de données réelles*

Les figures 6.22, 6.23, 6.24, 6.25, 6.26, 6.27 et 6.28 présentent des exemples de résultats pour chacune des voyelles. Dans chaque cas, la forme dessinée est celle qui possède l'erreur géométrique minimale. Ainsi, nous avons montré qu'il est possible à partir d'une distorsion fréquentielle d'obtenir des solutions proches de la forme de départ. Mais dans l'ensemble des solutions, il subsiste des solutions qui sont très éloignées.



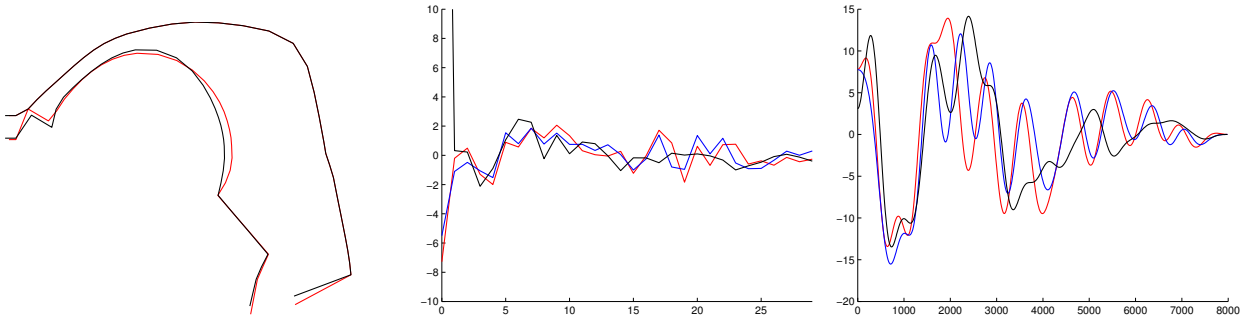
/e/

FIGURE 6.22 – *Exemple d'inversion à partir du signal réel pour la voyelle /e/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



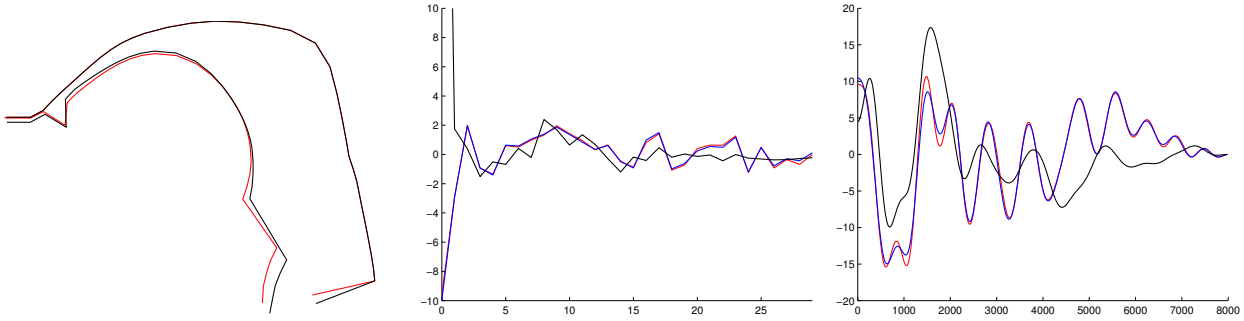
/ø/

FIGURE 6.23 – *Exemple d'inversion à partir du signal réel pour la voyelle /ø/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



/i/

FIGURE 6.24 – *Exemple d'inversion à partir du signal réel pour la voyelle /i/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*



/y/

FIGURE 6.25 – *Exemple d'inversion à partir du signal réel pour la voyelle /y/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

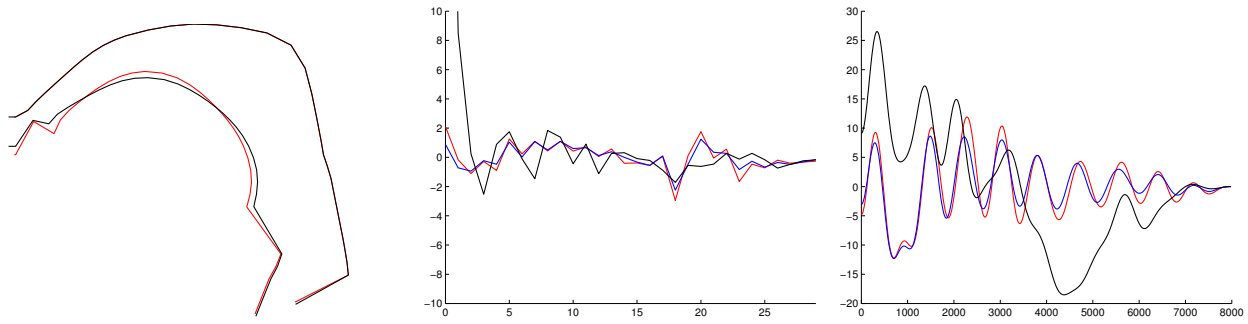


FIGURE 6.26 – *Exemple d'inversion à partir du signal réel pour la voyelle /ε/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

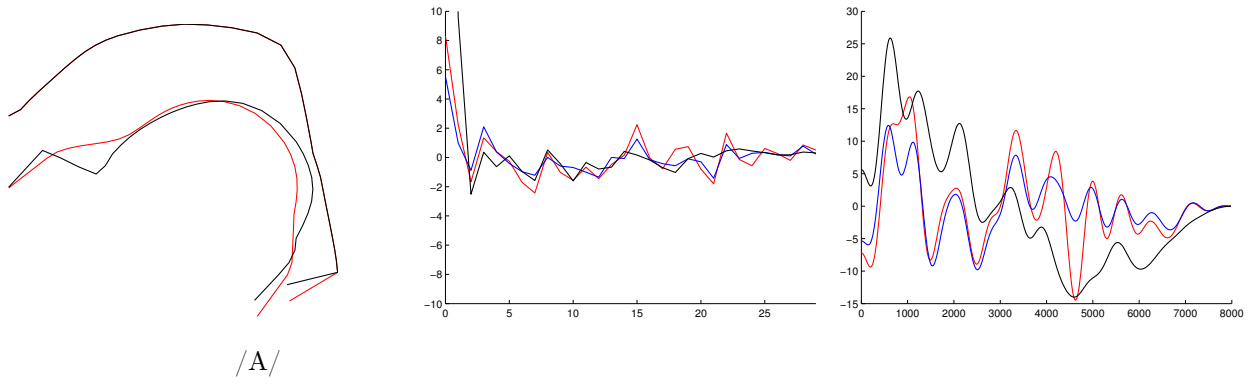


FIGURE 6.27 – *Exemple d'inversion à partir du signal réel pour la voyelle /a/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

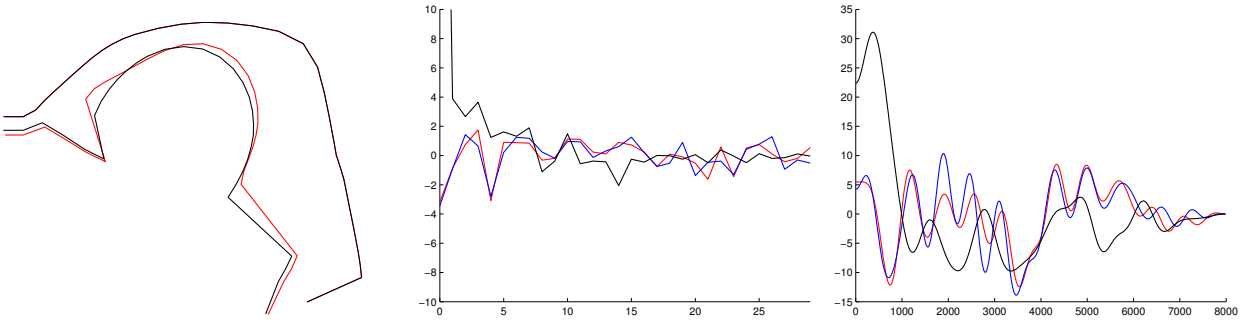


FIGURE 6.28 – *Exemple d'inversion à partir du signal réel pour la voyelle /u/. Courbe noire (forme à inverser), courbe rouge (meilleure solution avec l'erreur géométrique minimale) et courbe bleue (resynthèse des paramètres articulatoires obtenus par inversion).*

Le vecteur articulatoire qui donne l'erreur acoustique minimale n'est en général pas celui qui donne une erreur géométrique minimale. Dans le tableau 6.6, l'erreur géométrique correspondant au vecteur articulatoire donnant une erreur acoustique minimale est notée entre parenthèses à côté de l'erreur acoustique. Cette erreur est importante lorsque l'adaptation cepstrale n'est pas effectuée, or il serait intéressant que les vecteurs articulatoires présentant les erreurs géométriques les plus faibles aient également une erreur acoustique parmi les plus faibles. L'adaptation cepstrale permet de prendre en compte la pente spectrale et donc de réduire cette erreur géométrique. La figure 6.29 montre un exemple où la solution avec l'erreur acoustique est minimale avec ou sans adaptation cepstrale. On voit qu'il existe une grande différence entre les deux formes.

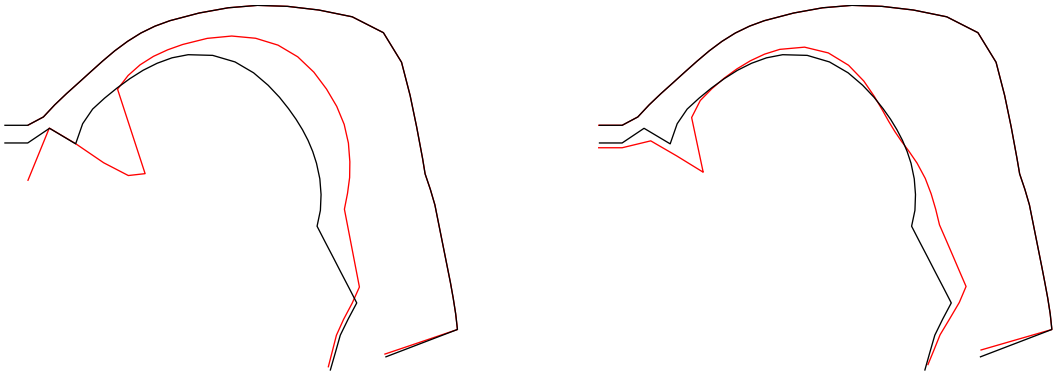


FIGURE 6.29 – *Les deux graphiques présentent les conduits vocaux originaux (courbe noire) et inversés (courbe rouge) avec l'erreur acoustique minimale sans adaptation cepstrale (graphique de gauche) et avec adaptation cepstrale (graphique de droite).*

L'inversion de données acoustiques issues du signal réel n'était pas du tout assurée. En effet,

les vecteurs cepstraux produits par le synthétiseur sont éloignés de ceux calculés sur le signal réel dû notamment à l'absence de source dans la synthèse articulatoire et aux erreurs liées au modèle articulatoire. Une adaptation des vecteurs cepstraux est donc nécessaire pour permettre d'accéder au codebook. Nous avons montré qu'il est possible de trouver un grand nombre de solutions inverses pour un vecteur acoustique donné et que nous retrouvons aussi des formes proches de la forme originale. Un des points clés qui n'est pas abordé ici est de savoir comment trouver la meilleure solution.

6.5 Modification du modèle

Notre méthode d'inversion d'analyse par synthèse utilisant les coefficients cepstraux comme vecteurs acoustiques donne d'excellents résultats lors de l'inversion de données synthétiques. Les méthodes permettant « d'adapter » le signal réel pour le rapprocher du signal synthétique nous ont permis de trouver des solutions fournissant des formes articulatoires proches de la forme originale. L'évaluation des résultats est réalisée en calculant la distance entre les formes géométriques fournies par le modèle et celles obtenues par inversion.

Maintenant nous souhaitons évaluer l'impact des erreurs liées au modèle articulatoire lui-même. Pour cela nous modifions légèrement le modèle et nous construisons le codebook associé. La modification du modèle consiste simplement à allonger ou raccourcir les cavités pharyngale et buccale. Nous avons choisi d'augmenter et de diminuer la taille des cavités de 10% pour réaliser nos tests. La figure 6.30 montre les deux configurations utilisées, une augmentation et une diminution de la taille du conduit. La première expérience concerne la normalisation de la longueur du conduit vocal.

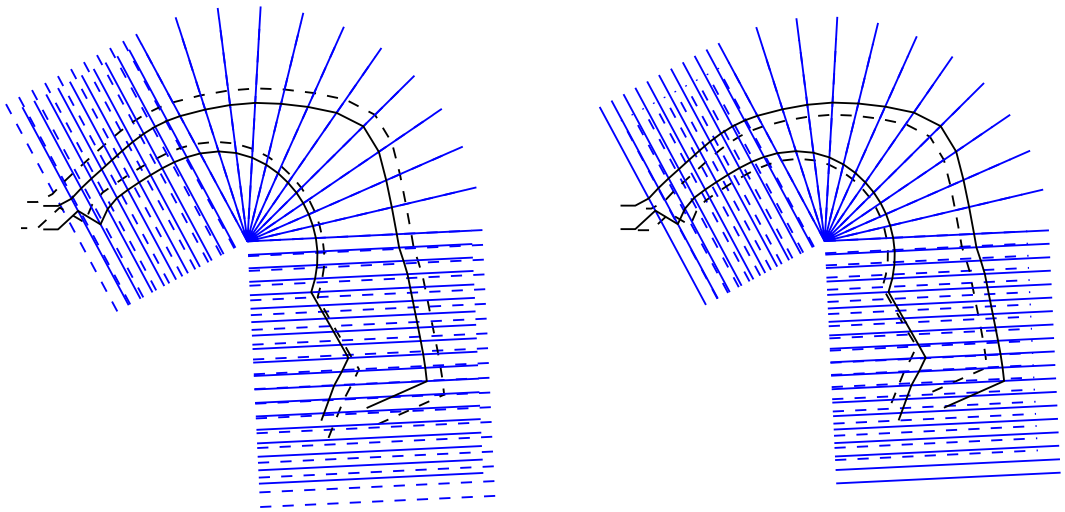


FIGURE 6.30 – *Modification du modèle original (courbe en trait plein). La courbe de gauche (tirets) représente une augmentation de 10% de la taille et la courbe (tirets) de droite une diminution de 10%.*

Nous voulons maintenant étudier l'effet de ces modifications sur les résultats de l'inversion. Pour cela, nous construisons différents codebooks pour les deux modèles issus d'une transformation du modèle original. Les caractéristiques des codebooks obtenus ainsi que les régions définies dans la section 5.4.1 sont présentées dans le tableau 6.7.

Comme on peut le voir il n'y a pas d'effets très notables sur les caractéristiques des codebooks construits après adaptation du modèle. On peut seulement remarquer que la position des zones relativement proches de la frontière change sans doute légèrement puisque le nombre d'hypercuboïdes obtenus a changé.

		Codebook n°1 (Volume minimal = 5 ; seuil acoustique = 5)			**Codebook n°2** (Volume minimal = 2 ; seuil acoustique = 2)		
		-10%	Original	+10%	-10%	Original	+10%
Totalité	Nombre d'hypercuboïdes	35549	35523	33037	177823	177981	169264
	Volume	190053,16	188953,03	177802,77	193279,47	192009,44	181655,06
	Erreur moyenne de resynthèse	1,48	1,48	1,46	1,13	1,15	1,12
Zone 1	Nombre d'hypercuboïdes	119	102	69	527	481	331
	Volume	508,31	435,68	294,73	562,77	513,63	332,11
	Erreur moyenne de resynthèse	2,49	2,54	2,64	2,00	2,04	1,96
Zone 2	Nombre d'hypercuboïdes	259	295	262	1961	1969	1972
	Volume	2101,57	2101,57	2097,30	2096,23	2102,64	2105,84
	Erreur moyenne de resynthèse	0,98	1,01	1,03	0,72	0,77	0,76
Zone 3	Nombre d'hypercuboïdes	133	148	137	855	880	880
	Volume	944,00	961,08	926,91	995,26	1009,13	964,29
	Erreur moyenne de resynthèse	0,95	0,92	1,00	0,64	0,65	0,67

TABLEAU 6.7 – *Caractéristiques des différents codebooks et zones construits pour le modèle diminué de 10%, original et augmenté de 10%.*

Les tableaux 6.8 et 6.9 présentent les erreurs obtenues par inversion des voyelles du codebook. Pour les données synthétiques, on constate que les erreurs acoustiques et géométriques sont plus grandes lorsque le modèle est modifié, ce qui était prévisible. Pour l'inversion des données réelles, on s'aperçoit que l'erreur acoustique est très proche de celle obtenue avec le modèle original. Pour

le codebook n°1, par l'exemple l'erreur acoustique minimale moyenne est de 3,69 pour le modèle original, de 3,83 pour un raccourcissement et de 3,89 pour un accroissement. Par contre l'erreur géométrique est plus importante. Ce qui signifie que lorsque le modèle est modifié, il est possible de trouver des solutions donnant un vecteur acoustique proche du vecteur à inverser mais en compensant la forme géométrique.

Codebook n°1

	Erreur minimale					
	-10%		Original		+10%	
Phonème	Géométrique	Acoustique	Géométrique	Acoustique	Géométrique	Acoustique
/e/	1,8	2,18	0,4	0,71	1,7	2,44
/ø/	1,6	2,00	0,4	0,68	1,5	1,70
/i/	1,9	2,17	0,4	0,56	2,0	2,67
/y/	2,0	2,60	0,5	0,96	1,5	2,58
/ɛ/	1,5	1,30	0,2	0,29	1,6	1,80
/a/	1,8	2,08	0,4	0,54	1,8	2,47
/u/	1,9	4,30	0,5	0,70	1,8	2,90
Moyenne	1,7	2,60	0,4	0,59	1,7	2,36

Codebook n°2

	Erreur minimale					
	-10%		Original		+10%	
Phonème	Géométrique	Acoustique	Géométrique	Acoustique	Géométrique	Acoustique
/e/	1,8	2,14	0,3	0,53	1,7	2,32
/ø/	1,6	1,83	0,3	0,59	1,4	1,59
/i/	1,8	2,35	0,3	0,38	2,0	2,60
/y/	1,9	2,47	0,4	0,79	1,5	2,45
/ɛ/	1,4	1,23	0,2	0,19	1,7	1,70
/a/	1,9	2,5	0,3	0,46	2,0	2,73
/u/	1,8	3,91	0,3	0,48	1,7	2,90
Moyenne	1,7	2,47	0,3	0,43	1,7	2,30

TABLEAU 6.8 – *Erreurs géométriques et acoustiques minimales moyennes pour chaque voyelle pour l'inversion de données synthétiques pour le modèle diminué de 10%, original et augmenté de 10%.*

Codebook n°1

	Erreur minimale					
	-10%		Original		+10%	
Phonème	Géométrique	Acoustique	Géométrique	Acoustique	Géométrique	Acoustique
/e/	1,9	3,74	0,9	3,59	1,7	3,87
/ø/	1,7	4,03	1,0	3,83	1,8	3,85
/i/	2,0	2,74	0,9	2,77	1,9	3,56
/y/	2,0	3,40	0,8	3,31	1,6	3,23
/ɛ/	1,6	3,98	1,1	3,97	1,7	4,36
/a/	2,9	4,65	2,1	4,54	2,9	4,79
/u/	2,1	4,19	1,1	4,03	2,2	3,87
Moyenne	2,0	3,83	1,0	3,69	1,9	3,89

Codebook n°2

	Erreur minimale					
	-10%		Original		+10%	
Phonème	Géométrique	Acoustique	Géométrique	Acoustique	Géométrique	Acoustique
/e/	1,8	3,74	0,8	3,63	1,6	3,96
/ø/	1,6	4,04	0,8	3,91	1,9	3,83
/i/	2,0	2,84	0,9	2,87	2,1	3,69
/y/	1,9	3,39	0,7	3,26	1,5	3,18
/ɛ/	1,6	4,00	1,2	4,01	1,7	4,50
/a/	3,1	4,37	2,5	4,30	3,2	4,60
/u/	2,0	4,34	1,4	4,33	2,2	4,05
Moyenne	1,9	3,88	1,0	3,82	1,9	4,02

TABLEAU 6.9 – *Erreurs géométriques et acoustiques minimales moyennes pour chaque voyelle pour l'inversion de données réelles après adaptation pour le modèle diminué de 10%, original et augmenté de 10%.*

La figure 6.31 montre un exemple d'inversion pour le modèle dont les cavités pharyngale et buccale sont raccourcies de 10%. On constate que pour la solution inverse la longueur des lèvres est légèrement plus grande de façon à compenser le raccourcissement du conduit.

La figure 6.32 montre un exemple d'inversion pour le modèle dont les cavités pharyngale et buccale sont augmentées de 10%. Lorsque l'on observe les fonctions d'aire associées aux conduits vocaux, on constate qu'elles sont très proches, ce qui signifie qu'il y a bien compensation pour que les vecteurs acoustiques correspondent.

Il s'agit de résultats préliminaires sur les conséquences d'une mauvaise adaptation du modèle. Nous étudierons plus en détail ces modifications dans un futur proche.

FIGURE 6.31 – *Exemple d'une solution d'inversion d'une voyelle (/y/) à partir du signal à un modèle où les cavités pharyngale et buccale sont raccourcies de 10%. Courbe rouge (solution inverse dans le modèle modifié) et courbe noire (forme de départ dans le modèle original).*

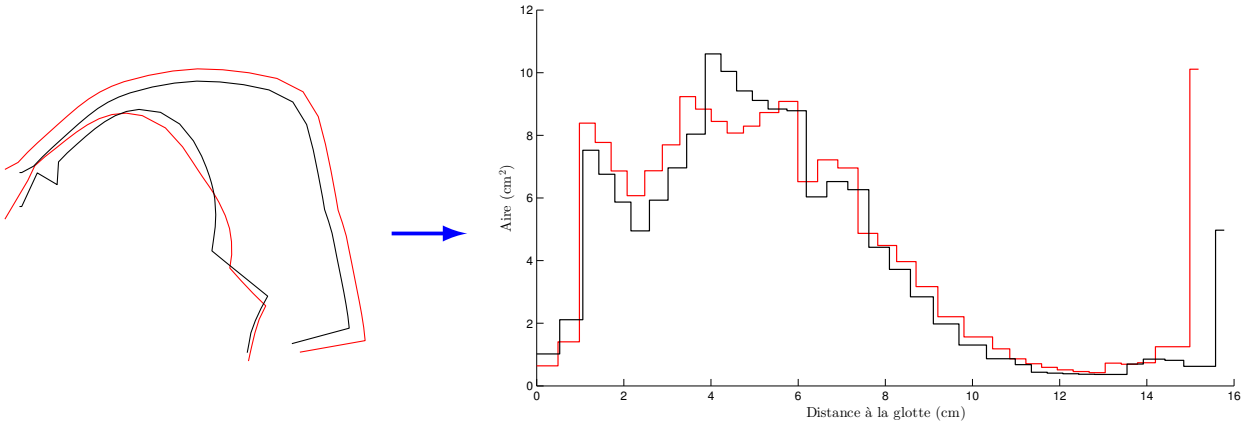


FIGURE 6.32 – *Exemple d'inversion avec le modèle où les cavités pharyngale et buccale sont augmentées de 10%. Les conduits vocaux (à gauche) et les fonctions d'aire (à droite) sont représentés en noir pour la forme initiale et en rouge pour la solution inverse qui minimise l'erreur acoustique.*

6.6 Conclusion

Notre objectif était d'étudier l'accès d'un codebook articulatoire à l'aide des coefficients cepstraux dans le cas favorable où les différences entre le modèle articulatoire et le conduit vocal du locuteur sont minimales. Cependant, nous insistons sur le fait que le modèle articulatoire est adapté

au locuteur mais que les détails très précis ne sont pas forcément représentés. Les expériences réalisées permettent de valider l'accès du codebook avec les coefficients cepstraux. Plus particulièrement, l'évaluation géométrique couvre l'ensemble de la langue mais pas uniquement la partie avant de la langue où les capteurs utilisés pour l'EMA ou les pastilles utilisées pour les micro-faisceaux de rayons X sont généralement collés. Cela signifie que nous ne forçons pas la correspondance du contour de la langue pour les points situés sur la partie avant de la bouche au détriment du réalisme de l'ensemble de la forme de la langue spécialement dans la région de la racine ce qui est souvent le cas pour d'autres méthodes d'inversion. Ce point est particulièrement important dans l'objectif d'utiliser l'inversion pour l'apprentissage d'une langue étrangère ou l'animation d'une tête parlante.

Nous continuerons d'étudier l'influence des disparités du modèle sur les résultats de l'inversion en dégradant artificiellement la qualité de la représentation du modèle avec le conduit vocal du locuteur. Ici nous avons étudié seulement un allongement et un raccourcissement de la longueur du conduit, on envisage d'étudier dans un futur proche plus précisément l'impact des erreurs liées à une adaptation du modèle articulatoire à un nouveau locuteur.

Conclusions et perspectives

Conclusion et perspectives

Mon travail de thèse a porté sur l'inversion acoustique-articulatoire à partir des coefficients cepstraux. La méthode utilisée est une approche d'analyse par synthèse qui nécessite l'utilisation d'un synthétiseur articulatoire. L'étape de synthèse utilise un modèle articulatoire qui permet de générer la forme du conduit vocal associée à une simulation acoustique pour générer le signal de parole.

Ici nous nous sommes placés dans un cadre contrôlé à la différence des autres études car nous disposons d'images cinéradiographiques de bonne qualité associées au signal acoustique. Le modèle articulatoire est construit spécifiquement pour le locuteur test afin de minimiser les erreurs géométriques. Sa construction est faite à partir des contours médio-sagittaux des principaux articulateurs visibles sur les images et nous avons développé différents outils permettant de réaliser le suivi semi-automatique des contours. Le contour de la langue, le principal articulateur, ne se prête pas au suivi automatique ou semi-automatique car il n'est pas toujours visible à cause du recouvrement par d'autres organes tels que les lèvres et à sa nature à se déformer fortement. Le contour a donc été donc tracé manuellement pour chacune des images où de la parole est prononcée. Nous avons réalisé une analyse statistique sur les contours afin d'extraire un petit nombre de paramètres pour contrôler le modèle. Nous avons donc construit un modèle spécifique qui permet d'approcher la partie avant de la langue plus finement que les modèles existants même en les adaptant au locuteur. Sept modes de déformation composent le modèle dont quatre permettent de contrôler la position et la forme de la langue.

Le modèle articulatoire nous permet ainsi de générer un vecteur acoustique synthétique associé à une forme de conduit vocal représentée par les sept paramètres du modèle. La difficulté est de comparer les vecteurs acoustiques issus de la synthèse à ceux calculés sur le signal naturel. Pour notre méthode d'inversion, nous avons choisi d'utiliser comme vecteurs acoustiques les coefficients cepstraux calculés sur le signal de parole. Généralement les formants sont utilisés comme vecteurs acoustiques mais leur détermination à partir des spectres est la source de nombreuses erreurs. Les coefficients cepstraux naturels intègrent la source qui excite le conduit vocal à la différence des coefficients cepstraux synthétiques. La comparaison directe des deux vecteurs acoustiques n'est donc pas possible. De plus, il subsiste des erreurs géométriques entre le modèle articulatoire et la forme réelle ce qui entraîne d'autres écarts entre les vecteurs cepstraux. Une de nos contributions porte sur le développement d'une adaptation des vecteurs cepstraux afin de limiter l'impact de ces erreurs lors de la recherche de solutions dans le codebook. La distorsion fréquentielle bilinéaire a été utilisée

pour réduire l'impact des différences géométriques liées au modèle articulatoire. Ensuite, une adaptation par transformation affine des coefficients cepstraux permet de réduire l'influence de la source d'excitation.

La comparaison entre les vecteurs cepstraux est nécessaire pour notre travail car nous effectuons une recherche dans un codebook construit à l'aide du synthétiseur articulatoire. En effet, la relation articulatoire-acoustique est représentée de façon compacte à l'aide d'un codebook organisé de façon à permettre la recherche rapide des solutions. Notre codebook se construit de façon récursive et se compose d'hypercuboïdes de différentes tailles. Dans chaque hypercuboïde, la relation articulatoire-acoustique est approchée linéairement. Grâce à notre étude sur la construction du codebook nous avons pu observer les différents comportements de la relation articulatoire-acoustique en fonction de la position dans l'espace articulatoire. Les régions situées près de la frontière sont plus difficilement linéarisables ce qui signifie que la relation articulatoire-acoustique n'y est par linéaire. Une étude plus fine est donc nécessaire pour modéliser ces zones. Nous ne nous sommes pas attardés sur ces zones car elles sont rarement atteintes par le locuteur. Il résulte qu'une petite partie de l'espace synthétisable n'est pas représenté dans le codebook. L'espace manquant dans le codebook ne pose pas de problème pour l'inversion de voyelles mais pourrait devenir sensible, critique lors de l'inversion de consonnes. Nous envisageons d'étudier ces zones plus précisément pour l'inversion qui n'est pas possible avec les formants.

Nous avons amélioré l'accès au codebook pour accélérer la recherche de solutions. Lors de l'inversion d'un vecteur cepstral, nous recherchons dans le codebook tous les hypercuboïdes susceptibles de contenir une solution. Le calcul d'une solution dans un hypercuboïde se fait par programmation quadratique car dans notre cas nous recherchons un vecteur articulatoire à sept paramètres à partir de vingt-neuf coefficients cepstraux. Le système à résoudre est donc sur-déterminé contrairement au cas des formants et il est donc impossible d'utiliser les mêmes méthodes. Le calcul effectif d'une solution est une opération relativement coûteuse en temps. Nous avons donc adopté plusieurs stratégies afin de sélectionner rapidement les hypercuboïdes susceptibles de fournir une solution. Nous avons utilisé les pics spectraux afin de filtrer les hypercuboïdes. Une organisation efficace du codebook en fonction du deuxième pic spectral nous a permis de réduire de façon importante le nombre d'hypercuboïdes. Un deuxième filtrage utilise un algorithme de programmation dynamique qui apparie les pics spectraux issus du vecteur cepstral à inverser avec ceux issus du codebook. Ce filtrage est robuste à la disparition de pics ou au contraire à l'apparition de pics. Ces filtrages ont permis de réduire le nombre d'hypercuboïdes explorés et donc d'améliorer la vitesse d'inversion. L'indexation suivant le deuxième pic et l'étape d'appariement sont suffisamment rapides pour ne pas pénaliser l'inversion.

L'ensemble des ces contributions a été testé sur les voyelles des quatre films aux rayons X du corpus. L'évaluation des résultats s'est effectuée en comparant les formes articulatoires calculées avec le modèle avec celles obtenues par inversion. L'évaluation géométrique couvre donc l'ensemble de la langue et pas uniquement la partie avant de la langue où les capteurs utilisés pour l'EMA ou les pastilles utilisées pour les micro-faisceaux de rayons X sont généralement collés. Cela signifie que nous ne forçons pas la correspondance du contour de la langue pour les points situés sur la partie avant au détriment du réalisme de l'ensemble de la forme de la langue spécialement dans la région de la racine de la langue ce qui arrive souvent dans d'autres méthodes d'inversion. Ce point

est particulièrement important dans l'objectif d'utiliser l'inversion pour l'apprentissage d'une langue étrangère ou l'animation d'une tête parlante.

Les différentes évaluations de notre méthode d'inversion portent sur le locuteur test utilisé pour construire le modèle. Nous avons commencé à étudier l'influence des disparités du modèle sur les résultats de l'inversion en dégradant artificiellement la qualité de la représentation du modèle avec le conduit vocal du locuteur. Dans cette thèse, nous avons examiné uniquement l'allongement et le raccourcissement de 10% de l'ensemble du conduit. Dans un futur proche, nous souhaitons étudier plus en détail ces influences en adaptant plus précisément le modèle à différents locuteurs.

Un des objectifs à plus long terme serait de modéliser le résidu du signal de parole qui n'est pas expliqué par le modèle de production sous la forme d'un processus stochastique. Le résidu serait destiné à prendre en compte les écarts entre le modèle d'analyse et le sujet mais aussi la source sonore qui excite le conduit vocal (vibration des cordes vocales ou bruit de friction). L'approche permettrait de combiner un modèle de production de la parole comme prédicteur, et une approche stochastique pour représenter l'erreur de prédiction et son évolution au cours du temps en fonction du type des sons articulés.

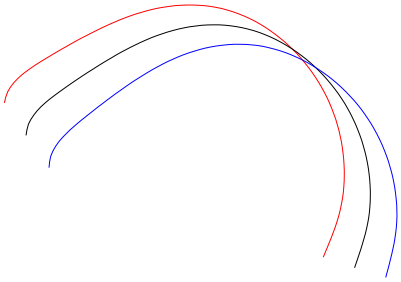
Annexes

Annexe A

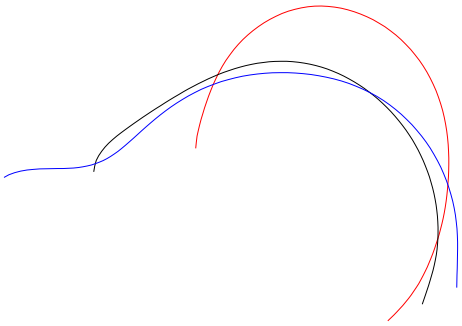
Composantes de la langue

Les différentes composantes en fonction des différentes approches et stratégies d'analyse sont présentées dans cette annexe. Pour toutes les composantes, nous avons :

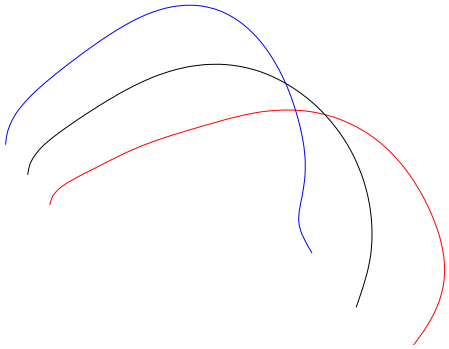
- la forme neutre en ligne noire,
- la forme correspondant à -3 écarts-types en ligne rouge et
- la forme correspondant à +3 écarts-types en ligne bleue.



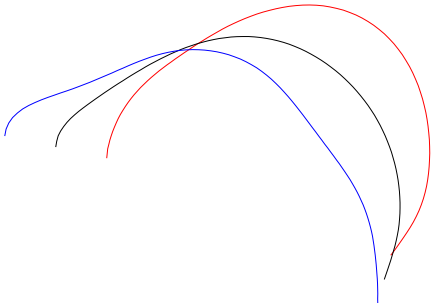
(a) Première composante de la mâchoire



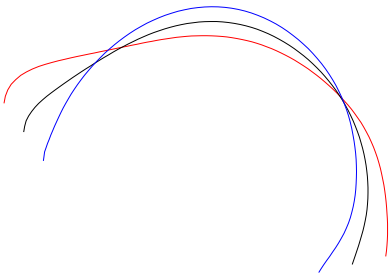
(b) Première composante de la langue (46,12%)



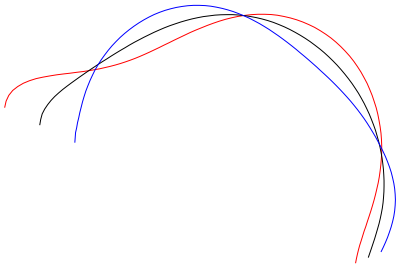
(c) Deuxième composante de la langue (28,84%)



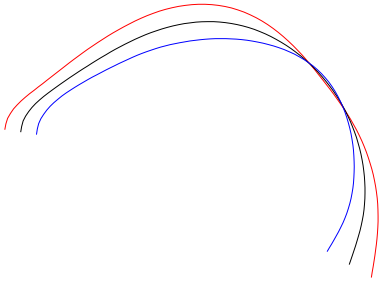
(d) Troisième composante de la langue (16,59%)



(e) Quatrième composante de la langue (4,02%)

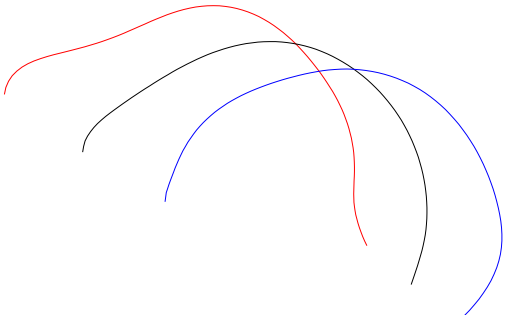


(f) Cinquième composante de la langue (1,92%)

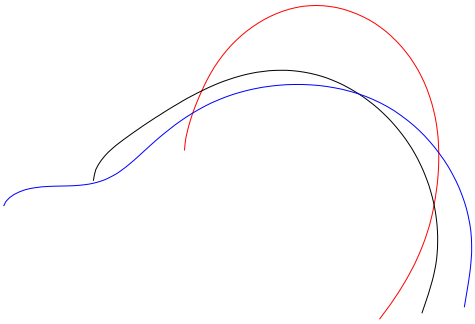


(g) Sixième composante de la langue (1,58%)

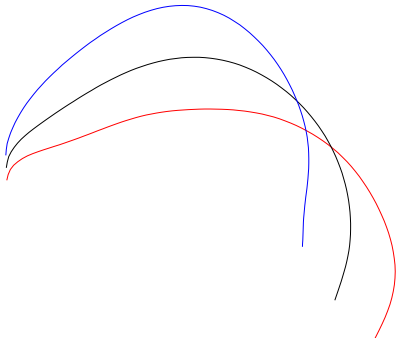
FIGURE A.1 – *Cascade ; coordonnées euclidiennes*



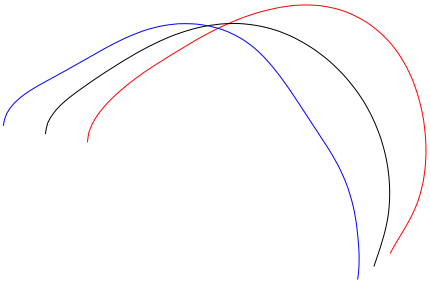
(a) Première composante de la mâchoire



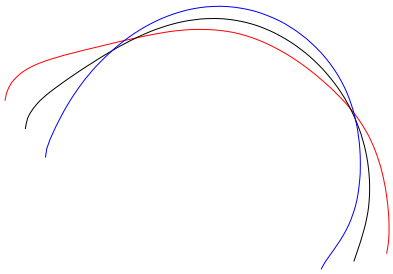
(b) Première composante de la langue (50,08%)



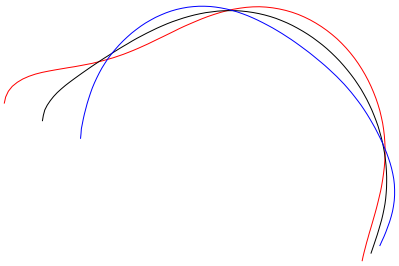
(c) Deuxième composante de la langue (24,38%)



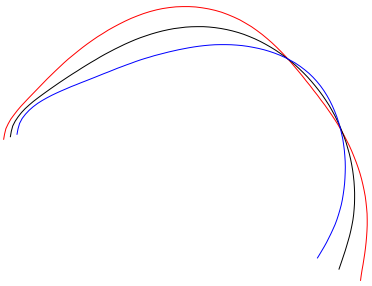
(d) Troisième composante de la langue (16,03%)



(e) Quatrième composante de la langue (4,52%)

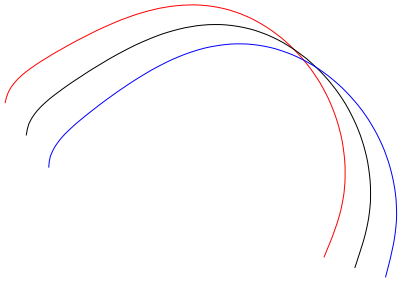


(f) Cinquième composante de la langue (2,18%)

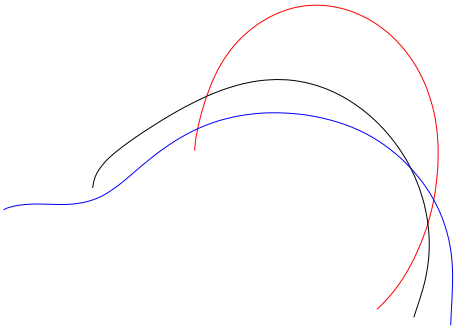


(g) Sixième composante de la langue (1,75%)

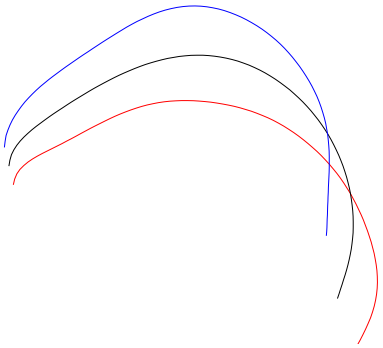
FIGURE A.2 – *Cascade + corrélation ; coordonnées euclidiennes*



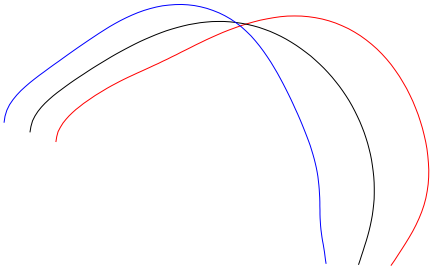
(a) Première composante de la mâchoire



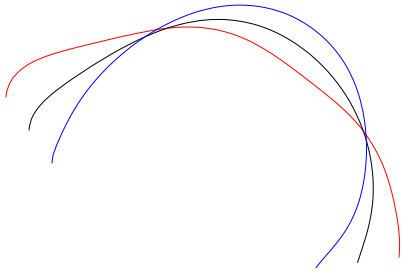
(b) Première composante de la langue (47,67%)



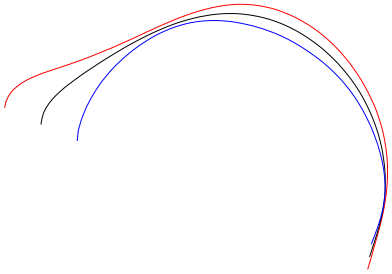
(c) Deuxième composante de la langue (26,10%)



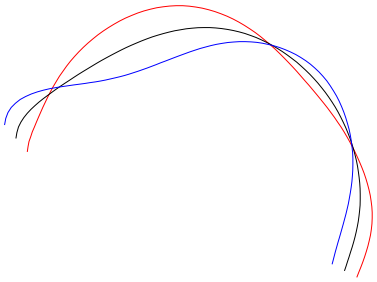
(d) Troisième composante de la langue (16,58%)



(e) Quatrième composante de la langue (5,28%)

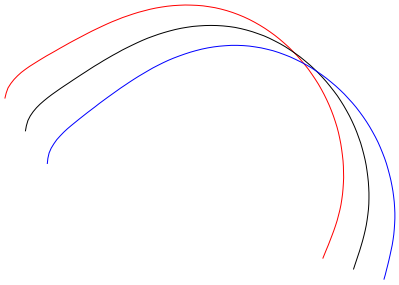


(f) Cinquième composante de la langue (1,70%)

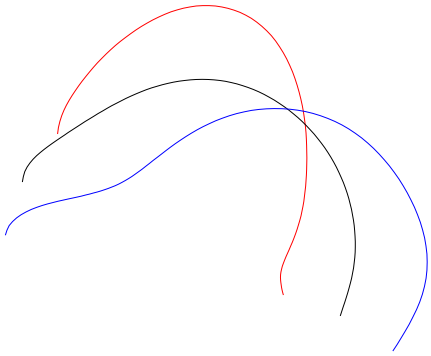


(g) Sixième composante de la langue (1,64%)

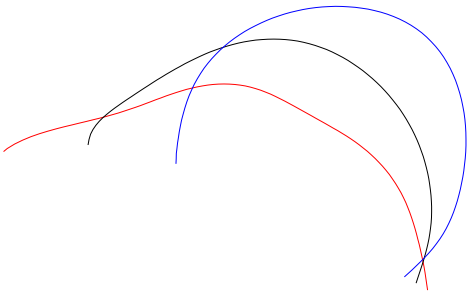
FIGURE A.3 – *Défaut ; coordonnées euclidiennes.*



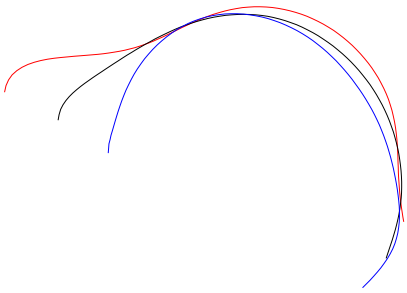
(a) Première composante de la mâchoire



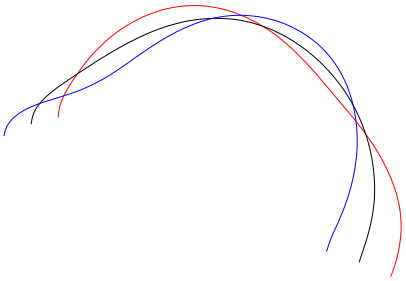
(b) Première composante de la langue (53,26%)



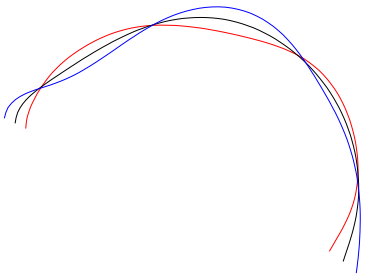
(c) Deuxième composante de la langue (32,16%)



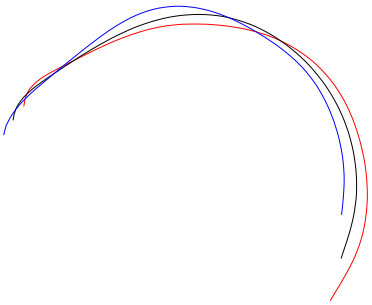
(d) Troisième composante de la langue (7,53%)



(e) Quatrième composante de la langue (4,62%)

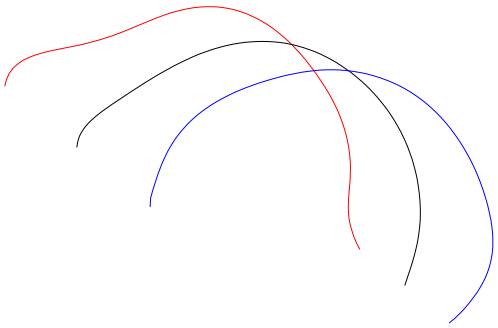


(f) Cinquième composante de la langue (0,96%)

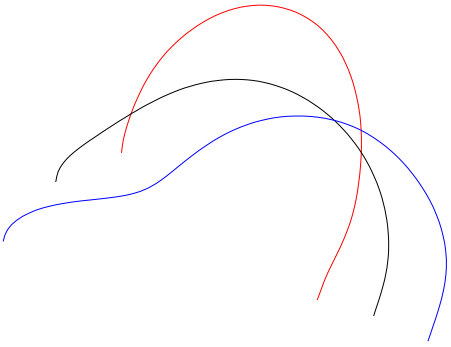


(g) Sixième composante de la langue (0,56%)

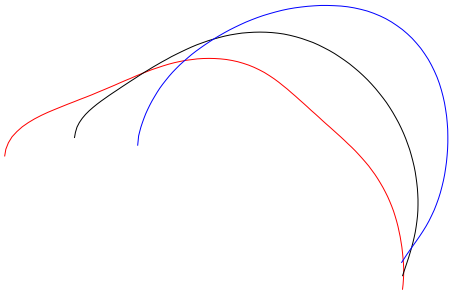
FIGURE A.4 – *Cascade ; distances + angles extrêmes.*



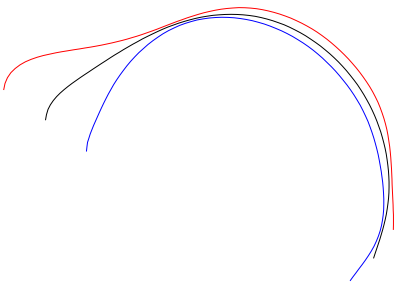
(a) Première composante de la mâchoire



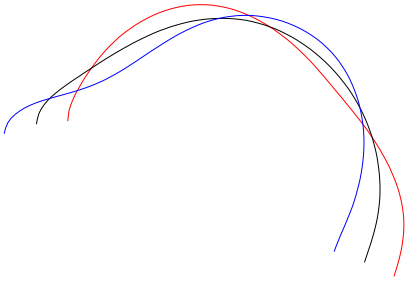
(b) Première composante de la langue (52,67%)



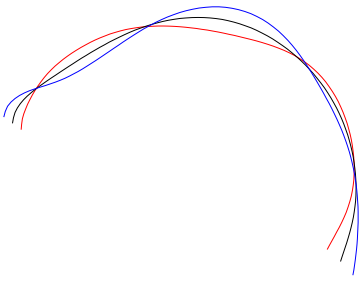
(c) Deuxième composante de la langue (29,86%)



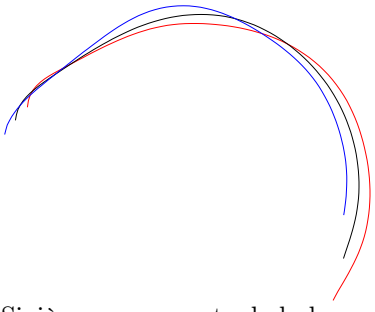
(d) Troisième composante de la langue (9,70%)



(e) Quatrième composante de la langue (5,03%)

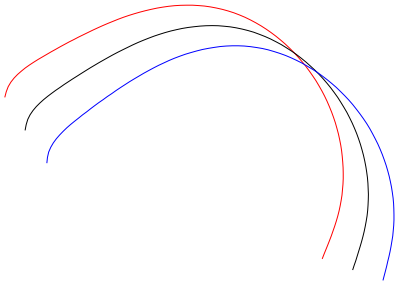


(f) Cinquième composante de la langue (1,10%)

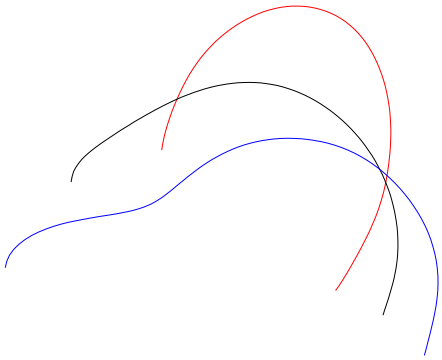


(g) Sixième composante de la langue (0,51%)

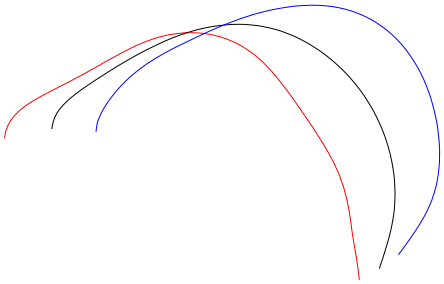
FIGURE A.5 – *Cascade + corrélation ; distances + angles extrêmes.*



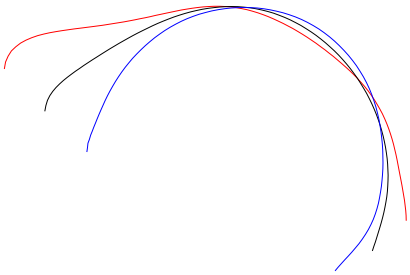
(a) Première composante de la mâchoire



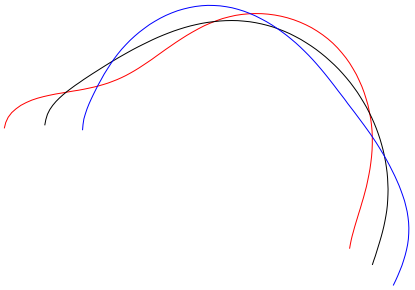
(b) Première composante de la langue (48,02%)



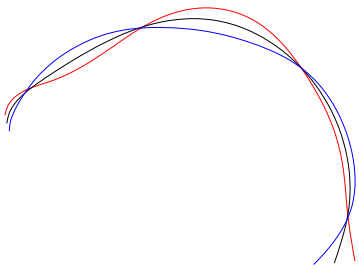
(c) Deuxième composante de la langue (31,63%)



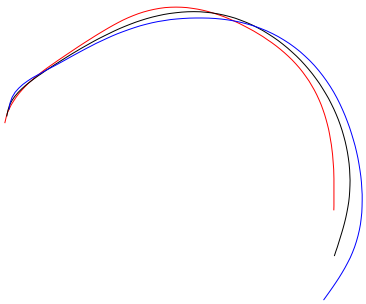
(d) Troisième composante de la langue (12,86%)



(e) Quatrième composante de la langue (5,03%)

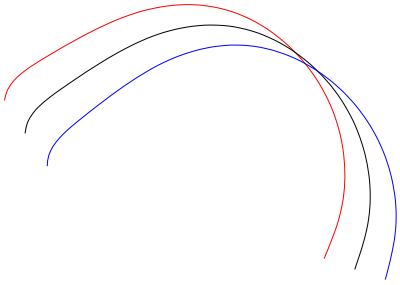


(f) Cinquième composante de la langue (0,96%)

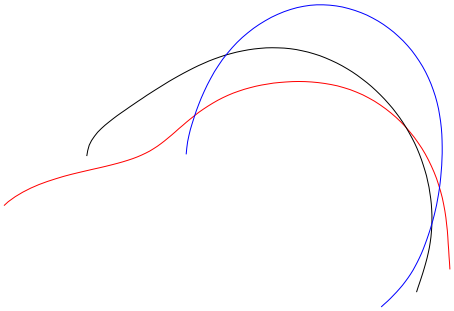


(g) Sixième composante de la langue (0,56%)

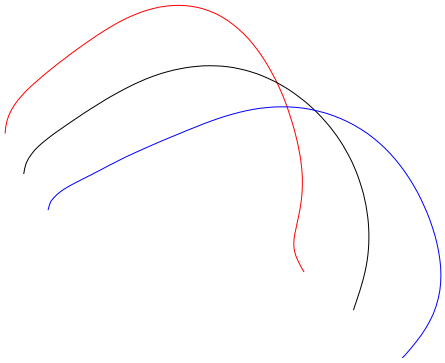
FIGURE A.6 – *Défaut ; distances + angles extrêmes.*



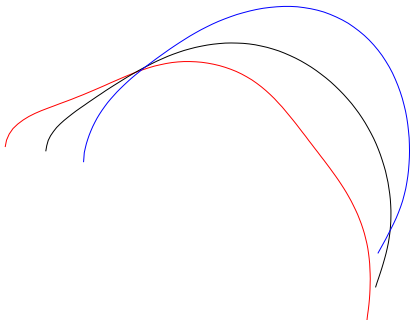
(a) Première composante de la mâchoire



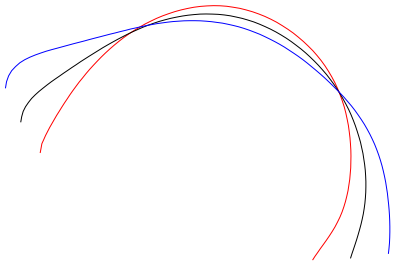
(b) Première composante de la langue (47,99%)



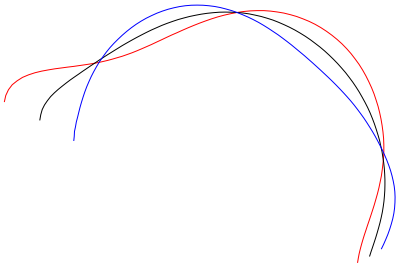
(c) Deuxième composante de la langue (29,05%)



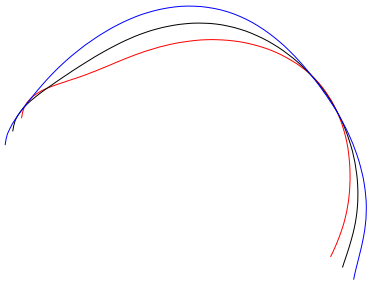
(d) Troisième composante de la langue (15,68%)



(e) Quatrième composante de la langue (3,30%)

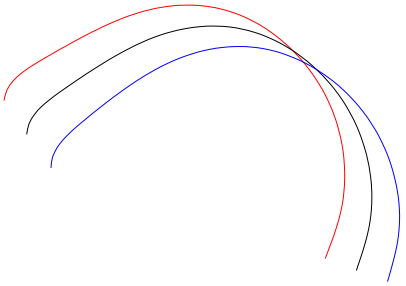


(f) Cinquième composante de la langue (1,98%)

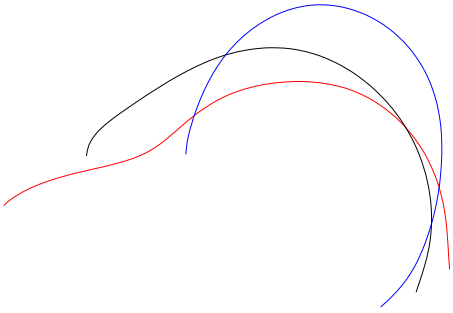


(g) Sixième composante de la langue (1,19%)

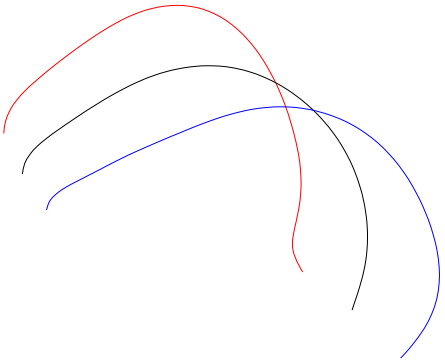
FIGURE A.7 – *Cascade ; coordonnées polaires.*



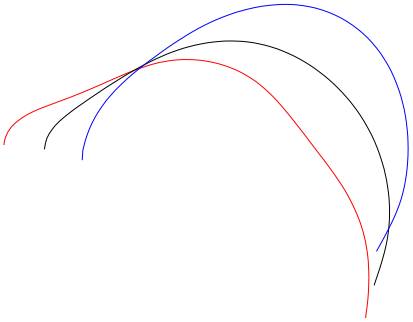
(a) Première composante de la mâchoire



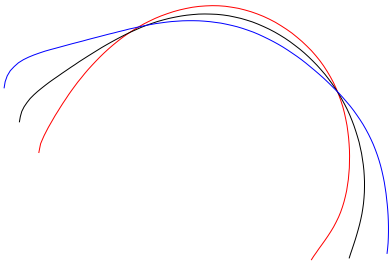
(b) Première composante de la langue (47,67%)



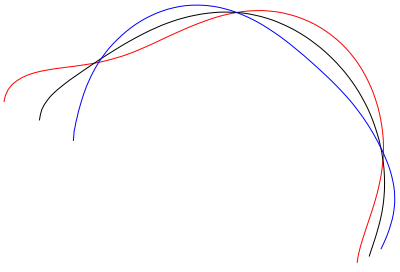
(c) Deuxième composante de la langue (25,74%)



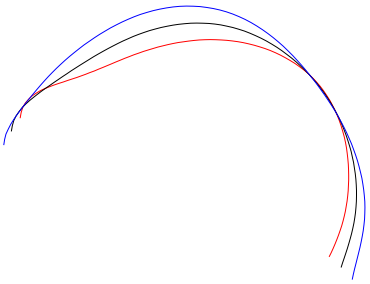
(d) Troisième composante de la langue (16,54%)



(e) Quatrième composante de la langue (5,84%)

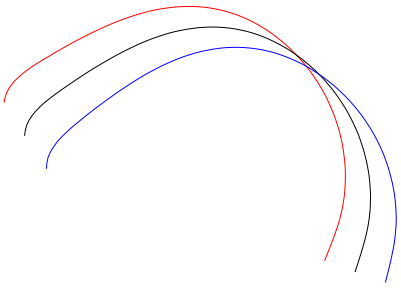


(f) Cinquième composante de la langue (2,11%)

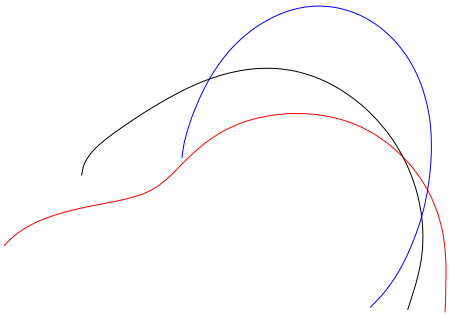


(g) Sixième composante de la langue (1,19%)

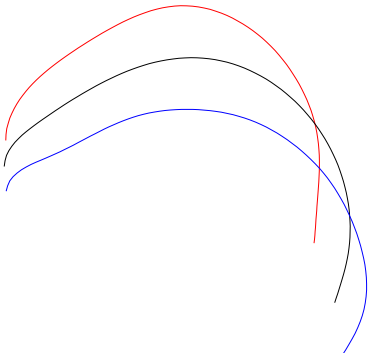
FIGURE A.8 – *Cascade + corrélation ; coordonnées polaires.*



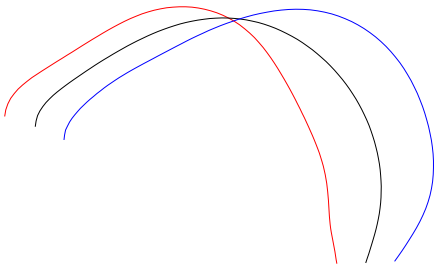
(a) Première composante de la mâchoire



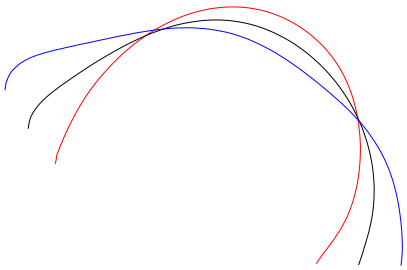
(b) Première composante de la langue (47,67%)



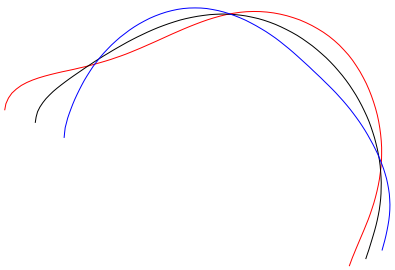
(c) Deuxième composante de la langue (25,74%)



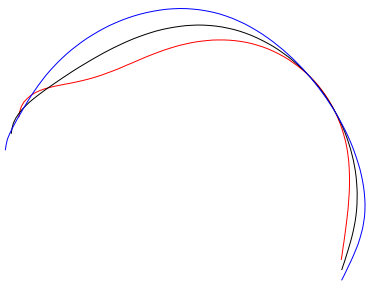
(d) Troisième composante de la langue (16,54%)



(e) Quatrième composante de la langue (5,84%)



(f) Cinquième composante de la langue (2,11%)



(g) Sixième composante de la langue (1,19%)

FIGURE A.9 – *Défaut ; coordonnées polaires.*

Annexe B

Construction de codebook

Les tableaux B.1, B.2 et B.3 présentent les caractéristiques des trois zones en faisant varier le volume minimal et le seuil acoustique. Pour chaque codebook, nous calculons le nombre d'hypercuboïdes, le volume total, le pourcentage de volume obtenu et l'erreur moyenne de resynthèse.

Zone 1

Volume minimal (subdivisions)	Seuil acoustique	Nombre d'hyper-cuboïdes	Volume total	Pourcentage de volume	Erreur moyenne de resynthèse
4,27 (16)	5	102	435,69	51,81%	2,54
	2	102	435,69	51,81%	2,54
	1	102	435,69	51,81%	2,54
	0,5	102	435,69	51,81%	2,54
2,13 (17)	5	228	486,95	57,90%	2,36
	2	228	486,95	57,90%	2,36
	1	228	486,95	57,90%	2,36
	0,5	228	486,95	57,90%	2,36
1,07 (18)	5	473	513,65	61,08%	2,05
	2	481	513,65	61,08%	2,04
	1	481	513,65	61,08%	2,04
	0,5	481	513,65	61,08%	2,04
0,53 (19)	5	949	528,60	62,79%	1,81
	2	989	528,06	62,79%	1,80
	1	989	528,06	62,79%	1,80
	0,5	989	528,06	62,79%	1,80
0,27 (20)	5	1894	550,75	65,49%	1,70
	2	2062	550,48	65,46%	1,68
	1	2062	550,48	65,46%	1,68
	0,5	2062	550,48	65,46%	1,68
0,13 (21)	5	3658	565,57	67,25%	1,61
	2	4230	564,64	67,14%	1,59
	1	4230	564,64	67,14%	1,59
	0,5	4230	564,64	67,14%	1,59
0,07 (22)	5	6771	574,79	68,35%	1,50
	2	8567	571,80	67,99%	1,44
	1	8567	571,80	67,99%	1,44
	0,5	8567	571,80	67,99%	1,44
0,03 (23)	5	12050	584,34	69,49%	1,36
	2	17369	579,77	68,94%	1,25
	1	17369	579,77	68,94%	1,25
	0,5	17369	579,77	68,94%	1,25

TABLEAU B.1 – *Caractéristiques de la zone 1 en fonction de différentes valeurs pour le seuil acoustique et la taille minimale d'un hypercuboïde (nombre d'hypercuboïdes, volume total, erreur moyenne de resynthèse).*

Zone 2

Volume minimal (subdivisions)	Seuil acoustique	Nombre d'hyper-cuboïdes	Volume total	Pourcentage de volume	Erreur moyenne de resynthèse
4,27 (16)	5	295	2101,57	97,14%	1,01
	2	492	2101,57	97,14%	0,96
	1	492	2101,57	97,14%	0,96
	0,5	492	2101,57	97,14%	0,96
2,13 (17)	5	392	2105,84	97,33%	0,99
	2	985	2103,71	97,24%	0,86
	1	985	2103,71	97,24%	0,86
	0,5	985	2103,71	97,24%	0,86
1,07 (18)	5	534	2106,91	97,38%	0,98
	2	1969	2102,64	97,19%	0,77
	1	1969	2102,64	97,19%	0,77
	0,5	1969	2102,64	97,19%	0,77
0,53 (19)	5	770	2113,85	97,70%	0,97
	2	3923	2108,51	97,46%	0,71
	1	3923	2108,51	97,46%	0,71
	0,5	3923	2108,51	97,46%	0,71
0,27 (20)	5	1163	2119,73	97,98%	0,96
	2	7687	2114,15	97,72%	0,66
	1	7919	2114,15	97,72%	0,66
	0,5	7919	2114,15	97,72%	0,66
0,13 (21)	5	1972	2121,88	98,08%	0,96
	2	14706	2116,65	97,83%	0,63
	1	15854	2116,75	97,84%	0,63
	0,5	15854	2116,75	97,84%	0,63
0,07 (22)	5	2175	2122,10	98,08%	0,98
	2	27380	2113,99	97,71%	0,56
	1	31667	2114,13	97,72%	0,55
	0,5	31667	2114,13	97,72%	0,55
0,03 (23)	5	3744	2124,66	98,20%	0,97
	2	48747	2111,40	97,59%	0,51
	1	62980	2101,61	97,14%	0,48
	0,5	62986	2097,96	96,97%	0,48

TABLEAU B.2 – *Caractéristiques de la zone 2 en fonction de différentes valeurs pour le seuil acoustique et la taille minimale d'un hypercuboïde (nombre d'hypercuboïdes, volume total, erreur moyenne de resynthèse).*

Zone 3

Volume minimal (subdivisions)	Seuil acoustique	Nombre d'hyper-cuboïdes	Volume total	Pourcentage de volume	Erreur moyenne de resynthèse
4,27 (16)	5	148	961,08	71,83%	0,92
	2	223	952,54	71,19%	0,82
	1	223	952,54	71,19%	0,82
	0,5	223	952,54	71,19%	0,82
2,13 (17)	5	242	1003,80	75,02%	0,88
	2	460	995,26	74,38%	0,73
	1	460	995,26	74,38%	0,73
	0,5	460	995,26	74,38%	0,73
1,07 (18)	5	398	1026,22	76,70%	0,84
	2	880	1009,13	75,42%	0,65
	1	945	1009,13	75,42%	0,65
	0,5	945	1009,13	75,42%	0,65
0,53 (19)	5	657	1052,92	78,69%	0,81
	2	1679	1031,56	77,09%	0,58
	1	1932	1031,56	77,09%	0,57
	0,5	1932	1031,56	77,09%	0,57
0,27 (20)	5	1110	1078,01	80,57%	0,80
	2	3239	1052,92	78,69%	0,54
	1	4851	1052,92	78,69%	0,52
	0,5	4851	1052,92	78,69%	0,52
0,13 (21)	5	1932	1096,97	81,98%	0,78
	2	6215	1071,62	80,09%	0,51
	1	7744	1071,36	80,07%	0,49
	0,5	8026	1071,36	80,07%	0,49
0,07 (22)	5	3420	1111,51	83,07%	0,76
	2	11353	1081,88	80,86%	0,45
	1	15006	1080,88	80,78%	0,41
	0,5	16191	1080,88	80,78%	0,41
0,03 (23)	5	5879	1124,83	84,07%	0,73
	2	19464	1099,25	82,15%	0,40
	1	28479	1097,06	81,99%	0,35
	0,5	32673	1090,59	81,51%	0,34

TABLEAU B.3 – *Caractéristiques de la zone 3 en fonction de différentes valeurs pour le seuil acoustique et la taille minimale d'un hypercuboïde (nombre d'hypercuboïdes, volume total, erreur moyenne de resynthèse).*

Annexe C

Appariement des pics spectraux par programmation dynamique

C.1 Rappel de l'algorithme

Soient $P = \big[p(m)\big] = p(1)\dots p(m)\dots p(M)$ et $Q = \big[q(n)\big] = q(1)\dots q(n)\dots q(N)$ deux ensembles correspondant aux pics de deux spectres, où $p(m)$ (resp. $q(n)$) correspond au $m^{ième}$ (resp. $n^{ième}$) pic avec M (resp. N) le nombre de pics de P (resp. Q).

Pour P et Q, on souhaite extraire un sous-ensemble de pics représenté par les séquences d'indices I et J.

$$I = \big[i(k)\big] = i(1)\dots i(k)\dots i(K) \text{ avec } K \leq M$$

$$J = \big[j(k)\big] = j(1)\dots j(k)\dots j(K) \text{ avec } K \leq N$$

I et J doivent préserver la monotonie de i et de j : $i(k) < i(k+1)$ et $j(k) < j(k+1)$.

Les séquences de pics $\bar{P}$ et $\bar{Q}$ correspondent aux séquences d'indices I et J.

$$\bar{P} = \big[p(i(k))\big] = p\big(i(1)\big)\dots p\big(i(k)\big)\dots p\big(i(K)\big)$$

$$\bar{Q} = \big[q(j(k))\big] = q\big(j(1)\big)\dots q\big(j(k)\big)\dots q\big(j(K)\big)$$

La détermination de i et j nécessite un critère qui minimise la distance entre deux pics de P et Q.

$$d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) = \Big|p\big(i(k)\big) - q\big(j(k)\big)\Big|; \quad \forall k \in K$$

Le critère global est défini par :

$$\min_{K,I,J} \sum_{k=1}^{K} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \tag{C.1}$$

$B\Big(p\big(i(k)\big)\Big) = b$ correspond au terme bonus avec b constant déterminé de façon ad hoc. Ce problème est résolu à l'aide de la programmation dynamique.

On définit la mesure partielle :

$$D(m,n) = \min_{i,j} \sum_{k=1}^{k^*} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \tag{C.2}$$

avec $i(k^*) = m$ et $j(k^*) = n$.

On décompose la somme en deux parties :

$$\begin{aligned} D(m,n) = \min_{i,j} \Bigg\{ & d\Big(p\big(i(k^*)\big), q\big(j(k^*)\big)\Big) - B\Big(p\big(i(k^*)\big)\Big) \\ & + \sum_{k=1}^{k^*-1} d\Big(p\big(i(k)\big), q\big(j(k)\big)\Big) - B\Big(p\big(i(k)\big)\Big) \Bigg\} \end{aligned} \tag{C.3}$$

$$\tag{C.4}$$

On pose $i(k^*-1) = l_1$ et $j(k^*-1) = l_2$, la formule de récursivité est donnée par :

$$D(m,n) = \min_{l_1<m, l_2<n} \Big\{ d\big(p(m), q(n)\big) - B\big(p(m)\big) + D(l_1, l_2) \Big\} \tag{C.5}$$

C.2 Résolution

La première étape consiste à calculer la matrice D et les matrices $indI$ et $indJ$ qui contiennent les antécédents.

Pour l_1 de 1 à M

Pour l_2 de 1 à N

Initialisation : $D(l_1, l_2) = 0$

$D(l_1, l_2) = d\Big(p\big(i(l_1)\big), q\big(j(l_2)\big)\Big) - B + \min_{ll_1<m, ll_2<n} D(ll_1, ll_2)$

On pose $\Delta = \min_{ll_1<m, ll_2<n} D(ll_1, ll_2) = D(\min ll_1, \min ll_2)$

$indI(l_1, l_2) = \min ll_1$

$indJ(l_1, l_2) = \min ll_2$

Finpour

Finpour

La seconde étape consiste à rechercher le minimum dans la matrice D. A partir des indices contenus dans les matrices $indI$ et $indJ$, on obtient les séquences I et J optimales.

Bibliographie

[Ace90] A. Acero. *Acoustical and environmental robustness in automatic speech recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1990.

[ACMT78] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *JASA*, 63(5) :1535–1555, May 1978.

[BBB01] D. Beautemps, P. Badin, and G. Bailly. Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America*, 109(5) :2165–2180, 2001.

[BBL95] D. Beautemps, P. Badin, and R. Laboissière. 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, 1995.

[BBR+02] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of the tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3) :533–553, 2002.

[BBV98] G. Bailly, Pierre Badin, and Anne Vilain. Inversion acoustique-articulatoire dynamique par codebook hypercuboique : premiers résultats. In *Journées d'Etudes sur la Parole - JEP'1998*, Martigny, Suisse, 1998.

[BF84] P. Badin and G. Fant. Notes on vocal tract computation. *Speech Transmission Laboratory–Quartely Progess Status Report*, 2-3/1984 :53–108, 1984.

[BML95] M.O. Berger, G. Mozelle, and Y. Laprie. Towards automatic extraction of tongue contours in X-ray images. In *Proceedings of the 9th Scandinavian Conference on Image Analysis*, pages 913–920, Upsala, Sweden, 1995.

[BPB92] L.-J. Boë, P. Perrier, and G. Bailly. The geometric vocal tract variables controlled for vowel production : proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20 :27–38, 1992.

[BYBBH09] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous. Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models. In *Interspeech 2009*, pages 2255–2258, Brighton, Royaume-Uni, September 2009. Département Parole et Cognition Département Parole et Cognition.

[Cal89] Calliope. Description acoustique. In *La parole et son traitement automatique*, chapter 3. Masson, Paris, 1989.

[Cha84] F. Charpentier. Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic non-linearities. *Speech Communication*, 3 :291–308, 1984.

[Cok76] C. H. Coker. A model of articulary dynamics and control. *Proceedings of the IEEE*, 64(4) :452–460, 1976.

[DH97] J. Dang and K. Honda. A physiological model of the tongue and jaw for simulating deformation in the midsagittal and parasagittal planes. *Journal of the Acoustical Society of America*, 102(5) :3167, 1997.

[EG96] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. *IEEE International Conference of Acoustics, Speech and Signal Processing*, 1 :346–349, 1996.

[EH90] J Edwards and K. S. Harris. Rotation and translation of the jaw during speech. *J Speech Hear Res*, 33(3) :550–562, 1990.

[Eng00] O. Engwall. A 3d tongue model based on mri data. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, Octobre, 2000.

[Eri07] C. Ericsdotter. Detail in vowel area functions. In *Proc of the 16th ICPhS*, pages 513–516, Saarbrücken, Germany, 2007.

[Fan70] G. Fant. Analytical constraints on the composition of speech spectra. In *Acoustic Theory of Speech Production, Second Printing*, pages 48–62. The Hague : Mouton & Co., 1970.

[FB06] J. Fontecave and F. Berthommier. Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data. In *Interspeech, Pittsburgh*, September 2006.

[Fla72] J. L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer-Verlag, 2nd ed, New York, 1972.

[GI83] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27 :1–33, 1983.

[GR97] A. Galván-Rodríguez. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives.* Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[GWTPP03] J.-M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan. A 3D dynamical biomechanical tongue model to study speech motor control. *Research Developments in Biomechanics*, 1 :49–64, 2003.

[HH04] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Trans. Speech and Audio Process.*, 12(2) :175–185, March 2004.

[HLG+96] J. Hogden, A. Löfqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics : new conclusions based on human data. *Journal of the Acoustical Society of America*, 100 :1819–1834, 1996.

[HS65] J. M. Heinz and K. N. Stevens. On the relations between lateral cineradiographs, area functions and acoustic spectra of speech. In *Proceedings of the 5th International Congress on Acoustics*, page A44., 1965.

[JM08] T.-T. M. Jackson and R. S. McGowan. Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels : Statistical considerations. *Journal of the Acoustical Society of America*, 123(1) :336–346, 2008.

[KIF75] S. Kiritani, K. Itoh, and O. Fujimura. Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *The Journal of the Acoustical Society of America*, 57(6) :1516–1520, 1975.

[KM75] S. Kiritani and K. Miyawaki. Computational model of the tongue. *Journal of the Acoustical Society of America*, 57 :S3A, 1975.

[LSS88] J. N. Larar, J. Schroeter, and M. M. Sondhi. Vector quantization of the articulatory space. *IEEE Trans. ASSP*, 36(12) :1812–1818, December 1988.

[Mae79] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[Mae82] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1 :199–229, 1982.

[Mae90] S. Maeda. Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[Man13] R. Mannell. Sound sources in the vocal tract. *http ://clas.mq.edu.au/acoustics/frequency/source.html*, consulté en février 2013.

[Mat99] B. Mathieu. *Modèles de production de parole et reconnaissance à partir d'automates.* PhD thesis, Université Henri Poincaré - Nancy I, 1999.

[McD98] J. McDonough. Speaker normalization with all-pass transforms. In *Technical Report No. 28, Center for Language Speech Processing, The Johns Hopkins University*, Baltimore, MD, USA, Septembre 1998.

[Mer67] P. Mermelstein. Determination of the vocal-tract shape from measured formant frequencies. *Journal of the Acoustical Society of America*, 41 :1283–1294, 1967.

[Mer73] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53 :1070–1082, 1973.

[MG76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech.* Springer-Verlag, Berlin Heidelberg New York, 1976.

[MKTH04] P. Mokhtari, T. Kitamura, H. Takemoto, and K. Honda. Evaluation of an lp-based method of inversion using mri-based vocal-tract area functions. In *Autumn Meeting of the Acoustical Society of Japan*, pages 237—238, Okinawa, Japan, 2004.

[ML97] B. Mathieu and Y. Laprie. Adaptation of Maeda's model for acoustic to articulatory inversion. In *Proceeding of Eurospeech 1997*, volume 4, pages 2015–2018, Rhodes, Greece, September 1997.

[MSS91] P. Meyer, J. Schroeter, and M. M. Sondhi. Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks. *IEEE Trans. ASSP*, 39(7) :1493–1502, 1991.

[MTTB12] R. S. McGowan, M. T-T.Jackson, and M. A. Berger. Analyses of vocal tract cross-distance to area mapping : An investigation of a set of vowel images. *The Journal of the Acoustical Society of America*, 131(1) :424–434, 2012.

[NNL+04] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *JASA*, 115(4) :1771–1776, April 2004.

[OJ72] A.V. Oppenheim and D.H. Johnson. Discrete representation of signals. *Proceedings of the IEEE*, 60(6) :681 – 691, june 1972.

[OL00] S. Ouni and Y. Laprie. Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire. In *Actes des Journées d'Étude sur la parole, Aussois*, June 2000.

[OL05] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *J. of the Acous. Soc. Am.*, 118(1) :444–460, 2005.

[Oun01] S. Ouni. *Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico-articulatoire.* Thèse de L'Université Henri Poincaré, Dec 2001.

[OVBG97] D. J. Ostry, E. Vatikiotis-Bateson, and P. L. Gribble. An examination of the degrees of freedom of human jaw motion in speech and mastication. *J Speech Lang Hear Res*, 40(6) :1341–1351, 1997.

[Ove62] J. E. Overall. Orthogonal factors and uncorrelated factor scores. *Psychological Reports*, pages 651–662, 1962.

[PA08] S. Panchapagesan and A. Alwan. Vocal tract inversion by cepstral analysis-by-synthesis using chain matrices. In *Proceeding of Interspeech 2008*, pages 2857–2860, 2008.

[PA11] S. Panchapagesan and A. Alwan. A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model. *The Journal of the Acoustical Society of America*, 129(4) :2144–2162, 2011.

[Pan08] S. Panchapagesan. *Frequency warping by linear transformation, and vocal tract inversion for speaker normalization in automatic speech recognition.* PhD thesis, University of California (Los Angeles), 2008.

[PBS92] P. Perrier, L.-J. Boë, and R. Sock. 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*, 35 :53–67, 1992.

[Per74] J. S. Perkell. *A physiologically-oriented model of tongue activity in speech production.* PhD thesis, Massachusetts Institute of Technology., 1974.

[PHT+92] G. Papcun, J. Hochberg, T.R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray micro- Beam data. *Journal of the Acoustical Society of America*, 92(2) :688–700, 1992.

[PL07] B. Potard and Y. Laprie. Compact representations of the articulatory-to-acoustic mapping. In *Proceedings of Interspeech 2007*, Antwerpen, Belgium, 27-31 August, 2007.

[PL09] B. Potard and Y. Laprie. A robust variational method for the acoustic-to-articulatory problem. In *10th Annual Conference of the International Speech Communication Association - INTERSPEECH 2009*, United Kingdom, Brighton, 2009.

[PLO04] B. Potard, Y. Laprie, and S. Ouni. Expériences d'inversion basées sur un modèle articulatoire. In *Journées d'Etudes sur la Parole - JEP'04 , Fès, Maroc*, April 2004.

[PMSN01] M. Pitz, S. Molau, R. Schlüter, and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. In *Proceeding of the EUROSPEECH 2001*, pages 2653–2656, 2001.

[Pot08a] B. Potard. *Inversion acoustique-articulatoire avec contraintes*. PhD thesis, Université Henri Poincaré - Nancy I, 2008.

[Pot08b] B. Potard. Inversion acoustique-articulatoire dynamique par codebook hypercuboique : premiers résultats. In *Journées d'Etudes sur la Parole - JEP'08*, Avignon, France, June 2008.

[PP97] Y. Payan and P. Perrier. Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Communications*, 22 (2/3) :185–205, 1997.

[Ric01] K. Richmond. Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech. In *Workshop on Innovation in Speech Processing, Institute of Acoustics*, pages 259–276, 2001.

[RJ93] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[RS78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

[SC95] J. Schoentgen and S. Ciocea. Direct calculation of the vocal tract area function from measured formant frequencies. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, volume 1, pages 745–748, Madrid, Spain, September, 1995.

[SH55] K. N. Stevens and A. S. House. Development of a quantitative description of vowel articulation. *JASA*, 27 :484–493, 1955.

[SHL+11] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Ferbach-Hecker, L. Ma, J. Busset, and J. Sturm. An X-ray database, tools and procedures for the study of speech production. In V.L. Gracco D.J. Ostry L. Ménard, S.R. Baum, editor, *Proceedings of the 9th International Seminar on Speech Production (ISSP2011)*, pages 41–48, Montréal, Canada, June 2011. Département Parole et Cognition de GIPSA-lab Département Parole et Cognition de GIPSA-lab.

[SLMD02] A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function transformations : Direct measurements of mid-sagittal distance and area with mri. *Speech Communication*, 36 :168–180, 2002.

[SLO98] V. Sanguineti, R. Laboissière, and D.J. Ostry. A dynamic biomechanical model for the neural control of speech production. *Journal of the Acoustical Society of America*, 103 :1615–1627, 1998.

[SMCJ99] C.H. Shadle, M. Mohammad, J.N. Carter, and P.J.B. Jackson. Multi-planar dynamic magnetic resonance imaging : new tools for speech research. In *ICPhS*, pages 623–626, 1999.

[SMP90] J. Schroeter, P. Meyer, and S. Parthasarathy. Evaluation of improved articulatory codebooks and codebook access distance measures. In *International Conference on Acoustics, Speech and Signal Processing*, pages 393–396, Albuquerque, NM, USA, April 1990.

[Sor92] V. N. Sorokin. Determination of vocal tract shape for vowels. *Speech Communication*, 11 :71–85, 1992.

[SS92] J. Schroeter and M. M. Sondhi. Speech coding based on physiological models of speech production. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 231–267. Dekker, New York, 1992.

[SSJ91] A. Soquet, M. Saerens, and P. Jospa. Acoustic-articulatory inversion based on a neural controller of a vocal tract model : further results. In O. Simula T. Kohonen, K. Mäkisara and J. Kangas, editors, *Artificial Neural Networks*, pages 371–376. North Holland : Elsevier, 1991.

[ST96] V.N. Sorokin and A.V. Trushkin. Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19 :105–118, 1996.

[TBT08] T. Toda, A. W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50 :215–227, 2008.

[TL99] G. Thimm and J. Luettin. Extraction of articulators in x-ray image sequences. In *Proc. EUROSPEECH*, pages 157–160, Budapest, september 1999.

[UMRR12] B. Uria, I. Murray, S. Renals, and K. Richmond. Deep architectures for articulatory inversion. In *Interspeech 2012*, Portland, Oregon, September 2012.

[Wes88] J. R. Westbury. Mandible and hyoid bone movements during speech. *Journal of Speech and Hearing Research*, 31(2) :405–416, 1988.

[WT95] R. Wilhelms-Tricarico. Physiological modeling of speech production : Methods for modeling soft-tissue articulators. *J. Acoust. Soc. Am.*, 97 :3805, 1995.

[ZNT11] H. Zen, Y. Nankaku, and K. Tokuda. Continuous stochastic feature mapping based on trajectory hmms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2) :417 –430, feb. 2011.

[ZR08] Le Zhang and S. Renals. Acoustic-articulatory modeling with the trajectory hmm. *Signal Processing Letters, IEEE*, 15 :245 –248, 2008.

Résumé

L'inversion acoustique-articulatoire de la parole consiste à récupérer la forme du conduit vocal à partir d'un signal de parole. Ce problème est abordé à l'aide d'une méthode d'analyse par synthèse reposant sur un modèle physique de production de la parole contrôlé par un petit nombre de paramètres décrivant la forme du conduit vocal : l'ouverture de la mâchoire, la forme et la position de la langue et la position des lèvres et du larynx. Afin de s'approcher de la géométrie de notre locuteur, le modèle articulatoire est construit à l'aide de contours articulatoires issus d'images cinéradiographiques présentant une vue sagittale du conduit vocal.

Ce synthétiseur articulatoire nous permet de créer une table formée de couples associant un vecteur articulatoire au vecteur acoustique correspondant. Nous n'utiliserons pas les formants (fréquences de résonance du conduit vocal) comme vecteur acoustique car leur extraction n'est pas toujours fiable provoquant des erreurs lors de l'inversion. Les coefficients cepstraux sont utilisés comme vecteur acoustique. De plus, l'effet de la source et les disparités entre le conduit vocal du locuteur et le modèle articulatoire sont pris en compte explicitement en comparant les spectres naturels à ceux produits par le synthétiseur car nous disposons des deux signaux.

Mots-clés: Inversion, acoustique, articulatoire, analyse par synthèse, coefficients cepstraux, modèle articulatoire

Abstract

The acoustic-to-articulatory inversion of speech consist in the recovery of the vocal tract shape from the speech signal. This problem is tackled with an analysis-by-synthesis method depending on a physical model of speech production controlled by a small number of parameters describing the vocal tract shape : the jaw opening, the shape and the position of the tongue and the position of lips and larynx. In order to approach the geometry of the speaker, the articulatory model is built with articulatory contours from cineradiographic images of the sagittal view of the vocal tract.

This articulatory synthesizer allows us to create a table made up with couples associating a articulatory vector with the corresponding acoustic vector. The formants (resonance frequency of the vocal tract shape) are not used as acoustic vector because their extraction is not always reliable causing errors during inversion. The cepstral coefficients are used as acoustic vector. Moreover, the source effect and the mismatch between the speaker vocal tract and the articulatory model are considered explicitly comparing the natural spectrum with those produced by the synthesizer because we have the both signals.

Keywords: Inversion, acoustic, articulatory, analysis-by-synthesis, cepstral coefficients, articulatory model