Nancy-Université
Université Henri Poincaré

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

Toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

➢ Contact SCD Nancy 1 : theses.sciences@scd.uhp-nancy.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

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

Inversion acoustique-articulatoire avec contraintes

THÈSE

présentée et soutenue publiquement le 23 octobre 2008

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Blaise Potard

Composition du jury

Rapporteurs :	Régine ANDRE-OBRECHT, Professeur, Université Paul Sabatier Pierre BADIN, Directeur de Recherche, CNRS
Examinateurs :	Noëlle CARBONELL, Professeur, Université Henri Poincaré Shinji MAEDA, Directeur de Recherche, CNRS Xavier RODET, Professeur, Université Paris VI
Directeur de thèse :	Yves LAPRIE, Directeur de Recherche, CNRS

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

Mis en page avec la classe thloria.

Remerciements

Je tiens tout particulièrement à remercier mon directeur de thèse, Yves Laprie, pour sa patience, ses conseils et ses encouragements. Je tiens également à remercier Shinji Maeda pour nous avoir fourni son modèle articulatoire et ses données, et pour les échanges scientifiques fructueux que nous avons pu avoir.

Je remercie profondément Martine Kuhlmann, la secrétaire de l'équipe, pour sa patience et sa gentillesse sans limites.

Je remercie ensuite tous les enseignants du département Informatique de l'Université Henri Poincaré – Nancy 1, pour leurs conseils, leur patience et leur soutien, et tout particulièrement Alain Mirgaux, Odile Mella et Vincent Colotte.

Enfin, je remercie ma famille et tous mes amis, pour leur soutien sans faille.

À mon père, ma mère, ma sœur Marguerite, mes frères Camille, Louis, Clément et Marin, loin de Nancy mais toujours près de mon cœur ;
À Delphine, Stéf, Matthias, Ben, Nico, Véro et tous les autres, parce qu'ils le valent bien ;
À tous les membres du Loria grâce à qui j'ai pu passer de très bons moments, tout particulièrement les occupants du bureau C130, les membres des équipes Parole, Magrit et Maia, et tous ceux que je ne cite pas mais qui se reconnaîtront.
Et à tous ceux qui sont partis.

Table des matières

Introduction

Chapitre 1
Préliminaires

1.1 Acoustique de la parole 1
1.1.1 Production de la parole 1
1.2 Synthèse articulatoire 4
1.2.1 Modélisation du conduit vocal 4
1.2.2 Passage de la coupe sagittale à la fonction d'aire 7
1.2.3 Simulation acoustique 9
1.3 Inversion acoustique-articulatoire 11
1.3.1 Problème mal posé 12
1.3.2 Principe de résolution 12
1.3.3 Résolution par introduction de contraintes 15
1.4 Conclusion 16

Partie I Inversion **17**

Introduction

Chapitre 2
Construction de codebook hypercuboïque

2.1 Paramétrisation acoustique et articulatoire 21
2.1.1 Modèle articulatoire de Maeda 21
2.1.2 Synthétiseur 22
2.1.3 Paramétrisation acoustique 23
2.1.4 Quelques résultats 23
2.2 Présentation de la structure hypercuboïdale 24

2.2.1 Choix de la structure 24
2.2.2 Définitions de la structure hypercuboïdale 24
2.2.3 Modélisation mathématique 25
2.3 Construction 27
2.3.1 Test de régularité 27
2.3.2 Utilisation des polynômes multivariés pour calculer l'approximation . 30
2.3.3 Seuillage acoustique 33
2.3.4 Subdivision 34
2.4 Réalisation 35
2.4.1 Calcul du vecteur acoustique 36
2.4.2 Cache de formants 37
2.4.3 Génération des points de test 39
2.5 Évaluation expérimentale 43
2.5.1 Resynthèse d'un vecteur articulatoire par interpolation 43
2.5.2 Valeurs optimales pour le calcul de la matrice jacobienne 44
2.5.3 Seuils de subdivision et précision acoustique 44
2.5.4 Temps de construction 46
2.5.5 Couverture de l'espace articulatoire 47

Chapitre 3
Inversion par codebook

3.1 La méthode d'inversion 51
3.1.1 Principe général 52
3.1.2 Restreindre l'ensemble d'hypercuboïdes à explorer 52
3.1.3 Résolution du système d'équation $P(X) = s$ 53
3.1.4 Échantillonnage de solutions 55
3.2 Quelques résultats 56
3.2.1 Inversion de voyelles isolées 56
3.2.2 Domaines acoustiques 59
3.2.3 Statistiques 60
3.2.4 Temps de calcul 65

Chapitre 4
Trajectoires articulatoires

4.1 Lissage non-linéaire 70
4.1.1 Critères sur la régularité de la trajectoire des articulateurs 70
4.1.2 Critère global à minimiser 71

4.1.3 Complexité . . . 73
4.2 Régularisation variationnelle . . . 73
4.3 Inversion dynamique . . . 74
4.3.1 Inversion de la synthèse acoustique . . . 75
4.4 Conclusion . . . 79

Partie II Contraintes **81**

Introduction

Chapitre 5
Contraintes phonétiques

5.1 Principe . . . 85
5.2 Domaines articulatoires . . . 85
5.2.1 Classification des phonèmes . . . 85
5.2.2 Transposition des contraintes phonétiques dans le modèle articulatoire 86
5.3 Domaines acoustiques . . . 87
5.3.1 Partitionnement de l'espace acoustique . . . 88
5.3.2 Données acoustiques . . . 89
5.4 Score phonétique . . . 90
5.5 Inversion avec contraintes phonétiques . . . 92
5.5.1 Construction de codebooks . . . 92
5.5.2 Inversion statique . . . 93
5.5.3 Inversion dynamique . . . 93
5.6 Exemples . . . 94

Chapitre 6
Évaluation des contraintes phonétiques

6.1 Correspondance articulatoire-acoustique . . . 99
6.2 Inversion statique . . . 101
6.2.1 Expériences d'inversion . . . 101
6.2.2 Analyse des résultats . . . 102
6.3 Inversion dynamique . . . 113
6.3.1 Inversion de transitions Voyelle-Voyelle . . . 113
6.3.2 Récapitulatif . . . 120
6.4 Conclusions et perspectives . . . 122

Chapitre 7
Contraintes visuelles

7.1 Origine 125
7.2 Données multimodales 126
7.2.1 Minicorpus 127
7.2.2 Corpus AL 128
7.2.3 Données du projet ASPI 128
7.3 Traitement des données 128
7.4 Implémentation des contraintes visuelles 129
7.4.1 Correspondance entre les marqueurs et le modèle articulatoire 130
7.4.2 Inversion avec contraintes visuelles 133
7.5 Expériences d'inversion 133
7.5.1 Comparaison des deux méthodes 133
7.5.2 Expériences d'inversion 135
7.6 Conclusion 137

Conclusions et perspectives **139**

Conclusions et perspectives

Annexe A
Corpus PB

A.1 Paramètres articulatoires 143

Annexes

Bibliographie **147**

Introduction

L'UNE des principales caractéristiques qui distinguent l'homme de l'animal est l'utilisation de la communication parlée. Bien que ne possédant pas l'appareil phonatoire le plus sophistiqué (certains oiseaux peuvent en effet réaliser des sons nettement plus élaborés), la parole et les mécanismes de production s'y rattachant restent parmi les manifestations les plus complexes et les moins bien connus du génie humain.

L'Homme n'a eu de cesse d'étudier ce phénomène, et depuis que les ordinateurs existent, de chercher à le recréer artificiellement. Depuis relativement longtemps, on a réalisé des modèles logiciels de l'appareil phonatoire, capables d'imiter la voix humaines en simulant les équations physiques de l'acoustique de façon plus ou moins simplifiée, et avec plus ou moins de succès. Ces modèles de synthèse peuvent parfois émettre des sons qui ressemble à s'y méprendre à une voix humaine réelle ; malheureusement, pour piloter fidèlement de tels modèles, il est nécessaire de disposer de données sur le mouvement des différents articulateurs, et sur la source sonore.

Il existe des techniques d'imagerie permettant d'obtenir des informations sur la position des articulateurs, mais aucune de ces techniques n'est parfaite : soit elles ne capturent que partiellement le conduit (images du visage du locuteur permettant de voir la position des lèvres et de la mâchoire, capteurs magnétiques à des positions précises à l'intérieur de la bouche, échographie...), soit n'ont pas une résolution temporelle suffisante pour capturer les subtilités de l'articulation de la parole naturelle (par exemple l'IRM, qui permet d'obtenir une vue tridimensionnelle du conduit vocal, mais nécessite des temps de pose prohibitifs, ou la répétition d'une même phrases des centaines de fois), soit encore sont franchement dangereuses (par exemple la cinéradiographie, où le locuteur subit des centaines de radiographies sur une durée de quelques secondes).

L'obtention logicielle des mouvements articulatoires à l'origine d'un son donné, ou *inversion acoustique-articulatoire*, voire tout simplement *inversion*, est une proposition alternative et complémentaire à ces techniques d'imagerie. Il s'agit en réalité d'un problème clé pour de nombreuses applications. Outre les modèles de synthèse vocale, ces données peuvent en effet être utiles dans de nombreux cas : la compression de parole pour la téléphonie, l'étude des langues, et surtout l'animation de têtes parlantes (pour l'industrie cinématographique et des loisirs, l'apprentissage des langues, l'aide aux personnes malentendantes...). Par ailleurs, les données articulatoires peuvent aussi trouver leur utilité dans des domaines aussi disparates que la reconnaissance automatique de la parole, la synthèse vocale par concaténation, la phonétique et la phonologie, la médecine...

De part sa position centrale dans tous les domaines liés au traitement automatique de la parole, et par sa myriade d'applications potentielles, l'inversion acoustique-articulatoire est l'un des points cruciaux de l'étude de la parole. D'ailleurs, très tôt, les pionniers de l'étude de la parole se sont intéressés à la modélisation articulatoire du conduit vocal (notamment Fant (Fant 1960), Stevens (Stevens & House 1955), Flanagan (Ishizaka & Flanagan 1972)...). Mais la véritable étude fondamentale du domaine de l'inversion, et également l'une des plus intéressantes, est le travail réalisé par Atal et ses collègues (Atal *et al.* 1978).

Malheureusement, les quatre années consacrées à l'étude de ce sujet ne me permettent pas

d'apporter une réponse aux nombreuses questions qui entourent ce domaine, et en premier lieu : l'inversion acoustique-articulatoire est-elle possible pour tous les sons de la parole ? Et, autrement plus problématique, est-ce que notre méthode fonctionne ?

L'une des difficultés majeures est en effet l'évaluation de l'inversion. Car, il faut se rendre à l'évidence, la quantité de données articulatoires *utilisables* est assez restreinte, les données étant souvent obtenues dans des conditions problématiques (systèmes invasifs, ou parole non spontanée), ou difficilement exploitables (enregistrement sonore non disponible ou non synchronisé...). Pour remédier à ce problème, l'équipe PAROLE du LORIA, et plusieurs partenaires européens, ont monté le projet européen ASPI (financé par le Programme IST de la Commission des Communautés Européennes, avec le numéro IST-2005-021324), dont l'un des objectifs est l'acquisition de données articulatoires selon des modalités bien définies, de manière à pouvoir évaluer de façon rigoureuse les différentes méthode d'inversion développées par chacun des partenaires. Malheureusement, la mise en place des systèmes d'acquisition a pris plus de temps que prévu, et par conséquent les données acquises n'ont pu que très partiellement être exploitées dans cette thèse.

Par conséquent, la comparaion avec des données réelles n'a été qu'assez rarement possible, la validation de la méthode d'inversion n'a été que partielle, et les résultats des expériences d'inversion présentées dans ce mémoire sont donc à considérer avec précaution.

Néanmoins, il apparaît que le travail effectué permet sur de nombreux points d'améliorer, sinon la fidélité des résultats aux trajectoires réelles (que l'on ne connaît en général pas), au moins la fiabilité, la rapidité, la stabilité du processus d'inversion. Les contributions sont de plusieurs ordres : une grande parties du travail a portée sur la méthode d'inversion elle-même, qui est est un prolongement des travaux effectués dans l'équipe, d'abord par Bruno Mathieu (Laprie & Mathieu 1998a), puis par Slim Ouni (Ouni 2001). Ce travail a paradoxalement été effectué à la fin de ma thèse, ce qui explique que les améliorations ne sont pas forcément toutes exploitées dans les autres parties.

L'autre grand axe de recherche a été l'élaboration de contraintes pour limiter l'espace de solutions de l'inversion. Les contraintes étudiées ici exploitent l'information contextuelle inhérente au processus de production de la parole.

L'hypothèse généralement retenue dans le cadre de l'inversion acoustique-articulatoire est que le processus de production cherche à minimiser la dépense énergétique, mais outre les problèmes liés à la modélisation du coût énergétique, il semble en réalité que le processus de production de la parole n'est pas seulement guidé par un principe de moindre effort, mais cherche également à suivre des caractéristiques articulatoires propres au dialecte parlé. L'hypothèse qui sous-tend la première classe de contraintes proposée ici repose sur un invariant articulatoire de la production des phonèmes d'une langue : on cherche à s'approcher au plus près, non du son correspondant au phonème, mais d'un patron articulatoire propre à celui-ci.

Certes, la compensation articulatoire, ou, en d'autres termes l'utilisation de configurations articulatoires distinctes pour la réalisation d'un même phonème en fonction du contexte, est une composante importante du processus de parole ; mais comme l'ont montré des études récentes (Qin & Carreira-Perpiñán 2007), elle est empiriquement peu exploitée : il semble qu'en parole spontanée, le plus important n'est pas de réaliser un patron vocalique de la façon la plus efficace énergétiquement, mais de la façon la plus proche d'une forme canonique.

C'est cette hypothèse – le processus de parole cherche à approcher au plus près des patrons articulatoires des phonèmes à prononcer – qui motive la première classe de contraintes ; nous supposons en outre que les patrons articulatoires ne sont pas seulement propres au locuteurs, mais similaires chez tous les locuteurs d'une langue donnée. Pour cette classe de contraintes, nous exploitons ainsi l'information contextuelle implicite : nous chercherons à déterminer les solutions

s'approchant le plus de la forme canonique propre à la langue et au phonème prononcé.

La deuxième classe de contraintes étudiée correspond à des contraintes dites visuelles : l'inversion ne porte plus simplement sur le son, mais exploite des données supplémentaires sur la position des articulateurs visibles, obtenues à partir d'images en stéréovision du visage du locuteur.

Ce mémoire est ainsi divisée en deux parties principales : la première présente le système d'inversion, et en particulier toutes les améliorations apportées à la méthode de construction de codebook hypercubique initiée par Slim Ouni (Ouni & Laprie 2005), la seconde présente les deux classes de contraintes : l'introduction et l'utilisation de contraintes dérivées de connaissances phonétiques génériques sur l'articulation des voyelles appelées « contraintes phonétiques », suivi d'expériences d'inversion « multimodale », utilisant en plus du son des données visuelles acquises par un système de stéréovision que j'ai contribué à développer.

Chapitre 1

Préliminaires

Cette partie présente succinctement les outils de simulation de la production de la parole que nous avons utilisés dans le cadre de l'inversion et nécessaires à la compréhension de cette thèse. Nous présenterons quelques prérequis concernant l'acoustique de la parole, la modélisation articulatoire du conduit vocal, et nous évoquerons rapidement les différentes techniques d'inversion acoustique-articulatoire.

1.1 Acoustique de la parole

Les ondes sonores sont des propagations de changements de pression, produits par la vibration des particules de l'air. La propagation de ces changements de pression est rapide : environ 340m.s^{-1}.

La parole est une forme de sons extrêmement complexe et élaborée, dont l'étude des propriétés acoustiques a véritablement été initiée par H. Helmholtz (von Helmholtz 1867). Il a cependant fallu attendre la fin du XIX^{e} siècle et l'élaboration de la transformée de Fourier pour que les méthodes modernes permettant de caractériser les différents sons de la parole apparaissent.

La transposition du signal temporel dans le domaine fréquentiel, aussi appelée analyse spectrale, permet en effet de caractériser visuellement chacune des classes des sons de la parole. Une analyse spectrale en bande étroite appliquée sur des sons de parole voisée permet de distinguer les harmoniques – des fréquences pour lesquelles l'intensité est nettement renforcée – de la fréquence fondamentale. Ces harmoniques sont dues à la vibration des cordes vocales, et leurs fréquences varient au cours du temps. L'analyse spectrale de la parole révèle aussi, en plus de la partie harmonique, une partie bruitée, liée à toutes les autres sources de son du conduit vocal (frictions, explosions, etc.). La partie harmonique domine nettement dans le cas des voyelles, la partie bruitée nettement pour la plupart des consonnes.

1.1.1 Production de la parole

Le système de production de la parole se décompose en trois parties :
- le système sous-glottique,
- le larynx,
- le système supra-glottique.

Le système sous-glottique est constitué des poumons et de la trachée. Il génère le flux d'air à l'origine du signal sonore. Ce flux d'air est ensuite modulé par le larynx et le système supra-glottique (les cavités du pharynx, de la bouche et éventuellement les fosses nasales) pour former

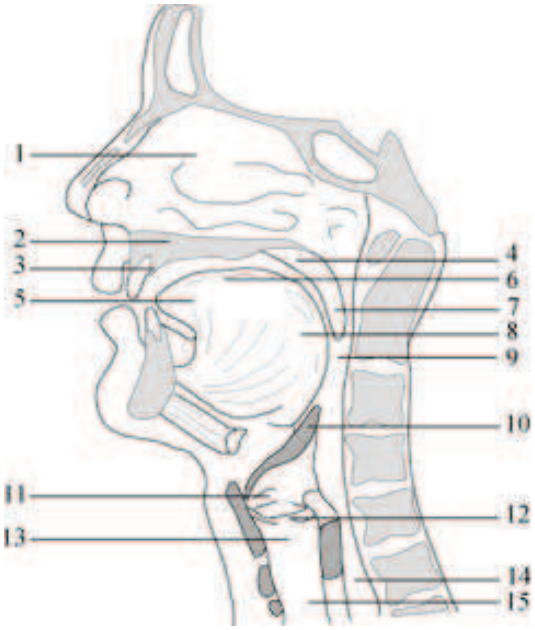


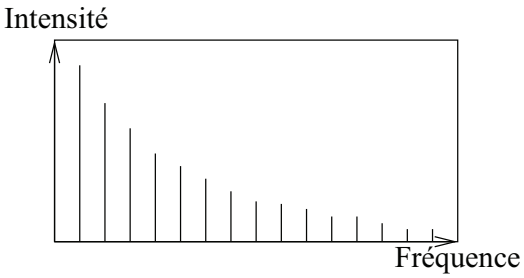
FIG. 1.1: *Schéma de l'anatomie du conduit vocal. 1 : fosses nasales, 2 : palais dur, 3 : mâchoire supérieure, 4 : palais mou (velum), 5 : pointe de la langue (apex), 6 : dos de la langue, 7 : luette, 8 : base de la langue (radix), 9 : pharynx, 10 : épiglotte, 11 : fausses cordes vocales, 12 : cordes vocales, 13 : larynx, 14 : œsophage, 15 : trachée. D'après Fant (Fant 1960).*

la parole.

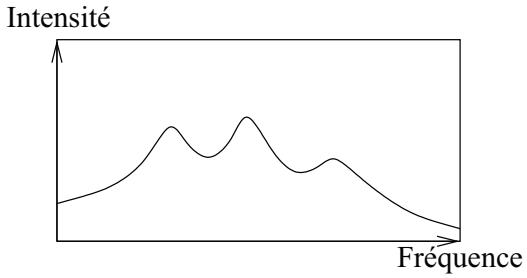
Le système de production de la parole comprend une ou plusieurs sources sonores, convoluées par un filtre (le système supra-glottique). Une source sonore est une interférence acoustique appliquée au flux respiratoire, et il y a deux origines principales. La première, appelée source voisée, est une modulation quasi-périodique du flux causé par les mouvements d'accolement et d'écartement rythmique de deux muscles du larynx appelés « cordes vocales » (cf. figure 1.1). Cette première source est caractéristique des voyelles, mais est aussi présente lors de la production de certaines consonnes. La seconde est causée par un retrécissement du conduit vocal créant une friction, la libération ou le blocage transitoire du flux et caractérise les consonnes. Le flux respiratoire utilisé pour la parole est en général une expiration contrôlée, mais il peut aussi s'agir dans de rares cas d'une inspiration.

La source voisée est créée par la vibration des cordes vocales, mais elle n'accompagne pas en permanence la parole. Elle a une forme spectrale du type représenté sur la figure 1.2a : il s'agit d'une série d'harmoniques dont l'intensité décroît avec la fréquence, multiples d'une fréquence fondamentale. La fréquence fondamentale est l'inverse de la période glottale, durée d'un cycle d'ouverture/fermeture des cordes vocales. La fréquence fondamentale détermine la hauteur de la voix.

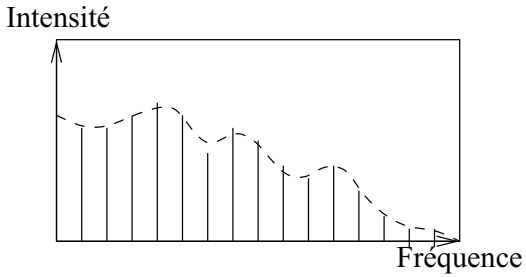
Le système supra-glottique (constitué des conduits vocal et nasal) agit comme un filtre sur la source, et est caractérisé par une fonction de transfert, typiquement de la forme représentée sur la figure 1.2b. Les pics d'intensité de cette fonction correspondent aux fréquences de résonance du conduit ; ces pics sont appelés, dans le cadre de la parole, *formants*, et sont d'une importance capitale. En effet, la fréquence, l'amplitude et la largeur de bande de ces différents pics permettent de caractériser les voyelles prononcées. D'autre part, ils donnent des contraintes assez fortes sur



(a) Spectre de la source vocale



(b) Fonction de transfert du conduit vocal



(c) Spectre du son émis

FIG. 1.2: *Schémas idéalisés de spectrogrammes de parole*

la forme et la position des cavités des conduits vocal et nasal.

Un spectre typique de son émis est représenté à la figure 1.2c (en pointillé, ce que l'on verrait en calculant un spectrogramme « bande large », en trait continu ce qu'on pourrait voir en calculant un spectrogramme « bande étroite »). On retrouve sur cette figure les pics vus précédemment ; les fréquences des formants restent relativement inchangées, mais leurs autres caractéristiques sont alterées : amplitudes et largeurs de bande sont modifiées.

On considère en général que la connaissance de la fréquence des trois premiers formants est suffisante pour discriminer toutes les voyelles, et même que la connaissance de la fréquence des deux premiers suffit dans la majorité des cas. Dans notre application, on ne retiendra du signal de parole que la fréquence des trois premiers formants. En effet, il se trouve que sans données supplémentaires sur la source, il est difficile d'en extraire d'autres indications quant à la forme du conduit vocal : la forme du spectre acoustique de la source dépend beaucoup de l'effort vocal (plus l'effort est important, et plus la pente de l'enveloppe spectrale de la source est faible), et donc également le spectre acoustique du son émis. On peut ainsi émettre des sons de spectres d'aspects très différents à partir d'une même forme de conduit vocal. Par contre, une même forme de conduit donnera toujours, à peu de choses près, les mêmes fréquences formantiques.

Les caractéristiques des formants sont difficiles à déterminer de manière fiable. Ainsi le suivi de formants, qui cherche simplement à déterminer de façon automatique les fréquences de ces formants, est un problème toujours ouvert, et les techniques les plus avancées (telles que celles de Laprie (Laprie 2004) ou Deng (Deng *et al.* 2006)) font toujours des erreurs dans de nombreux cas. Les amplitudes et largeurs de bande des formants sont encore plus difficiles à déterminer avec précision, et leur détermination nécessite une connaissance de la source. Pour ces raisons, et parce que notre étude se limite aux voyelles non-nasales, les vecteurs acoustiques que nous manipulerons seront essentiellement des triplets des trois premières fréquences formantiques pour les signaux de parole naturels. Pour les signaux synthétiques, nous calculerons les fréquences de

résonance de la fonction de transfert. Nous considérons donc qu'il y a une bonne correspondance entre ces fréquences de résonance et les fréquences des formants.

1.2 Synthèse articulatoire

Le système de production de parole humaine a été largement étudié dans le but de produire une voix artificielle qui ressemble le plus possible à une voix naturelle. Différentes modélisations de l'appareil phonatoire, plus ou moins élaborées, ont été proposées et implémentées tout au long des XX^{e} et XXI^{e} siècles. Pendant longtemps, l'imitation fidèle du fonctionnement de l'appareil phonatoire humain apparaissait comme la meilleure façon de faire de la synthèse de parole réaliste, mais depuis une vingtaine d'années, avec le développement de la synthèse par concaténation de segments non-uniformes (Sagisaka 1988) qui donne d'excellents résultats (Black & Campbell 1995; Beutnagel *et al.* 1999), cette approche a été quelque peu délaissée.

En revanche, après avoir atteint les limites de la synthèse par concaténation, qui nécessiterait des corpus gigantesques pour corriger ses quelques défauts, on constate depuis environ 2 ans un très net regain d'intérêt vers la synthèse purement articulatoire (Birkholz 2007), et la synthèse mixte utilisant des informations articulatoires en plus d'enregistrement sonores pour la synthèse de parole, comme par exemple les travaux de Pfitzinger (Pfitzinger 2005), ou encore le projet COUGAR (King & Richmond 2005) au CSTR (*Centre for Speech Technology Research*) de l'Université d'Édimbourg.

Un synthétiseur articulatoire comporte deux parties principales :

1. Un modèle du conduit vocal ; en général, il s'agit d'un dispositif permettant de décrire n'importe quelle forme de conduit vocal à partir d'un nombre réduit de paramètres.
2. Une simulation acoustique permettant de générer le « son » connaissant le conduit vocal.

1.2.1 Modélisation du conduit vocal

De nombreuses modélisations du conduit vocal, plus ou moins élaborés, sont décrites dans la littérature. Nous ne ferons pas ici une revue exhaustive des différents modèles, mais simplement une revue des différentes classes de modèles, en présentant brièvement l'un des plus représentatifs de chaque classe.

1.2.1.1 Modèles à fonction d'aire

Il est admis depuis longtemps (Fant 1960) que l'élément le plus important pour l'acoustique de la parole est la *fonction d'aire* du conduit vocal, c'est-à-dire la donnée de l'aire de la section transversale du conduit vocal le long de la courbe médiane du conduit, de la glotte à l'ouverture des lèvres. Il est ainsi admis implicitement que le conduit vocal est assimilable à un conduit droit de section variable ; il y a cependant assez peu d'estimations pratiques des erreurs introduites par cette hypothèse. Sondhi (Sondhi 1986) a montré que l'angle influait très peu sur les fréquences des résonances pour un conduit de section fixe. Ciocea (Ciocea 1997) a notamment montré que la forme précise du conduit vocal et l'angle du coude entre les cavités orales et pharyngales n'avaient que peu d'influence sur la fonction de transfert du conduit, tout du moins pour les fréquences au-dessous de 4kHz.

Modèles à trois paramètres Parmi les plus anciens modèles de fonction d'aire, on trouve des modèles à trois paramètres : l'un proposé par Stevens et House (Stevens & House 1955), dont une extension est utilisée par Atal (Atal *et al.* 1978), et un autre proposé par Fant (Fant 1960), tous deux il y a environ 50 ans. Les trois paramètres correspondent respectivement à la position de la constriction, à l'aire à la constriction, et à « l'ouverture » des lèvres (en fait le rapport entre la hauteur et l'aire de l'ouverture des lèvres). Malgré la simplicité de ces modèles, ils permettent de représenter schématiquement les configurations articulatoires des voyelles.

Concaténations de tubes L'une des façons les plus simples est de décrire le conduit vocal comme une succession de petits tubes. Dans ce type de modèles, la fonction d'aire est entièrement décrite par un petit nombre de couples (longueur, aire de la section) décrivant chaque segment de « tuyau ». Cette modélisation n'est pas à proprement parler articulatoire. Elle permet de décrire un conduit humain, mais présente plusieurs inconvénients : d'une part, il est nécessaire de manipuler un grand nombre de paramètres pour avoir un modèle suffisamment souple pour épouser tous les types de configurations. D'autre part il n'y a aucune garantie qu'une fonction d'aire donnée corresponde effectivement à un conduit vocal humain, ce qui peut s'avérer problématique, notamment pour l'inversion acoustique-articulatoire.

L'un des modèles à fonction d'aire les plus évolués est celui de Schoentgen et Ciocea (Schoentgen & Ciocea
Comme les segments utilisés sont coniques, il est possible d'obtenir une fonction d'aire continue. Le nombre de segments peut être très important, et le système est capable d'adapter automatiquement la longueur du conduit modélisé au locuteur.

1.2.1.2 Modèles articulatoires

Une autre classe de modèles du conduit vocal cherche à approcher une représentation fidèle du conduit vocal ; et le plus souvent, la visualisation la plus simple que l'on puisse en avoir : la coupe médio-sagittale. Il s'agit encore à l'heure actuelle de la seule représentation pour laquelle on dispose de données de bonne qualité et en assez grand quantité, et qui fournisse suffisamment d'informations pour reconstituer assez fidèlement l'acoustique. On a donc rapidement cherché à construire des modèles permettant de représenter les différentes coupes réalisables par un humain, si possible contrôlés par un petit nombre de paramètres. On voit également apparaître depuis quelques années des modélisations plus complètes du conduit vocal.

Modèles géométriques La forme du conduit vocal est modifié par le mouvement des articulateurs tels que la mâchoire, la langue, les lèvres, le larynx. L'une des façons naturelles de construire un modèle articulatoire est d'utiliser ces articulateurs comme commandes. Tout l'art de la modélisation articulatoire est d'étudier et de représenter la morphologie complexe des articulateurs et leur déformation d'une façon simple tout en conservant une image précise des éléments pertinents pour la synthèse acoustique de la parole.

Plusieurs modèles décrivant le conduit vocal comme une combinaison de formes géométriques simples pilotée par un petit nombre de paramètres ont ainsi été construits. L'un des plus réussis, et très utilisé aujourd'hui, est le modèle de Mermelstein (Mermelstein 1973), lui-même étant une extension du modèle de Coker (Coker 1973).

Ces modèles permettent de décrire les configurations du conduit vocal des voyelles et consonnes à partir d'un petit nombre de paramètres ; ils souffrent cependant de plusieurs défauts. D'une part, les valeurs à donner pour les commandes articulatoires ne sont pas forcément faciles à trouver : pour déterminer les paramètres idéaux qui représentent une configuration de conduit vocal donnée obtenue par exemple par rayons X ou IRM, il est nécessaire d'effectuer une opération

d'optimisation. D'autre part, ces modèles ont été élaborés de façon ad hoc, parfois à partir d'images réelles, mais également d'expertise humaine et d'intuition. Il est par conséquent difficile d'évaluer leur pertinence.

Modèles issus d'analyses factorielles Une alternative à l'approche géométrique (liée à l'expertise humaine), est l'élaboration de modèles à l'aide d'analyses factorielles sur des données articulatoires réelles. Il apparaît en effet qu'il existe beaucoup de redondance dans les caractéristiques des formes de conduits vocaux, et il est donc parfaitement envisageable d'y appliquer des analyses factorielles pour obtenir des modèles qui décrivent la majeure partie de la variabilité à l'aide d'un petit nombre de paramètres orthogonaux ou/et non-corrélés.

L'un des modèles les plus connus est le modèle de Maeda (Maeda 1979; Maeda 1990). Il décrit un conduit vocal complet à partir de trois modèles indépendants pour les lèvres, la langue, et le larynx. On peut en effet considérer que ces trois articulateurs modifient le conduit vocal de façon indépendante, bien qu'ils soient tous les trois influencés par la position de la mâchoire inférieure. L'analyse factorielle utilisée par Maeda pour traiter les données articulatoires se devait d'être suffisamment souple pour rendre compte de cette particularité et soustraire l'influence de la mâchoire sur les autres articulateurs. En effet, sa position peut être aisément déterminée sur les radiographies en mesurant l'écart entre les incisives supérieures et inférieures. Une analyse en composantes principales n'étant pas adaptée, Maeda (Maeda 1979) a utilisé une analyse en composantes orthogonales arbitraires (proposée par Overall (Overall 1962)), que l'on appelle aussi analyse en composantes principales guidée, pour soustraire l'influence de la mâchoire. Chaque zone du conduit vocal (lèvres, langue, larynx) est alors étudiée indépendamment. Pour chacune des zones, des paramètres de contrôle sont obtenus par une analyse en composantes principales sur les données décorrélées de l'influence de la mâchoire, en retenant suffisamment de composantes pour expliquer l'essentiel de la variance. Le nombre de paramètres nécessaires est variable dans chaque zone ; pour la zone du larynx, un paramètre suffit ; pour les lèvres, les données analysées sont l'ouverture verticale des lèvres, l'ouverture horizontale (ou étirement) des lèvres, et la protrusion. Deux paramètres intrinsèques (en plus de la mâchoire) ont été retenus pour décrire ces données : l'ouverture verticale et la protrusion ; l'ouverture horizontale des lèvres est déduite des deux autres paramètres ; pour la langue, trois paramètres supplémentaires sont nécessaires pour décrire 96% de la variance des radiographies, soit un total de 7 paramètres (voir figure 1.3).

Par ailleurs, il est possible d'adapter le modèle articulatoire à des locuteurs différents : deux paramètres d'élongation des conduits oral et pharyngal permettent, dans une certaine mesure, d'adapter la forme du conduit à un nouveau locuteur. Ces paramètres influent uniformément sur les dimensions des deux conduits, mais en jouant habilement avec, il est possible d'établir un modèle capable de former des fonctions d'aire correspondant aux réalisations acoustiques d'un locuteur différent. Galván-Rodrigez a établi une méthode semi-automatique d'adaptation permettant à partir des fréquences formantiques des voyelles d'un locuteur donné d'établir les coefficients d'élongation (Galván-Rodrigez 1997). Le défaut de cette méthode est qu'elle suppose que la réalisation d'une voyelle donnée utilise une configuration articulatoire unique et indépendante du locuteur.

Plus récemment, Badin et al. ont proposé un modèle articulatoire également issu d'une analyse factorielle, mais basé sur des IRM 3D et des vidéos des différents phonèmes du Français (Badin *et al.* 2002).

Modèles biomécaniques Certains modèles articulatoires cherchent à modéliser bien plus que la forme du conduit vocal : les modèles biomécaniques permettent de prendre en compte

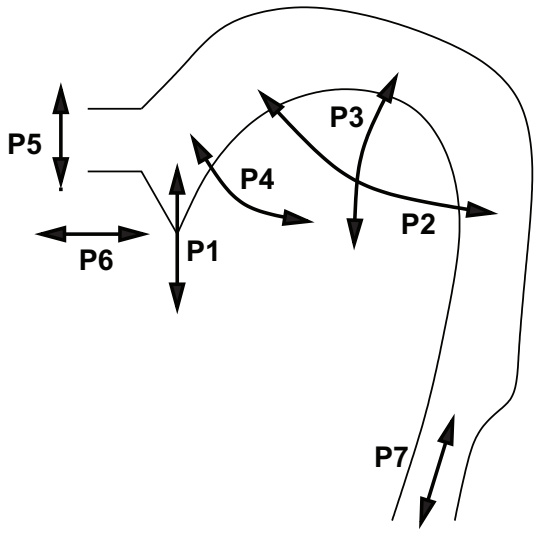


FIG. 1.3: *Les sept paramètres du modèle de Maeda : la mâchoire (ou jw) P1, l'ouverture verticale (lh) P5 et la protrusion des lèvres (lp) P6, la position du corps de la langue (tb) P2, la forme de la langue (ts) P3, un dernier terme contrôlant la pointe de la langue (tt) P4, et enfin la hauteur du larynx (lx) P7.*

la dynamique du système de production de parole, grâce à une modélisation des structures musculaires par des ressorts et des masses. L'un des premiers modèles de ce type est celui de Perkell (Perkell 1974). Il simplifiait considérablement le processus de production. Des modèles beaucoup plus élaborés ont été proposés depuis, tel le modèle de l'ICP proposé par Gérard et al. (Gérard *et al.* 2003), qui modélise la langue en trois dimensions. Les inconvénients, outre la lourdeur des calculs nécessaires à leur utilisation, sont le nombre très important de paramètres de commande et la difficulté de les déterminer ; ce qui les rend peu pratiques, tout du moins aujourd'hui, pour l'inversion acoustique-articulatoire.

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

En théorie, l'acoustique du conduit vocal, que ce soit sous la forme de l'onde sonore pour une simulation temporelle, ou sous la forme d'une fonction de transfert pour une simulation fréquentielle, peut être calculée à partir de sa représentation géométrique en trois dimensions. Actuellement, les simulations acoustiques tridimensionnelles ne sont pas particulièrement convaincantes, probablement à cause de la difficulté d'appliquer un maillage approprié pour le conduit vocal pour une méthode à éléments finis, et de l'imprécision de la connaissance de la géométrie du conduit vocal. Par ailleurs, ces méthodes de calcul sont particulièrement gourmandes en temps de calcul, prenant généralement plusieurs heures de calcul pour quelques millisecondes de parole. Une étude récente due à Ramsay et Shaddle (Ramsay & Shadle 2006) étudiant la formation des turbulences dans le cas des fricatives, à l'aide d'une simulation précise d'un modèle de flux visqueux incompressible tridimensionnel, prend ainsi 14 jours pour simuler deux millisecondes de parole sur un cluster de 16 optérons.

Pour ces raisons, les modèles de calcul classiques de propagation unidimensionnelle utilisant la fonction d'aire du conduit vocal semblent toujours valables. Pour améliorer la fidélité acoustique à moindre coût, il est envisageable également de passer à une modélisation à deux dimensions, en faisant des hypothèses simplificatrices telle qu'une symétrie axiale du conduit vocal (Hélie 2002).

La plupart des modèles articulatoires vus précédemment ont en commun de ne modéliser qu'une coupe médio-sagittale du conduit vocal, et pour passer au conduit vocal complet, il est nécessaire de disposer d'un modèle de passage pour retrouver la troisième dimension à partir de cette représentation bidimensionnelle du conduit. En pratique, si l'on néglige la forme précise des coupes et que l'on se contente d'un modèle à fonction d'aire, le procédé est simplifié mais reste loin d'être évident. Le modèle le plus utilisé est le modèle « alpha-beta » proposé par Heinz et Stevens (Heinz & Stevens 1965).

Dans ce modèle, l'aire transversale $A(x)$ du conduit pour une position x le long de la courbe médiane du conduit vocal est déduite à partir de la « distance sagittale » $d(x)$, c'est-à-dire la distance entre les parois antérieure et postérieure du conduit vocal. Ces auteurs convertissent la distance $d(x)$ en une aire $A(x)$ à l'aide d'une fonction puissance :

$$A(x) = \alpha(x) * d^{\beta(x)}(x), \tag{1.1}$$

où $\alpha(x)$ et $\beta(x)$ sont des paramètres dont les valeurs dépendent de la position x le long de la ligne médiane du conduit vocal. La courbe médiane du conduit vocal est déterminée en calculant l'intersection du conduit vocal avec une grille semi-polaire, en calculant les centres des segments reliant les parois antérieure et postérieure du conduit (cf. figure 1.4).

Comme la forme d'une section transversale de conduit est complexe, les valeurs de α et β doivent être déterminées de façon empirique. D'autre part, Perrier et al. ont montré que pour les valeurs importantes de d, l'aire du conduit vocal différait de façon importante de la valeur prédite par l'équation 1.1 (Perrier *et al.* 1992). Ceci étant, cette erreur n'est pas dramatique dans notre cas, car il apparaît que les fréquences des trois premiers formants sont peu sensibles aux petites variations dans la fonction d'aire pour les aires importantes (Ericsdotter 2007), ce qui rend la conversion α-β utilisable lorsque l'on se contente des fréquences des premiers formants – ce qui est suffisant pour l'étude des voyelles.

Il est nécessaire de souligner que cette modélisation – coupe médio-sagittale, modèle de passage, fonction d'aire – du conduit vocal souffre de nombreuses imperfections. D'une part, l'approximation grossière du conduit vocal à l'aide d'une fonction d'aire n'est acoustiquement pertinente que pour les fréquences inférieures à 4kHz. Au-dessus de cette fréquence, les modes transverses de l'onde sonore ne peuvent plus être négligés, car la longueur d'onde devient du même ordre de grandeur que la largeur du conduit : si on suppose qu'un conduit vocal fait au plus 4cm de large, soit 0.08m pour un aller-retour, et sachant que le son se propage dans l'air à 350m.s^{-1}, la fréquence maximale pour laquelle on n'aura pas de résonance latérale est d'environ $350/0.08 \approx 4.3\text{kHz}$. D'autre part, la détermination de $d(x)$ n'est pas non plus évidente. En effet, on suppose dans les modèles à fonction d'aire que le modèle de propagation est une onde plane, et il est par conséquent nécessaire de mesurer $d(x)$ dans la direction du front d'onde ; or il s'avère que mesurer le long d'une grille semi-polaire comme cela est fait par exemple dans le modèle de Maeda n'est pas toujours idéal. Pour un tuyau courbé de 90° avec une courbe douce, on peut raisonnablement supposer que le front d'onde se déplace orthogonalement à la courbe des centres géométriques des sections. Idéalement, il faudrait donc déterminer la courbe milieu du conduit vocal, et mesurer $d(x)$ comme la longueur du segment orthogonal à cette courbe coupant les parois du conduit. Des études fondées sur cette approche (Maeda 1972; Goldstein 1980) semblent montrer que les longueurs de conduit obtenues sont légèrement plus courtes que lorsqu'on utilise une grille semi-polaire, et que l'on peut de cette façon obtenir une simulation acoustique qui permette d'approcher de façon plus fidèle les formants mesurés. Ces études auraient cependant besoin d'être confirmées sur de plus gros corpus de données.

Il existe des modélisations plus élaborées qui prédisent l'aire des coupes transversales à partir de la distance sagittale de façon plus précise, notamment celle de Badin et al. (Badin *et al.* 2005).

Il semble toutefois, selon l'étude de Ericsdotter (Ericsdotter 2007), que pour la simulation des voyelles, les erreurs acoustiques liées au modèle de passage en lui-même sont négligeables devant celles dues aux autres facteurs d'erreur ; l'utilisation de mesures précises de la fonction d'aire sur des données réelles ne semble en effet pas améliorer de manière significative la fidélité des formants synthétiques aux formants mesurés.

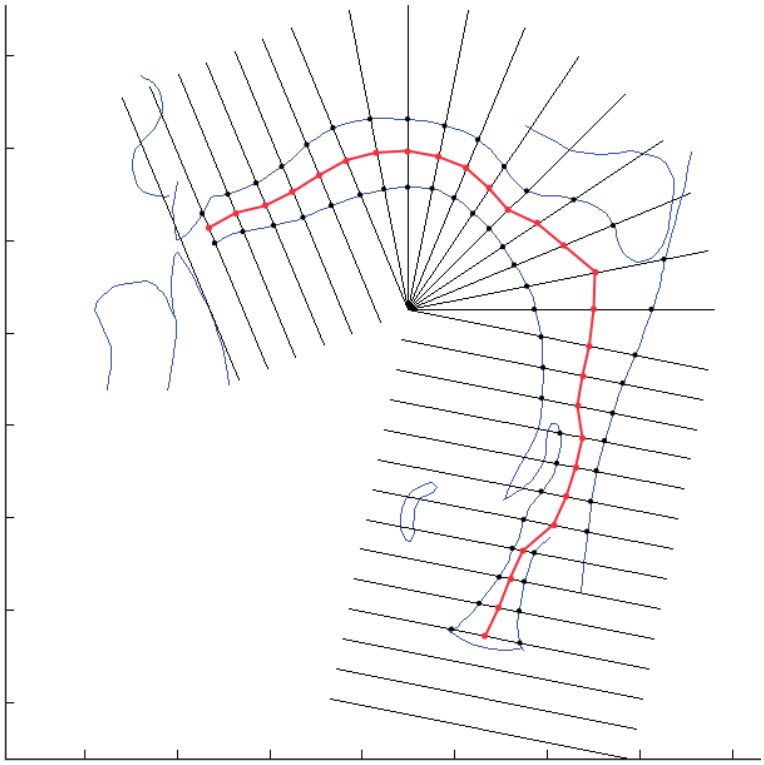


FIG. 1.4: *Détermination de la courbe médiane du conduit vocal avec une grille semi-polaire. La courbe médiane (en rouge) est construite en reliant les centres des segments les parois antérieure et postérieure du conduit.*

1.2.3 Simulation acoustique

L'écoulement de l'air dans un conduit est régi par les équations de Navier-Stokes :

$$\rho\left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v}\right] = -\nabla p + \eta\nabla^2\mathbf{v} + (\zeta + \frac{\eta}{3})\nabla(\nabla\,\mathbf{v}), \tag{1.2}$$

où ρ est la densité du fluide, ζ et η leurs coefficients de viscosité, supposés ne pas dépendre de la température ni de la pression du fluide, $\mathbf{v}$ est sa vitesse volumique et p sa pression. Cette équation est extrêmement complexe, mais nous pouvons fort heureusement la simplifier considérablement.

Dans le cas de la production de parole, l'ordre de grandeur des vitesses à considérer est tel que l'air peut être considéré comme un fluide incompressible, et en faisant l'hypothèse de la conservation de la masse et d'un processus adiabatique, l'équation 1.2 peut se ramener à :

$$\Delta p - \frac{1}{c^2}\partial_t^2 p = 0, \tag{1.3}$$

où p désigne la pression dans le conduit vocal, et c la célérité du fluide (environ 340m.s^{-1} pour l'air à 20°C).

Par ailleurs, comme nous l'avons évoqué précédemment, le conduit vocal est essentiellement un conduit étroit, et par conséquent le mode principal de résonance est lié à la longueur du conduit, la transversalité ne pouvant provoquer que des modes de résonance de fréquence élevée, que l'on peut négliger dans le cas des voyelles. On peut ainsi considérer que l'on est en présence d'une onde se déplaçant le long du conduit vocal, et la génération et la propagation des sons peuvent ainsi être décrites par l'équation unidimensionnelle suivante, dite de Webster :

$$\frac{1}{A(x)}\frac{\partial}{\partial x}\left[A(x)\frac{\partial P(x)}{\partial x}\right] + \frac{1}{c^2}\frac{\partial^2 P(x)}{\partial t^2} = 0 \tag{1.4}$$

où $A(x)$ désigne l'aire de l'isophase de pression $P(x)$ (x désignant l'abscisse le long de l'axe du conduit vocal). Cette équation est linéaire (donc relativement simple à étudier), et est valable pour n'importe quelle forme d'onde à symétrie axiale d'axe x. Dans notre cas, comme nous nous intéressons essentiellement aux voyelles, pour lesquelles le conduit est relativement large, il est raisonnable de supposer en outre que l'onde est une onde plane, orthogonale au conduit vocal. Ce dernier est modélisé comme un tuyau rectiligne de section variable, donnée par la fonction d'aire.

En discrétisant le tuyau de section variable en une concaténation de tubes de section constante, et en modélisant les conditions aux limites (i.e. les pertes au niveau de la paroi du conduit), nous pouvons obtenir la fonction de transfert du conduit vocal relativement simplement. Un des outils les plus utilisés à cet effet est l'analogie entre l'acoustique et l'électricité : en effet, la pression et le débit volumique dans le conduit vocal suivent les mêmes équations différentielles que la tension et l'intensité dans une ligne électrique (Flanagan 1972). Une section de conduit uniforme peut être ainsi modélisée par un petit élément de ligne électrique avec pertes (cf. figure 1.5b). Dans cette analogie, R correspond aux pertes dues à la friction visqueuse de l'air sur les parois du tube, C correspond à la « capacité acoustique », c'est-à-dire la compressibilité de l'air, L est l'« inductance acoustique » due à l'inertie de l'air, enfin G permet de modéliser les pertes thermiques, dont nous ne tiendrons pas compte non plus. Les valeurs de ces éléments sont les suivantes :

$$L = \frac{\rho}{A}, C = \frac{A}{\rho c^2}$$

En cherchant les solutions à (1.4) sous la forme d'ondes monochromatiques, c'est-à-dire de la forme : $\phi(x,t) = \psi(x)e^{j\omega t}$, nous avons pour chaque tube :

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\omega^2}{c^2}\psi = 0 \tag{1.5}$$

Maintenant, en posant $y = G + j\omega C$ et $z = R + j\omega L$, ainsi que $\gamma^2 = yz = -\frac{\omega^2}{c^2}$ (γ est appelée constante de propagation), nous pouvons obtenir comme solution, en régime continu, pour chaque section, une relation linéaire de la forme :

$$\begin{pmatrix} P_s \\ U_s \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \times \begin{pmatrix} P_e \\ U_e \end{pmatrix},$$

où P_e et P_s désignent respectivement la pression à l'entrée et à la sortie de la section étudiée, et U_e et U_s le débit volumique à l'entrée et à la sortie de cette même section. En désignant par l la

longueur de cette section, et par A l'aire, les éléments de la matrice s'écrivent :

$$\begin{aligned} A &= \cosh \gamma l \\ B &= -\sqrt{\frac{z}{y}} \sinh \gamma l \\ C &= -\sqrt{\frac{y}{z}} \sinh \gamma l \\ D &= \cosh \gamma l \end{aligned}$$

Par ailleurs, chacun des tubes du conduit vocal pouvant être ainsi modélisé par une matrice de la forme $T_i = \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix}$, on obtient, pour une succession de tubes, une fonction de transfert de la forme :

$$\begin{pmatrix} P_l \\ U_l \end{pmatrix} = T_n \cdot T_{n-1} \cdots T_1 \cdot T_0 \begin{pmatrix} P_g \\ U_g \end{pmatrix}$$

La fonction de transfert globale du conduit vocal est simplement le produit des matrices T_i, il s'agit donc d'une matrice 2×2, dont les composantes A, B, C, D vérifient :

$$P_g = AP_l + BU_l, U_g = CP_l + DU_l$$

La pression acoustique étant nulle à la sortie des lèvres, on a $P_l = 0$, et par conséquent la fonction de transfert du conduit $\frac{U_l}{U_g}$ est donnée par :

$$\frac{U_l}{U_g} = \frac{1}{D}$$

Les formants correspondant aux résonances du conduit, leurs fréquences sont celles qui annulent D. Il suffit donc, pour trouver les fréquences des formants, de trouver les solutions ω_i de l'équation :

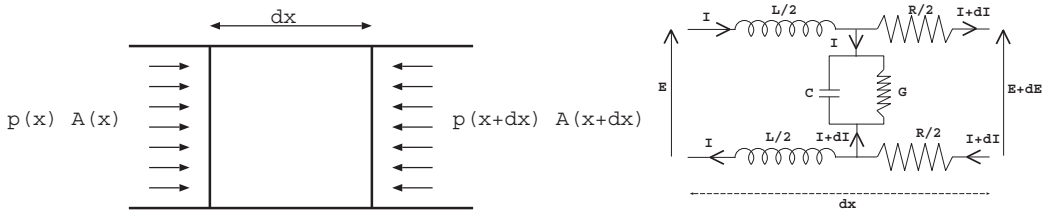
$$D(\omega) = 0.$$

Les ω_i étant des longueurs d'onde, les fréquences des formants se déduisent rapidement ainsi : $f_i = \frac{c}{\omega_i}$.

En utilisant l'analogie électrique, il est également possible de traiter le conduit nasal : il suffit de considérer que le conduit nasal est un tube acoustique branché en parallèle avec le tube correspondant au conduit oral et au tube représentant le pharynx. Cependant, dans le cadre de cette thèse nous ne traiterons pas des sons nasalisés, qui compliquent l'inversion acoustique-articulatoire.

1.3 Inversion acoustique-articulatoire

L'inversion acoustique-articulatoire désigne le problème consistant à retrouver la forme du conduit vocal (ou la position des articulateurs) à partir du signal de parole émis. L'objectif est de réaliser l'*inverse* de la synthèse articulatoire, que nous avons présentée dans la section précédente, d'où le terme d'*inversion*. En pratique, le problème est simplifié en paramétrisant les espaces acoustique et articulatoire, en réduisant le signal acoustique en un vecteur plus simple à manipuler, par exemple les caractéristiques des premiers formants pour l'étude des voyelles, et le domaine articulatoire en décrivant la forme du conduit vocal à partir d'un vecteur de taille réduite, par exemple un vecteur de contrôle d'un modèle articulatoire. Le problème de l'inversion se réduit alors à l'étude d'une relation de l'espace des vecteurs acoustiques vers l'espace des vecteurs articulatoires.



(a) Tube de section constante.

(b) Élement de ligne électrique avec perte.

FIG. 1.5: *Analogie acoustique ↔ électricité.*

1.3.1 Problème mal posé

Un problème est dit *bien posé* s'il admet une solution (condition d'existence), si celle-ci est unique (condition d'unicité) et si elle est stable (condition de stabilité). Si un problème ne vérifie pas au moins l'une de ces conditions, il est dit *mal posé*.

De façon générale, les problèmes d'inversion du conduit vocal sont mal posés. La première condition – l'existence – dépend de la fidélité du modèle physique de production de parole. Il est parfois très difficile voire impossible de montrer qu'un modèle est capable de produire tous les vecteurs acoustiques d'un signal de parole (Sorokin *et al.* 2000), même si pour des vecteurs simples, comme le triplet des trois premières fréquences formantiques, cela est envisageable.

À propos de la seconde propriété (unicité), il est bien connu qu'il existe une infinité de fonctions d'aire pouvant produire le même ensemble de formants (Atal *et al.* 1978). Par ailleurs, il a été observé à plusieurs reprises que cette non-unicité était exploitée par le biais d'articulations compensatoires, par exemple chez les ventriloques ou imitateurs, ou dans des expériences d'articulation perturbée (Lindblom *et al.* 1979; Savariaux & Orliaguet 1995), mais aussi dans des conditions normales d'élocution (Qin & Carreira-Perpiñán 2007). Si cette variabilité de l'articulation pour produire un son donné est avérée, il semble cependant qu'en pratique, l'articulation compensatoire est assez peu exploitée dans des conditions normales d'élocution : C. Qin (Qin & Carreira-Perpiñán 2007) a observé sur les sujets de son étude que seuls 5% des vecteurs acoustiques avaient des antécédents non-uniques.

Enfin, la propriété de stabilité (une petite perturbation de l'entrée ne devant entraîner qu'une petite perturbation de la solution) n'est également pas toujours assurée, en fonction de la méthode d'inversion utilisée.

Pour toutes ces raisons, l'inversion acoustique-articulatoire est considérée comme un problème mal posé.

1.3.2 Principe de résolution

L'inversion acoustique-articulatoire a été largement étudiée dans le passé (Schroeder 1967; Mermelstein 1967; Atal *et al.* 1978; Schroeter & Sondhi 1994; Sorokin *et al.* 2000), et un certain nombre d'approches ont été proposées afin d'atténuer la nature mal posée du problème. Un inventaire presque exhaustif de ces différentes méthodes peut être trouvé dans un rapport du projet européen ASPI (Maeda *et al.* 2006) ; nous ne présenterons pas ici un inventaire aussi détaillé, mais une classification rapide des différentes méthodes d'inversion proposées.

1.3.2.1 Méthodes basées sur des données d'« apprentissage » (réelles ou synthétiques)

Ces méthodes nécessitent une quantité importante de données, qui sont malheureusement rarement disponibles en quantité ou qualité suffisante. Pour résoudre le problème des données manquantes, une solution simple consiste à les créer artificiellement, par exemple à l'aide d'un synthétiseur articulatoire. Une autre solution couramment utilisée consiste à n'étudier qu'une représentation partielle du conduit vocal, par exemple la position de marqueurs électromagnétiques, qu'il est possible d'obtenir en quantité importante.

Méthodes par tabulation Les méthodes d'inversion utilisant des tables de correspondance vecteur acoustique ↔ vecteur articulatoire sont appelées méthodes par tabulation, ou par *codebooks*. Ces tables peuvent être construites à partir de données obtenues de façon artificielle, c'est-à-dire à l'aide d'un système de synthèse articulatoire, ou alors sur des données réelles où données articulatoires et acoustiques sont enregistrées en parallèle.

Le principe de la méthode d'inversion est alors simple : un vecteur acoustique étant donné, on recherche dans la table les valeurs qui en sont proches, et on renvoie les vecteurs articulatoires correspondants. L'objectif essentiel de ces méthodes est de réaliser une couverture *adéquate* des espaces articulatoire et acoustique, ce qui en fonction de l'application visée peut désigner des choses très diverses.

Différentes techniques sont utilisées afin d'obtenir cette couverture adéquate, d'organiser de façon efficace la table, et s'affranchir des zones inexplorées de l'espace articulatoire. Une organisation astucieuse de la table permet de retrouver rapidement les vecteurs articulatoires qui produisent un vecteur acoustique proche d'un vecteur acoustique donné.

Ce type d'approche est parmi les plus anciens, et reste également parmi les plus prometteurs, même si, comme pour la plupart des méthodes, les résultats obtenus sont rarement à la hauteur des attentes. Le premier à utiliser ce type d'approche est Atal (Atal *et al.* 1978), à une époque où les performances des machines étaient très éloignées de ce que l'on connaît actuellement. Certains choix techniques faits à l'époque ne sont plus guère pertinents, mais une grande partie des résultats présentés par Atal sont toujours exploités dans les méthodes d'inversion par codebook actuelles ; ce travail reste ainsi à bien des égards une étude fondamentale pour l'inversion acoustique-articulatoire par codebook.

Parmi les nombreuses méthodes d'inversion utilisant des codebooks obtenus à l'aide d'un synthétiseur articulatoire, différents types d'échantillonnage de l'espace articulatoire ont été proposés :

- échantillonnage régulier (Atal *et al.* 1978),
- échantillonnage aléatoire (Schroeter & Sondhi 1992; Boë *et al.* 1992),
- interpolation à partir de vecteurs racines (Larar *et al.* 1988; Sorokin & Trushkin 1996),
- échantillonnage adaptatif (Charpentier 1984; Sorokin & Trushkin 1996; Ouni & Laprie 2000; Potard & Laprie 2007).

De même, plusieurs modélisations de la relation acoustique ⇒ articulatoire ont été proposées pour s'affranchir des zones manquantes. La méthode la plus simple est de considérer que la fonction est constante dans un petit voisinage autour du vecteur acoustique (Atal *et al.* 1978; Larar *et al.* 1988; Schroeter & Sondhi 1992). Une autre méthode consiste à considérer un comportement linéaire autour des vecteurs acoustiques (Atal *et al.* 1978; Charpentier 1984; Sorokin & Trushkin 1
Ouni & Laprie 2000; Potard *et al.* 2004). Certains encore emploient des modélisations polynômiales (Potard & Laprie 2007) ou stochastiques (Laboissière 1992; Hogden *et al.* 1996; Hiroya & Honda 2004
Richmond 2001).

Une propriété essentielle des méthodes utilisant des données réelles par rapport à celles utilisant des données artificielles est de réduire de façon substantielle les problèmes liés à la non-unicité. Il a en effet été observé de façon empirique qu'en élocution normale la non-unicité *pratique* ne concerne qu'un nombre très réduit de phonèmes (Qin & Carreira-Perpiñán 2007). Les méthodes utilisant un synthétiseur articulatoire sont confrontées à la non-unicité *théorique*, qui est nettement plus importante que celle observée en pratique.

Ces méthodes permettent de trouver rapidement des vecteurs articulatoires qui donnent la bonne image acoustique, ce qui peut suffire pour certaines applications. Mais s'il s'agit de guider un élève qui apprend l'articulation d'un son (par exemple dans le cadre de l'apprentissage d'une langue étrangère), ou dans l'optique d'animer d'une tête parlante à l'intention de malentendants, la correspondance acoustique est loin d'être suffisante, et il est bien souvent nécessaire d'effectuer un travail conséquent pour résoudre le problème de non-unicité.

Différentes approches ont été proposées pour cela : traiter le problème en amont, par exemple en restreignant volontairement l'espace articulatoire à explorer, ou en augmentant la dimension du vecteur acoustique (et par conséquent limiter considérablement le problème de non-unicité), traiter le problème immédiatement en introduisant des contraintes statiques (par exemple la minimisation de la distance à la position neutre), ou bien traiter le problème en aval en plaçant des contraintes sur les trajectoires articulatoires.

Méthodes utilisant un apprentissage Les méthodes fondées sur un apprentissage statistique de type réseaux de neurones ou Modèle de Markov caché sont une variante relativement courante des méthodes à codebooks. La table est simplement remplacée par une « boîte noire » associant à un vecteur acoustique un vecteur articulatoire, ou à une séquence de vecteurs acoustiques une séquence de vecteurs articulatoires pour les modèles dynamiques. Les données utilisées lors de l'apprentissage sont parfois basées sur un modèle articulatoire (Atal & Rioul 1989; Soquet *et al.* 1990; Papcun *et al.* 1992), mais sont désormais le plus souvent basées sur des données réelles (Hiroya & Honda 2004; Toda *et al.* 2004; Richmond 2006).

1.3.2.2 Inversion « directe »

Les données articulatoires existantes étant rarement satisfaisantes (soit en quantité insuffisante, soit en qualité insuffisante, soit propres à un locuteur et ne pouvant pas être adaptées simplement, soit encore ne pouvant pas être utilisées pour l'application particulière étudiée), beaucoup d'auteurs ont cherché à développer des méthodes d'inversion qui ne nécessitent pas de données articulatoires.

Les premières méthodes de ce type sont basées sur une étude de Mermelstein (Mermelstein 1967) sur la relation entre fréquences propres et fonction d'aire d'un conduit vocal sans pertes, ces fréquences propres correspondant aux fréquences des formants observables sur un spectre de parole.

D'autres méthodes exploitent la pseudo-linéarité locale de la relation de l'articulatoire vers l'acoustique sans passer par l'utilisation d'un codebook, notamment la méthode développée par Schoentgen et Ciocea (Schoentgen & Ciocea 1997), qui utilise une optimisation pour déterminer un vecteur articulatoire ayant pour image un triplet de fréquences formantiques donné. Ces méthodes aboutissent à une solution unique sur une séquence en ajoutant des contraintes de pseudo-énergie sur les positions des paramètres articulatoires.

Les approches par réseaux de neurones ou par modélisation stochastique nécessitent une quantité très importante de données pour avoir des résultats fiables, et ne fonctionnent en général pas avec un autre locuteur que celui d'apprentissage. Par ailleurs, on ne dispose pas de données sur le conduit vocal complet en quantité suffisante, et ces méthodes font généralement

leur apprentissage sur des données partielles du conduit vocal obtenues à l'aide de différentes techniques d'acquisition : EMA, rayons X Micro-Beam, articulographe, échographie..., et ne permettent donc pas l'animation d'une tête parlante complète, ce qui est notre objectif.

Pour cette raison, nous avons choisi d'utiliser une méthode d'analyse par synthèse utilisant un modèle articulatoire complet du conduit vocal, en l'occurrence le modèle de Maeda (Maeda 1979; Maeda 1990). Pour garantir une exploration aussi complète que possible de l'espace des solutions, nous utilisons un codebook construit en explorant l'intégralité de l'espace articulatoire, et reposant sur une modélisation localement polynômiale de la relation articulatoire vers acoustique.

1.3.3 Résolution par introduction de contraintes

L'objectif de l'inversion est de retrouver l'évolution de la forme du conduit vocal à l'origine d'un signal de parole donné. Or, comme nous l'avons remarqué précédemment, les modèles de synthèse articulatoire utilisés généralement ne permettent d'obtenir qu'une approximation du signal acoustique réellement produit par une forme de conduit vocal donnée. Par ailleurs, il est également fort complexe d'extraire l'information propre au conduit vocal du signal acoustique original : même dans le cas d'un enregistrement non bruité, il reste nécessaire d'employer une paramétrisation permettant d'extraire l'information acoustique propre au conduit vocal lui-même, et les techniques de traitement du signal permettant de le faire ne sont malheureusement pas parfaites.

Les fréquences des premiers formants sont à peu de choses près les seuls paramètres que l'on puisse estimer de façon fiable à partir d'un synthétiseur sans modèle de source. Elles fournissent une description phonétiquement pertinente des voyelles. Pour ces raisons, elles constituent un candidat de choix pour la paramétrisation du signal acoustique dans le cadre de l'inversion acoustique-articulatoire. Malheureusement, le nombre de formants pouvant être obtenus de façon fiable est assez réduit, puisque les hypothèses formulées dans le cadre des synthétiseurs classiques ne garantissent de bons résultats que pour les fréquences inférieures à 4kHz, ce qui ne permet de décrire fidèlement que les fréquences des trois ou quatre premiers formants. La taille réduite du vecteur acoustique accentue considérablement la non-unicité des solutions de l'inversion.

Il est par conséquent nécessaire d'introduire des contraintes supplémentaires afin de réduire la taille des ensembles de solutions à considérer. En pratique, de nombreux types de contraintes ont été proposés et étudiées. Parmi les contraintes introduites, on trouve notamment :

- La surdétermination du vecteur acoustique. Un vecteur acoustique de dimension plus importante que celle du vecteur articulatoire est utilisé ; il n'y a alors qu'au plus une solution exacte (Charpentier 1984).
- Restrictions sur les aires maximales et minimales des sections transversales du conduit (Sorokin *et al.* 20
- Minimisation de la distance à la forme neutre (Yehia & Itakura 1996).
- Maximisation de la continuité spatiale de la fonction d'aire (Yehia & Itakura 1996).
- Constance du volume du conduit vocal (Soquet *et al.* 1991).
- Maximisation de la continuité de l'évolution temporelle des formes de conduit (Flanagan *et al.* 1980).
- Minimisation de la variation temporelle des paramètres articulatoires (Flanagan *et al.* 1980).
- Minimisation du travail musculaire (Sorokin 1992)...

Nous présenterons plus en détail, dans la deuxième partie de cette thèse, deux autres types de contraintes, exploitant des informations extraites du contexte de la parole ; contraintes « phonétiques » dans un premier temps, imposant des contraintes articulatoires relatives au phonème prononcé, puis contraintes « visuelles », exploitant des informations sur les articulateurs visibles, à partir d'images vidéos en stéréovision.

1.4 Conclusion

Notre méthode d'inversion repose sur les méthodes de Mathieu (Mathieu 1999) et Ouni (Ouni 2001) développées au sein de l'équipe PAROLE du LORIA. La méthode que nous avons élaborée au cours de cette thèse complète le système existant sous deux formes : d'une part, par une amélioration substantielle des performances de la méthode (nous obtenons une meilleure précision, avec des codebooks plus compacts, et dans des temps plus courts). Ces améliorations constituent l'objet de la première partie de la thèse. D'autre part, dans une deuxième partie, par l'introduction de contraintes originales exploitant « l'information contextuelle » :

- information contextuelle implicite dans les chapitres 5 et 6, par l'introduction de contraintes phonétiques, c'est-à-dire de contraintes articulatoires spécifiques au phonème reconnu sur le signal de parole inversé,
- information contextuelle explicite dans le chapitre 7, où des informations relatives aux articulateurs visibles obtenues de manière automatique sont utilisés en complément de l'information acoustique.

Première partie

Inversion

Introduction

Dans cette partie, la méthode d'inversion proprement dite est présentée. Comme nous l'avons dit précédemment, notre méthode s'inspire largement de la méthode développée par Slim Ouni dans sa thèse (Ouni 2001), tout en l'améliorant considérablement. La grande force de la méthode de Ouni est de permettre une représentation exhaustive et sous une forme compacte de la relation acoustique ⇒ articulatoire.

Par rapport à la méthode originale, diverses modifications notables ont été apportées : une modélisations plus fine de la relation articulatoire ⇒ acoustique, une précision accrue des calculs ainsi que diverses améliorations algorithmiques permettant d'accélérer les calculs.

Dans cette partie sont présentés les différents modules composant l'infrastructure de notre méthode d'inversion. Le chapitre 2 présente la méthode de construction de « codebook hypercubique » telle qu'utilisée par Slim Ouni, ainsi que les différentes améliorations apportées au niveau de la structuration et de la modélisation des données. Le chapitre 3 présente la méthode d'inversion statique par codebook. Enfin, le chapitre 4 présente les modules pour l'inversion dynamique.

Chapitre 2

Construction de codebook hypercuboïque

Introduction

Notre but est de représenter l'ensemble de la relation articulatoire ⇒ acoustique (notée par la suite Ar ⇒ Ac) de façon compacte. Pour cela, nous allons réaliser un pavage de l'espace articulatoire en petits éléments, où la relation de articulatoire vers l'acoustique peut être évaluée très rapidement. Comme nous l'avons évoqué précédemment, les études de (Fant 1970; Atal *et al.* 1978; Sorokin & Trushkin 1996) montrent toutes que la relation Ar ⇒ Ac est naturellement localement linéaire.

2.1 Paramétrisation acoustique et articulatoire

Notre méthode de tabulation est générique, et ne dépend aucunement du modèle articulatoire ou du synthétiseur utilisé – à vrai dire, avec quelques légères modifications, elle pourrait même être utilisée sur un corpus de données réelles. Les espaces articulatoires et acoustiques seront donc considérés de façon très générique comme étant des espaces vectoriels réels, et on supposera simplement disposer d'une application f allant d'un sous-domaine de l'espace articulatoire vers un sous-domaine de l'espace acoustique.

Cela étant, dans toutes nos applications, nous utiliserons comme espace articulatoire l'espace des paramètres de contrôle du modèle articulatoire de Maeda (Maeda 1990), et comme espace acoustique l'espace des caractéristiques des premiers formants (essentiellement fréquences, mais parfois aussi largeurs de bande et amplitude). Revenons donc rapidement sur le modèle articulatoire, sur le synthétiseur articulatoire intégré, et sur notre paramétrisation acoustique.

2.1.1 Modèle articulatoire de Maeda

Établi à partir d'une analyse statistique de données cinéradiographiques, le modèle articulatoire de Maeda (Maeda 1979; Maeda 1990) est l'un des modèles les plus utilisés pour modéliser le conduit vocal. Il décrit un conduit vocal complet à partir de trois modèles indépendants pour les lèvres, la langue, et le larynx.

Maeda (Maeda 1979) a établi son modèle à partir de données cinéradiographiques d'une locutrice française native, issues de l'Institut de Phonétique de Strasbourg(Bothorel *et al.* 1986). À l'aide d'une analyse en composantes orthogonales arbitraires (proposée par Overall (Overall 1962)),

aussi appelée analyse en composantes principales guidée, il a construit un modèle de la coupe médio-sagittal du conduit vocal, contrôlé par 7 paramètres. Un premier paramètre (jw) contrôle l'ouverture de la mâchoire ; chaque zone du conduit vocal (lèvres, langue, larynx) ne dépend que de ce paramètre et de paramètres intrinsèques. La zone du larynx est controlée par un paramètre ; les lèvres sont décrites par deux paramètres intrinsèques : l'ouverture verticale (lh) et la protrusion (lp) ; pour la langue, trois paramètres intrinsèques sont utilisés, correspondant à la position du corps de la langue (tb), forme de la langue (ts) et pointe de la langue (tt). Ces 7 paramètres (voir figure 2.1) permettent de décrire 96% de la variance des radiographies.

Ce modèle décrit le conduit vocal de la locutrice de référence, mais il est possible d'adapter le modèle articulatoire à des locuteurs différents : deux paramètres d'élongation des conduits oral et pharyngal permettent, dans une certaine mesure, d'adapter la forme du conduit à un nouveau locuteur. Ces paramètres influent uniformément sur les dimensions des deux conduits, mais en jouant habilement avec, il est possible d'établir un modèle capable de former des fonctions d'aire correspondant aux réalisations acoustiques d'un locuteur différent. Galván-Rodrigez a établi une méthode semi-automatique d'adaptation permettant à partir des fréquences formantiques des voyelles d'un locuteur donné d'établir les coefficients d'élongation (Galván-Rodrigez 1997). Le défaut de cette méthode est qu'elle suppose que la réalisation d'une voyelle donnée utilise une configuration articulatoire unique et indépendante du locuteur.

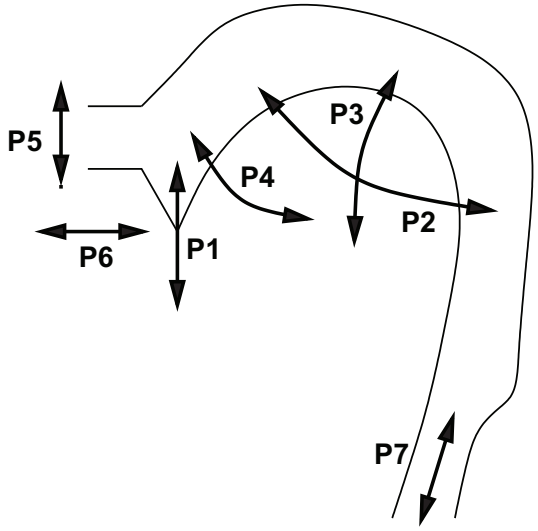


FIG. 2.1: *Les sept paramètres du modèle de Maeda : la mâchoire (ou jw) P1, l'ouverture verticale (lh) P5 et la protrusion des lèvres (lp) P6, la position du corps de la langue (tb) P2, la forme de la langue (ts) P3, un dernier terme contrôlant la pointe de la langue (tt) P4, et enfin la hauteur du larynx (lx) P7.*

2.1.2 Synthétiseur

Le modèle articulatoire ne décrit que la coupe medio-sagittale du conduit. Pour déterminer la fonction de transfert de celui-ci il est d'abord nécessaire de retrouver sa troisième dimension (cf. Section 1.2.2), ce qui permet d'en déduire sa fonction d'aire. Un synthétiseur également dû à Maeda (Maeda 1972) permet alors de déterminer la fonction de transfert correspondante (cf. Section 1.2.3).

F_i	$d(P_1)$	$d(P_2)$	$d(P_3)$	$d(P_4)$	$d(P_5)$	$d(P_6)$	$d(P_7)$
F_1	0.157	0.193	0.105	0.071	0.188	0.028	0.014
F_2	0.047	0.155	0.060	0.050	0.108	0.019	0.034
F_3	0.030	0.070	0.027	0.035	0.032	0.009	0.030
$\sum F_i$	0.234	0.418	0.192	0.155	0.327	0.056	0.077

Les nombreuses hypothèses simplificatrices (régime stationnaire, onde plane, etc.) ainsi que l'approximation permettant de retrouver la troisième dimension introduisent cependant des erreurs, et la fonction de transfert ne peut être considérée comme valable que pour les fréquences inférieures à 4kHz, et pour les conduits sans constriction. Par conséquent, dans l'état actuel, il ne peut guère être utilisé que pour l'étude des voyelles.

2.1.3 Paramétrisation acoustique

Le synthétiseur fréquenciel actuel n'inclut pas de modèle de source glottique et calcule simplement la fonction de transfert du conduit vocal. Il est donc nécessaire, pour espérer avoir une correspondance entre les signaux acoustiques naturels et synthétiques, d'utiliser une paramétrisation acoustique capable d'éliminer l'influence de la source du signal, ou qui y soit insensible. Par ailleurs, il est également souhaitable de choisir une paramétrisation robuste au bruit, et qui ne soit pas trop sensible aux défauts du synthétiseur.

Pour toutes ces raisons, nous avons préféré dans la très grande majorité de nos expériences utiliser les fréquences des premiers formants comme vecteurs acoustiques ; on peut en effet espérer avoir une correspondance entre les premiers formants et les premières résonnances de la fonction de transfert. On désignera d'ailleurs souvent par la suite, par abus de langage, les résonnances de la fonction de transfert et leur manifestation dans le signal de parole, sous le nom de « formants ». Généralement, on se limitera aux fréquences des 3 premiers formants, qui sont les seuls à pouvoir être déterminés de façon fiable sur le signal de parole.

2.1.4 Quelques résultats

Une première expérience relativement facile à réaliser est de calculer l'influence des différents paramètres sur l'acoustique. Nous présentons ici l'influence moyenne de chacun des paramètres. Cette moyenne est calculée sur l'ensemble de l'espace articulatoire « utile[1] ». Pour chaque fréquence formantique, et pour chaque composante j d'une vecteur articulatoire α, nous calculons $d(P_j) = \int_\alpha |\frac{\Delta F_i(\alpha_j)}{F_i(\alpha)} d\alpha|$ sur l'ensemble de l'espace articulatoire, $\Delta F_i(\alpha_j)$ étant la variation de fréquence formantique correspondant à une variation d'une unité du paramètre j considéré. La table 2.1.4 présente ces résultats.

L'examen de cette table nous indique que l'influence acoustique de chaque paramètre est très variable. Les paramètres P6 et P7 (respectivement protrusion des lèvres et hauteur du larynx) ont ainsi une influence très faible comparée à celle des autres paramètres. Les deux paramètres ayant le plus d'influence acoustique sont P2 et P5, respectivement la position du dos de la langue et l'ouverture des lèvres.

[1] C'est-à-dire l'ensemble des formes des conduits pour lesquels le locuteur de référence produit une image acoustique, i.e. les conduits sans occlusion.

2.2 Présentation de la structure hypercuboïdale

La méthode élaborée dans le cadre de cette thèse est une généralisation de la méthode à « codebook hypercubique » élaborée par Slim Ouni dans le cadre de sa thèse (Ouni 2001). Notre méthode est en effet plus générale ; elle permet de représenter l'espace articulatoire non plus sous la forme d'hypercubes – c'est-à-dire la généralisation dans un espace à M dimensions du carré, – mais sous la forme de M-parallélotopes droits – c'est-à-dire la généralisation dans un espace à M dimensions du rectangle.

L'objet d'un codebook pour l'inversion acoustique-articulatoire est de parvenir à représenter l'intégralité de la relation Ar $\Rightarrow$ Ac de façon compacte et précise. L'apport de la méthode initiée par Ouni a été important : il s'agissait de la seule méthode représentant l'intégralité de la relation Ar $\Rightarrow$ Ac avec une précision acoustique homogène. Sorokin (Sorokin *et al.* 2000) utilise également une méthode à codebook avec une précision acoustique homogène, mais il ne prend en compte qu'une partie de l'espace articulatoire. Les autres méthodes à base de codebooks (Schroeter et Larar (Larar *et al.* 1988), Charpentier (Charpentier 1984), ...) ne garantissent pas une précision acoustique homogène.

Nous présentons dans cette thèse des améliorations à la méthode d'Ouni portant essentiellement sur 3 points : la structure de représentation (hypercube dans le cas d'Ouni), la modélisation mathématique de la fonction (linéaire dans le cas d'Ouni), et enfin sur l'évaluation numérique de certains éléments essentiels de la méthode (test de linéarité, choix des points...). Ces différentes améliorations donnent une plus grande flexibilité et une bien meilleure robustesse au système d'inversion.

2.2.1 Choix de la structure

Dans cette partie, nous cherchons à déterminer une structure qui nous permette d'explorer de façon exhaustive la totalité de l'espace utile des paramètres contrôlant le modèle articulatoire que nous utilisons, celui de de Maeda. Comme nous l'avons vu à la section 1.2.1.2, ce modèle articulatoire est contrôlé par 7 paramètres, chacun de ces paramètres variant typiquement dans l'intervalle $[-3;3]$. On peut ainsi considérer que cet espace articulatoire est contenu dans un hypercube de dimension 7 et de côté 6. La structure retenue pour mener à bien cette exploration se doit d'être simple tout en permettant de représenter l'ensemble de la relation étudiée de façon compacte et de préférence non redondante (ce que ne permettent pas l'utilisation de boules, par exemple). La structure de base retenue par Slim Ouni était l'hypercube ; nous avons pour notre part choisi d'utiliser le parallélotope droit, ou *hypercuboïde*, qui est une structure plus puissante tout en restant relativement simple. Cette structure alliée à une méthode de construction adaptée permet également d'éliminer toute redondance.

2.2.2 Définitions de la structure hypercuboïdale

Dans la suite, on se place dans le cadre très général où les solutions recherchées peuvent être décrites par un vecteur réel à M dimensions, $M > 0$; l'espace vectoriel correspondant, ou plutôt sa restriction à un sous-ensemble de vecteurs utiles, sera désigné sous le nom d'*espace articulatoire*. Dans nos applications, M sera systématiquement égal à 7, mais il est important de souligner que notre méthode pourrait être appliquée à n'importe quel modèle articulatoire.

Par convention, et par soucis de concision et de cohérence avec les dénominations utilisées par Slim Ouni, on désignera par *hypercuboïde* plutôt que par M-parallélotope droit les éléments du codebook. Un codebook constitué d'hypercuboïdes sera dénommé *codebook hypercuboïque*.

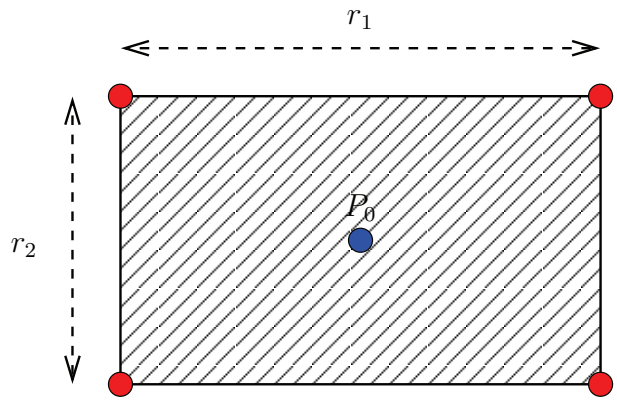


FIG. 2.2: *Hypercuboïde de centre P_0 et de rayon $\vec{r}$ dans un espace à 2 dimensions.*

De façon formelle, un M-parallélotope droit H_c est un ensemble de vecteurs x de $\mathbb{R}^M$ vérifiant la propriété suivante : il existe une base orthonormée B de $\mathbb{R}^M$ et un vecteur $\vec{r}$ de $\mathbb{R}^{+*M}$ tels que, si on note par $x_1, ..., x_M$ la projection des vecteurs x dans la base B, on ait :

$$H_c = \{x \in \mathbb{R}^M \ | \forall i \in \{1..M\} \ |x_i| \leq r_i\}$$

Le vecteur nul dans la base B, que l'on notera P_0, sera appelé le *centre* de l'hypercuboïde.

Par ailleurs, un hypercuboïde de $\mathbb{R}^M$ est entièrement déterminé par la donnée d'une base orthonormée B (soit deux vecteurs de $\mathbb{R}^M$, un pour la position de l'origine de la base, et un autre pour son orientation), et d'un vecteur $\vec{r}$, que l'on appellera le *rayon* de l'hypercuboïde. Par la suite, nous ne mentionnerons plus la base B, car son orientation sera toujours la même par construction, et les coordonnées de l'origine sont identiques à celle du centre P_0 de l'hypercuboïde exprimées dans la base canonique. Les hypercuboïdes que nous manipulerons seront ainsi entièrement déterminées par la donnée de P_0 et $\vec{r}$:

$$H_c(P_0, \vec{r}) = \{x \in \mathbb{R}^M \ | \forall i \in \{1..M\} \ |(x - P_0)_i| \leq r_i\} \tag{2.1}$$

On remarquera qu'un hypercube au sens de Ouni est simplement un hypercuboïde dont les coefficients du rayon sont tous égaux. La figure 2.2.2 illustre le concept d'hypercuboïde dans un espace à 2 dimensions.

2.2.3 Modélisation mathématique

En général, les méthodes à base de tables utilisent des fonctions extrêmement simples pour modéliser localement des relations complexes. Dans le cadre de l'étude de l'inversion acoustique-articulatoire, deux grand types de fonctions ont ainsi été utilisés pour modéliser le comportement local : des fonctions constantes par morceaux (Atal *et al.* 1978; Schroeter & Sondhi 1992; Boë *et al.* 1992), ou des fonctions linéaires par morceaux (Charpentier 1984; Sorokin & Trushkin 1996; Ouni & Laprie 2000). Nous avons développé dans le cadre de cette thèse une méthode plus générale utilisant des polynômes multivariés, dont ces deux types de fonctions ne sont en définitive que des cas particuliers.

Cette classe de fonctions n'étant pas souvent utilisée dans la communauté du traitement automatique de la parole, nous nous permettons, pour simplifier la compréhension du lecteur, de la présenter brièvement (pour plus de détails, voir par exemple (Birkhoff 1979)).

2.2.3.1 Polynômes multivariés

Un *polynôme multivarié*, de variables $x_1, \ldots, x_M$ d'un anneau A, noté $P(x_1, \ldots, x_M)$, est une somme de monômes ; un *monôme multivarié*, $Q(x_1, \ldots, x_M)$ étant un terme de la forme :

$$c.{x_1}^{k_1} {x_2}^{k_2} \cdots {x_M}^{k_M},$$

où $c \in A$ est le coefficient du monôme, et pour tout i de $1, \ldots, M$, $k_i \in \mathbb{N}$ est l'exposant de la variable x_i.

Le *degré* d'un monôme est défini comme la somme des exposants, $\sum_{i=1}^{M} k_i$. Le degré d'un polynôme est défini, comme d'habitude, comme le maximum des degrés des monômes qui le composent.

Au monôme $Q(x_1, \ldots, x_M)$, on associe classiquement une *fonction monôme* $q : A^M \mapsto A$, qui à tout vecteur $V \in A^M$ de valeurs $(v_1, \ldots, v_M)$ pour les variables $x_1, \ldots, x_M$ associe la valeur $q(v_1, \ldots, v_M) \in A$ définie ainsi : $c.{v_1}^{k_1} {v_2}^{k_2} \cdots {v_M}^{k_M}$. De la même façon, pour tout polynôme multivarié, on peut associer une fonction polynôme comme la fonction somme des fonctions monômes associées aux monômes qui le composent.

On définit X comme la somme de monômes $x_1 + x_2 + \cdots + x_M$. Pour $n \in \mathbb{N}$, on définit également X^n ainsi : $(x_1 + x_2 + \cdots + x_M)^n$. Nous avons la relation suivante :

$$X^n = \sum_{1 \leq i_1, i_2, \ldots, i_n \leq M} x_{i_1} x_{i_2} \cdots x_{i_n} \tag{2.2}$$

Si A est un anneau commutatif (ce qui sera toujours le cas pour nous), beaucoup des termes de la somme peuvent être groupés ; on peut ainsi récrire la formule 2.2 en utilisant la formule du multinôme :

$$(x_1 + x_2 + \cdots + x_M)^n = \sum_{k_1, k_2, \ldots, k_M} \binom{n}{k_1, k_2, \ldots, k_M} x_1^{k_1} x_2^{k_2} \cdots x_M^{k_M}$$

La somme est effectuée sur toutes les séquences d'indices entiers positifs $k_1, \ldots, k_M$ tels que $\sum_{i=1}^{M} k_i = n$; les nombres

$$\binom{n}{k_1, k_2, \ldots, k_M} = \frac{n!}{k_1! k_2! \cdots k_M!}$$

sont appelés les coefficients multinômes.

Dans la suite, l'anneau A sera considéré comme un corps. Le nombre de monômes distincts dans cette somme est égal à $\binom{M+n-1}{n}$. Pour simplifier l'écriture, on peut ordonner ces monômes, par exemple en utilisant l'ordre lexicographique inverse sur les vecteurs $k_1, \ldots, k_M$.

Pour alléger les notations, les polynômes multivariés seront dans la suite notés simplement $P(X)$ au lieu de $P(x_1, x_2, \ldots, x_M)$. Il faut noter, qu'en fonction du contexte, X peut-être vu comme une variable de l'anneau produit A^M, comme le polynôme $P(x_1, x_2, \ldots, x_M) = x_1 + x_2 + \cdots + x_M$, ou encore comme la fonction polynôme $A^M \mapsto A : (x_1, \cdots, x_M) \to x_1 + \cdots + x_M$. Tout polynôme de degré n peut s'écrire ainsi :

$$P(X) = A_0 + A_1 X + A_2 X^2 + \cdots + A_n X^n.$$

Dans cette expression, A_0 est un élément de A, $A_1 = (A_{1,1}, \ldots, A_{1,M}) \in A^M$ est le vecteur de coefficients de X, i.e.

$$A_1 X = A_{1,1} x_1 + A_{1,2} x_2 + \cdots + A_{1,n} x_n.$$

De la même façon, $\forall m \in \{1, \ldots, n\}, A_m \in R^{\binom{M+m-1}{m}}$ est le vecteur de coefficients de X^m, avec

$$A_m X^m = \sum_{\sum_{i=1}^{M} k_i = n} A_{m,w} \binom{n}{k_1, k_2, \ldots, k_M} x_1^{k_1} x_2^{k_2} \cdots x_M^{k_M},$$

où w est un indice pour le vecteur $k_1, k_2, \ldots, k_M$.

2.3 Construction

La relation à modéliser pourra ainsi dans notre méthode être décrite localement par un polynôme multivarié ; nous allons à présent expliciter comment l'espace articulatoire peut être découpé en sous-relations locales, ou en d'autres termes, la façon dont notre codebook est construit.

Le principe de la méthode de construction du codebook est très simple et réalise naturellement un pavage de l'espace articulatoire : elle consiste en une analyse récursive de plus en plus fine de la relation Ar $\Rightarrow$ Ac locale. Pour résumer, un hypercuboïde est analysé de la façon suivante : on calcule une approximation (linéaire ou polynomiale) de la relation locale à partir d'un ensemble de points de l'hypercuboïde, puis on y effectue un test pour estimer la proximité de l'approximation avec le comportement réel de la relation dans cet hypercuboïde. Si l'approximation est suffisamment proche de la réalité, l'hypercuboïde est conservé. Sinon, ce dernier est décomposé en sous-hypercuboïdes qui seront à leur tour analysés, et ainsi de suite.

La méthode de construction met ainsi en œuvre trois éléments essentiels : le calcul d'une approximation locale, un test de régularité, et une méthode de subdivision. Pour des raisons didactiques et historiques, test de régularité et approximation sont présentés simultanément.

2.3.1 Test de régularité

L'un des points cruciaux de cette méthode est la réalisation d'un test rapide et efficace qui permette de déterminer si localement la relation est suffisamment conforme au comportement attendu. L'un des éléments à prendre en compte est le temps prohibitif que prend le calcul d'une image d'un vecteur articulatoire ; il est ainsi nécessaire de limiter au maximum le nombre d'évaluations à effectuer dans un hypercuboïde donné. Plusieurs approches ont été envisagées : la première est le test utilisé par Ouni, c'est-à-dire un test sur tous les segments reliant deux sommets de l'hypercuboïde ; une deuxième approche consiste à comparer l'image réelle de certains points du cube (en l'occurrence les sommets) à l'image prédite par l'approximation. Pour cette seconde approche, deux méthodes de calcul d'approximation ont été développées : la première, uniquement valable pour les approximations linéaires, utilise le calcul de la matrice jacobienne en un point particulier (le centre) de la structure, la seconde calcule une matrice d'approximation linéaire « optimale » pour l'hypercuboïde, que l'on généralise facilement à une approximation polynomiale de degré quelconque.

2.3.1.1 Test « classique »

Le test de linéarité développé par Ouni est assez rapide et consiste à étudier localement le comportement de la relation aux extrémités de l'hypercuboïde considéré. Pour tous les segments reliant deux sommets de la structure (des hypercubes dans le cas d'Ouni), la linéarité locale est considérée à l'aide du point milieu du segment : l'image interpolée linéairement à partir des sommets et l'image synthétisée à partir du modèle articulatoire sont comparées, et si la différence

entre les deux images est inférieure à un seuil prédéfini ϵ, la relation articulatoire-acoustique est considérée comme linéaire dans la structure.

De façon plus formelle, si on note Ar l'espace articulatoire, et Ac l'espace acoustique, M et N étant les dimensions respectives de l'espace articulatoire et de l'espace acoustique, alors on désigne par f la relation $f : \text{Ar} \rightarrow \text{Ac}$ représentant le synthétiseur articulatoire. On désigne par S_{Hc} l'ensemble des sommets d'un hypercuboïde Hc. Soit $P_i, P_j \in S_{\text{Hc}}$ deux sommets de l'hypercuboïde, on désigne par F_i et $F_j \in \text{Ac}$ les vecteurs images : $F_i = f(P_i)$ et $F_j = f(P_j)$.

Le test de linéarité entre les deux sommets s'écrit ainsi :

$$d\left(\frac{F_i + F_j}{2}, f\left(\frac{P_i + P_j}{2}\right)\right) \leq \epsilon,$$

où d est une fonction de distance de $\text{Ac} \times \text{Ac}$ vers $\mathbb{R}^+$, ou, en d'autre termes, une mesure de la similitude entre deux vecteurs acoustiques ; la forme pratique de cette fonction sera discutée à la section 2.3.3.

La complexité de ce test de linéarité est assez élevé : en effet, nous avons 2^M sommets, et le nombre de segments différents reliant deux sommets est élevé : il y a $2^{M-1} \times (2^M - 1)$ segments différents. L'hypercuboïde ne sera considéré comme linéaire que si tous ces tests réussissent.

Le nombre de points différents à synthétiser est un peu plus faible : il faut synthétiser 2^M sommets, et α_M milieux de segments différents. On peut montrer simplement[2] que $\alpha_M = 3^M - 2^M$. Au total, on doit donc synthétiser 3^M vecteurs distincts dans chaque hypercuboïde.

En pratique, l'espace articulatoire étant de dimension 7, cela fait 8128 segments à considérer, soit 2187 appels au synthétiseur, ce qui est déjà un nombre très élevé.

Ce test, bien qu'assez artificiel (une très grande partie de « l'intérieur » reste totalement inexplorée) est fiable, mais n'est pas optimal, car il nécessite l'évaluation d'un nombre prohibitif de points. Par ailleurs, cette méthode ne permet pas de déterminer avec quelle précision les solutions de l'inversion pourront être générées. Nous avons donc développé des approches différentes.

2.3.1.2 Test à partir de la matrice jacobienne

Cette deuxième méthode nécessite l'évaluation d'un nombre de points nettement plus faible que la précédente, tout en garantissant un test fiable. Elle consiste à calculer une matrice jacobienne de la relation $\text{Ar} \Rightarrow \text{Ac}$ autour d'un point précis ; à partir de cette matrice et de l'image de ce point, il est ainsi possible, en faisant l'hypothèse de linéarité, de déterminer l'image de n'importe quel point de la structure. Il suffit alors de comparer les images estimées grâce à cette méthode aux images réelles calculées à l'aide du synthétiseur, pour vérifier si l'hypothèse de linéarité est effectivement vérifiée. En d'autres termes : soit $P_0 \in \text{Hc}$ un point de l'hypercuboïde. On note $F_0 = f(P_0)$ l'image de ce point, et Δf_{P_0} la matrice jacobienne en ce point, c'est-à-dire la matrice des dérivées partielles selon chaque composante articulatoire, calculées en ce point :

[2] Idée : considérer les vecteurs à M dimensions pour lesquels chaque composante prend 1 des 3 états $\{-1, 0, 1\}$. Considérer la bijection canonique de ces vecteurs avec les points particuliers (sommets, centres) de l'hypercube. Observer que tous les vecteurs qui contiennent au moins 1 zéro correspondent aux centres de segments, ceux qui n'en contiennent aucun sont les sommets. La combinatoire est triviale pour trouver le résultat.

$$\Delta f_{P_0} = \begin{bmatrix} \frac{\partial f_1}{\partial \alpha_1}(P_0) & \frac{\partial f_1}{\partial \alpha_2}(P_0) & \cdots & \frac{\partial f_1}{\partial \alpha_M}(P_0) \\ \frac{\partial f_2}{\partial \alpha_1}(P_0) & \frac{\partial f_2}{\partial \alpha_2}(P_0) & \cdots & \frac{\partial f_2}{\partial \alpha_M}(P_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial \alpha_1}(P_0) & \frac{\partial f_N}{\partial \alpha_2}(P_0) & \cdots & \frac{\partial f_N}{\partial \alpha_M}(P_0) \end{bmatrix}$$

L'hypothèse de linéarité nous permet de faire l'approximation suivante :

$$\forall P_x \in \mathrm{Hc}, f(P_x) \approx F_0 + (P_x - P_0).\Delta f_{P_0}$$

Le test de linéarité s'écrit alors :

$$d\left(F_0 + (P_x - P_0).\Delta f_{P_0}, f\left(P_x\right)\right) \leq \epsilon,$$

Cette méthode repose sur la façon dont les solutions de l'inversion vont être générées : en effet, lors de l'inversion, on utilise l'hypothèse de linéarité dans la structure pour calculer l'image d'un vecteur articulatoire ; et, comme en le verra au chapitre 3, cela se fait à l'aide de la matrice jacobienne calculée au centre de l'hypercuboïde. En vérifiant la linéarité ainsi, on a une idée plus précise de l'erreur que l'on commet lors de l'inversion.

Cette méthode peut être utilisée avec un nombre quelconque de points, ces points pouvant être choisis aléatoirement à l'intérieur de l'hypercuboïde. En pratique, pour limiter le temps des calculs, nous nous contentons généralement de tester les 128 sommets de l'hypercuboïde, pour lesquels le calcul de l'image a l'avantage d'être « factorisable » (cf. section 2.4.2).

La fiabilité[3] de ce test en n'utilisant que les sommets de l'hypercuboïde est équivalente à celle du test « classique » : dans les deux cas, on obtient des codebooks dont le taux de points pour lesquels l'erreur de resynthèse dépasse le seuil acoustique, est inférieur à 0.1%.

2.3.1.3 Minimisation de l'erreur d'interpolation

Une dernière méthode consiste à généraliser la méthode précédente à une interpolation polynomiale de degré supérieur ou égal à 1, en optimisant le calcul de la matrice d'interpolation de façon à minimiser l'erreur commise lors de la resynthèse.

La complexité des calculs devient cependant vite rédhibitoire pour les degrés élevés. Par ailleurs, bien qu'il soit relativement simple d'appliquer une régression polynomiale de degré quelconque de la relation Ar $\Rightarrow$ Ac, il est nettement plus compliqué d'exploiter par la suite les données du codebook lors de l'inversion, puisqu'utilisant une modélisation non-linéaire, on ne peut plus utiliser les méthodes classiques d'algèbre linéaire. Cette approche nous permet toutefois de déterminer le modèle permettant d'approcher au mieux le comportement réel de la relation Ar $\Rightarrow$ Ac, et par ailleurs, de réaliser un synthétiseur articulatoire à codebook très précis.

Comme nous le verrons dans la section 2.5, il semble que l'interpolation polynomiale de degré 2 permette d'obtenir une excellente précision acoustique tout en occupant un espace nettement plus réduit que l'interpolation linéaire. Les degrés supérieurs améliorent encore ces résultats (mais occupent davantage d'espace disque), il semblerait donc que l'interpolation polynomiale soit la plus à même de représenter fidèlement la relation Ar $\Rightarrow$ Ac.

[3] C'est-à-dire avec quelle précision la linéarité locale est évaluée ; pour cette mesure deux éléments entrent en ligne de compte : le taux de points synthétisés dépassant le seuil, et le nombre de subdivisions dont on aurait pu se passer.

La figure 2.3 illustre cette idée : sur cette figure est représentée en trait fin continu une fonction (mono-dimensionnelle) que l'on cherche à approcher, sur un intervalle borné, par une approximation linéaire, et plusieurs façons de l'approcher au mieux. La première (tirets noirs épais) représente l'approximation à partir de la tangente (correspondant à la matrice jacobienne en multi-dimension), calculée au centre. On constate que l'erreur (flèches verticales entre la courbe et son approximation) est très importante sur les extrémités de la zone. La deuxième approximation (ligne tiret-point rouge) correspond à une manière particulière de calculer une valeur approchée de la tangente (ou la matrice jacobienne), qui sera explicitée à la section 2.4.3.1, et qui permet de beaucoup mieux approcher la fonction. Enfin, la ligne tiret-point-point bleue correspond à l'approximation minimisant l'erreur d'interpolation (au sens de la norme absolue), et qui a donc la plus faible erreur absolue. On peut observer au passage qu'il ne s'agit pas forcément de l'approximation qui minimise l'erreur moyenne.

2.3.2 Utilisation des polynômes multivariés pour calculer l'approximation

Considérons un hypercuboïde Hc de l'espace des paramètres articulatoires. Pour chaque composante $F_i, 1 \leq i \leq N$ du vecteur acoustique, nous souhaitons trouver le polynôme $P(X)$ (de degré n) qui décrit au mieux la relation propre à la composante F_i, $f_i : \mathrm{Ar}^M \mapsto \mathrm{Ac}$ dans la structure. En d'autres termes, nous cherchons à déterminer un polynôme $P(X)$ qui minimise l'erreur d'approximation dans la relation $\{f_i(X) = P(X)\}$ pour $X \in \mathrm{Hc}$.

Ceci peut être fait relativement simplement en résolvant numériquement un système d'équations de la forme $\{f_i(X_j) = P(X_j)\}$ pour un grand nombre de vecteurs articulatoires $X_j = (x_{j1}, x_{j2}, \cdots, x_{jM})$. Un grand nombre de vecteurs articulatoires, au moins autant que de coefficients à déterminer, sont choisis dans l'hypercuboïde ; leurs images acoustiques réelles sont déterminées grâce au synthétiseur. Après réécriture du système $f_i(X_j) = P(X_j)$ comme une équation linéaire de coefficients inconnus, le système d'équations surdéterminé peut être résolu par une Décomposition en Valeurs Singulières (ou SVD), qui a la propriété intéressante de minimiser l'erreur au sens des moindres carrés sur l'ensemble des points considérés (Golub & Loan 1989). Cette méthode est cependant sensible à l'ensemble de points utilisé pour l'approximation.

Une autre approche, qui généralise celle présentée à la section 2.3.1.2, consiste à calculer une approximation de Taylor d'ordre n autour d'un point particulier de l'hypercuboïde. En réalité, bien que l'approximation soit meilleure autour du point considéré, elle est moins bonne en moyenne dans l'ensemble de l'hypercuboïde, la précision acoustique que l'on obtiendrait en utilisant ce type de polynôme serait ainsi moins bonne qu'avec la méthode à base de SVD. D'autre part, cela rend la méthode dépendante du choix du point où le développement de Taylor est calculé. Enfin, l'étape de discrétisation utilisée lors du calcul de l'image acoustique à partir d'un vecteur articulatoire, rend le calcul des dérivées partielles sensible à des erreurs numériques importantes, particulièrement pour les termes de degré élevé. Pour finir, en utilisant le schéma précédent et un ensemble de points bien choisi, il est toujours possible de « simuler » l'approximation de Taylor en n'importe quel point : il suffit de choisir un ensemble de points concentrés autour du centre d'approximation désiré.

Soit C_i le vecteur composé de tous les coefficients du polynôme $P(X)$:

$$C_i = (\underbrace{A_0}_{1} | \underbrace{A_1}_{M} | \cdots | \underbrace{A_n}_{\binom{M+n-1}{n}}).$$

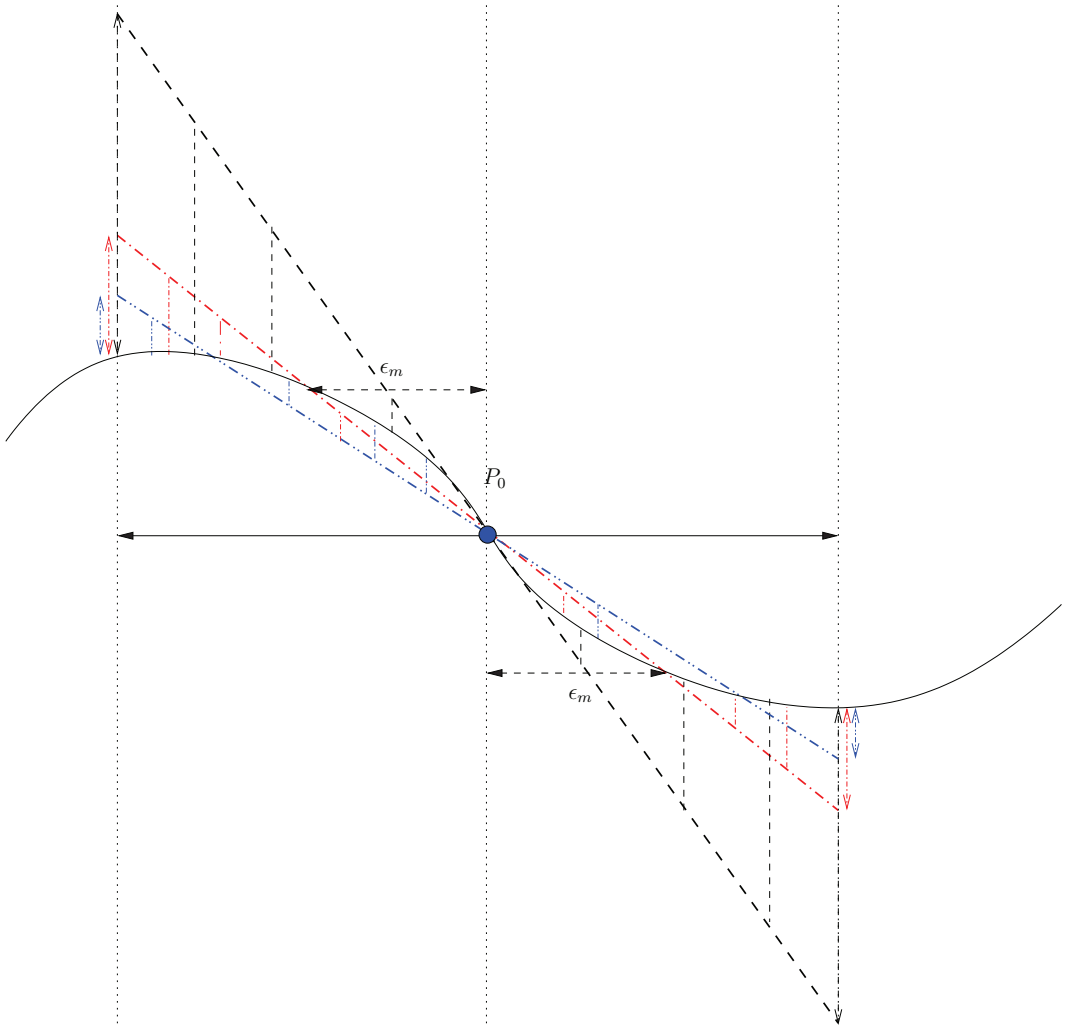


FIG. 2.3: *Différentes méthodes d'approximation linéaire et erreurs correspondantes (flèches verticales) : la courbe que l'on cherche à approcher est celle en trait fin continu, et trois types d'approximation sont présentés : approximation par la tangente au centre (tirets noirs), approximation par calcul décentré de la tangente au centre (tiret-point rouge), approximation par minimisation de l'erreur (tiret-point-point bleu).*

Soit W_j le vecteur composé des vecteurs ${X_j}^k$:

$$W_j = (\underbrace{{X_j}^0}_{1} | \underbrace{{X_j}^1}_{M} | \cdots | \underbrace{{X_j}^n}_{\binom{M+n-1}{n}}) =$$

$$\Bigg(1 \mid x_{j1}, \ldots, x_{jM} | \cdots | {x_{j1}}^n, \ldots,$$

$$\binom{n}{k_1, k_2, \ldots, k_M} {x_{j1}}^{k_1} {x_{j2}}^{k_2} \cdots {x_{jM}}^{k_M}, \ldots, {x_{jM}}^n\Bigg)$$

L'équation $f_i(X_j) = P(X_j)$ peut ainsi être récrite sous la forme de l'équation linéaire $f_i(X_j) = W_j.C_i$. Enfin, désignons par B_i le vecteur contenant tous les $f_i(X_j)$, et W la matrice contenant tous les W_j. Nous obtenons alors le système d'équations linéaires suivant :

$$W.C_i = B_i \tag{2.3}$$

Pour que ce système soit surdéterminé, le nombre m de vecteurs articulatoires à évaluer doit être au moins égal au nombre de coefficients distincts, soit :

$$m \geq \sum_{k=0}^{n} \binom{M+k-1}{k} = \binom{M+n}{n}$$

Ce système surdéterminé peut être « résolu » facilement à l'aide de la SVD, et le polynôme solution trouvé présente la propriété intéressante de minimiser l'erreur d'approximation au sens des moindres carrés (Golub & Loan 1989). En répétant le processus pour chacune des composantes du vecteur acoustique, nous pouvons ainsi obtenir une approximation polynomiale optimale pour la resynthèse.

Bien entendu, il ne faut pas perdre de vue que cette approximation est certes optimale, mais au sens des moindre carrés, et sur un ensemble de points réduit, alors que l'on désirait initialement trouver une approximation optimale au sens de la norme acoustique qui nous intéresse – on cherche à minimiser $\max d(f(X_j), P(X_j))$ – et valable pour l'hypercuboïde complet. Il est cependant bien connu (Lefèvre & Zimmermann 2004) que, d'une part, cette approximation au sens de la norme euclidienne est bien meilleure que ce que l'on avait précédemment en utilisant l'approximation de Taylor, et d'autre part qu'avec un échantillonnage adapté cette méthode permet d'obtenir un polynôme ayant des résultats proches de ceux du polynôme optimal au sens de la norme qui nous intéresse (aussi appelé polynôme minimax). En effet, il a été observé que le polynôme de Chebyshev optimal au sens des moindres carrée était souvent proche du polynôme minimax. Il existe également un algorithme, dû à Remez (Remez 1934), permettant de calculer ce polynôme minimax, mais celui-ci n'est pas vraiment envisageable dans notre cas, du fait de son coût élevé en nombre d'appels au synthétiseur articulatoire.

En pratique, les vecteurs X_j utilisés dans le calcul ne sont pas les vecteurs articulatoires absolus, mais sont les coordonnées relatives au centre de l'hypercuboïde, de façon à assurer une bonne précision de l'approximation polynomiale de degré élevé. Le point de référence choisi, qui est dans notre cas systématiquement le centre géométrique de l'hypercuboïde, a une légère influence sur la précision acoustique de la resynthèse à partir de l'approximation polynomiale, mais cette influence peut être considérée négligeable pour des hypercuboïdes de taille réduite.

Le choix des ensembles de points à utiliser pour échantillonner la relation Ar $\Rightarrow$ Ac dans un hypercuboïde pour construire l'approximation reste l'un des points les plus délicats de la

méthode. Pour le moment, les points utilisés pour calculer le polynôme d'interpolation, peuvent être les points « génériques » (centres et sommets), qui ont l'avantage d'avoir un coût d'évaluation pratiquement gratuit, des points aléatoires, un échantillonnage à la Chebyshev, ou toute combinaison de ces ensembles. Il s'agit en définitive d'un compromis entre le temps de calcul d'une part, et la qualité de l'interpolation polynomiale d'autre part. En pratique, pour l'interpolation de degré inférieur ou égal à 3, nous utilisons généralement les sommets de l'hypercuboïde, et les points milieux des segments reliant deux sommets. Ces points présentent plusieurs avantages. D'une part, leur « coût » de synthèse est faible, car leur calcul est « factorisable ». En effet ces points sont partagés par de nombreux hypercuboïdes (cf. explications détaillées à la section 2.4.2). D'autre part, étant principalement situés sur les frontières d'hypercuboïdes, ils permettent de favoriser une transition relativement lisse entre les différents hypercuboïdes. Pour les degrés élevés, un échantillonnage aléatoire ou de Chebychev peut être utilisé, mais en pratique nous n'avons guère eu l'occasion de tester ces cas de façon approfondie, les calculs devenant excessivement lourds.

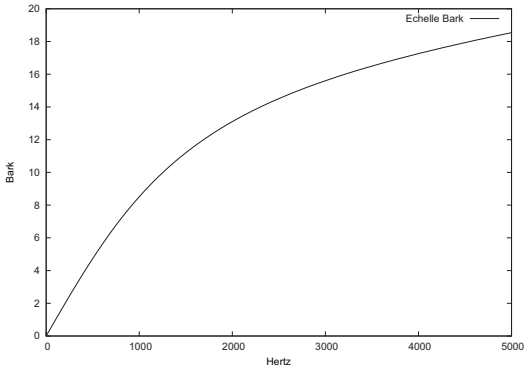
2.3.3 Seuillage acoustique

La distance d utilisée dans le test de régularité peut être définie de plusieurs façons. En général, elle est définie de façon à imposer un seuil différent sur chacun des formants, ces seuils pouvant être choisis expérimentalement ou arbitrairement ; par exemple, les seuils peuvent être fixés à 50Hz pour F1, 75Hz pour F2 et 100Hz pour F3.

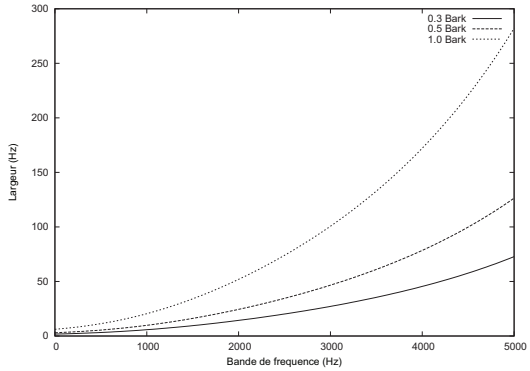
Cependant, l'échelle Hertz n'est pas forcément la meilleure échelle pour avoir un test de régularité qui soit réaliste d'un point de vue perceptif, c'est-à-dire que les seuils retenus correspondent aux seuils de sensibilité du système auditif périphérique humain.

En effet, la résolution fréquentielle de l'oreille n'étant pas linéaire (Zwicker & Feldtkeller 1981), il est nécessaire d'utiliser une transformation qui permette d'approcher le comportement de l'oreille humaine, pour obtenir une résolution acoustique perceptivement homogène pour le seuil de linéarité. Dans cette optique, l'échelle psychoacoustique Bark est souvent utilisée dans le traitement de la parole.

La figure 2.4b donne la valeur en Hertz correspondant à un seuil fixe en Bark, en fonction de la bande de fréquence considérée, pour trois seuils : 0.3, 0.5 et 1.0 Bark.



(a) Échelle psychoacoustique Bark.



(b) Largeur de bande en Hertz correspondant à un seuil fixe en bark, en fonction de la bande de fréquence considérée.

En pratique, les seuils acoustiques que nous utilisons lors de la construction de codebook sont

généralement exprimés en Bark.

2.3.4 Subdivision

Le dernier point clé de l'algorithme est la façon dont la subdivision se déroule dans l'espace articulatoire. L'une des qualités requises pour un bon codebook est sa concision : la façon dont la subdivision se déroule doit pouvoir assurer la concision tout en permettant de garder une représentation fidèle de la relation, c'est-à-dire qui supprime les zones non-synthétisables, tout en conservant le plus grand nombre possible de régions synthétisables.

En assurant une meilleure précision acoustique qu'auparavant, les méthodes d'interpolation vues précédemment permettent de limiter le nombre de subdivisions dues à la non-linéarité. Malheureusement, à elles seules, elles ne permettent pas de limiter les subdivisions le long des frontières de l'espace articulatoire, qui représentent la majorité des subdivisions. Nous allons présenter dans cette partie l'algorithme de subdivision utilisé par Ouni, puis différents algorithmes et améliorations développés pour tirer partie de la nouvelle structure hypercuboïdale.

2.3.4.1 Subdivision « classique »

Ouni subdivisait simplement chaque hypercube en divisant chacune des M dimensions de l'espace articulatoire par 2, ce qui faisait qu'il était systématiquement nécessaire de parcourir 2^M sous-hypercubes dès que le test de linéarité échouait (soit 128 pour le modèle de Maeda). Cette méthode était très coûteuse en temps de calcul et en espace occupé, puisque dès que l'on voulait améliorer la résolution articulatoire du codebook d'un niveau de subdivision supplémentaire, on multipliait le temps de calcul par environ 100, et également par environ 100 la taille du codebook. Nécessairement, cela ralentissait considérablement les accès au codebook, et par conséquent le processus d'inversion. D'autre part, on pouvait constater que parmi les 2^M sous-hypercubes d'un hypercube donné, beaucoup partageaient des comportements très proches et auraient donc pu être regroupés.

2.3.4.2 Subdivision dirigée

C'est ce constat qui est à la base de la méthode proposée ici : la subdivision ne se fera plus dans les M directions de l'espace articulatoire, mais dans une seule direction à la fois (c'est-à-dire, le long de l'hyperplan passant par le centre et normal à la direction), la direction étant déterminée de façon à diviser au minimum dans les directions pour lesquelles l'interpolation polynomiale est fidèle à la relation Ar $\Rightarrow$ Ac.

Au moins deux façons de faire peuvent être envisagées : déterminer la direction qui maximise la « non-linéarité » et diviser dans cette direction, ou déterminer le demi-espace qui minimise la « non-linéarité », et diviser dans la direction normale à ce demi-espace. Aucune de ces méthodes n'est parfaite : intuitivement, la première tente de déterminer la direction qui a le plus de chance de devoir être subdivisée, et risque de subdiviser artificiellement des cubes, alors que la seconde, qui tente de déterminer la zone qui a le moins de chance de devoir être subdivisée, et ainsi de conserver des hypercuboïdes le plus gros possibles, risque de ne pas déterminer la direction qui minimise le nombre de cubes.

La détermination de la direction qui maximise l'erreur d'interpolation n'est pas un processus évident, surtout si l'on utilise une approximation polynomiale optimale. En effet, le principe même de la méthode d'optimisation est de répartir l'erreur le plus uniformément possible, et par conséquent de faire en sorte que les irrégularités soient réparties suivant toutes les directions. Cet inconvénient peut être éliminé en mesurant l'erreur d'interpolation, non par rapport au

polynôme d'interpolation optimal, mais par rapport au développement de Taylor au centre. En effet, le développement de Taylor, au centre de l'hypercuboïde, permet de réellement mesurer de façon équitable dans chaque direction, l'écart par rapport à la condition idéale. En pratique, la détermination de la direction se fait, pour chaque direction, en parcourant l'ensemble des points de test (qui ne sont pas forcément les mêmes que ceux ayant permis de calculer l'interpolation optimale), en calculant l'erreur par rapport à l'interpolation, et en calculant la moyenne de cette erreur, pondérée par la distance de la projection du point sur l'axe considéré au centre, divisée par la distance du point au centre.

En d'autres termes :

$$\forall i, 1 \leq i \leq M, \epsilon(i) = \frac{\sum_{j=1}^{m} \frac{|X_{j,i}|}{\|X_j\|} \times d(f(X_j), P_T(X_j))}{\sum_{j=1}^{m} \frac{|X_{j,i}|}{\|X_j\|}},$$

et la direction maximisant l'erreur est alors le $i \in \{1...M\}$ qui maximise $\epsilon(i)$.

La détermination du « demi-espace » (ou demi-hypercuboïde) minimisant l'erreur d'interpolation se fait de façon un peu similaire. Comme précédemment, l'erreur d'interpolation se calcule par rapport au développement de Taylor au centre ; en pratique, la détermination du demi-espace le plus régulier se fait en considérant l'ensemble des points de tests, en calculant l'erreur d'interpolation, puis la moyenne de cette erreur, pondérée par l'inverse de la distance du point au centre.

En d'autres termes :

$$\forall i, 1 \leq i \leq 2 \times M, \widetilde{\epsilon(i)} = \frac{\sum_{X_j \in E_i} \frac{d(f(X_j), P_T(X_j))}{|X_j|}}{\sum_{X_j \in E_i} \frac{1}{|X_j|}},$$

où E_i désigne le demi-espace d'indice i, et la direction minimisant l'erreur est alors celle du demi-espace E_i, $i \in \{1...2 \times M\}$ qui minimise $\widetilde{\epsilon(i)}$. On remarquera qu'il existe deux demi-espaces par direction de l'espace articulatoire.

2.3.4.3 Subdivision multiple

Une dernière idée qui pourrait permettre d'améliorer encore les choses est de ne pas se contenter de diviser par deux les dimensions, mais de les diviser par un nombre entier quelconque, en regroupant si possible les zones pour lesquelles la linéarité est déjà assurée. Cette granularité plus fine doit permettre d'obtenir un codebook encore plus compact, mais risque de ralentir encore davantage la construction. Nous avons implémenté cette possibilité, mais guère eu le temps de tester son efficacité. Les tests préliminaires indiquent que les gains de place ne sont pas extraordinaires.

2.4 Réalisation

Dans cette partie, nous allons décrire quelques éléments techniques de la mise en œuvre informatique de la méthode. Nous présentons ici la façon dont est calculée l'image acoustique par le synthétiseur articulatoire de Maeda, une optimisation importante de la méthode qui permet de

gagner un temps considérable dans la construction de codebooks, et enfin une étude de l'influence du choix des ensembles de points utilisés pour le calcul des coefficients polynomiaux sur la précision acoustique du codebook.

2.4.1 Calcul du vecteur acoustique

Le calcul du vecteur acoustique est, à plusieurs titres, un élément clé de notre méthode d'inversion. Le système de synthèse acoustique utilisé dans le modèle de Maeda a déjà été décrit à la section 1.2 pages 4–11, nous ne reviendrons donc pas dessus. Nous ne discutons dans cette partie que de la façon dont est déterminé le *vecteur acoustique* proprement dit, en l'occurrence les différentes caractéristiques des formants (fréquence, largeur de bande, amplitude). Nous avons explicité précédemment comment la fonction de transfert du conduit vocal était calculée ; pour calculer les caractéristiques des formants, il s'agit de déterminer les caratéristiques des pics de cette fonction[4]. La fonction de transfert associe, à une fréquence donnée, l'amplitude du signal. La détermination des pics de la fonction peut ainsi se faire en discrétisant la fonction et en déterminant les maxima locaux, par exemple par dichotomie (méthode du « peak-picking »).

En pratique, de façon à déterminer des formants aussi proches que possible de ceux que l'on déterminerait sur un signal réel, nous préférerons évaluer les formants non comme les pics du spectre de la fonction de transfert, mais en les estimant par une LPC effectuée sur le spectre discrétisé. La LPC nous permet de déterminer les caractéristiques des formants de façon relativement précise, d'éliminer les pics qui ne seraient pas décelables dans le signal acoustique réel, et de séparer les formants très proches (qui parfois ne donnent lieu qu'à un seul pic). Il est cependant nécessaire d'évaluer les erreurs liées à la discrétisation du spectre et à la LPC elle-même, en comparant les fréquences déterminées par cette méthode par rapport à la méthode de peak-picking, qui détermine la fréquence exacte des pics (à 0,01 Hz près).

Nous avons ainsi calculé l'erreur commise liée à l'utilisation de la LPC, en fonction du nombre de points utilisés pour la discrétisation du spectre, et de l'ordre de la LPC. Nous calculons un spectre discret de l'image acoustique entre 0 et 5000Hz, avec un échantillonnage régulier.

La figure 2.4 donne un exemple de spectre d'image acoustique, ainsi que du spectre LPC correspondant (pour une LPC d'ordre 16 effectuée sur le spectre discret, comportant 500 points). On peut constater que sur cette figure les quatre premiers formants sont parfaitement capturés.

Les figures 2.5a, 2.5b, 2.5c montrent les erreurs RMS obtenues pour respectivement F1, F2 et F3 en faisant varier l'ordre de la LPC entre 8 et 20, et le nombre de points utilisés pour discrétiser le spectre entre 50 et 3000.

On constate que les meilleurs résultats sont obtenus avec une LPC d'ordre 15 et autour de 500 points, avec des erreurs RMS respectivement de l'ordre de 2Hz, 1Hz et 3Hz. Pour calculer ces scores, les erreurs « grossières » (i.e. formants dédoublés ou confondus) n'ont pas été prises en compte. En d'autres termes, on n'a mesuré que la similitude entre le calcul de formants par peak-picking et par LPC pour les cas où les deux méthodes détectent des formants similaires. L'un des résultats les plus intéressants de l'étude est que l'on peut observer que le nombre de points utilisés pour le calcul du spectre n'a plus guère d'influence au-delà de 300 points, ce qui signifie qu'en définitive la méthode par LPC n'a pas une complexité trop importante, puisque le

[4] Il est bon de rappeler que pour calculer un signal acoustique réel, la fonction de transfert doit être convoluée par une source sonore, glottique dans le cas des voyelles ; par conséquent, sans information sur la forme de la source glottique, seules les fréquences des formants peuvent être déterminées de façon fidèle, les amplitudes et – dans une moindre mesure – les largeurs de bande étant modifiées par la source glottique. Lors de l'inversion d'un signal acoustique réel, nous ne prendrons ainsi généralement en compte que les informations sur la fréquence des formants.

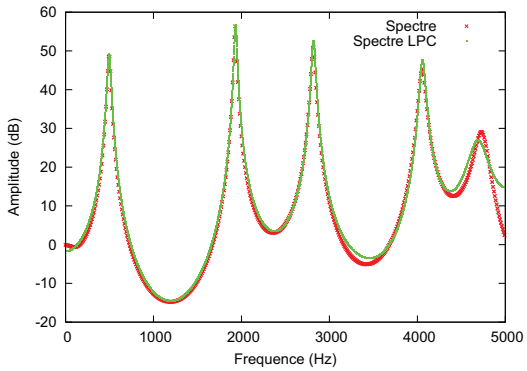


FIG. 2.4: *Spectre de l'image acoustique d'un vecteur articulatoire, et spectre LPC correspondant. La fréquence est en Hertz, l'amplitude en dB.*

nombre de points nécessaire pour la discrétisation donnant un résultat optimal est assez faible. En l'occurrence, elle est même légèrement plus rapide que notre méthode de peak-picking, qui nécessite environ 500 points pour déterminer les formants à 0,01 Hz près.

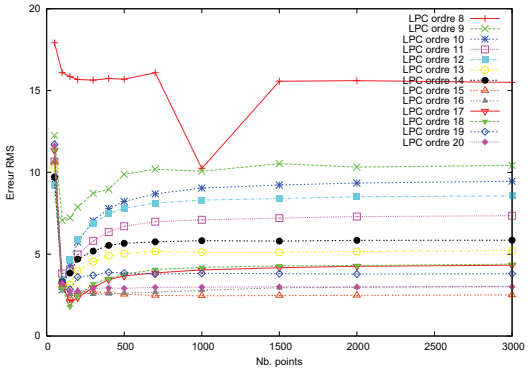
2.4.2 Cache de formants

Le calcul des fréquences formantiques reste cependant très coûteux en temps, il est donc souhaitable de limiter au maximum le nombre d'appels au synthétiseur, et en particulier faire en sorte de ne pas avoir à recalculer plusieurs fois l'image d'un point. Nous avons ainsi mis en œuvre un cache (fondé sur une organisation arborescente des vecteurs, et l'utilisation d'une table de hachage) pour accélérer ce calcul : si l'image d'un vecteur articulatoire a été calculée précédemment, on trouvera très rapidement son image dans le cache ; sinon on calcule son image à l'aide du synthétiseur.

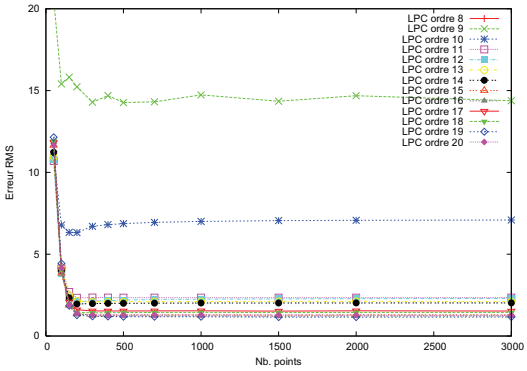
Les gains des temps sont conséquents : en effet, beaucoup de points appartenant à plusieurs hypercuboïdes sont calculés un grand nombre de fois ; en particulier, chaque sommet d'un hypercuboïde appartient également à $2^M - 1$ autres hypercuboïdes, il y a ainsi un gain potentiel en temps très important. De la même façon, les milieux des segments reliant deux sommets de l'hypercuboïde appartiennent également – pour la majorité d'entre eux – à plusieurs hypercuboïdes.

Il est facile de voir que chaque sommet d'un hypercuboïde appartient à (au maximum) 2^M hypercuboïdes, et chaque hypercuboïde contenant 2^M sommets, cela fait que dans le cas idéal on peut obtenir avec un cache un coût asymptotique de $\frac{2^M}{2^M} = 1$ pour le calcul des images des sommets d'un hypercuboïde donné.

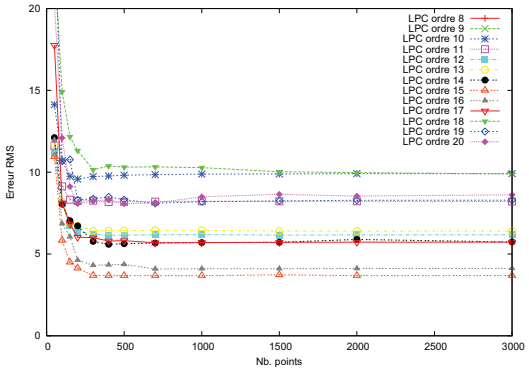
Le coût asymptotique optimal du calcul des images des milieux des segments reliant deux sommets de l'hypercuboïde est un peu plus délicat à obtenir ; en décrivant les coordonnées – dans une base adéquate – des points remarquables (sommets, milieux) de l'hypercuboïde grâce à un code ternaire $E = \{-1; 0; 1\}$, on voit facilement que E^M est exactement égal à l'ensemble des points remarquables de l'hypercuboïde. On note $F = \{-1; 1\}$. L'ensemble des sommets est l'ensemble F^M, i.e. l'ensemble des vecteurs de E^M ne contenant pas de coordonnée égale à 0, et l'ensemble des milieux est alors l'ensemble des vecteurs de E^M contenant au moins une coordonnée à 0. On voit alors trivialement que le nombre de sommets est exactement $|F^M| = 2^M$, et le nombre de milieux est ainsi $|E^M| - |F^M| = 3^M - 2^M$. Une autre façon de le voir est de le considérer comme l'union pour k variant de 1 à M des vecteurs ayant k coordonnées égales à 0.



(a) Erreur RMS (Hz) pour F1.



(b) Erreur RMS (Hz) pour F2.



(c) Erreur RMS (Hz) pour F3.

FIG. 2.5: *Mesure de la différence entre formants calculés par « peak-picking » et par LPC. Les images de 10000 vecteurs articulatoires choisis aléatoirement sont calculées par chacune des deux méthodes (en faisant varier l'ordre et le pas de discrétisation du spectre pour la LPC), et la différence entre les deux est quantifiée en calculant la moyenne quadratique des différences (erreur RMS).*

On voit facilement qu'il y en a exactement $C_k^M * 2^{M-k}$. Le nombre de milieux est ainsi :

$$\sum_{k=1}^{M} C_k^M * 2^{M-k} = \sum_{k=0}^{M} C_k^M * 2^{M-k} - 2^M = 3^M - 2^M.$$

On peut évaluer facilement le nombre d'hypercubes auquel appartient un milieu en fonction du nombre de ses composantes non nulles p : il s'agit de 2^p. Le poids de chaque type de milieux est ainsi $C_k^M * \frac{2^{M-k}}{2^p} = C_k^M$ (car $p = M-k$). On peut ainsi déterminer le poids des points milieux ; il s'agit de :

$$\sum_{k=1}^{M} C_k^M = 2^M - 1.$$

Au total, dans un hypercuboïde le poids des points remarquables (sommets et milieux) passe ainsi de 3^M sans cache à 2^M avec cache, ce qui est un gain important. Asymptotiquement, pour le modèle articulatoire de Maeda ($M = 7$), on peut ainsi diviser le temps de calcul par 17 pour un niveau de subdivision donné. Bien entendu, le gain de vitesse réel est plus faible. Le gain de vitesse réel mesuré est de l'ordre de 10 pour un codebook avec les paramètres que nous utilisons habituellement.

2.4.3 Génération des points de test

Pour les tests de régularité, l'utilisation de différents ensembles de points peut être envisagée : le centre, les points utilisés pour le calcul de la matrice jacobienne, les sommets de l'hypercuboïde, les milieux des segments reliant deux sommets de l'hypercuboïde, et enfin des points choisis de façon pseudo-aléatoire.

2.4.3.1 Centre de l'hypercuboïde et matrice jacobienne

Le centre de l'hypercuboïde a l'avantage d'être le seul point à égale distance (euclidienne) de tous les sommets. Il occupe un rôle central dans notre méthode, d'autant plus que dans notre cas, nous l'utilisons systématiquement pour générer les solutions de l'inversion (contrairement à e.g. Ouni ou Sorokin). Par définition, la matrice jacobienne est la matrice des dérivées partielles de la fonction en un point donné (ici le centre de l'hypercuboïde), c'est-à-dire la matrice $N \times M$ suivante :

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

et dans un petit voisinage de x_0, la relation suivante est vérifiée :

$$f(x) \approx f(x_0) + \Delta f_{x_0} \times (x - x_0)$$

La matrice jacobienne est rarement calculée de façon exacte, mais une valeur approchée de celle-ci, que l'on notera J, est généralement évaluée numériquement en évaluant la fonction dans un petit voisinage du point considéré. Mais du fait de la façon dont est évaluée l'image acoustique considérée (discrétisation du spectre suivie d'une LPC), la résolution acoustique locale est assez mauvaise, donc la méthode traditionnelle de calcul de la matrice jacobienne n'est pas la mieux adaptée. De plus, comme nous l'avons évoqué précédemment, l'interpolation linéaire en utilisant la matrice jacobienne n'est pas celle qui donne les meilleurs résultats en général, mais l'une des façons de compenser cette erreur est « d'écarter » les points utilisés pour le calcul du centre de la structure.

Différentes méthodes ont ainsi été testées pour tenter de déterminer le calcul qui nous permette d'obtenir la meilleure précision acoustique lors de la resynthèse. Deux types de calculs ont ainsi été expérimentés :

1. Calcul de la matrice jacobienne à une distance constante du centre

Pour $j \in \{1, \ldots, M\}$, on note δ_j le vecteur de $\mathbb{R}^M$ dont toutes les composantes sont nulles, sauf la j^e qui est égale à 1. Un certain ϵ_c ayant été choisi a priori, on calcule alors la matrice jacobienne de la façon suivante :

$$\forall i \in \{1, \ldots, M\}, \forall j \in \{1, \ldots, N\}, J_{i,j} = \frac{f_i(x_0 + \epsilon_c.\delta_j) - f_i(x_0 - \epsilon_c.\delta_j)}{2\epsilon_c}$$

Ce calcul correspond à l'une des évaluations numériques classiques pour le calcul de la matrice jacobienne. Mais dans notre cas, il semble probable que l'ϵ_c qui nous permette d'obtenir les meilleurs résultats sera relativement grand, alors que l'on prend habituellement une valeur très petite. Le ϵ_c optimal peut être déterminé expérimentalement en construisant différents codebooks avec différentes valeurs pour ce paramètre, et en évaluant le nombre d'hypercuboïdes qui ont été générés pour représenter une même zone de l'espace, ainsi que le volume total du codebook (c'est-à-dire la somme des volumes des hypercuboïdes), qui donne une indication de la couverture de l'espace articulatoire dans le codebook. La précision acoustique de la resynthèse est également prise en compte.

Le tableau 2.1 donne une idée des résultats obtenus.

ϵ_c	# HC	Volume	Δ F1	σ F1	Δ F2	σ F2	Δ F3	σ F3
0.001	8007	825.1	11.0	9.0	13.5	11.2	16.5	13.2
0.002	7070	877.3	9.4	7.4	13.0	10.4	19.1	15.2
0.005	5924	893.8	9.3	7.2	13.2	10.5	19.8	15.6
0.01	6093	896.9	9.2	7.1	13.2	10.5	19.4	15.2
0.02	6280	897.2	9.2	7.0	13.1	10.4	19.2	15.1
0.05	6433	903.3	9.2	7.1	13.1	10.4	19.4	15.3
0.10	6880	918.0	9.2	7.1	13.0	10.4	19.2	15.2
0.20	7512	930.0	9.0	6.9	12.9	10.3	19.2	15.2
0.30	5534	903.7	9.0	6.8	13.2	10.5	19.6	15.5
0.40	3359	883.0	9.5	7.1	13.9	10.9	20.2	15.8
0.50	2746	858.4	10.8	8.1	14.3	11.1	20.1	15.7
0.60	1897	671.0	7.0	4.9	11.8	9.0	20.6	16.2
0.70	1375	624.4	7.6	5.3	12.2	9.2	20.9	16.4
0.80	1002	551.2	7.5	5.2	12.4	9.3	21.3	16.7
0.90	794	486.9	8.1	5.6	13.3	9.8	21.4	16.6
1.00	394	464.6	10.3	6.8	16.6	12.1	23.3	17.5

TAB. 2.1: *Nombre de morceaux, volume, et précision acoustique pour différentes valeurs de ϵ_c. La précision acoustique est estimée en resynthétisant aléatoirement 100000 points du codebook et en comparant l'image estimée à partir du jacobien à celle obtenue à partir du synthétiseur articulatoire. Pour les trois premières fréquences formantiques, la moyenne et l'écart-type de l'erreur sont présentés.*

2. Calcul de la matrice jacobienne à une distance proportionnelle au rayon de l'hypercuboïde

Une autre façon simple de calculer la matrice jacobienne consiste à évaluer localement la fonction non pas à une distance constante du centre de l'hypercuboïde, mais à une distance ϵ_m proportionnelle à son rayon. Ainsi, on peut imaginer que la matrice calculée permettra d'épouser plus fidèlement la fonction f dans tout l'hypercuboïde, plutôt que simplement dans un voisinage du centre ne changeant pas de taille. Il est cependant probable que pour des valeurs importantes de ϵ_m la fiabilité du test de régularité en souffre, et que la précision acoustique résultante s'en ressente. En reprenant les mêmes notations que précédemment, la formule de calcul de la matrice jacobienne devient la suivante :

$$\forall i \in \{1, \ldots, M\}, \forall j \in \{1, \ldots, N\}, J_{i,j} = \frac{f_i(x_0 + \epsilon_m.r_j.\delta_j) - f_i(x_0 - \epsilon_m.r_j.\delta_j)}{2\epsilon_m.r_j}$$

Le ϵ_m optimal peut comme précédemment être déterminé expérimentalement en construisant différents codebooks avec différentes valeurs pour ce paramètre, avec les mêmes modalités que précédemment.

Le tableau 2.2 présente les résultats obtenus.

L'interprétation de ces deux tableaux n'est pas simple. En effet, plusieurs critères doivent être pris en compte :

1. D'une part, le « volume » de l'espace articulatoire contenu dans le codebook. Celui-ci varie très peu, sauf pour les valeurs élevées de ϵ_c, pour lesquelles le volume s'écroule. On constate également (ce qui est plus surprenant) que le volume *augmente* pour les valeurs élevées de ϵ_m. Un volume très inférieur à la valeur optimale est une indication que le paramétrage est particulièrement inefficace : on a perdu une grande partie de l'espace articulatoire défini, principalement les zones situées près des frontières. La précision acoustique de ces codebooks n'est alors plus comparable à celle des autres : si on supprime les cubes les plus difficiles à représenter, la tâche se trouve nécessairement simplifiée. Pour que les codebooks soient comparables entre eux, le premier critère à respecter est qu'ils représentent le même espace articulatoire.
2. Le deuxième critère est la précision acoustique du codebook. Nous cherchons à obtenir une resynthèse qui soit la plus fidèle possible ; cette précision acoustique est représentée sous la forme de 6 valeurs : l'erreur RMS moyenne, ainsi que l'écart-type de cette erreur, sur les fréquences des trois premiers formants.
3. Enfin, et c'est probablement ce qui nous intéresse le plus en général, la concision du codebook. Celle-ci est matérialisée par le nombre d'hypercuboïdes contenus dans le codebook. Plus ce nombre est petit, et plus le codebook est concis.

Une façon plus visuelle de présenter les choses est de représenter les courbes d'erreur en fonction des ϵ, mais la mesure de l'erreur elle-même ne permet pas de rendre compte de la concision du codebook. Une autre mesure plus significative est l'erreur pondérée par une mesure de la densité du codebook : globalement, la précision sur chaque composante du vecteur acoustique augmente d'un facteur 2 lorsque l'on augmente la résolution articulatoire d'un facteur 2 pour chaque dimension de l'espace articulatoire, c'est-à-dire qu'un gain d'un facteur 2 pour la précision acoustique nécessite de multiplier le nombre d'hypercuboïdes par 2^M, soit 128. En d'autres termes, pour tenir compte de la densité du codebook, une mesure homogène de l'erreur est la suivante :

$$e_j = \Delta F_j \times \sqrt[7]{\frac{\#Hc}{v}}$$

ϵ_m	# HC	Volume	Δ F1	$\sigma F1$	Δ F2	σ F2	Δ F3	σ F3
0.001	7930	817.1	10.4	8.4	13.6	11.2	17.2	13.6
0.002	7253	851.6	9.4	7.3	13.5	10.7	19.9	15.7
0.005	5796	892.6	9.3	7.1	13.2	10.5	19.7	15.6
0.01	5889	896.0	9.2	7.0	13.3	10.6	19.5	15.4
0.02	6251	897.1	9.2	7.1	13.2	10.5	19.4	15.3
0.05	6301	898.8	9.3	7.1	13.1	10.5	19.5	15.4
0.10	6162	901.2	9.3	7.1	13.1	10.4	19.5	15.4
0.20	6778	911.8	9.2	7.1	13.0	10.3	19.0	15.0
0.25	6828	918.3	9.1	7.0	13.1	10.5	18.9	14.9
0.30	6819	918.6	9.1	6.9	13.1	10.4	18.8	14.9
0.35	6536	919.1	9.1	6.9	12.9	10.2	19.0	15.0
0.40	6081	920.4	9.1	6.9	13.2	10.5	19.1	15.1
0.45	5603	946.4	9.0	6.8	13.5	10.6	19.8	15.5
0.50	5559	950.3	9.1	6.8	13.7	10.7	20.1	15.8
0.55	4998	954.0	9.2	7.0	14.0	10.9	20.5	16.0
0.60	4799	957.6	9.5	7.2	14.1	10.9	20.8	16.2
0.70	4647	963.4	10.8	8.3	15.6	12.0	22.0	16.8
0.80	4302	967.9	10.8	8.1	20.1	15.5	25.2	18.7
0.90	4126	971.6	8.3	5.9	21.2	16.2	27.5	20.2
1.00	3743	975.5	8.6	6.0	21.9	16.5	29.7	21.6
1.20	3150	981.3	8.5	5.8	21.5	15.9	33.1	23.3
1.50	3946	978.7	8.5	5.8	19.2	13.9	33.2	22.0

TAB. 2.2: *Nombre de morceaux, volume, et précision acoustique pour différentes valeurs de ϵ_m. La précision acoustique est estimée en resynthétisant aléatoirement 100000 points du codebook et en comparant l'image estimée à partir du jacobien à celle obtenue à partir du synthétiseur articulatoire. Pour les trois premières fréquences formantiques, la moyenne et l'écart-type de l'erreur sont présentés.*

La figure 2.6 représente e_1, e_2 et e_3 en fonction de respectivement ϵ_c et ϵ_m. Les valeurs pour lesquelles le codebook résultant n'a pas un volume suffisant ont été supprimées.

La mesure pondérée de l'erreur nous permet de nous rendre compte de la précision propre due au paramétrage. Deux tendances se dégagent de ce graphique. Premièrement, on constate que, comme on pouvait s'y attendre, le calcul de la matrice jacobienne à une distance proportionnelle au rayon donne des résultats légèrement meilleurs que le calcul à une distance constante, et ce quelle que soit la situation. Mais la différence n'est pas vraiment spectaculaire. Par ailleurs, on constate que l'erreur pondérée atteint un minimum dans chaque cas : pour le calcul à distance constante, ce minimum est atteint autour de 0.3, alors que pour le paramétrage à distance proportionnelle au rayon, le minimum est atteint autour d'un facteur de 0.45. À noter qu'à partir d'un facteur multiplicatif de 0.5, la matrice jacobienne est en fait calculée *à l'extérieur* de l'hypercuboïde.

À titre de comparaison, Ouni calculait la matrice jacobienne avec $\epsilon_m = 0.2$. En utilisant un $\epsilon_m = 0.45$, on a un gain de densité d'environ 26%, et une erreur $e = \bar{e_j}$ sensiblement équivalente (en réalité, en très légère diminution, de 0,8%). Le gain en place dû à l'optimisation du calcul de la matrice jacobienne est donc appréciable.

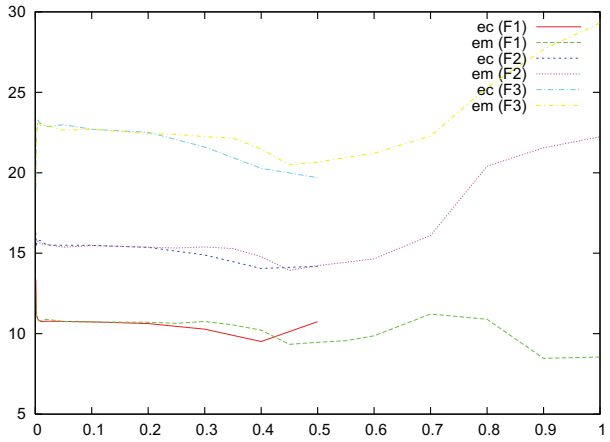


FIG. 2.6: *Représentation graphique de l'erreur acoustique pondérée pour les trois premiers formants (en Hertz) en fonction des valeurs respectives de ϵ_c et ϵ_m.*

2.5 Évaluation expérimentale

2.5.1 Resynthèse d'un vecteur articulatoire par interpolation

La resynthèse, à l'aide d'un codebook, de l'image acoustique d'un vecteur articulatoire choisi au hasard est conceptuellement simple. Il s'agit d'identifier l'hypercuboïde auquel appartient ce vecteur, et à partir de l'interpolation polynomiale locale trouver l'image. Calculer l'image d'un polynôme en un point est très simple et prend un temps relativement court ; la principale difficulté pour une resynthèse efficace est donc de déterminer efficacement l'hypercuboïde qui convient. L'algorithme naïf consistant à parcourir l'ensemble du codebook en testant successivement chacun des hypercuboïdes est bien entendu à exclure pour des codebooks contenant plusieurs milliers d'hypercuboïdes. L'algorithme simple et efficace que nous utilisons pour trouver rapidement le bon hypercuboïde est très classique : nous discrétisons l'espace articulatoire en petits hypercubes, et nous mettons dans chacun de ces hypercubes une référence vers tous les hypercuboïdes qui ont une intersection non vide avec cet hypercube (un hypercuboïde assez grand peut ainsi être référencé dans plusieurs hypercubes).

L'avantage de notre algorithme est de ne pas du tout dépendre de la structure des données du codebook, il est aussi bien adapté aux hypercuboïdes que, par exemple, à des parallélotopes quelconques ou des boules. Grâce à une organisation arborescente de ces hypercubes, on peut déterminer immédiatement dans quel hypercube il est nécessaire de chercher la bonne structure permettant de synthétiser l'image du vecteur articulatoire. Il suffit ensuite de parcourir la liste des hypercuboïdes candidats, et de déterminer celui qui convient. Plus la discrétisation de l'espace articulatoire pour la construction des hypercubes sera fine, et plus les listes à parcourir seront courtes, mais en contrepartie l'arbre des hypercubes occupera une place plus importante et mettra plus de temps à être construit. Avec une discrétisation peu précise de l'espace articulatoire (chaque cube fait une unité de volume), on arrive déjà à une détermination très rapide du bon hypercuboïde. La resynthèse est alors environ 1000 fois plus rapide qu'en utilisant le synthétiseur articulatoire.

Cette resynthèse de vecteurs acoustiques par interpolation comparée à la synthèse « réelle » à l'aide du modèle articulatoire nous permet de définir une mesure de la précision acoustique d'un codebook. Dans notre étude, nous calculons différentes erreurs de resynthèse : l'erreur absolue

moyenne, l'erreur RMS, l'écart-type de l'erreur, ainsi que le maximum de l'erreur.

2.5.2 Valeurs optimales pour le calcul de la matrice jacobienne

Comme nous avons pu le constater, la méthode de calcul de la matrice jacobienne influence considérablement la taille, mais aussi la précision acoustique du codebook résultant.

Nous avons déjà présenté les résultats et la paramétrisation optimale pour le calcul de la matrice jacobienne à la section 2.4.3.1. La paramétrisation qui donne les meilleurs résultats est le calcul de la matrice jacobienne à une distance proportionnelle au rayon, et, de façon plus étonnante, à une distance importante du centre $\epsilon_m = 0.45$ (donc en réalité, pratiquement sur les faces de l'hypercuboïde, chacune étant à une distance de $\frac{r_j}{2}$ du centre, i.e. $\epsilon_m = 0.5$).

2.5.3 Seuils de subdivision et précision acoustique

L'un des enjeux majeurs de la construction du codebook est d'aboutir à une « base de données » représentant la relation Ar $\Rightarrow$ Ac de façon concise (pour avoir une inversion rapide) et fidèle (pour avoir une inversion précise). Ces deux objectifs étant pratiquement contradictoires, il est nécessaire d'établir un compromis entre la précision acoustique, le nombre de niveaux de subdivision maximal dans l'espace articulatoire, et enfin la taille du codebook résultant. Le temps de calcul nécessaire pour construire le codebook peut également entrer en ligne de compte, mais ce n'est pas essentiel puisque l'on n'a besoin de ne le faire en principe qu'une fois par locuteur.

Le niveau de subdivision maximal doit être limité : en effet, à la frontière de l'espace articulatoire synthétisable, la subdivision pourrait se poursuivre à l'infini, alors que ce sont a priori des zones moins fondamentales pour l'inversion, et probablement plus turbulentes, donc plus difficilement linéarisables. En l'espèce, cela revient à fixer un seuil de discrétisation de l'espace articulatoire. Il s'agit donc de déterminer un seuil de subdivision limite qui nous permette d'obtenir un codebook relativement compact, tout en conservant une bonne résolution de l'espace articulatoire. Un compromis est déterminé de façon expérimentale : sur une petite partie de l'espace articulatoire, on génère plusieurs codebooks avec différentes résolutions articulatoires et seuils acoustiques, et on mesure le volume défini, le nombre d'hypercuboïdes et la précision acoustique moyenne de la resynthèse.

Le tableau 2.3 présente les résultats obtenus sur un $1/128^{\text{e}}$ de l'espace articulatoire. On impose différentes valeurs pour le volume d'un hypercuboïde minimal, et différents seuils acoustiques, et on détermine le nombre d'hypercuboïdes, le volume total de l'espace contenu dans le codebook, et les erreurs RMS (calculées à partir de l'évaluation de 100000 points par codebook) sur les trois premières fréquences formantiques.

On constate que lorsque le volume minimal d'un hypercube diminue, l'espace articulatoire représenté au sein du codebook augmente. On constate également que plus la marge d'erreur admise est importante, et plus on arrive à représenter un espace articulatoire important en un petit nombre d'hypercuboïdes. On constate que l'erreur mesurée (en Hertz) est conforme aux attentes : l'erreur pour un seuil de 0,5 Bark est environ la moitié de celle pour 1 Bark, et celle pour 0,3 Bark est environ un tiers de celle pour 1 Bark. De façon assez intéressante, on constate que l'erreur pour F1 est plus importante que celle pour F2, ce qui indique que F1 a, dans la zone considérée, un comportement moins linéaire que F2. Enfin, on constate que le nombre d'hypercuboïdes nécessaire change assez peu entre 1 et 0,5 Bark, alors qu'il augmente de façon considérable pour 0,3 Bark, ce qui semble indiquer que 0,5 Bark est un bon compromis, puisque la précision acoustique est sensiblement meilleure pour une pénalité assez faible au niveau de la concision du codebook.

Vol. minimal	Seuil acoustique	Nb. Hc.	Vol. total	$\Delta F1$	$\Delta F2$	$\Delta F3$
1	1.0	118	673.826660	13.382	8.573	22.251
	0.5	237	630.043945	6.592	4.670	13.606
	0.3	217	378.026367	4.589	3.380	8.849
0.5	1.0	157	694.650146	13.158	8.540	22.142
	0.5	330	679.699951	6.449	4.490	13.659
	0.3	384	467.727539	4.501	3.385	9.055
0.25	1.0	182	701.324341	13.133	8.387	21.863
	0.5	351	685.039307	6.506	4.570	13.577
	0.3	456	487.750122	4.540	3.443	9.135
0.1	1.0	772	780.079834	12.186	7.904	21.202
	0.5	919	761.125122	6.142	4.263	13.192
	0.3	925	549.419678	4.354	3.301	9.136
0.05	1.0	1423	823.609009	11.694	7.580	20.927
	0.5	1659	810.661133	6.072	4.301	13.120
	0.3	3571	725.763428	4.193	2.997	8.723
0.025	1.0	2296	852.767517	11.705	7.760	20.787
	0.5	2749	846.696777	5.959	4.143	13.071
	0.3	5637	794.688843	4.108	2.897	8.584
0.01	1.0	3968	880.622253	11.200	7.320	20.275
	0.5	4728	879.733887	5.840	4.083	12.924
	0.3	8678	845.283325	4.015	2.889	8.639
0.005	1.0	5432	892.859802	11.186	7.103	20.163
	0.5	6290	892.741577	5.845	4.016	12.910
	0.3	12646	878.393921	3.960	2.814	8.480

TAB. 2.3: *Caractéristiques du codebook (nombre d'hypercuboïdes, volume total, précision acoustique) en fonction du seuil acoustique et du volume minimal d'un hypercuboïde.*

On constate enfin que le volume défini augmente progressivement vers la valeur limite (que l'on a déterminée empiriquement comme étant égal à 1062,0 pour la zone de l'espace articulatoire explorée) au fur et à mesure que le volume minimal d'un hypercuboïde diminue (voir figure 2.7), et que les écarts de volume total entre les différents seuils acoustiques diminuent petit-à-petit, ce qui indique que les zones non incluses dans le codebook ne sont pas des zones non régulières acoustiquement, mais uniquement des zones situées à la frontière de l'espace articulatoire (il suffit qu'un hypercuboïde de taille minimale comporte un sommet dans la zone interdite pour qu'il soit rejeté du codebook). On remarque que le volume de l'espace articulatoire non représenté est important : même avec un volume minimal de 0.005, le codebook ne représente encore que 84% du total. Avec un volume minimal de 0.0005, on trouverait que le volume n'est que de 90% du total, et le nombre d'hypercuboïdes 10 fois plus important.

Ce phénomène est essentiellement lié à la dimension importante de l'espace articulatoire : les volumes et les nombres d'hypercuboïdes à manipuler croissent de manière très rapide en dimension 7. Nous présenterons à la section 2.5.5 une façon de remédier à ce problème.

En définitive, on constate ici que le volume minimal a plus d'influence sur la concision du codebook résultant que le seuil acoustique[5], et que le nombre d'hypercuboïdes est à peu de chose près inversement proportionnel à ce volume minimal. Pour ne pas avoir un codebook de taille trop importante, on choisit généralement un volume minimal de 0.05, associé à un seuil acoustique de 0.5 Bark.

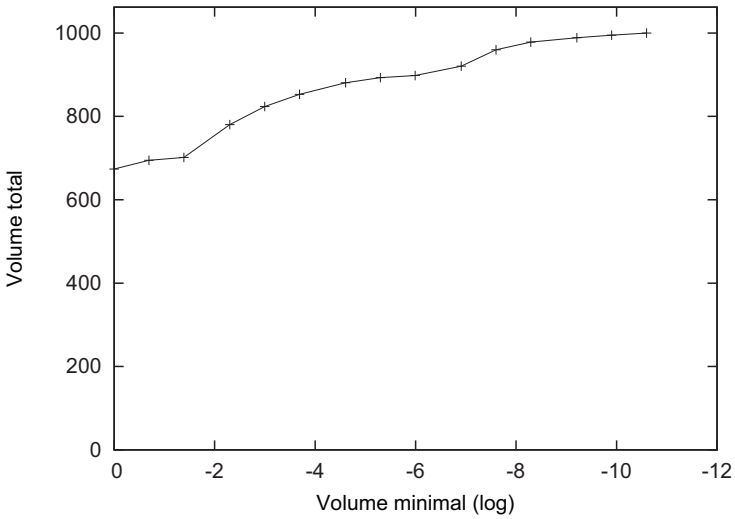


FIG. 2.7: *Volume total d'un codebook partiel en fonction du volume minimal d'un hypercuboïde (échelle logarithmique pour le volume minimal).*

2.5.4 Temps de construction

Le cache de formants décrit à la section 2.4.2 permet d'accélérer considérablement le temps de construction du codebook. En effet, le temps d'accès au cache est négligeable devant le calcul de l'image d'un vecteur articulatoire. Pour calculer le facteur d'accélération réel que cela apporte, il

[5] Cette remarque n'est valable que pour le domaine de valeurs exploré pour le seuil acoustique ; en réalité, si l'on impose des seuils acoustiques vraiment petits, le nombre d'hypercuboïdes nécessaires croît exponentiellement.

suffit de construire un même codebook en activant ou désactivant le cache. Pour un codebook dont on ne calcule que les sommets, le facteur d'accélération mesuré est d'environ 11.4 (soit 1140%), et pour un codebook dont on synthétise tous les points remarquables le facteur d'accélération est de 7.96. On constate également qu'avec la méthode avec cache, on arrive à calculer un codebook utilisant tous les points remarquables plus rapidement que la méthode sans cache n'utilisant que les sommets, alors que les nombres d'appels réels au synthétiseur devraient être au mieux asymptotiquement très semblables. Deux raisons peuvent être avancées pour expliquer ce résultat :

- Lors de l'exploration des zones non définies de l'espace articulatoire, le nombre d'appels au synthétiseur nécessaire pour déterminer si un hypercuboïde est défini est sensiblement le même dans les deux cas ; le nombre d'appels réels est donc nécessairement nettement plus réduit dans le cas de la méthode avec cache. Comme les zones non définies représentent la majeure partie de l'espace articulatoire, on peut estimer que la méthode avec cache gagne un temps considérable dans ces zones. En réalité, même si les zones non définies représentent la majeure partie de l'espace articulatoire, leur exploration prend dans tous les cas un temps relativement faible, car le calcul de leur image acoustique est nettement plus rapide que pour les vecteurs acoustiques définis. Sur le total du temps de calcul, l'exploration des zones non définies prend ainsi moins de 10% du temps total. Les gains sur ces zones n'expliquent donc qu'une partie du gain de temps.
- Une autre raison, plus intéressante, et que l'utilisation d'un plus grand nombre de points permet à l'heuristique de détermination de la direction de subdivision d'être plus efficace : le codebook est plus compact tout en représentant un volume (légèrement) plus large. Le nombre d'appels réels au synthétiseur est nettement plus faible dans la version avec cache.

2.5.5 Couverture de l'espace articulatoire

Il reste à vérifier que les codebooks créés grâce à cette méthode permettent de réaliser de bonnes couvertures de l'espace articulatoire « utile », c'est-à-dire celui où il existe des images acoustiques pertinentes.

La figure 2.8 illustre ce problème de couverture : il s'agit ici d'une coupe réelle de l'espace articulatoire de Maeda, représentant les zones synthétisables à partir du synthétiseur articulatoire (en pointillés) et les zones synthétisables à partir du codebook (rectangles, correspondant aux coupes des hypercuboïdes). La zone en gris clair correspond à la zone interdite (c'est-à-dire l'espace des vecteurs articulatoire n'ayant pas d'image acoustique définie). On constate sur cette figure que les zones synthétisables mais absentes du codebook (les zones en pointillés sur gris clair) représentent tout de même un volume non négligeable ; pour remédier à ce problème, il faudrait utiliser une granularité plus fine, mais les temps de calcul s'en ressentiraient, ainsi que l'espace occupé par le codebook. Par ailleurs, on peut également observer que la plus grande partie des subdivisions est due, non à une non-linéarité de la relation Ar $\Rightarrow$ Ac à proprement parler, mais au suivi de la frontière.

Ces zones non synthétisables à partir du codebook sont situées aux limites de l'espace articulatoire définies, elles sont donc vraisemblablement peu importantes pour l'inversion. Leur volume est cependant conséquent : avec les réglages que l'on utilise habituellement, environ 23% de l'espace articulatoire utile est absent du codebook. Avec les réglages utilisés habituellement par Ouni, c'était 37% de l'espace articulatoire utile qui était manquant. Les régions concernées ont beau être proches de la frontière, et par conséquent vraisemblablement moins pertinentes, il était tout de même important de concevoir une méthode qui permette d'améliorer la couverture sans trop dégrader la précision acoustique des codebooks.

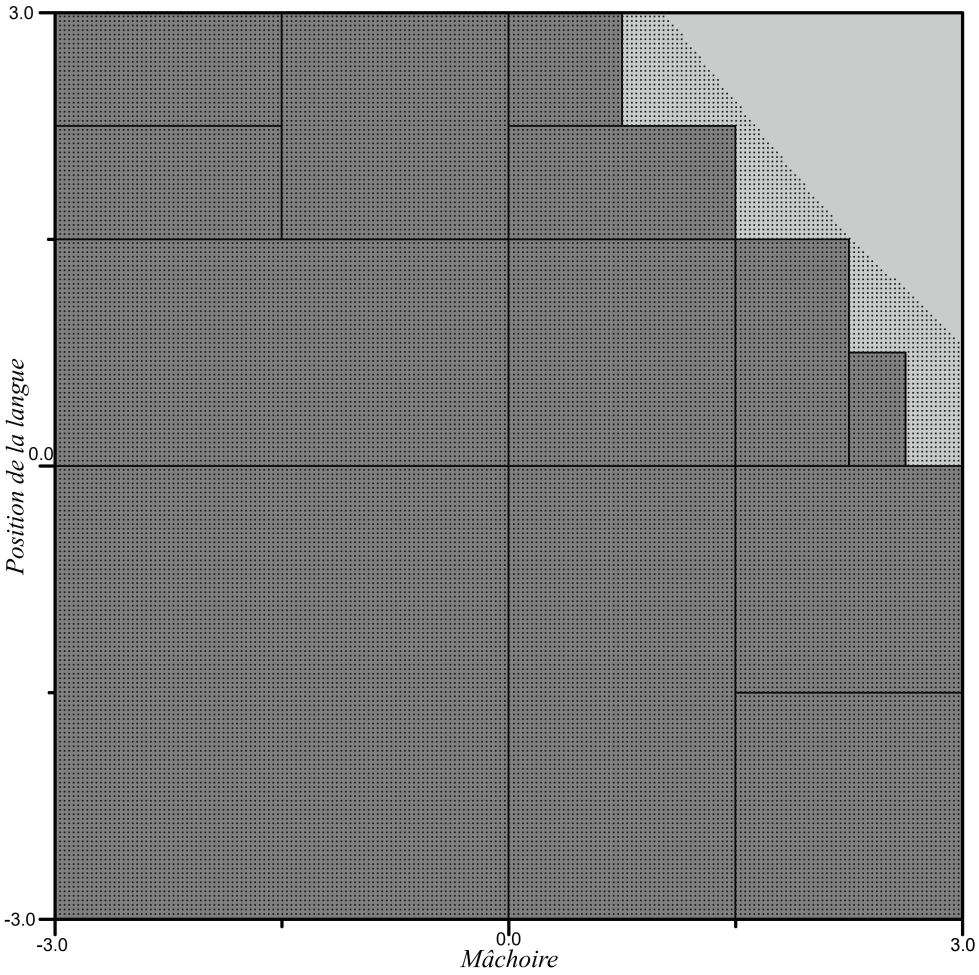
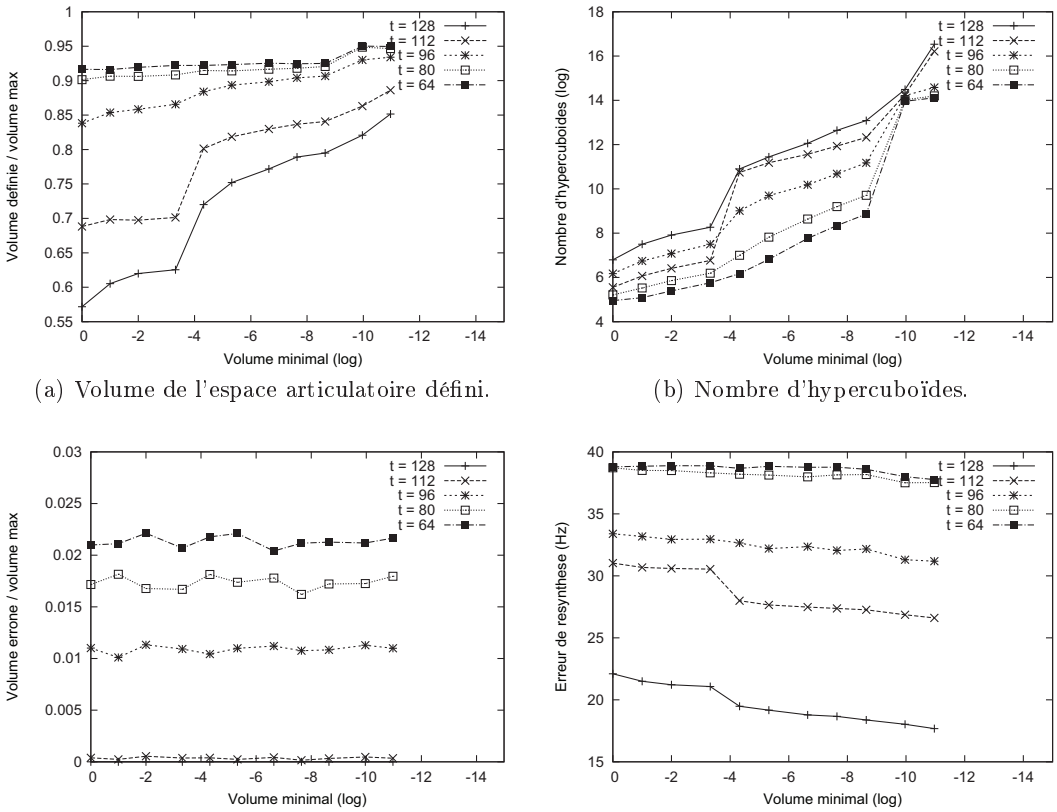


FIG. 2.8: *Coupe selon deux composantes (mâchoire et position de la langue) de l'espace articulatoire de Maeda. Les rectangles en gris foncé correspondent aux coupes des hypercuboïdes, la zone pointillée à l'espace synthétisable (chaque point représente un vecteur articulatoire ayant une image acoustique). Les deux paramètres varient entre -3 et +3.*

Lorsque l'on étudie en détail le comportement à la frontière, on constate qu'une grande partie des hypercuboïdes rejettés sont pratiquement entièrement définis : il n'y a souvent que quelques sommets dont l'image acoustique n'est pas définie ; il est par conséquent dommage de les rejeter. Nous avons donc intégré une modification dans l'algorithme permettant d'accepter les hypercuboïdes ayant « suffisamment » de sommets définis, c'est-à-dire un nombre de sommets définis supérieur à un seuil fixé au préalable. On impose également que la matrice jacobienne soit entièrement calculable ; le paramètre ϵ_m optimal trouvé précédemment, calculant la matrice jacobienne très près des bords de l'hypercuboïde, n'est plus ici le meilleur choix. Pour cette expérience, nous avons fixé $\epsilon_m = 0.3$.

En dehors de cette condition, les modalités de l'expérience sont les mêmes qu'à la section 2.5.3, en imposant un seuil de précision acoustique de 1 Bark. Différents codebooks ont été construits, en faisant varier le seuil du nombre minimal de sommets définis, désigné par t dans la suite, et le volume minimal de l'hypercuboïde. Différentes caractéristiques des codebooks résultants ont été calculées : le volume défini (c'est-à-dire le volume de l'espace articulatoire contenu dans le



(a) Volume de l'espace articulatoire défini.

(b) Nombre d'hypercuboïdes.

(c) Volume de l'espace articulatoire erroné.

(d) Erreur de resynthèse.

FIG. 2.9: *Caractéristiques des codebooks en fonction du volume minimal pour différentes valeurs du seuil t.*

codebook pour lequel il existe une image acoustique), le volume erroné (c'est-à-dire le volume de l'espace articulatoire contenu dans le codebook pour lequel il n'existe pas d'image acoustique), le nombre d'hypercuboïdes, et l'erreur de resynthèse acoustique. Ces caractéristiques sont représentées, en fonction du volume minimal d'un hypercuboïde, sur la figure 2.9.

La figure 2.9a présente le rapport du volume de l'espace articulatoire défini contenu dans le codebook par le volume de l'espace articulatoire défini total, en fonction du volume minimal d'un hypercuboïde (échelle logarithmique), et du nombre t de sommets définis minimal variant entre 64 (la moitié des sommets est définie) et 128 (la totalité des sommets est définie). La figure 2.9b présente le nombre d'hypercuboïdes du codebook (échelle logarithmique), en fonction du volume minimal d'un hypercuboïde et de t. La figure 2.9c présente le rapport du volume de l'espace articulatoire erroné contenu dans le codebook par le volume de l'espace articulatoire défini total, en fonction du volume minimal d'un hypercuboïde et de t. Enfin, la figure 2.9d présente la « précision acoustique » des codebooks, c'est-à-dire l'erreur moyenne commise lorsqu'on synthétise l'image acoustique à l'aide de celui-ci. La moyenne des moyennes des erreurs sur les trois premiers formants est présentée ici pour les différentes valeurs de t et en fonction du volume minimal.

On constate sur ces figures que le paramètre t influe considérablement sur le volume de l'espace articulatoire défini (figure 2.9a), et ce dès le début : le volume est supérieur à 90% même

pour un volume minimal important pour les valeurs faibles de t ($t = 80$ ou $t = 64$). On voit de plus que pour ces valeurs le codebook est aussi nettement plus compact : il y a environ 4 fois moins d'hypercubes nécessaires pour décrire les codebooks avec $t = 64$ qu'avec $t = 128$.

Il apparaît donc que cette simple modification améliore considérablement la compacité des codebooks ; mais il y a tout de même un prix à cette modification, présenté sur la figure 2.9c : ces codebooks contiennent en effet des zones de l'espace articulatoire qui ne devraient pas avoir d'image acoustique. Cela peut avoir des conséquences fâcheuses pour l'inversion, car il devient alors possible de trouver comme solution des vecteurs articulatoires n'ayant pas d'image acoustique en utilisant de tels codebooks. On constate cependant que le volume de l'espace articulatoire qui ne devrait pas se trouver dans le codebook est nul pour $t = 128$, négligeable pour $t = 112$, et reste modéré pour les valeurs plus faibles de t (et ne dépend pas du volume minimal). On remarque également que l'erreur acoustique de synthèse par codebook augmente lorsque t diminue ; l'erreur restant toutefois inférieure au seuil acoustique imposé. Cette erreur diminue légèrement lorsque le volume minimal augmente.

Cette légère modification permet ainsi d'obtenir des codebooks réalisant une couverture pratiquement complète de l'espace articulatoire, et ce même en utilisant des hypercuboïdes de volume minimal important. En fonction de l'objectif, il peut être préférable d'utiliser un seuil t de valeur élevée (par exemple $t = 112$) si l'on ne veut pas avoir de fausses solutions pour l'inversion, ou des seuils t de valeur faible si l'on désire avoir des codebooks offrant une bonne couverture, très compacts et rapides à construire. Il faut toutefois remarquer que si l'on souhaite une meilleure précision acoustique, le facteur principale de subdivision devient la non-régularité acoustique, et les gains de place deviennent nettement moins importants.

Conclusion

Nous avons présenté et réalisé au cours de cette thèse une méthode de construction de codebooks permettant de réaliser des codebooks ayant une excellente précision acoustique lors de la resynthèse, ce qui nous permet d'obtenir une synthèse par codebook extrêmement rapide et précise. Il faut signaler que ce travail a essentiellement été effectué en fin de thèse, et que l'effet des diverses améliorations apportés ne pourra pas toujours être mesuré dans les parties suivantes de ce manuscrit. Par ailleurs, il est également important de souligner que l'espace articulatoire synthétisable par le codebook n'est pas strictement le même que celui synthétisable à l'aide du synthétiseur articulatoire ; en pratique, avec les réglages habituels, l'espace synthétisable par le codebook représente environ la moitié de l'espace synthétisable total, c'est-à-dire que la moitié des formes articulatoires qui ont une image acoustique ne sont pas dans le codebook. Ces vecteurs articulatoires correspondent aux formes ayant une constriction étroite, proche d'une fermeture complète du conduit, donc essentiellement aux consonnes. Cela n'a pas trop de conséquence pour l'étude présente, puisque nous nous sommes concentrés sur l'étude de l'inversion des voyelles, mais cela pourrait être un problème si l'on voulait utiliser la même méthode pour l'inversion des consonnes (fricatives notamment).

Chapitre 3

Inversion par codebook

Dans ce chapitre nous décrivons la méthode d'inversion statique utilisée, ou, en d'autres termes, la manière dont nous déterminons, à l'aide du codebook, un ensemble de vecteurs articulatoires dont l'image est proche d'un vecteur acoustique donné.

La méthode d'inversion par codebook que nous utilisons est très semblable aux autres méthodes d'inversion fondées sur des approximations linéaires d'ordre 1.

Schroeter et Sondhi (Schroeter & Sondhi 1992) utilisent une méthode classique par discrétisation et tabulation : ils discrétisent l'espace acoustique des trois premiers formants en petits cuboïdes, et rangent les images acoustiques de vecteurs articulatoires choisis aléatoirement dans chacun de ces cuboïdes. L'inversion d'un vecteur acoustique consiste alors simplement à choisir le cuboïde qui le contient.

Atal (Atal *et al.* 1978) utilise une méthode mixte : une première approximation des solutions de l'inversion était calculée grâce à une méthode classique par discrétisation et tabulation, puis ces solutions étaient améliorées en considérant une approximation linéaire dans un petit voisinage.

Sorokin (Sorokin *et al.* 2000) utilise une méthode comparable à celle de Schroeter pour obtenir une première approximation, couplée à une méthode d'optimisation utilisant le codebook comme synthétiseur articulatoire rapide.

Notre méthode repose sur la méthode de Ouni (Ouni & Laprie 2001) d'exploration de l'espace nul ; certaines modifications que nous y avons apportées sont liées à la structure différente de notre codebook, les autres améliorent la précision, la robustesse, ou la rapidité de la méthode.

Nous présentons dans cette partie la méthode générale d'inversion par codebook, c'est-à-dire sans l'adjonction de contraintes particulières ; celle-ci sera légèrement modifiée pour la prise en compte des contraintes phonétiques(cf. partie II) ou visuelles (cf. partie 6.4).

3.1 La méthode d'inversion

Contrairement à la très grande majorité des méthodes existantes, notre méthode d'inversion ne cherche pas à déterminer une solution unique, mais tous les candidats potentiels vérifiant la contrainte acoustique à un échantillonnage près. Il est ensuite possible de déterminer parmi ces candidats la solution qui vérifie au mieux un certain critère reposant sur différentes contraintes (par exemple des contraintes de type biodynamique portant sur l'écart à la forme neutre). Ces contraintes sont appliquées dans les étapes ultérieures de l'inversion, de façon à ne pas supprimer prématurément des solutions localement mauvaises, mais qui pourraient s'avérer globalement très bonnes pour une inversion dynamique.

3.1.1 Principe général

Le principe utilisé pour l'inversion est relativement simple : étant donné un vecteur acoustique s, on cherche à déterminer l'ensemble des vecteurs articulatoires v dont l'image par le synthétiseur articulatoire est proche de s. Plus formellement, on cherche un sous-ensemble représentatif de $\{v \in \mathrm{Ar} | F(v) = s \pm \Delta s\}$. En réalité, il est difficilement envisageable d'utiliser le synthétiseur articulatoire réel à cette étape, nous utilisons donc l'approximation du codebook.

Le codebook ne donnant qu'une approximation de la fonction dans un petit voisinage, une première étape consiste à déterminer dans l'ensemble du codebook les hypercuboïdes susceptibles de contenir des solutions. Ensuite, en utilisant l'approximation locale :

$$F(X) \approx P(X),$$

où $P(X)$ désigne le polynôme d'interpolation approchant la relation dans l'hypercuboïde considéré, l'inversion consiste à résoudre :

$$P(X) = s$$

ou, en d'autres termes,

$$P(X) - s = 0. \tag{3.1}$$

La formulation du problème est simple : il nous suffit de résoudre l'équation précédente, qui est un système de M équations (non-linéaires) à N inconnues.

Par ailleurs, dans le cas où P est un polynôme de degré 1, le système devient linéaire, et la résolution peut se faire simplement grâce aux méthodes classiques d'analyse linéaire. Dans notre cas, si nous cherchons à inverser des triplets de fréquences formantiques, $N = 7$ et $M = 3$, nous obtenons ainsi un système sous-déterminé ; le sous-espace vectoriel des solutions est a priori de dimension $N - M = 4$, et il est facile d'en déterminer une base.

La grande difficulté est de déterminer, dans ce sous-espace vectoriel, l'ensemble des solutions valables, c'est-à-dire celles qui se situent dans l'hypercuboïde où l'on a considéré l'approximation : en effet, il s'agit de calculer l'intersection d'un espace vectoriel à 4 dimensions et d'un parallélotope droit à 7 dimensions, ce qui est extrêmement difficile dans le cas général.

La solution retenue est très semblable à celle proposée par Ouni : on commence par déterminer une base du sous-espace vectoriel des solutions, puis on réalise un échantillonnage des solutions, en s'aidant de programmation linéaire pour borner la taille de l'espace à explorer.

Différents types d'échantillonnage peuvent être envisagés. Ouni réalisait un échantillonnage régulier qui était le même dans chaque hypercube. Nous avons choisi de réaliser un échantillonnage aléatoire, le nombre de points générés pouvant être proportionnel aux dimensions de l'hypercuboïde, ou contrôlé de façon à obtenir un même nombre de solutions pour chacun des vecteurs acoustiques.

3.1.2 Restreindre l'ensemble d'hypercuboïdes à explorer

La rapidité de l'inversion est intrinsèquement liée au nombre d'hypercuboïdes à explorer, l'étape de programmation linéaire étant la plus coûteuse en temps de calcul. Il est par conséquent intéressant de pouvoir déterminer le plus tôt possible les hypercuboïdes n'ayant pas de solution, pour ne pas avoir à les explorer. Ouni (Ouni 2001) explorait tous les hypercubes ; nous proposons un algorithme très simple qui permet d'éliminer rapidement un nombre important de mauvais candidats.

L'algorithme que nous utilisons pour éliminer très rapidement un ensemble d'hypercuboïdes sans solution repose sur un constat très simple : l'image acoustique d'un hypercuboïde est toujours incluse dans un parallélépipède rectangle aux côtés parallèles aux axes, dans l'espace des

trois premières fréquences formantiques. Ce parallélépipède rectangle se déduit facilement à partir de l'image des différents sommets de l'hypercuboïde : il s'agit du plus petit parallélépipède qui contient tous ces sommets. Cette propriété permet facilement de concevoir un algorithme rapide permettant d'éliminer un grand nombre d'hypercuboïdes n'ayant aucune solution : l'espace est discrétisé en petits cubes acoustiques dans lesquels on place une référence (sous la forme d'une liste chaînée) à tous les hypercuboïdes candidats dont l'image acoustique a une intersection non vide avec le cube. Cela permet en moyenne de diviser par 20 le nombre d'hypercuboïdes à considérer, pour un surcoût en mémoire raisonnable et un coût en temps de traitement divisé par environ trois[6] ; le surcoût lié au parcours de la table est négligeable.

3.1.3 Résolution du système d'équation $P(X) = s$

Il y a deux cas à envisager : le cas où P est un polynôme de degré égal à 1, et le cas où P est un polynôme de degré strictement supérieur.

Polynôme de degré 1

Dans le cas où P est un polynôme de degré 1, la résolution de 3.1 est relativement simple, puisque P est alors simplement une application affine. L'équation peut s'écrire sous la forme $P(X) = A.X + P_0 = s$, où A est une matrice $M \times N$ et P_0 un vecteur de dimension M, soit au final :

$$A.X = b, \tag{3.2}$$

où $b = s - P_0$.

La résolution de l'équation 3.2, qui est une équation linéaire, est simple. Dans le cas général, le rang de la matrice A est égale à la dimension de l'espace acoustique, soit, dans notre cas, 3. La dimension du noyau de A, noté $Ker(A)$, i.e. l'espace vectoriel $X|A.X = 0$, est de $7-3 = 4$. Toute solution de l'équation 3.2 peut s'écrire sous la forme $X_0 + v$, où X_0 est un vecteur particulier tel que $A.X_0 = b$, et $v \in Ker(A)$.

Nous résolvons numériquement l'équation 3.2 en utilisant la décomposition en valeurs singulières (plus connue sous le sigle SVD (Golub & Loan 1989)) qui fournit une solution particulière X_0 ayant la propriété de minimiser la distance – au sens des moindres carrés – au centre de l'hypercuboïde, et une base orthonormale de $Ker(A)$. De plus, son calcul est assez rapide.

La donnée de X_0 et de la base de $Ker(A)$ permet de reconstituer l'ensemble des solutions de 3.2. Il reste à sélectionner uniquement les solutions qui sont contenues dans l'hypercuboïde. Il serait bien utile pour cela de déterminer un critère qui nous permette de déterminer rapidement s'il existe au moins un point dans l'hypercuboïde ; on pourrait penser que la propriété particulière de X_0 de minimiser la distance euclidienne au centre puisse s'avérer utile dans ce cas. En l'occurrence, la seule chose que l'on puisse dire, c'est que si X_0 est dans l'hypercuboïde, alors il existe des solutions dans l'hypercuboïde. La réciproque est fausse ; pour que le point X_0 permette de déterminer immédiatement s'il existe au moins un point dans l'hypercuboïde, il faudrait qu'il minimise la distance au centre, mais au sens de la norme $||_r$ définie ainsi : $|x|_r = \max_{1 \leq i \leq N} \frac{|x_i|}{r_i}$ dont on vérifie aisément qu'elle est bien une norme si $\forall\, i, r_i > 0$. Malheureusement, il n'est pas du tout simple de déterminer un tel point, la seule méthode envisageable étant la programmation

[6] Les hypercuboïdes n'ayant pas de solutions étant éliminés relativement rapidement par la programmation linéaire par rapport à un hypercuboïde avec solution, le traitement d'un hypercuboïde sans solution est, même avec l'étape de programmation linéaire, environ 10 fois plus rapide que pour un hypercuboïde contenant des solutions.

linéaire, qui prend un temps important. Néanmoins, la méthode basée sur la SVD nous permet de déterminer un ensemble de solutions de la forme :

$$S^* = \{X_s | X_s = X_0 + \sum_{j=1}^{4} \lambda_j v_j\}, \tag{3.3}$$

où $\{v_j\}$ est la base du noyau déterminée par SVD et les λ_j les coordonnées de la projection dans $Ker(A)$ du point X_s exprimées dans cette base.

Polynôme de degré strictement supérieur à 1

Le cas où P est un polynôme de degré élevé est nettement plus difficile à gérer, puisque le système à résoudre n'est plus linéaire. Une solution simple pour contourner la difficulté est de se ramener au cas précédent en subdivisant l'hypercuboïde et linéarisant localement la relation. On perd l'intérêt d'utiliser des polynômes de degré élevé, mais on peut en revanche contrôler très précisément l'erreur commise.

Soit P^* une approximation de P au premier degré en un point x_0 donné de l'hypercuboïde. On construit P^* comme le développement de Taylor à l'ordre 1 de P en x_0. Dans un sous-hypercuboïde Hc' de rayon r' l'erreur commise est $\max_{x \in \mathrm{Hc}'} |P^*(x) - P(x)|$. Cette erreur est facilement majorée par $\|A\|$:

$$\|A\| = \sum_{i=2}^{n} |r'|^i \, \|A_i\|_\infty, \tag{3.4}$$

où les A_i correspondent aux vecteurs de coefficients des monômes d'ordre i du polynôme P (cf. 2.2.3), et la norme $\|\|_\infty$ à la norme matricielle infinie, c'est-à-dire à la somme des valeurs absolues des coefficients pour une matrice ligne (ce que sont les A_i). En effet, pour tout $x \in \mathrm{Hc}'$,

$$P^*(x) - P(x) = P^*(x) - \sum_{i=0}^{n} A_i.x^n \tag{3.5}$$

$$= -\sum_{i=2}^{n} A_i.x^n \quad \text{car } P^*(x) = A_0 + A_1.x. \tag{3.6}$$

Par conséquent,

$$|P^*(x) - P(x)|_\infty = \left| \sum_{i=2}^{n} A_i.x^n \right|_\infty \tag{3.7}$$

$$\leq \sum_{i=2}^{n} \|A_i\|_\infty \times |x|_\infty^n. \tag{3.8}$$

Or, pour $x \in \mathrm{Hc}'$, $|x|_\infty \leq |r'|_\infty$. D'où le résultat.

$\|A\|$ est un polynôme (sur $\mathbb{R}$) du n^{e} degré en $y = |r'|$. Nous désignons par $Q^*(y)$ ce polynôme. Il est alors simple de déterminer un rayon r' nous permettant de garantir une erreur acoustique inférieure à un seuil donné ϵ : il suffit de déterminer l'unique[7] racine positive de l'équation $Q^*(y) = \epsilon$.

[7] La solution est unique car $Q^*(y)$ étant un polynôme à coefficients positifs de degré $n > 1$, il est strictement croissant sur $\mathbb{R}^+$.

En pratique, dans les cas que nous avons testés (i.e. polynômes de degré inférieur ou égal à 4), un rayon r' dont toutes les composantes sont inférieures ou égales à 0.2 permet de garantir une erreur inférieure à 10Hz pour chaque formant.

Cette discrétisation des hypercuboïdes se fait « au vol », c'est-à-dire uniquement lorsque l'on cherche à trouver des solutions dans un hypercuboïde donné. Par conséquent, la place occupée sur le disque ne s'en ressent pas. Par ailleurs, la constructions des sous-hypercuboïdes est rapide à réaliser ; en effet, le vecteur de coefficients est simplement : $(A_0 + A_1.x_0|A_1)$.

Il nous reste à déterminer les points de S^* qui se trouvent dans l'hypercuboïde, i.e. $S = S^* \cap \mathrm{Hc}$.

3.1.4 Échantillonnage de solutions

Comme nous l'avons évoqué précédemment, il n'existe pas de méthode connue permettant de calculer $S = S^* \cap \mathrm{Hc}$ de manière formelle. La seule solution à notre disposition est donc de réaliser un échantillonnage de l'espace des solutions.

Cet échantillonnage peut se faire « assez simplement » en utilisant un algorithme de programmation linéaire. En effet, l'appartenance de X_s à Hc se traduit simplement par :

$$\forall i, -\frac{r_i}{2} \leq X_{si} \leq \frac{r_i}{2},$$

ou, sous forme vectorielle :

$$-\frac{r}{2} \leq X_s \leq \frac{r}{2},$$

ou encore :

$$-\frac{r}{2} \leq X_0 + V.\lambda \leq \frac{r}{2}. \tag{3.9}$$

Le système 3.9 définit un polytope de dimension 4 (ou 4-polytope). À notre connaissance, personne ne sait déterminer S de manière formelle. Slim Ouni (Ouni & Laprie 2000) a présenté un algorithme d'échantillonnage pour ce 4-polytope :

1. Par programmation linéaire, déterminer le plus petit hypercuboïde (H') de dimension 4 qui contient le 4-polytope.
2. Échantillonner l'hypercuboïde de dimension 4 et vérifier l'appartenance de chacun des points à l'hypercuboïde Hc de dimension 7.

L'hypercuboïde H' est construit en déterminant successivement les valeurs minimales et maximales des λ_j vérifiant l'équation 3.9. Au total, cela donne 8 systèmes linéaires à résoudre : 4 maximisations des λ_j ($j = 1..4$), et 4 minimisations des λ_j ($j = 1..4$). On obtient ainsi 2 valeurs différentes pour chacun des quatre λ_j, dont les diverses combinaisons nous donnent les 2^4 sommets de l'hypercuboïde en remplaçant les λ_j dans l'équation 3.3. L'hypercuboïde H' est alors échantillonné, et pour chacun des points l'appartenance à l'hypercuboïde Hc est vérifiée.

Plusieurs types d'échantillonnage sont possibles : un échantillonnage régulier ou aléatoire, avec une densité du maillage constante ou variable. Ouni utilisait un maillage régulier de densité constante, c'est-à-dire que le nombre de points du maillage était toujours identique quelles que soient les tailles de Hc ou H'.

Le nombre de points et la forme de l'échantillonnage a son importance : il va conditionner directement l'étape de recherche de trajectoires par lissage non-linéaire. Un nombre trop important de solutions trop proches les unes des autres ralentira considérablement l'inversion ; un nombre trop faible risque de conduire à la sélection d'une trajectoire très éloignée de la réalité. Pour notre part, nous avons préféré un échantillonnage aléatoire et une densité moyenne proportionnelle à la

taille de Hc. L'échantillonnage aléatoire évite les effets de seuillage généralement observés lors de l'utilisation de grilles régulières ; la densité de points proportionnelle au volume de Hc permet de garantir une distance articulatoire homogène entre les solutions issues de chaque hypercuboïde.

3.2 Quelques résultats

Dans cette section, nous présentons quelques exemples d'expériences d'inversion statique. Nous effectuons quelques expériences d'inversion de voyelles isolées, ainsi qu'une détermination des espaces acoustiques atteignables à partir du modèle articulatoire et d'un codebook typique.

3.2.1 Inversion de voyelles isolées

Le but essentiel d'un système d'inversion est de déterminer un sous-ensemble des solutions permettant la réalisation d'un son donné, qui dans notre cas, de par les limitations de notre modèle articulatoire, sera forcément une voyelle. Cette partie illustre l'inversion de plusieurs voyelles du français pour notre locutrice de référence, PB.

Les sons inversés sont simplement des triplets de fréquences formantiques. De façon à pouvoir faire une comparaison où n'interviennent pas les défauts liés au modèle articulatoire ou au suivi de formants, les vecteurs acoustiques inversés sont des vecteurs *synthétiques*, qui ont été générés en synthétisant, à l'aide du modèle articulatoire, différentes formes du conduit vocal extraites du livre de Bothorel (Bothorel *et al.* 1986), dont PB est l'un des sujets. L'inversion a ensuite été effectuée sur ces vecteurs acoustiques, et les formes obtenues ont pu ainsi être comparées à l'originale.

Ce type d'expérience est courant, et présente l'avantage de ne nécessiter aucune adaptation du modèle articulatoire. Le principal inconvénient est qu'il ne présente qu'une efficacité *théorique* de la méthode, en prenant des hypothèses difficilement satisfiables : d'une part, on suppose que l'on sait faire une adaptation parfaite du modèle articulatoire au locuteur à inverser – ce qui est loin d'être réaliste, puisque l'on ne parvient déjà pas à adapter parfaitement le modèle articulatoire au locuteur de référence ayant permis son élaboration – d'autre part, la détermination de certaines des composantes des vecteurs acoustiques peut être très problématique (notamment pour la largeur de bande et l'amplitude des formants). Les résultats que l'on obtient avec des vecteurs acoustiques contenant les amplitudes et les largeurs de bande des formants démontrent simplement un potentiel intéressant de cette méthode dans un cadre un peu éloigné d'une application réelle.

Les figures 3.2a, 3.2b, 3.2c, 3.2d, 3.2e présentent le résultat de l'inversion des vecteurs acoustiques synthétiques (limités aux trois premières fréquences formantiques) obtenues à partir de vecteurs articulatoires déterminées à partir de radiographies réelles de PB extraites du livre de Bothorel et al. (Bothorel *et al.* 1986). Il s'agit des 5 voyelles /a,e,i,u,y/ pris dans des contextes variés. La table 3.2.1 présente plus précisément les formants synthétiques inversés, ainsi que le contexte des formes inversés. Nous avons également à notre disposition le signal acoustique enregistré au moment de la radiographie, mais celui-ci est de très mauvaise qualité (cf. figure 3.1). Précisons à nouveau que les vecteurs acoustiques synthétiques sont différents de ceux que l'on peut mesurer sur le signal acoustique correspondant.

Un examen rapide de la table 3.2.1 nous permet de formuler plusieurs remarques.

1. Concernant le *nombre* de solutions trouvées lors de l'inversion. Le nombre de solutions trouvées est très variable d'un phonème à l'autre. Le phonème qui a le plus grand nombre de solutions est, sans surprise, le /a/, avec 140456 solutions. Le phonème qui en a le moins

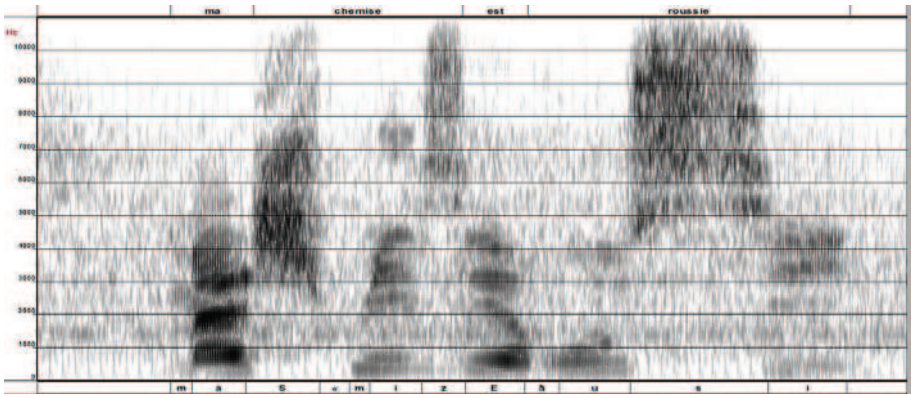


FIG. 3.1: *Spectrogramme à large bande de la première phrase de la locutrice PB. On constate que le signal acoustique est très bruité, avec notamment un renforcement spectral marqué autour de 1500Hz.*

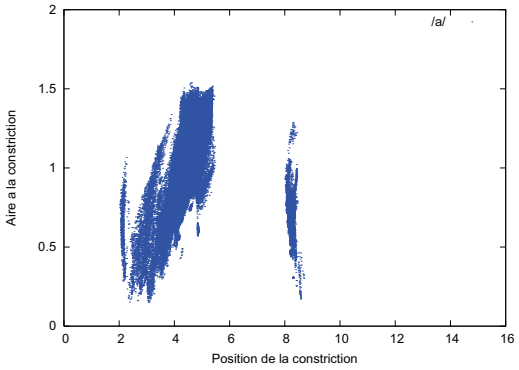
Voy.	Contexte	$F1$	$F2$	$F3$	Nb. de sol.	$\Delta F1$	$\Delta F2$	$\Delta F3$
a	tab**a**	806 (788)	1536 (1445)	2641 (2365)	140456	17.6	17.9	21.1
e	dyg**e**	465 (467)	2285 (2278)	2890 (2862)	92094	11.8	30.7	48.2
i	ab**i**	355 (365)	2291 (2307)	3442 (3314)	46626	25.2	20.5	62.9
u	ptik**u**	370 (387)	981 (829)	2307 (2411)	6217	14.5	31.8	13.8
y	ãbig**y**	335 (376)	2066 (1916)	2531 (2427)	33237	12.6	44.6	42.7

TAB. 3.1: *Un ensemble de voyelles du français pour la locutrice PB, leur contexte, les trois premières fréquences formantiques du vecteur acoustique synthétisé à partir de la forme de conduit (entre parenthèses les valeurs correspondantes mesurées sur l'enregistrement sonore), le nombre de solutions trouvées, et l'erreur moyenne entre les formants originaux et l'image des vecteurs articulatoires trouvés par inversion.*

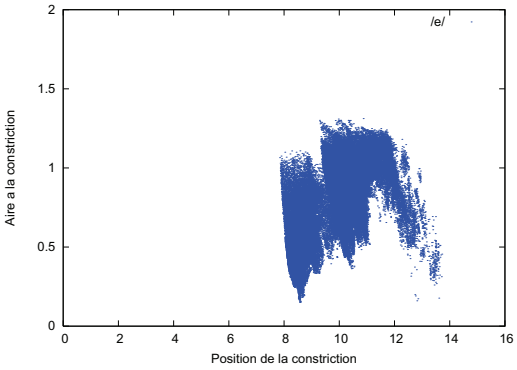
est le /u/, avec 6217 solutions. Rappelons que l'échantillonnage des solutions a été effectué de façon à générer un nombre de solutions proportionnel à la taille de l'espace des solutions, et, par conséquent, les nombres de solutions trouvées donnent réellement une indication sur la densité d'un phonème donné dans le codebook. Par ordre décroissant de densité, les phonèmes sont ainsi classés : /a,e,i,y,u/. Le /u/ étant notablement plus faible que les autres phonèmes, on peut ici soupçonner un défaut au niveau du codebook ; en effet, le /u/ présente le double inconvénient d'être situé en marge de l'espace articulatoire (deux des paramètres articulatoires permettant de réaliser le /u/ sur l'image de conduit que l'on cherche à retrouver sont en effet égaux à la valeur maximale admise pour les paramètres articulatoires lors de la construction du codebook) et d'avoir une constriction maximale d'aire réduite, donc proche des limites admises pour une voyelle. Il est donc probable[8] qu'une partie de l'espace articulatoire permettant la réalisation du vecteur acoustique du /u/ n'est pas incluse dans le codebook.

2. Concernant la *qualité* des solutions. On constate une grande disparité de l'erreur acoustique

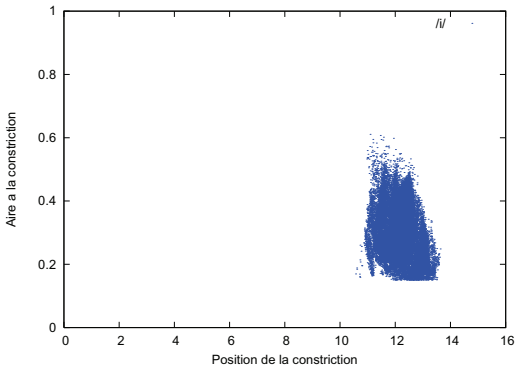
[8] Une étude préliminaire, non développée ici, indique que la densité des solutions pour cette voyelle est multipliée par un facteur 3 par rapport aux autres voyelles en agrandissant l'espace articulatoire de deux unités pour chaque paramètre, et par un facteur 6 en assouplissant légèrement la condition sur la constriction minimale admise.



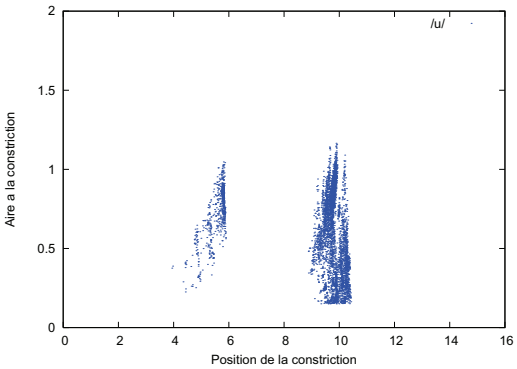
(a) Voyelle /a/



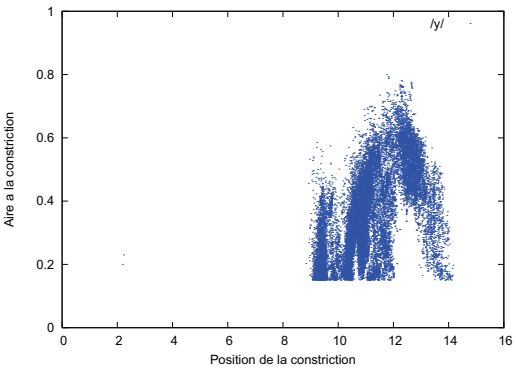
(b) Voyelle /e/



(c) Voyelle /i/



(d) Voyelle /u/



(e) Voyelle /y/

en fonction du phonème étudié ; pour F1, toutes les voyelles ont une qualité correcte, sauf le /i/, qui a une erreur RMS située autour de 25Hz, ce qui est pratiquement le double de l'erreur des autres voyelles. Pour F2, on constate une bien plus grande disparité : deux voyelles (/a/ et /i/) sont autour de 20Hz, 2 autres autour de 30Hz (/e/ et /u/), et enfin le /y/ a une erreur de 45Hz. Pour F3, l'erreur varie de moins de 15Hz pour /u/ à plus de 60Hz pour /i/.

3.2.2 Domaines acoustiques

Dans cette partie, nous présentons deux éléments : en premier lieu, « l'étendue acoustique » du modèle articulatoire de Maeda, ou, en d'autres termes, l'espace des sons atteignables dans l'espace formantique (restreint dans les figures au plan F1-F2 pour une lecture plus aisée), représenté sous la forme du triangle vocalique classique de la phonétique. En second lieu, nous présentons l'étendue et la densité acoustique que l'on peut atteindre par inversion en utilisant notre codebook : pour chaque triplet de formants, une inversion est réalisée, et le nombre de points générés est proportionnel à la taille de l'espace articulatoire des solutions possibles. Ce diagramme donne une indication sur « la facilité » de réaliser chaque zone de l'espace acoustique. Comme précédemment, l'espace acoustique représenté est restreint à l'espace F1-F2.

Dans chaque cas, un diagramme correspondant à un locuteur et un autre correspondant à une locutrice sont présentés.

3.2.2.1 Domaine acoustique du modèle articulatoire

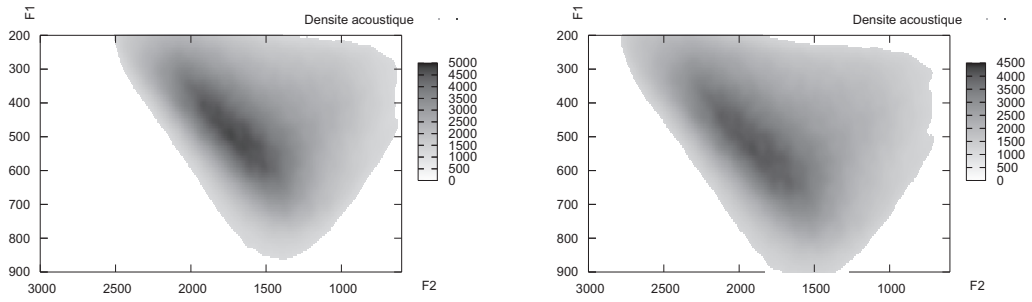
Les figures 3.2a et 3.2b ont été réalisées en générant de manière aléatoire un grand nombre (10^6) de vecteurs articulatoires et en calculant leur image acoustique. Pour une répartition plus naturelle des vecteurs articulatoires, chaque composante des vecteurs articulatoires générés suit une loi gaussienne centrée réduite[9]. Pour une lecture plus simple, ces images sont représentées dans le plan F1-F2 plutôt qu'en 3 dimensions dans l'espace F1-F2-F3. L'espace acoustique discrétisé par un pas de 5Hz sur F1 et F2 est « coloré » en niveaux de gris en fonction du nombre de vecteurs articulatoires ayant une image située dans un rayon de 10Hz autour du point considéré : plus le nombre[10] de vecteurs acoustiques est élevé, plus le point résultant est foncé.

L'observation de ces deux figures nous montre que les capacités acoustiques du modèle articulatoire pour chacun des locuteurs normalisés (c'est-à-dire pour lesquels des paramètres d'élongations pharyngale et orale nominaux ont été utilisés) sont tout à fait conformes à ce que l'on peut trouver dans la littérature : l'espace acoustique forme une sorte de triangle dans l'espace F1-F2 (communément appelé triangle vocalique), pour la locutrice l'espace acoustique est un peu plus étendu que pour le locuteur masculin, et on observe un léger décalage vers les fréquences plus élevées. Une chose également intéressante à prendre en compte est la densité des solutions : celle-ci donne une indication sur le nombre de configurations de vecteurs articulatoires permettant d'atteindre un même vecteur acoustique. On constate que les zones ayant une densité

[9] Cette loi gaussienne est simulée en utilisant la forme polaire de la transformation de Box-Muller. L'utilisation d'une loi gaussienne centrée réduite pour générer des vecteurs articulatoires pour le modèle de Maeda paraît pertinente, du fait de la construction du modèle à partir d'une analyse en composantes arbitraires (Overall 1962) à partir de données réelles, et du fait que les paramètres de contrôle ont été centrés et réduits. On peut effectivement observer que lors de la projection des données articulatoires originales dans le modèle, pratiquement tous les paramètres suivent une loi gaussienne (cf. Annexe A). Il est ainsi justifié d'utiliser un générateur ayant de telles caractéristiques pour « simuler » une articulation réelle.

[10] En réalité, chaque vecteur articulatoire situé dans le cercle contribue de façon inversement proportionnelle à sa distance au centre : si on appelle $l(x,p)$ la distance d'un vecteur acoustique x au point p considéré, alors chaque vecteur contribue à hauteur de $1 - l(x,p)^2/r^2$ à la valeur du point, où r est le rayon du cercle.

vraiment importante sont assez restreintes : la zone de densité maximale a une forme d'ellipse, centrée autour du point (500, 1600) pour le locuteur masculin, et autour de (550, 1900) pour la locutrice.



(a) Locuteur masculin.

(b) Locuteur féminin

FIG. 3.2: *Domaine acoustique pour un locuteur et une locutrice. L'abscisse correspond à F2 (Hertz), l'ordonnée à F1 (Hertz). La couleur indique la densité des images acoustiques, suivant une échelle linéaire.*

3.2.2.2 Densité des solutions dans le codebook

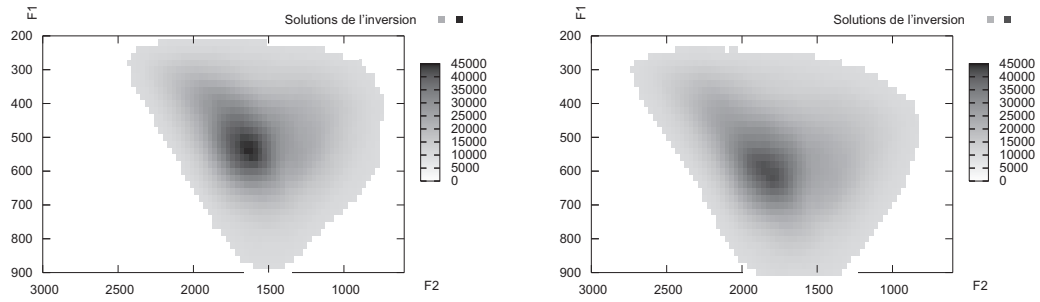
Les figures 3.3a et 3.3b ont été obtenues en effectuant une inversion sur chaque zone de l'espace acoustique avec un pas d'échantillonnage de 10, 20, et 30Hz pour respectivement F1, F2 et F3. Ce graphique donne une indication de l'espace des vecteurs acoustiques atteignables à partir du codebook. La comparaison de ces figures avec les figures 3.2a et 3.2b vues précédemment nous permet de constater que le codebook couvre à peu de choses près l'intégralité de l'espace acoustique atteignable par les locuteurs de référence. On constate également qu'il y a une variation importante dans les densités de solutions trouvées lors de l'inversion : aux points de densité maximale pour le locuteur masculin ou la locutrice (situés aux mêmes endroits que sur les figures respectivement 3.2a et 3.2b), la densité est environ 100 fois plus élevée que dans la zone correspondant au /u/, c'est-à-dire autour de (315, 760) et (311, 804) respectivement pour un locuteur et une locutrice, légèrement en dehors du « triangle ».

3.2.3 Statistiques

Nous présentons dans cette partie différentes statistiques sur l'inversion à l'aide de codebooks. Plusieurs types de codebooks ont été étudiés : les codebooks classiques utilisant l'interpolation linéaire avec des vecteurs acoustiques limitées aux trois premières fréquences formantiques, mais aussi avec des vecteurs acoustiques plus complexes, utilisant les largeurs de bande et amplitudes des formants, et nous avons également construit des codebooks utilisant une interpolation à l'aide de polynômes de degré strictement supérieur à 1.

3.2.3.1 Précision acoustique

Pour chacun de ces types de codebooks, la précision acoustique de la resynthèse a été évaluée en générant un grand nombre de vecteurs articulatoires (100000) de manière aléatoire, en



(a) Locuteur masculin.

(b) Locuteur féminin

FIG. 3.3: *Inversion de l'espace acoustique pour un locuteur et une locutrice. L'abscisse correspond à F2 (Hertz), l'ordonnée à F1 (Hertz). La couleur indique la densité des images acoustiques, suivant une échelle linéaire.*

calculant pour chacun des points les images acoustiques à l'aide du codebook et du synthétiseur articulatoire, et en comparant ces deux images. Pour chacun de ces codebooks, nous pouvons ainsi calculer différentes caractéristiques de l'erreur de resynthèse de l'image acoustique à partir du codebook : l'erreur absolue moyenne, l'erreur RMS, l'erreur maximale et l'écart-type. Ces valeurs sont calculées indépendamment pour chaque composante du vecteur acoustique. Le tableau 3.2a présente les caractéristiques des différents codebooks comparés.

Le tableau 3.2b donne les résultats correspondants à un codebook classique, avec un vecteur acoustique de taille variable, le test de linéarité ne portant que sur les fréquences des trois premiers formants, avec une précision de un Bark pour chaque composante.

Le tableau 3.2c donne les résultats pour un codebook construit dans les mêmes conditions, mais avec un polynôme d'interpolation de degré 4.

Pour illustrer l'erreur de resynthèse, nous présentons l'erreur commise lors de l'inversion le long d'une grille régulière sur les deux premiers formants (F1, F2) : nous inversons tous les points de coordonnées $(i \times 100, j \times 100)$ dans l'espace acoustique (F1, F2), et nous resynthétisons tous les vecteurs articulatoires trouvés. La figure 3.4 présente la densité des solutions en fonction des coordonnées (F1,F2) suivant une échelle de densité logarithmique de façon à bien visualiser la dispersion des solutions.

Le codebook utilisé est CBPB, c'est-à-dire un codebook de degré 1, d'une précision acoustique de 1.0 Bark, pour la locutrice de référence. La figure est constituée de petits nuages situés autour des points inversés, qui indiquent la dispersion acoustique des solutions de l'inversion. On peut constater sur cette figure que la dispersion le long de l'axe F2 est assez homogène (bien que si l'on observe attentivement, on peut observer que le nuage est légèrement décentré vers les fréquences basses), mais qu'en revanche la dispersion suivant l'axe F1 est essentiellement dirigée vers les fréquences basses. On constate ainsi que l'inversion à l'aide de ce codebook a une légère tendance à surestimer la valeur de F1.

3.2.3.2 Amélioration de la précision

Une méthode simple qui permet d'améliorer la précision acoustique pour un coût relativement faible est la suivante : calculer l'image acoustique du vecteur articulatoire, et corriger l'erreur à l'aide d'une descente de gradient. Si le gradient provient de l'hypercuboïde local, la correction

Nom	Nb. Hc.	Volume	Deg.	Dim. Ac.
30	47437	78271.3	1	9
31	28972	79587.9	4	9

(a) Caractéristiques des codebooks utilisés : Nombre d'hypercubes, volume total, degré des polynômes utilisés, dimension du vecteur acoustique.

Ac	Err. moy.	Err. RMS	Err. Max	Écart-type
F1	6.242	9.201	80.153	6.760
F2	9.736	15.623	159.147	12.219
F3	10.781	19.160	301.979	15.839
Bw F1	1.903	2.690	42.893	1.902
Bw F2	2.558	4.746	95.245	3.998
Bw F3	6.665	25.809	861.521	24.934
Am F1	0.399	0.502	4.141	0.305
Am F2	0.495	0.638	4.627	0.402
Am F3	0.524	0.711	11.272	0.480

(b) Caractéristiques de la resynthèse pour un codebook de degré 1 pour chacune des composantes du vecteur acoustique (fréquence, largeur de bande et amplitude pour les trois premiers formants) : Moyenne de l'erreur, erreur RMS, erreur « maximale », et écart-type.

Ac	Err. moy.	Err. RMS	Err. Max	Écart-type
F1	0.597	1.352	54.947	1.213
F2	1.089	1.840	30.849	1.484
F3	2.178	6.688	338.922	6.323
Bw F1	0.241	0.417	11.306	0.340
Bw F2	0.537	1.176	20.366	1.047
Bw F3	2.943	16.211	703.144	15.942
Am F1	0.037	0.075	2.968	0.065
Am F2	0.051	0.093	2.326	0.078
Am F3	0.076	0.181	10.468	0.164

(c) Caractéristiques de la resynthèse pour un codebook de degré 4.

TAB. 3.2: *Caractéristiques des codebooks utilisés pour l'inversion, et de l'erreur commise lors de la synthèse articulatoire à partir de ces codebooks.*

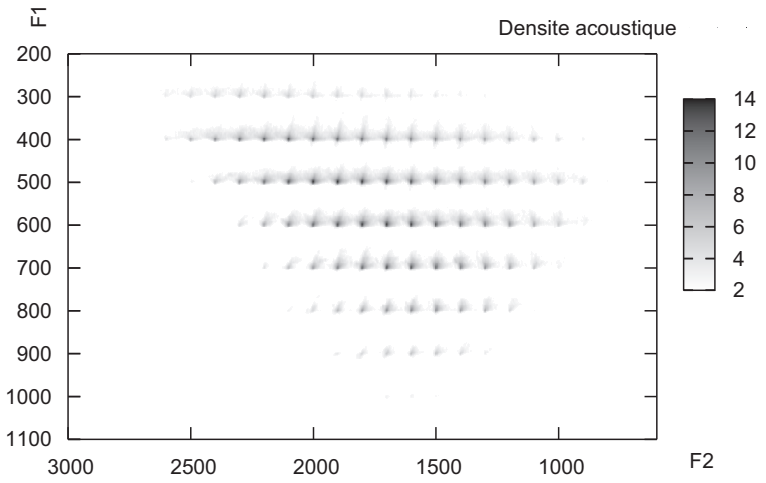


FIG. 3.4: *Dispersion des solutions pour une inversion le long d'une grille régulière de l'espace des deux premières fréquences formantiques. L'abscisse correspond à F2 (en Hertz), l'ordonnée à F1 (en Hertz). La couleur indique la densité de solutions (suivant une échelle logarithmique).*

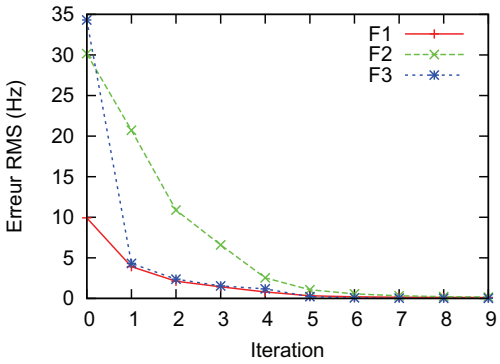
coûte essentiellement le prix d'un appel au synthétiseur articulatoire par échantillon ; la correction n'est bien entendu pas parfaite, puisque le gradient n'est pas tout à fait correct.

La figure 3.5a montre l'amélioration apportée par ce procédé sur une séquence d'un millier de points en fonction du nombre d'itérations : en 10 itérations, la diminution de l'erreur RMS est supérieure à 99% pour chaque fréquence formantique. À titre de comparaison, en suivant le même principe mais en calculant le gradient à l'aide du synthétiseur (cf. figure 3.5b), la correction demande moins d'itérations (il n'y a besoin que de 2 itérations pour atteindre une diminution de l'erreur supérieure à 99%, en 4 itérations l'erreur est nulle à $0.001Hz$ près), mais le nombre d'appels au synthétiseur nécessaire est beaucoup plus important : le calcul du gradient seul nécessite 14 appels au synthétiseur, donc au total 15 appels au synthétiseur par échantillon et par itération.

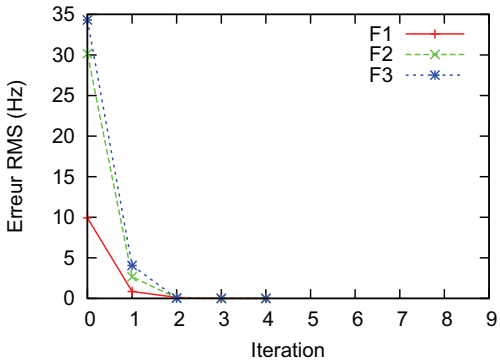
3.2.3.3 Précision articulatoire

Il est également intéressant d'étudier l'influence des différentes modalités de l'inversion statique sur la précision articulatoires des résultats. Dans l'expérience qui suit, nous étudions plus particulièrement l'influence du nombre de points générés sur la « proximité articulatoire » des solutions à la solution originale.

Nous quantifions la distance entre deux formes de conduit vocal de deux façons différentes : une *distance géométrique* (notée d_1) d'une part –basée sur la projection de la forme de conduit vocal sur la grille de Maeda,– consistant en une moyenne quadratique des distances entre points analogues, et une *distance articulatoire* (notée d_2) d'autre part, qui est simplement la distance quadratique moyenne entre deux vecteurs articulatoires.



(a) Correction acoustique « mixte».



(b) Correction acoustique en utilisant uniquement le synthétiseur.

En d'autres termes :

$$d_1(X,Y) = \sqrt{\frac{\sum_{j=1}^{N} |P(X)_j - P(Y)_j|^2}{N}},$$

où P désigne l'opérateur de projection d'un vecteur articulatoire vers la grille de Maeda, N est le nombre de points de la grille, et $P(X)_j$ désigne l'un des projetés du vecteur articulatoire X sur la grille, et

$$d_2(X,Y) = \sqrt{\frac{\sum_{i=1}^{7} (X_i - Y_i)^2}{7}},$$

où X et Y sont deux vecteurs articulatoires.

Dans l'expérience qui suit, une séquence de vecteurs articulatoires « réels » a été utilisée : il s'agit de la première phrase de la locutrice PB de l'étude de (Bothorel *et al.* 1986). Des vecteurs acoustiques de différentes tailles – de 3 à 5 fréquences formantiques – ont été générés en calculant les images acoustiques de cette séquence de vecteurs articulatoires[11]. L'inversion a ensuite été effectuée sur ces vecteurs acoustiques, et pour chacun de ces vecteurs acoustiques, on a déterminé le vecteur articulatoire minimisant la distance (respectivement d_1 ou d_2) au vecteur articulatoire original parmi l'ensemble des solutions de l'inversion statique.

Dans le tableau 3.3 nous présentons la moyenne sur la séquence de la « proximité articulatoire » minimale à l'original, au sens des deux normes d_1 et d_2, en fonction de la densité de solutions générées, et pour différentes tailles de vecteur acoustique. La première colonne « nbp » correspond aux valeurs données au paramètre contrôlant la densité des solutions à générer (il correspond environ au nombre de solutions à générer par unité de volume articulatoire). La colonne « sols » correspond au nombre de solutions effectivement générées, la colonne d_1 à la distance géométrique (en cm), la colonne d_2 à la distance articulatoire.

On retrouve expérimentalement le comportement qu'on pouvait prévoir théoriquement : il est nécessaire de multiplier la densité des points générés par un facteur de 2^7 pour diminuer la distance minimale d'un facteur 2.

[11] À noter que des vecteurs acoustiques peuvent être générés, à l'aide du synthétiseur utilisé, pour toutes les configurations articulatoires ; par soucis de réalisme, nous ne prenons en considération que les configurations de conduits sans occlusion, c'est-à-dire essentiellement les voyelles et les fricatives.

nbp	sols	d_1 (cm)	d_2
100	4597	0.139	0.611
200	8710	0.127	0.561
500	19358	0.108	0.489
1000	34584	0.094	0.431
2000	62188	0.089	0.412
5000	134978	0.081	0.370
10000	246487	0.076	0.351
20000	457638	0.072	0.334
50000	1052800	0.066	0.305

(a) inversion sur 3 formants

nbp	sols	d_1 (cm)	d_2
100	596	0.161	0.684
200	1142	0.140	0.604
500	2581	0.104	0.484
1000	4801	0.095	0.445
2000	8687	0.092	0.441
5000	19332	0.083	0.408
10000	35199	0.077	0.369
20000	65307	0.071	0.347
50000	149901	0.066	0.318

(b) inversion sur 4 formants

nbp	sols	d_1 (cm)	d_2
100	157	0.166	0.743
200	327	0.145	0.719
500	764	0.122	0.566
1000	1430	0.105	0.491
2000	2631	0.108	0.508
5000	5873	0.098	0.463
10000	10895	0.093	0.451
20000	20340	0.085	0.418
50000	47086	0.085	0.398

(c) inversion sur 5 formants

TAB. 3.3: *Précision géométrique (d_1) et articulatoire (d_2) minimale des solutions de l'inversion statique pour des vecteurs acoustiques de tailles diverses. La première colonne (« nbp ») est la valeur donnée au paramètre contrôlant la densité des solutions à générer. La colonne « sols » correspond au nombre de solutions effectivement générées.*

Cette expérience permet également d'obtenir une indication sur le volume de l'espace des solutions : passer de 3 à 4 formants fait diminuer la taille de cet espace d'un facteur 7 environ, alors que passer de 4 à 5 formants ne le fait diminuer que d'un facteur 3.

Pour finir : en théorie, si le vecteur articulatoire que l'on cherche à trouver est présent dans le codebook, la distance minimale à la solution optimale ne devrait dépendre que de la densité des points générés, indépendamment de la taille du vecteur acoustique. On semble avoir pratiquement ce comportement si l'on ne considère que l'inversion sur 3 ou 4 formants. Cependant, cela n'est plus le cas pour 5 formants : cela indique que certains des vecteurs articulatoires que l'on cherche à trouver sont en réalité absents du codebook.

3.2.4 Temps de calcul

La méthode générale que nous utilisons à cette étape, c'est-à-dire une SVD suivie d'une programmation linéaire, puis d'un échantillonnage à pas fixe ou aléatoire, est très gourmande en temps de calcul. En pratique, en utilisant toutes les étapes, la méthode prend environ 30 secondes pour inverser chaque échantillon de parole, sur une machine récente. C'est l'un des gros points faibles de cette méthode ; il paraît donc important de rechercher diverses méthodes qui puissent

permettre d'accélérer ce calcul, peut-être au détriment de la qualité de l'inversion.

Quelques optimisations peuvent être faites qui accélèrent le calcul sans pénalité au niveau de la qualité : une première optimisation simple concerne le précalcul des matrices de SVD : en effet, celles-ci ne dépendent que de la matrice jacobienne des hypercuboïdes, il n'est donc pas nécessaire de les calculer plus d'une fois. Le gain de temps n'est cependant appréciable que si l'on inverse de longues séquences de parole et que certains hypercuboïdes reviennent souvent.

Une autre optimisation, celle-ci influant beaucoup sur la qualité, est le développement d'une heuristique très rapide en remplacement de l'étape de programmation linéaire, qui est, et de très loin, l'étape qui occasionne le plus de calculs. Au fond, cette étape ne sert qu'à borner l'espace à échantillonner, il suffit donc de déterminer un algorithme qui nous permette de trouver des bornes relativement bonnes en un temps très court. Pour le moment, nous nous contentons d'utiliser l'algorithme le plus simple possible : en l'occurrence, des bornes fixes ne dépendant pas de l'hypercuboïde. Il est évident que ce type de borne va détériorer de façon significative la qualité de l'inversion, mais les gains de vitesse sont plus qu'appéciables : cela permet de diviser le temps de calcul par un facteur d'environ 50. Il est tout de même souhaitable de vérifier que l'inversion ne s'en trouve pas trop perturbée.

Nous avons donc comparé, pour une séquence articulatoire « réelle », les résultats de l'inversion du signal synthétique par chacune de ces deux modalités : la méthode « classique » sans optimisation, et la méthode avec optimisation. L'important, pour cette première étape, est qu'il existe, parmi l'ensemble des solutions, un vecteur articulatoire solution « proche » du vecteur articulatoire réel. Nous mesurons donc dans chacun des cas la distance moyenne (au sens de deux distances différentes).

cas	#HC	# sol	d_1	d_2
inversion rapide	3459	4193	0.0561	0.4289
inversion normale	5670	5705	0.0618	0.0.3937

TAB. 3.4: *Comparaison de la qualité des résultats de l'inversion rapide et de ceux de la méthode normale suivant plusieurs critère : nombre d'hypercubes détectés comme contenant des solutions, nombre de solutions effectivement trouvées, et distance (suivant deux métriques différentes) à la solution originale. La distance d_1 est une distance géométrique, la distance d_2 la distance euclidienne sur les vecteurs articulatoires.*

Les tests ne sont bien sûr pas exhaustifs, puisque ces résultats ne concernent que l'inversion d'une cinquantaine de vecteurs articulatoires, et dans un seul type de conditions (un seul codebook testé, génération de points proportionnelle au volume de l'espace des solutions), mais il semble que les résultats soient globalement assez proches de ceux obtenus avec la méthode traditionnelle.

Les bornes fixes ont été déterminées de façon automatique, en cherchant à maximiser le nombre moyen d'hypercuboïdes présentant des solutions sur une séquence acoustique contenant des formants pour l'ensemble des voyelles du français. Il est probable que l'on pourrait améliorer relativement facilement la qualité de la méthode rapide en utilisant des modélisations plus fines des bornes.

Conclusion

Ce chapitre présentait la première étape de l'inversion acoustique articulatoire, c'est-à-dire la génération d'un ensemble discret de vecteurs articulatoires dont l'image acoustique est proche d'un vecteur donné. Cette méthode dérive de la méthode introduite par Ouni (Ouni 2001). Nous avons présenté dans ce chapitre diverses améliorations de cette méthode, et avons évalué précisément son efficacité. Une amélioration importante – mais dont l'exploration doit être poursuivie – est l'élimination de l'étape de programmation linéaire, ce qui permet d'accélérer considérablement le processus.

Chapitre 4

Trajectoires articulatoires

Nous avons vu que l'inversion statique est un problème mal contraint. Nous allons nous intéresser dans cette partie à l'inversion dynamique, c'est-à-dire celle de segments acoustiques. Par rapport à l'inversion statique, dont la principale contrainte porte sur la proximité de l'image acoustique avec les données à inverser, l'inversion de segments ajoute une contrainte de type temporel : en effet, le mouvement des articulateurs doit être régulier, en particulier continu, et par conséquent cela induit une contrainte de lissage sur les mouvements des articulateurs. Cette contrainte de lissage des trajectoires ne permet cependant pas d'obtenir une trajectoire unique, et de nombreux types de contraintes sur les trajectoires temporelles des articulateurs ont été présentés dans la littérature, en particulier par Sorokin (Sorokin *et al.* 2000).

Les contraintes sur la régularité des trajectoires articulatoires sont de plusieurs ordres : la plus simple est la condition de continuité des trajectoires, et on peut par exemple rechercher la trajectoire qui minimise l'effort des articulateurs. Mais on peut également imposer des conditions similaires sur les différentes dérivées des trajectoires des articulateurs : vitesse (dérivée première), accélération (dérivée seconde), voire secousse (dérivée troisième). Cette dernière est la plus pertinente d'un point de vue physiologique, et est très utilisée pour modéliser les trajectoires des membres humains ; l'une des publications les plus citées à ce sujet est celle de Flash et Hogan (Flash & Hogan 1985) qui présente une confirmation expérimentale d'une minimisation de la secousse dans des cas particuliers de mouvements des bras. Voir également Engelbrecht (Engelbrecht 2001) pour une revue plus générale des différentes modélisations de la cinématique des mouvements humains.

D'autres contraintes utilisées fréquemment s'inspirent de la nature même des articulateurs de la parole et de leurs propriétés physiologiques. En particulier, un critère couramment utilisé est une pénalisation des trajectoires s'éloignant des positions neutres des articulateurs (Perkell 1974). Un autre critère souvent évoqué fait intervenir la notion de commande articulatoire et de planification, stipulant que le nombre de commandes programmables en un temps donné est limité, ce qui pourrait être implémenté en pénalisant les changements rapides de direction des mouvements des articulateurs ; mais à notre connaissance personne n'est réellement parvenu à implémenter ce type de contrainte.

La recherche de trajectoires elle-même s'effectue dans notre cas en deux phases : une première phase de lissage non-linéaire recherche parmi les solutions obtenues à l'aide du codebook une trajectoire initiale ; une seconde phase de régularisation variationnelle (Laprie & Mathieu 1998b) permet d'améliorer cette trajectoire initiale, tant au niveau de la proximité acoustique des solutions que de la régularité de la courbe.

4.1 Lissage non-linéaire

Notre algorithme de lissage non-linéaire est dérivé de l'algorithme de Ney (Ney 1983). Il s'agit d'un algorithme de programmation dynamique avec la capacité supplémentaire de ne pas forcément choisir une solution à chaque instant. Cette propriété nous permet de pallier les vecteurs acoustiques aberrants qui peuvent parfois survenir lorsque l'on fait du suivi de formants de façon automatique, surtout lorsque le signal est bruité, ou les imperfections du codebook (en particulier dans le cas de consonnes lorsque l'on cherche à inverser des séquences VCV).

En d'autres termes, l'objectif de l'algorithme est de choisir, parmi l'ensemble S des solutions possibles, $S = \{S(0)...S(t)...S(T)\}$ (où $S(t)$ est l'ensemble des solutions trouvées à l'instant t), une trajectoire $s = \{\alpha_{j(0)}...s_{j(k)}...\alpha_{j(K)}\}$, où $K \leq T$, j est une fonction positive entière strictement croissante $0 \leq j(k) < j(k+1) \leq T$, et $\alpha_{j(k)}$ est un vecteur articulatoire élément de $S_{j(k)}$.

Le but de cet algorithme est de choisir, parmi les solutions trouvées grâce au codebook, une solution qui optimise un critère généralement basé sur un critère de régularité locale de la trajectoire. Nous verrons dans les parties suivantes de cette thèse d'autres critères basés sur des contraintes issues de connaissances humaines, puis sur une approche multimodale de l'inversion. Le critère de régularité peut prendre plusieurs formes ; il est en général basé sur une vision biodynamique des paramètres articulatoires du modèle. Généralement, on considère que chaque articulateur α_i possède une certaine masse m_i, et on peut alors chercher à minimiser l'énergie dépensée grâce aux formules habituelles de mécanique du solide. Cependant, ce genre d'approche est difficilement défendable, car les modèles utilisés, bien que généralement basés sur des données réelles, ont souvent des commandes articulatoires déterminées grâce à des analyses statistiques ou géométriques de données articulatoires ; les seuls modèles qui pourraient se prévaloir de telles contraintes sont les modèles biomécaniques, qui ont un nombre de commandes très élevé, et en général très peu d'informations sur l'activation musculaire ou même simplement sur la dynamique réelle des articulateurs correspondants.

Le modèle de Maeda est également basé sur une étude statistique de données réelles, mais conduite dans l'optique d'obtenir des commandes articulatoires qui correspondent le mieux possible aux articulateurs classiques des phonéticiens, tout en ayant un nombre restreint de paramètres de commande. Ces commandes ne sont pas pour autant assimilables à des commandes d'objets physiques, bien qu'elles s'en approchent davantage que les commandes d'autres modèles, tel celui de Mermelstein (Mermelstein 1973). Il est par conséquent difficile de justifier l'utilisation d'une contrainte basée sur la minimisation d'une pseudo-énergie potentielle du système, qui ne correspondrait pas à une réalité physiologique.

4.1.1 Critères sur la régularité de la trajectoire des articulateurs

Les critères que nous utilisons ici se basent ainsi simplement sur une régularité des trajectoires des articulateurs. La régularité peut être envisagée à plusieurs niveaux : on peut envisager de minimiser l'écart à la position neutre des articulateurs (i.e. $\int |\alpha(t)|dt$), le « mouvement » des articulateurs (i.e. $\int \left|\frac{d\alpha(t)}{dt}\right| dt$), la « vitesse » ($\int \left|\frac{d^2\alpha(t)}{dt^2}\right| dt$), et de façon générale les combinaisons des différentes dérivées de la position :

$$\sum_{n=0}^{D} \beta_n \int \left|\frac{d^n\alpha(t)}{dt^n}\right| dt, \tag{4.1}$$

où les β_n sont des coefficients de pondération pour les différentes dérivées, et D le degré de dérivation maximal envisagé.

Schoentgen (Maeda *et al.* 2006) considère, tout comme une bonne partie des biomécaniciens (Flash & Hogan 1985; Engelbrecht 2001) que le critère le plus pertinent physiologiquement est celui de la « secousse », c'est-à-dire celui correspondant à la dérivée troisième de la position. Dans notre cas, nous utilisons une combinaison linéaire des différents critères (dérivées de la position aux ordres 0 à 3).

L'inversion s'effectuant à des instants discrets, les intégrales sur le temps sont remplacées par des sommes discrètes, et les dérivées en un point donné sont évaluées à partir des valeurs voisines par les schémas classiques de l'analyse numérique, c'est-à-dire en approchant la fonction par un polynôme d'interpolation sur le support le plus proche possible du point à évaluer. En pratique, nous utilisons les formules décentrées « à gauche », c'est-à-dire :

$$f^{(n)}(t_{m+n}) = P_n^{(n)}\{f(t_m), f(t_{m+1}), \ldots, f(t_{m+n})\}(t_{m+n}),$$

où $f^{(n)}$ désigne la dérivée n^{e} d'une fonction f, et P_n est le polynôme d'interpolation de degré n déterminé sur le support donné. En pratique, n sera dans notre cas toujours inférieur ou égal à 3.

L'utilisation de la formule décentrée à gauche est rendue nécessaire pour conserver les hypothèses nécessaires à l'application de la programmation dynamique ou du lissage non-linéaire. Le calcul des dérivées successives est en revanche assez peu précis en utilisant les formules décentrées. La dérivée n^{e} du polynôme de degré n se calcule facilement grâce à la formule des différences divisées : si on désigne par $y_0, \ldots, y_i, \ldots, y_n$ les images respectives par f de $t_m, \ldots, t_{m+i}, \ldots, t_{m+n}$, alors on peut calculer récursivement :

$$\begin{aligned} [y_\nu] &:= y_\nu & \nu &= 0, \ldots, n \\ [y_\nu, \ldots, y_{\nu+j}] &:= \frac{[y_{\nu+1}, \ldots y_{\nu+j}] - [y_\nu, \ldots y_{\nu+j-1}]}{t_{m+\nu+j} - t_{m+\nu}} & j &= 1, \ldots, n-1, \nu = 0, \ldots, n-j. \end{aligned}$$

En utilisant les polynômes d'interpolation de Newton, on en déduit immédiatement la dérivée n^{e} du polynôme d'interpolation de degré n :

$$f^{(n)}(t_{m+n}) = n! * [y_0, \ldots, y_n].$$

4.1.2 Critère global à minimiser

En se basant uniquement sur les critères de régularité des trajectoires des articulateurs vues précédemment, et en discrétisant temporellement l'équation 4.1, on obtient une fonction de coût global, c'est-à-dire correspondant à une séquence articulatoire complète, de la forme suivante :

$$C(\alpha, T) = \sum_{i=1}^{T} (t_i - t_{i-1}) \sum_{d=0}^{\min(i,D)} \beta_d \left| d! * [\alpha(i-d), \ldots, \alpha(i)] \right|_{\text{Ar}},$$

où les β_d sont des coefficients de pondération, $||_{\text{Ar}}$ une norme sur les vecteurs articulatoires, et D le degré maximum de dérivation que l'on souhaite considérer (typiquement égal à 3). Pour simplifier l'écriture, on désigne par $c(\alpha, i)$ le coût local, i.e. :

$$c(\alpha, i) = (t_i - t_{i-1}) \sum_{d=0}^{\min(i,D)} \beta_d \left| d! * [\alpha(i-d), \ldots, \alpha(i)] \right|_{\text{Ar}}.$$

Ainsi,

$$C(\alpha, T) = \sum_{i=1}^{T} c(\alpha, i).$$

La minimisation de cette fonction de coût peut être réalisée par une forme de programmation dynamique, car elle présente une propriété de sous-structure optimale (ou, en d'autre termes, les problèmes à l'instant $i+1$ peuvent être résolus à partir de la solution des problèmes à l'instant i).

En effet, si on désigne par $\mathrm{A}(i), i \geq D$ un vecteur de $S(i-D) \times \ldots \times S(i)$, et par $m(i, \mathrm{A}(i))$ le coût minimum d'une trajectoire articulatoire se finissant par $\mathrm{A}(i)$, i.e. :

$$m(i, \mathrm{A}(i)) = \min_{\alpha \in S(0) \times \ldots \times S(i-D-1)} C((\alpha, \mathrm{A}(i)), i),$$

alors on a la relation suivante :

$$\begin{aligned}
m(i+1, \mathrm{A}(i+1)) &= \min_{\alpha \in S(0) \times \ldots \times S(i-D)} C((\alpha, \mathrm{A}(i+1)), i+1) \\
&= \min_{\alpha \in S(0) \times \ldots \times S(i-D)} (C((\alpha, \mathrm{A}(i+1)), i) + c((\alpha, \mathrm{A}(i+1)), i+1)) \\
&= \min_{\alpha \in S(0) \times \ldots \times S(i-D-1)} \left(\min_{y_{i-D} \in S(i-D)} C((\alpha, y_{i-d}, \mathrm{A}(i+1)), i) \right) + c(\mathrm{A}(i+1), i+1) \\
&= c(\mathrm{A}(i+1), i+1) + \min_{y_{i-D} \in S(i-D)} m(i, (y_{i-D}, y_{i-D+1}, \ldots, y_i)).
\end{aligned}$$

On voit donc que les minima de la fonction de coût à un instant donné peuvent se calculer à partir des minima de l'instant précédent. Le minimum global est simplement :

$$\min_{\mathrm{A}(T) \in S(T-D) \times \ldots \times S(T)} m(T, \mathrm{A}(T))$$

Il est cependant nécessaire de parcourir et de calculer, pour chaque instant i, l'ensemble des $m(i, \mathrm{A}(i))$. La complexité du calcul d'un $m(i, \mathrm{A}(i))$ particulier est de $|S(i-D)|$, ce qui donne, pour calculer l'ensemble des $m(i, \mathrm{A}(i))$, une complexité de $\Pi_{k=0}^{D} |S(i-k)|$.

Ceci correspondait à l'analyse de la fonction de coût dans le cas simple où l'on sélectionne une solution à chaque instant. Pour le lissage non linéaire, la formule de la fonction de coût à minimiser devient :

$$C(j, \alpha) = \sum_{k=1}^{K} (t_{j(k)} - t_{j(k-1)}) \sum_{d=0}^{D} \beta_d \left| d! * [\alpha(j(k-d)), \ldots, \alpha(j(k))] \right|_{\mathrm{Ar}}$$

L'objectif consiste à trouver une fonction positive strictement croissante j et un ensemble de solutions α qui permettent de minimiser la fonction de coût C. Bien entendu, cette fonction de coût n'est pas suffisante : pour que l'algorithme de Ney fonctionne, il est nécessaire d'ajouter un bonus pour favoriser la sélection du maximum d'instants, sinon la solution optimale que l'on trouve est systématiquement la solution triviale qui ne sélectionne aucun instant.

Nous ajoutons donc un bonus strictement positif B pour chaque instant préservé dans la trajectoire ; nous obtenons ainsi une fonction de coût de la forme :

$$C(j, \alpha) = \sum_{k=1}^{K} (t_{j(k)} - t_{j(k-1)}) \left(\sum_{d=0}^{D} \beta_d \left| d! * [\alpha(j(k-d)), \ldots, \alpha(j(k))] \right|_{\mathrm{Ar}} - B \right)$$

Il reste à déterminer les différents poids à donner à chacun des paramètres : D, B, et les β_d. Ainsi qu'à préciser la forme de la norme $||_{\mathrm{Ar}}$, qui transforme un vecteur articulatoire en un scalaire. En général, faute de données précises, nous utilisons simplement la norme euclidienne, mais il pourrait être envisagé d'accorder un poids plus fort pour les composantes qui correspondent aux paramètres articulatoires les plus significatifs, tel que celui correspondant aux mouvements de la mâchoire.

La valeur de D, comme nous l'avons déjà évoqué, est inférieure ou égale à 3, et est en pratique souvent limitée à 1. En général, seul $\beta_D = 1$, les autres sont égaux à 0. B est déterminé de façon à ne pénaliser que les mouvements très brusques.

4.1.3 Complexité

L'un des éléments à prendre en compte pour déterminer le degré de dérivation maximal est la complexité résultante de l'algorithme. En effet, même sans parler de sélection des instants – propre au lissage non-linéaire, – comparé à un algorithme de programmation dynamique classique, la complexité théorique de notre algorithme est importante, surtout si le niveau maximal de dérivation D est élevé. Si on désigne par M un majorant du nombre de solutions à chaque instant, il est nécessaire de parcourir M^D solutions pour chaque forme, soit au total une complexité globale en $O(M^{D+1} * T)$, T étant le nombre d'instants. Le lissage non-linéaire augmente encore considérablement la complexité, puisque l'on passe alors dans le cas général à une complexité en $O((M * T)^{D+1} * T)$. En pratique, on n'utilise pas le cas le plus général de lissage non-linéaire, on restreint le parcours à une fenêtre temporelle variable (dont la taille dépend du nombre de solutions dans le voisinage de l'instant présent), et diverses optimisations techniques permettent de limiter encore davantage la complexité réelle.

4.2 Régularisation variationnelle

L'algorithme de lissage non-linéaire permet, parmi un sous-ensemble discret des solutions possibles, de retrouver la trajectoire minimisant un certain critère. Néanmoins, la trajectoire trouvée n'est pas la meilleure dans l'absolu ; il ne s'agit que d'une approximation d'une trajectoire idéale, la qualité de l'approximation dépendant de la précision acoustique et de la densité des points générés lors de la première étape de l'inversion. La trajectoire trouvée à l'issue de l'étape de lissage non-linéaire est simplement une trajectoire initiale, que l'on améliore ensuite grâce à un algorithme d'optimisation reposant sur le calcul variationnel (Laprie & Mathieu 1998b).

La trajectoire initiale présente plusieurs défauts : d'une part, la précision acoustique de la trajectoire retrouvée est au mieux celle du codebook, et d'autre part la densité des points est nécessairement assez faible pour que l'algorithme de lissage non-linéaire conserve un temps d'exécution raisonnable.

L'algorithme de régularisation variationnelle optimise comme précédemment une fonction de coût basée sur la régularité de la trajectoire, en ajoutant un critère de proximité acoustique des images. La précision ne dépend plus de celle du codebook, car l'algorithme exploite directement le modèle articulatoire. Le critère lié à la régularité de la trajectoire que l'on prend est, par soucis de cohérence, en général le même que celui utilisé pour sélectionner la trajectoire discrète, mais il est également possible de prendre un critère légèrement différent.

La fonction de coût à minimiser est ainsi de la forme suivante :

$$I = \sum_{t_0}^{t_f} |f(t) - F(\alpha(t))| + \lambda \sum_{t_0}^{t_f} C(\alpha(t), \alpha'(t), \ldots, \alpha^{(n)}(t)),$$

le paramètre λ permettant de contrôler le compromis entre la qualité acoustique et la qualité des trajectoires articulatoires.

La fonction I est minimisée grâce au calcul variationnel, qui est un processus itératif. L'algorithme améliore la solution initiale itérativement jusqu'à obtenir un minimum ; il n'y a malheureusement aucune garantie que ce minimum local soit le minimum global recherché. La solution initiale a donc une importance capitale, et a tout intérêt a être aussi proche que possible du minimum global. Celle issue du lissage non-linéaire est un bon candidat, puisqu'elle réalise un minimum global (mais dans un espace des solutions discrétisé et avec une erreur acoustique relativement importante) : on peut espérer que la solution trouvée par le lissage non-linéaire soit effectivement proche du minimum global réel, même si on n'en a aucune garantie dans le cas général. Par ailleurs, en supposant que l'on parvienne à trouver effectivement le minimum global, on n'a aucune garantie qu'il s'agisse de la véritable trajectoire articulatoire du locuteur.

L'algorithme de lissage par calcul variationnel tel que présenté par Bruno Mathieu dans sa thèse (Mathieu 1999) a été légèrement modifié. Il supposait en effet implicitement que les instants pour lesquels on dispose d'une solution initiale sont équirépartis, et avec une fréquence d'échantillonnage constante. Cette hypothèse était un peu contradictoire avec le lissage non-linéaire, et pouvait être problématique dans les zones où le vecteur acoustique étudié n'était pas défini (par exemple un vecteur acoustique composé des trois premières fréquences formantiques dans le cas des fricatives sourdes). Nous avons donc modifié l'algorithme de façon à permettre l'utilisation d'un échantillonnage non régulier.

4.3 Inversion dynamique

L'inversion dynamique est la composante la plus intéressante de l'inversion. La principale difficulté de notre méthode (comme de la majorité des méthodes d'inversion) est que l'on manque de références articulatoires auxquelles comparer les résultats. Dans de nombreux cas, cette difficulté est ignorée en ne mesurant pas l'écart par rapport à une référence articulatoire, mais tout simplement en ne prenant en compte que l'écart du signal acoustique resynthétisé par rapport au signal acoustique de départ. Cela permet notamment d'éliminer les problèmes liés à l'inexactitude du synthétiseur articulatoire (la synthèse des vecteurs articulatoires mesurés sur les radiographies ne permet pas d'obtenir exactement les vecteurs acoustiques mesurés), ainsi que le fait que l'inversion acoustique articulatoire est un problème mal posé : il existe une infinité de configurations de conduits vocaux permettant d'obtenir un vecteur acoustique donné. Nous désignerons cette façon d'inverser la version « faible » de l'inversion. Un autre problème, légèrement plus simple que le problème général, est l'inversion de la synthèse acoustique de trajectoires articulatoires. Cela permet de ne pas souffrir des insuffisances du système de synthèse. Nous appellerons ce problème l'inversion « moyenne ». Nous nous intéressons ici surtout à la version « forte » du problème : nous tentons de retrouver les configurations articulatoires initiales à partir du signal sonore réel.

Combinées à l'inversion statique vues dans le chapitre 3, les étapes de lissage non-linéaire et de régularisation variationnelle nous permettent d'obtenir des trajectoires articulatoires uniques et lisses. Nous présentons dans cette section quelques expériences d'inversion dynamiques. Il est en général difficile de déterminer si les trajectoires trouvées sont conformes à la réalité, car on ne dispose pas, le plus souvent, de données articulatoires très précises.

Pour les premières expériences présentées, nous disposons de la référence (des données articulatoires acquises en cinéradiographie). Pour éviter toute perturbation liées au modèle, nous n'inversons pas le signal naturel, mais un signal artificiel généré à partir des données articulatoires. Nous étudierons l'influence de la taille du vecteur acoustique sur la solution trouvée.

Les autres expériences réalisées concernent des données pour lesquelles on dispose de beaucoup moins d'informations. Dans le premier cas, il s'agira du même sujet que dans l'expérience précédente, mais en travaillant cette fois-ci sur les données acoustiques réelles. Dans le deuxième cas, il s'agira d'un tout autre sujet, pour lequel on dispose d'informations sur la position de marqueurs électromagnétiques.

4.3.1 Inversion de la synthèse acoustique

Nous avons utilisé pour cette expérience les paramètres articulatoires de la locutrice PB, gracieusement fournis par Shinji Maeda (cf. Annexe 1), et obtenus à partir de données cinéradiographiques de l'Institut de Phonétique de Strasbourg (Bothorel *et al.* 1986). Nous avons synthétisé, à l'aide du synthétiseur intégré au modèle articulatoire, les 5 premiers formants correspondant à ces paramètres articulatoires, et nous avons pratiqué l'inversion sur un nombre décroissant de fréquences formantiques. Certaines configurations articulatoires correspondant à des fermetures et le modèle de synthèse utilisé, statique, ne permettant pas d'y associer une image acoustique, l'inversion ne peut être effectuée que pour les configurations du conduit vocal sans occlusion.

Dans cette expérience, nous quantifions la distance entre deux formes de conduit vocal de deux façons différentes : une *distance articulatoire* (notée d_1) d'une part, qui est simplement la distance quadratique moyenne entre deux vecteurs articulatoires, et une *distance géométrique* (notée d_2) d'autre part –basée sur la projection de la forme de conduit vocal sur la grille de Maeda,– consistant en une moyenne quadratique des distances entre points analogues.

En d'autres termes :

$$d_1(X,Y) = \sqrt{\frac{\sum_{i=1}^{7}(X_i - Y_i)^2}{7}},$$

où X et Y sont deux vecteurs articulatoires, et

$$d_2(X,Y) = \sqrt{\frac{\sum_{j=1}^{N}|P(X)_j - P(Y)_j|^2}{N}},$$

où P désigne l'opérateur de projection d'un vecteur articulatoire vers la grille de Maeda, N est le nombre de points de la grille, et $P(X)_j$ désigne l'un des projetés du vecteur articulatoire X sur la grille.

4.3.1.1 Inversion sur 5 formants

Nous présentons ici les résultats obtenus en effectuant l'inversion de la première phrase du corpus, « Ma chemise est roussie ». Les vecteurs acoustiques sont de dimension 5 (il s'agit des 5 premières fréquences formantiques), et le pas d'échantillonnage est 20ms. Rappelons que les configurations articulatoires correspondant à un conduit fermé n'ont pas d'image acoustique.

La figure 4.1a présente les résultats de l'inversion pour le paramètre articulatoire correspondant à l'ouverture des lèvres. Sur cette figure sont représentées : la trajectoire articulatoire originale (croix vertes), les solutions de l'inversion statique (points noirs), la trajectoire obtenue à l'issue du lissage non-linéaire (étoiles bleues), et enfin la trajectoire obtenue après régulation

variationnelle. On constate sur cette figure que la similitude entre les courbes est presque parfaite... même les segments n'ayant pas d'image acoustique (notamment [20-170] et [340-420]) ne posent pas de problème. On constate simplement un décrochement non négligeable (une erreur d'une demi-unité) autour de 500ms. On observe ici également certaines limitations du codebook : pour certains vecteurs acoustiques, on ne trouve ici aucun antécédent (notamment aux instants 20 et 120). Avant l'étape de régulation variationnelle, des « solutions » pour ces vecteurs sont réintroduites en interpolant linéairement à partir des solutions trouvées pour les instants adjacents.

Si la trajectoire articulatoire trouvée pour le paramètre d'ouverture des lèvres est très satisfaisant, il n'en est pas de même pour tous les paramètres. Les paramètres les plus pertinents pour la phrase prononcée sont ici généralement fidèles, sauf pour le paramètre de protrusion des lèvres, qui est assez mauvais. Ici, il semble que l'erreur soit en partie liée à la mesure du paramètre lui-même, les mouvements importants des lèvres d'une trame sur l'autre sur les formes de conduit associées ne paraissant pas réalistes. D'autre part, il a été observé (Mawass *et al.* 2000) que dans certaines zones de l'espace articulatoire, la variation du paramètre de protrusion a une influence négligeable sur l'acoustique, et on peut observer que son effet sur les trois premiers formants est effectivement négligeable par rapport aux autres paramètres (cf. section 2.1.4). Néanmoins, la trajectoire retrouvée est très éloignée de l'original, ce qui indique un défaut de la méthode d'inversion.

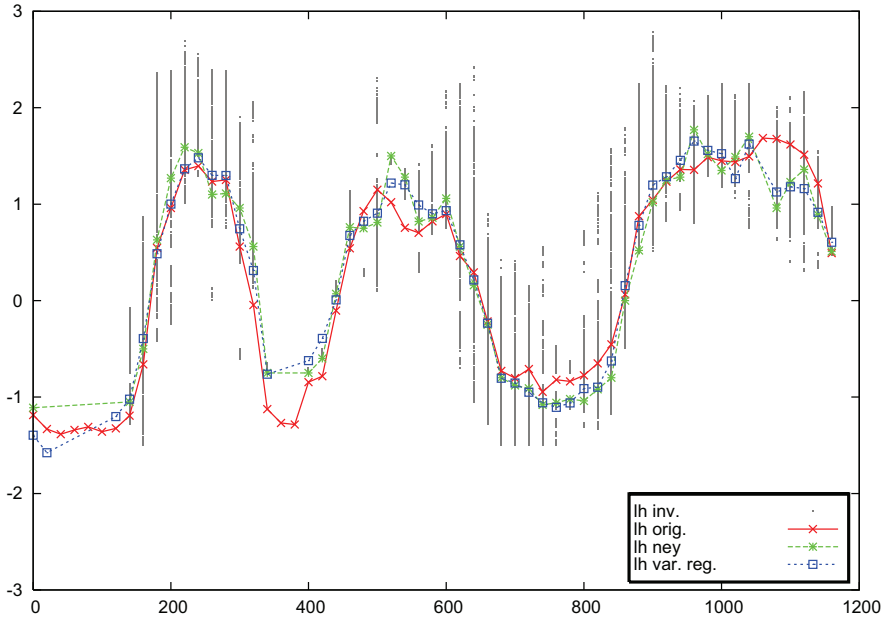
Le tableau 4.1 présente les résultats de l'inversion de façon plus quantitative. Nous y présentons les distances d_1 et d_2 moyennes pour les différents étapes de l'inversion : l'inversion statique (*Inv.*), le lissage non-linéaire (*Ney*) et la régulation variationnelle (*Var.*). Dans le cas de l'inversion statique, il s'agit de la moyenne des distances minimales (c'est-à-dire que pour chaque instant on a déterminé la forme de conduit minimisant chacune des distances parmi toutes les solutions de l'inversion statique). On constate que parmi les solutions de l'inversion statique on a des solutions très proches de l'original, mais qu'elles ne sont malheureusement pas toujours retenues lors du lissage non-linéaire. La régularisation variationnelle augmente de façon importante la fidélité à l'original, mais on voit que la solution finale est toujours assez loin de la solution originale, et est nettement moins bonne que la meilleure des solutions de l'inversion statique. Cela indique qu'avec de meilleures contraintes on pourrait probablement encore améliorer les résultats. On constate également que l'erreur géométrique est très faible : l'erreur moyenne est de l'ordre de 1mm.

On observe également que la solution trouvée par régulation variationnelle présente des trajectoires très fidèles à l'original pour chacun des paramètres (sauf pour la protrusion des lèvres, et dans une moindre mesure pour le paramètre contrôlant la pointe de la langue, non présenté ici), et systématiquement meilleures que celles trouvées par lissage non-linéaire.

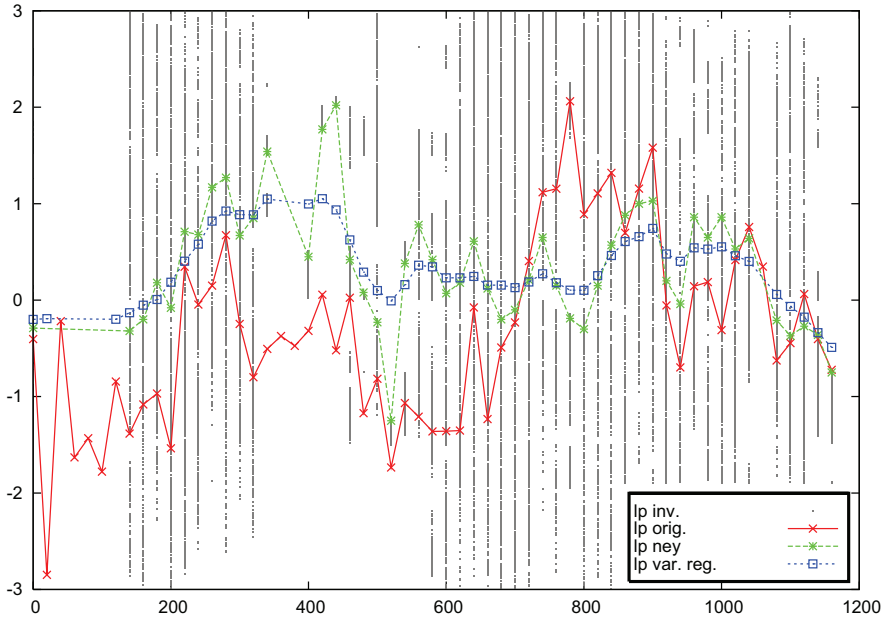
4.3.1.2 Inversion sur 4 ou 3 formants

La même expérience que précédemment est répétée, mais en utilisant comme vecteur acoustique les 4 ou 3 premiers formants.

Le tableau 4.2 présente les erreurs RMS moyennes sur les paramètres articulatoires pour différentes expériences d'inversion. Les paramètres de l'inversion statique ont été choisis de façon à ce que la distance minimale à la solution optimale soit la même dans chaque cas (cf. section 3.2.3). La colonne *Inv.* correspond à l'erreur articulatoire moyenne sur les solutions issues de l'inversion statique à partir du codebook. La colonne *Opt.* correspond à la moyenne sur toute la séquence de la distance articulatoire minimale entre les solutions de l'inversion statique et la trajectoire originale. La colonne *Ney* correspond à l'erreur moyenne mesurée à l'issue du lissage



(a) Ouverture des lèvres



(b) Protrusion des lèvres

FIG. 4.1: *Paramètres d'ouverture et de protrusion des lèvres, original et retrouvé par inversion, pour la première phrase du corpus de PB, « Ma chemise est roussie ». Les points noirs éparpillés correspondent aux solutions de l'inversion statique, la trajectoire articulatoire originale est représentée par des croix rouges, la trajectoire articulatoire trouvée par lissage non-linéaire par des étoiles vertes, celle trouvée après régulation variationnelle par des carrés bleus. Les ordonnées sont en unités de paramètre articulatoire, l'abscisse en ms.*

Err.	d_1	d_2 (cm)	jw	tp	lh	lp
Inv.	0.32	0.07	1.37	0.59	0.54	1.72
Ney	0.53	0.12	0.60	0.32	0.26	1.10
Var.	0.43	0.10	0.38	0.19	0.21	1.06

TAB. 4.1: *Distance quadratique moyenne aux données articulatoires originales pour les solutions de l'inversion statique, la solution issue du lissage non linéaire, et celle issue de la régulation variationnelle (pour le cas de l'inversion statique, il s'agit de la moyenne des distances minimales). d_1 est exprimé en unité de paramètre articulatoire, d_2 en cm. On présente également la distance quadratique moyenne pour les 4 paramètres du modèle de Maeda les plus pertinents : ouverture de la mâchoire, position de la langue, ouverture et protrusion des lèvres.*

NF	Inv.	Opt.	Ney	Liss. ac.	Comb.
3	1.46	0.49	0.69	0.68	0.51
4	1.21	0.42	0.65	0.61	0.38
5	1.02	0.39	0.45	0.42	0.30

TAB. 4.2: *Distance articulatoire à l'originale pour différents types de solutions, en fonction de la taille du vecteur acoustique.*

non-linéaire, la colonne *Liss. ac.* correspond à l'erreur moyenne mesurée suite à une optimisation de la contrainte acoustique sur la trajectoire issue du lissage non-linéaire, et enfin la colonne *Comb.* correspond à l'erreur moyenne mesurée à l'issue de la régulation variationnelle avec une combinaison de contraintes acoustiques et dynamiques.

Plusieurs remarques peuvent être formulées :

1. La distance articulatoire moyenne des solutions issues de l'inversion statique avec l'originale décroît avec la dimension du vecteur acoustique, ce qui est tout à fait le comportement attendu. Il est à noter que même avec 5 formants, la distance moyenne est toujours d'une unité, ce qui montre l'étendue des phénomènes compensatoires. Cette mesure globale masque cependant de profondes disparités entre les différents paramètres articulatoires.
2. La distance articulatoire minimale décroît également lorsque la dimension du vecteur articulatoire augmente. Il s'agit de la borne inférieure que l'on peut espérer atteindre à l'aide du lissage non-linéaire (avec les paramètres utilisés pour cette expérience... on peut bien entendu augmenter la précision en augmentant le nombre de points générés lors de l'inversion par codebook).
3. Le lissage acoustique de la solution issue du lissage non-linéaire des solutions de l'inversion statique à l'aide du codebook se contente d'éliminer l'erreur acoustique liée à l'utilisation du codebook ; l'erreur acoustique passe alors de quelques dizaines de Hz à moins de 0.01 Hz. Il s'agit simplement d'une régularisation variationnelle avec des poids nuls pour les termes articulatoires. On peut noter une réduction non négligeable de l'erreur articulatoire, sans que cette réduction soit considérable. Cela semble indiquer que les gains que l'on peut espérer d'une meilleure précision acoustique du codebook sont certainement relativement faibles.
4. La dernière colonne indique l'erreur à l'issue de la régulation variationnelle effectuée avec une combinaison de contraintes acoustiques et dynamiques. On constate cette fois ci une

amélioration substantielle de la précision, en particulier sur l'inversion avec un vecteur acoustique de dimension 4.

On constate qu'avec 4 formants, les résultats sont au final pratiquement aussi bons que ceux obtenus avec 5 : les erreurs au niveau de l'inversion générale et de Ney sont nettement plus importantes qu'avec 5 formants, mais la régulation variationnelle permet de réduire considérablement cette erreur. On observe en revanche qu'avec 3 formants la solution finale reste assez éloignée de l'originale.

Dans tous les cas, on constate qu'en particulier lorsque l'on utilise un nombre restreint de formants, la contrainte acoustique est loin d'être suffisante pour retrouver les trajectoires articulatoires initiales.

4.4 Conclusion

Les algorithmes de lissages non-linéaires de de régulation variationnelle utilisés permettent de faire de l'inversion de séquences de parole et trouvent une solution unique. Dans des conditions idéales, on parvient à trouver des trajectoires articulatoires très proches de l'originale lorsque l'on dispose d'un vecteur acoustique de dimension élevée. Malheureusement, les hypothèses formulées (adéquation parfaite entre le modèle articulatoire et le locuteur inversé, obtention de vecteurs acoustiques de dimension élevée sans aucune erreur) ne sont guère réalistes. Lorsque l'on s'écarte de ce cadre idéal, on s'aperçoit rapidement que les contraintes sur la dynamique des paramètres articulatoires ne sont plus suffisantes. La nécessité de développer des contraintes alternatives apparaît ainsi très clairement.

Deuxième partie

Contraintes

Introduction

Comme nous l'avons vu précédemment, la contrainte acoustique est insuffisante pour garantir l'existence d'un unique vecteur articulatoire solution de l'inversion : un vecteur acoustique de dimension élevée est trop sensible aux défauts du synthétiseur articulatoire ou au bruit dans le signal de parole, un vecteur acoustique de dimension trop faible caractérise mal la forme du conduit vocal. Il est donc nécessaire d'utiliser d'autres contraintes.

Les contraintes généralement utilisées sont d'ordre biomécanique, et une solution unique est déduite en minimisant une fonction de coût représentant la pseudo-énergie du système. Or, comme nous l'avons déjà montré, ces contraintes ne sont pas toujours appropriées, et ne correspondent pas forcément à la réalité des trajectoires articulatoires que l'on cherche à retrouver : d'une part, l'utilisation de ce type de contrainte pour les modèles articulatoires non biomécaniques (tel celui de Maeda, que nous utilisons) est difficilement justifiable ; d'autre part, le critère à minimiser fait encore débat dans la communauté des dynamiciens même pour les mouvements les plus simples (Engelbrecht 2001), et concernant les mouvements de parole, on ne dispose pour le moment, d'aucune validation empirique de cette hypothèse. Il est donc souhaitable d'étudier des types de contraintes alternatifs. Dans cette thèse, deux types de contraintes, exploitant l'information contextuelle (implicite ou explicite), sont présentés.

Le premier type de contraintes étudié ici se base sur l'invariabilité articulatoire empirique : il a en effet été observé depuis fort longtemps par les phonéticiens que certaines caractéristiques articulatoires sont nécessaires pour articuler certains sons ; par exemple, la fermeture labiale pour les consonnes /p,b,m/, ou la protrusion des lèvres pour les voyelles /y,u/ du français. Cette notion de « caractéristiques articulatoires nécessaires » permet de définir, en fonction des phonèmes, des zones articulatoires ayant une plus forte probabilité d'apparaître.

Nous avons ainsi élaboré une méthode qui attribue, à chaque forme de conduit, un score de confiance en fonction de la distance aux caractéristiques articulatoires attendues pour le son correspondant.

Le chapitre 5 présente la méthode et la façon dont elle s'intègre au processus d'inversion, le chapitre 6 présente différentes expériences effectuées à l'aide de ces contraintes phonétiques dans l'optique de les évaluer.

Nous présenterons dans le chapitre 7 un autre type de contraintes : il s'agit de contraintes exploitant l'information contextuelle visuelle.

5

Contraintes phonétiques

5.1 Principe

La méthode présentée ici consiste à utiliser des connaissances phonétiques pour attribuer un score de confiance aux formes articulatoires trouvées par l'inversion. En substance, cela revient à donner un meilleur score aux formes les plus « conformes » du point de vue phonétique compte tenu du son prononcé. la méthode développée fait appel à deux modules essentiels.

Chaque phonème peut être décrit en fonction de critères articulatoires phonétiquement pertinents. Ces critères sont traduits en tant que contraintes sur les paramètres articulatoires du modèle utilisé et définissent des domaines articulatoires privilégiés. On disposera ainsi, pour chaque phonème, d'un ensemble, ou *classe*, de contraintes.

Un deuxième module, dit acoustique, va déterminer la classe de contraintes à appliquer en fonction du son que l'on souhaite inverser, ce qui, dans notre cas, peut se ramener à déterminer le phonème prononcé.

Chaque solution trouvée par l'inversion se voit alors attribuer un score en fonction de la distance du vecteur articulatoire au domaine privilégié associé au phonème reconnu.

5.2 Domaines articulatoires

L'élément le plus important de notre méthode est l'élaboration des classes de contraintes phonétiques à appliquer lors de l'inversion d'un son donné. Nous nous intéresserons ici exclusivement à l'utilisation de contraintes statiques. Les contraintes développées ne prendront donc pas en compte le contexte acoustique temporel – mais cela n'empêchera pas de pouvoir employer ces contraintes pour l'inversion dynamique. Les contraintes étant basées sur des études phonétiques seront définies relativement à la plus petite unité de parole distinctive, c'est-à-dire le phonème. Par ailleurs, de par la nature même des vecteurs acoustiques utilisés – les premiers formants – et des caractéristiques du synthétiseur articulatoire utilisé, notre étude se limitera aux voyelles non nasales du français.

5.2.1 Classification des phonèmes

Dans le cas particulier des voyelles, quatre caractéristiques phonétiques essentielles sont généralement retenues : l'ouverture de la bouche, l'étirement et la protrusion des lèvres, et la position du dos de la langue.

Les domaines articulatoires utilisées dans notre méthode sont essentiellement issues d'une classification des phonèmes du français selon trois de ces quatre caractéristiques. L'étirement des lèvres n'étant pas exprimable dans notre modèle articulatoire, nous ne l'avons pas pris en compte. Cette classification a été élaborée avec l'aide d'Anne Bonneau, à partir d'ouvrages de phonétique de référence (Marchal 1980; Ladefoged 2005). Elle a originellement été élaborée dans l'optique de développer un modèle de coarticulation labiale, et a été validée grâce à des données acquises en stéréovision sur une dizaine de locuteurs français natifs (Robert *et al.* 2005).

Le tableau 5.2.1 présente la classification des 10 voyelles non nasales du français selon ces trois critères. D correspond à la position du dos de la langue, O à l'ouverture de la bouche, et P à la protrusion des lèvres. Le codage est relativement simple : plus le chiffre suivant la lettre est élevé, plus la valeur de la contrainte associée est élevée. Par exemple P1 correspond à une protrusion très faible, tandis que P4 correspond à une protrusion très forte. Cette classification rend compte de l'articulation moyenne de locuteurs français.

On peut remarquer que pour le lieu principal d'articulation (qui correspond à D dans le cas des voyelles), le domaine des valeurs possibles est un sous-domaine des valeurs acceptables pour les consonnes (de 1 pour /p,b,m/ à 9 pour /ʁ, ɹ/). La position 1 correspond à une articulation labiale, 2 à une position dentale, 3 alvéo-dentale, 4 alvéolaire, 5 post-alvéolaire, 6 palatale, 7 post-palatale, 8 vélaire, 9 uvulaire. D ne varie qu'entre 6 et 8 pour les voyelles.

Voyelle	Dos de la langue	Ouverture	Protrusion
i	D6	O1	P1
e	D6	O2	P1
ɛ	D6	O3	P1
a	D7	O4	P1
y	D6	O1	P4
ø	D6	O2	P3
œ	D6	O3	P2
u	D8	O1	P4
o	D8	O2	P3
ɔ	D8	O3	P2

TAB. 5.1: *Classification des voyelles du français selon trois critères phonétiques. D correspond à la position du dos de la langue, O à l'ouverture de la bouche, et P à la protrusion des lèvres.*

5.2.2 Transposition des contraintes phonétiques dans le modèle articulatoire

Pour la majorité des modèles articulatoires, la transposition de caractéristiques phonétiques en des paramètres du modèle peut être assez complexe. Dans le cas du modèle de Maeda (Maeda 1979) en revanche, les paramètres peuvent facilement s'interpréter comme des articulateurs au sens phonétique. Par conséquent, l'expression des contraintes phonétiques sous la forme de paramètres articulatoires est très simple : la protrusion des lèvres et la position du dos de la langue sont déjà des paramètres du modèle, et l'ouverture des lèvres est une combinaison linéaire de deux paramètres (la position de la mâchoire, et l'ouverture intrinsèque des lèvres).

En réalité, l'expression de cette dernière contrainte utilise également, dans notre modèle, la position du dos de la langue, de façon à prendre en compte des effets compensatoires décrits dans (Maeda 1990) : Maeda a observé que pour les voyelles non arrondies /i,a,e/, la position du dos de la langue et l'ouverture de la mâchoire avaient des effets parallèles sur l'image acoustique, et par

conséquent se compensaient mutuellement. Il a également observé que cet effet compensatoire était réellement utilisé par ses sujets de tests. De plus, la direction de compensation ne semblait pas dépendre de la voyelle prononcée : il y avait une corrélation linéaire

$$Tp + \gamma Jw = \text{ Constante},$$

où Tp est la position du dos de la langue, Jw est la position de la mâchoire et γ est le coefficient directeur, qui est le même pour /a/ et /i/. Les autres voyelles n'ont pas été étudiées, car il n'y en avait pas assez d'occurrences dans sa base de données ciné-radiographique. Maeda a observé cette compensation chez ses deux sujets (mais les coefficients de corrélation étaient bien entendu différents). Comme nous travaillons sur une partie de ses données, nous avons repris, pour l'inversion des données de la locutrice PB, le coefficient que Maeda a trouvé expérimentalement (approximativement égal à 0,66). Cet effet compensatoire permettait à Maeda d'expliquer la majeur partie de la variabilité articulatoire intra-locuteur de /a/ et /i/.

L'ouverture de la bouche est ainsi donnée par la relation suivante :

$$\mathrm{O} = \min(Tp + \gamma Jw, Lh),$$

où Lh correspond à l'ouverture des lèvres.

La position du dos de la langue est donnée par

$$\mathrm{D} = Tp$$

et la protrusion des lèvres par

$$\mathrm{P} = Lp.$$

Les paramètres du modèle articulatoire varient typiquement entre -3 et $+3$ unités. Les valeurs des contraintes correspondant à chaque cible articulatoire ont été ajustées en utilisant une représentation visuelle du conduit modèle de façon à obtenir des formes de conduit canoniques correctes. Le tableau 5.2.2 présente les valeurs cibles pour chacune des contraintes.

Dos de la langue	Ouverture de la bouche	Protrusion des lèvres
-1.5	-2.0	-1.0
0.0	-1.0	-0.5
1.5	0.0	0.5
	0.5	1.5

TAB. 5.2: *Valeurs des cibles pour chacune des contraintes phonétiques. La première colonne contient les valeurs respectives de D6, D7, D8, la deuxième les valeurs de O1, O2, O3, O4 et enfin la troisième P1, P2, P3, P4.*

5.3 Domaines acoustiques

La deuxième étape de la méthode consiste à déterminer la classe de contraintes phonétiques à appliquer, en fonction du signal de parole inversé. Ces contraintes vont s'appliquer à tous les vecteurs articulatoires solutions de l'inversion des vecteurs acoustiques extraits de ce signal de parole. En d'autres termes, il s'agit d'élaborer un modèle qui à un vecteur acoustique (ou une séquence de vecteurs acoustiques) associe une classe de contraintes articulatoires (ou une séquence de classes de contraintes). Un tel modèle sera désigné par la suite sous le terme de *modèle*

acoustique. De façon plus générale, un tel modèle permet également d'associer, indépendamment d'un signal de parole, une classe de contraintes phonétiques à un vecteur articulatoire arbitraire, en la déterminant à partir de l'image acoustique de celui-ci.

L'intérêt d'utiliser un module distinct pour déterminer la classe de contraintes phonétiques à appliquer permet à la méthode d'avoir une bien plus grande flexibilité. Nous n'avons malheureusement pas eu l'occasion d'explorer cet aspect en profondeur, mais une modification du modèle acoustique pourrait, par exemple, permettre d'appliquer des contraintes relatives à *l'intention articulatoire*, plutôt qu'à la réalisation acoustique reconnue ; en effet, il est courant, surtout en élocution rapide, de ne pas atteindre les cibles acoustiques relatives au phonème à prononcer – sans que cela pose de problème du point de vue perceptif – avec cependant une modification de l'articulation dans l'intention d'atteindre les cibles articulatoires correspondantes. On ne peut rendre compte de ce phénomène si l'on applique les contraintes phonétiques en supposant que l'objectif articulatoire du locuteur correspond au phonème reconnu dans le signal sonore.

L'utilisation d'un moteur de reconnaissance phonétique tel qu'utilisé classiquement en reconnaissance automatique de la parole ne permettrait pas de déterminer de façon appropriée la classe de contraintes phonétique à appliquer, c'est pour quoi nous ne nous en servons pas. Nous présentons ici un modèle acoustique qui consiste en une partition de l'espace acoustique des fréquences des trois premiers formants en voyelles. Il s'agit d'un modèle acoustique extrêmement simple et assez rudimentaire, mais qui a l'avantage de permettre de contrôler très précisément la méthode, et d'utiliser les mêmes vecteurs acoustiques que pour notre système d'inversion ; il ne s'agit cependant pas de la contribution la plus importante des contraintes phonétiques.

5.3.1 Partitionnement de l'espace acoustique

Pour des raisons pratiques, il est souhaitable de définir un modèle acoustique fondé sur les mêmes vecteurs acoustiques que ceux utilisés pour l'inversion. En effet, si on lie la classe de contraintes à appliquer à un vecteur articulatoire à l'image acoustique de celui-ci, il est nécessaire d'utiliser le synthétiseur articulatoire pour calculer cette image acoustique, ce qui prend un temps conséquent. Mais si le modèle acoustique est basé sur les mêmes vecteurs acoustiques que ceux inversé, et que l'on considère un vecteur articulatoire solution de l'inversion, son image acoustique étant – théoriquement – son antécédent, on peut se passer d'utiliser le synthétiseur articulatoire pour déterminer son image.

Les vecteurs acoustiques que nous utilisons pour l'inversion étant généralement composés des fréquences des trois premiers formants, nous avons défini un modèle de sélection qui est une partition de l'espace des trois premières fréquences formantiques en sous-domaines. Plusieurs modèles de partitionnement ont été testés :

- Un diagramme de Voronoï des valeurs moyennes des fréquences des voyelles.
- Un diagramme de Voronoï des valeurs moyennes des fréquences des voyelles, pondéré par l'écart-type correspondant pour chacune des fréquences formantiques (cf. figure 5.1).

Les diagrammes de Voronoï peuvent être vus comme une partition au plus proche voisinage d'un ensemble de points : tout vecteur de l'espace sera rattaché au point le plus proche au sens d'une certaine métrique. Dans le premier cas, la métrique utilisée est une simple distance euclidienne (notée D_1). Dans le deuxième cas, la métrique utilisée est distance euclidienne pondérée dans chaque direction par l'écart-type de la fréquence formantique correspondante (notée D_2). En d'autres termes :

$$D_1(F,V) = \sqrt{\sum_{i=1}^{3}(F_i - f_i(V))^2},$$

où F désigne un vecteur acoustique, V désigne une voyelle, et $f(V)$ le vecteur acoustique correspondant aux valeurs moyennes de ses fréquences formantiques (cf. tableau 5.3), et :

$$D_2(F,V) = \sqrt{\sum_{i=1}^{3} \left(\frac{F_i - f_i(V)}{\sigma_i(V)} \right)^2},$$

où $\sigma(V)$ désigne le vecteur acoustique correspondant aux écarts-types des fréquences formantiques de la voyelle V.

Dans l'idéal, les données acoustiques à utiliser pour cette modélisation devraient être spécifiques au locuteur ; nous ne disposons malheureusement pas de suffisamment de données sur notre locutrice de référence pour faire cela. Du reste, nous tenons également à obtenir un modèle aussi générique que possible. Les données que nous utilisons sont issues d'une des plus complètes études sur les sons du français, réalisée par Lonchamp (Lonchamp 1984). Les centres des voyelles ainsi que les écarts-types[12] dépendent du genre du locuteur et sont spécifiques au français. De plus amples informations sur les diagrammes de Voronoï peuvent être trouvées dans Aurenhammer (Aurenhammer & Klein 1999).

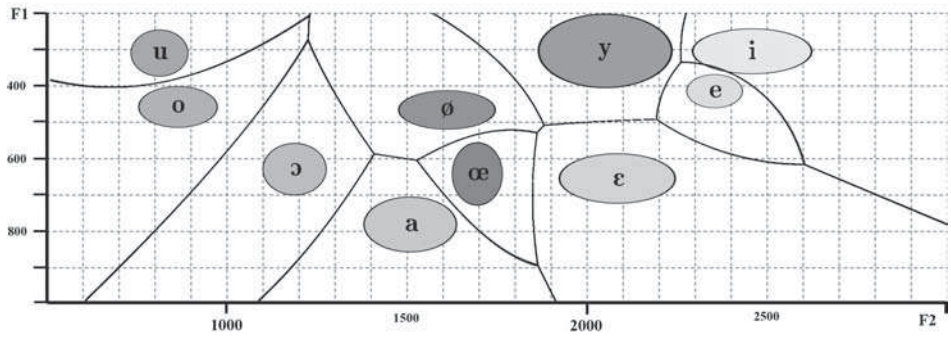


FIG. 5.1: *Partition de l'espace acoustique en utilisant un diagramme de Voronoï pondéré sur les données des voyelles, projeté dans l'espace F1/F2. Les poids sont les écarts-types de chacune des fréquences formantiques*

5.3.2 Données acoustiques

Les données acoustiques sont issues de (Calliope 1989), définies à partir des données de 10 locuteurs et 9 locutrices natifs. Il s'agit d'une des études les plus complètes pour les sons du français. De cette étude, nous n'avons retenu que les voyelles non nasales du français, c'est-à-dire les voyelles /a,e,ɛ,i,o,ɔ,u,y,œ,ø/. Le tableau 5.3 présente les données utilisées : pour chaque voyelle on donne la moyenne et l'écart-type des fréquences des trois premiers formants.

[12] À noter que l'on a priori besoin, pour notre modèle acoustique, des moyennes des écarts-types pour chaque locuteur (en d'autre termes, la moyenne des variabilités intra-locuteur) ; malheureusement, seuls les écarts-types des moyennes formantiques (en d'autre termes, la variabilité inter-locuteurs) sont disponibles, et à notre connaissance il n'existe pas d'étude du français qui dispose de cette donnée ; nous faisons donc ici l'hypothèse audacieuse que les variabilités intra- et inter-locuteurs sont très liées.

Voy.	Fréquence (Hz)			Écart-type (Hz)		
	F1	F2	F3	σF1	σF2	σF3
i	306	2456	3389	42	111	168
e	417	2351	3128	31	52	115
ɛ	660	2080	2954	46	108	156
a	788	1503	2727	51	86	174
ɔ	634	1180	2690	48	59	198
o	461	855	2756	38	73	240
u	311	804	2485	43	53	284
y	305	2046	2535	68	124	139
ø	469	1605	2581	36	90	148
œ	647	1690	2753	58	47	155

(a) Locutrice

Voy.	Fréquence (Hz)			Écart-type (Hz)		
	F1	F2	F3	σF1	σF2	σF3
i	308	2064	2976	34	134	147
e	365	1961	2644	31	119	107
ɛ	530	1718	2558	49	132	103
a	684	1256	2503	47	32	131
ɔ	531	980	2399	39	60	116
o	383	793	2283	22	63	126
u	315	764	2027	43	59	136
y	300	1750	2120	37	121	182
ø	381	1417	2235	44	106	113
œ	517	1391	2379	42	94	91

(b) Locuteur

TAB. 5.3: *Moyenne et écart-type des fréquences des trois premiers formants pour les voyelles du français. D'après (Calliope 1989).*

5.4 Score phonétique

Après avoir choisi un modèle de partitionnement de l'espace acoustique, il nous reste à expliquer comment un « score phonétique » – c'est-à-dire, une évaluation numérique de la pertinence phonétique – peut être associé à chaque solution de l'inversion. En résumé, chaque vecteur acoustique, en fonction de la région de l'espace acoustique à laquelle il appartient, est attaché à un « domaine articulatoire idéal » (défini par les valeurs des contraintes du tableau 5.2.1). À chaque vecteur articulatoire V généré par l'inversion (dont l'image par le synthétiseur articulatoire est donc très proche de ce vecteur acoustique), on peut ainsi associer un « score phonétique », en fonction de la distance de V au « domaine idéal ».

Une façon simple d'évaluer cette distance consiste à calculer la norme du vecteur défini par V et par sa projection orthogonale sur le domaine (la projection est bien unique car les domaines sont convexes). Nous préférons en réalité calculer un score relatif à chaque type de contrainte (position du dos de la langue, ouverture de la bouche, protrusion des lèvres), qui est encore plus simple à calculer et plus flexible : on peut en effet facilement choisir de privilégier certains types de contraintes.

Le calcul effectif du score dépend de deux variables : la valeur cible de la contrainte considérée $\theta(v,t)$, où v est la voyelle, et t le type de contrainte ; et une marge $\sigma(v,t) > 0$, qui ensemble définissent un intervalle de validité pour la contrainte : $I(v,t) = [\theta(v,t) - \sigma(v,t); \theta(v,t) + \sigma(v,t)]$.

Si le calcul de la contrainte de type t pour la voyelle V – la façon dont les contraintes sont calculées à partir d'un vecteur articulatoire a été explicitée à la section 5.2.2 – conduit à une valeur dans l'intervalle $I(v,t)$, alors on lui attribue un score parfait de 1 pour cette contrainte. Sinon, on lui donne un score positif inférieur à 1 décroissant exponentiellement en fonction de la distance à $I(v,t)$. Le score final est simplement une combinaison linéaire des 4 contraintes de façon à obtenir des scores dans l'intervalle $[0;1]$. Dans notre modèle actuel, toutes les contraintes ont un poids égal.

$\theta(v,t)$ correspond à la valeur objective moyenne définie à partir du tableau de classification. σ correspond à une marge d'erreur, difficile à déterminer. Cette marge doit rendre compte de deux

éléments : l'importance de la contrainte pour la réalisation du phonème, ainsi que la diversité des réalisations observables. Dans les deux cas, il est relativement difficile de déterminer une valeur précise pour chaque phonème.

Il est cependant envisageable de le faire de façon statistique pour les types de contraintes pour lesquelles on dispose de suffisamment de données articulatoires. En particulier, il est tout à fait envisageable de le faire à partir de données vidéo pour les deux contraintes phonétiques « visibles » (ouverture de la bouche et protrusion des lèvres) en effectuant une analyse statistique sur des données visuelles, par exemple celles de Robert et al. (Robert *et al.* 2005) pour 10 locuteurs français. De la même façon, il serait envisageable d'étudier la position de la langue à partir de données EMA sur un nombre important de locuteurs : un petit nombre de marqueurs est suffisant pour déterminer la position avant-arrière de celle-ci. Dans notre cas, les valeurs de $\sigma(v,t)$ ont été déterminées manuellement et varient entre 0.1 et 0.6 en fonction de la voyelle et du type de contrainte considéré. Le tableau 5.4 récapitule les valeurs des contraintes pour chacune des voyelles.

Voyelle	D	O	P
i	-2.0 ± 0.2	-2.0 ± 0.2	-1.0 ± 0.3
e	-1.5 ± 0.2	-1.5 ± 0.3	-1.0 ± 0.3
ɛ	-1.5 ± 0.3	0.0 ± 0.2	-1.0 ± 0.3
a	0.0 ± 0.5	0.5 ± 0.5	-1.0 ± 0.3
y	-1.5 ± 0.2	-1.0 ± 0.3	1.5 ± 0.1
ø	-1.5 ± 0.2	-0.5 ± 0.3	0.5 ± 0.2
œ	-1.5 ± 0.3	0.0 ± 0.2	-0.5 ± 0.3
u	1.5 ± 0.3	-1.0 ± 0.3	1.5 ± 0.1
o	1.5 ± 0.2	-0.5 ± 0.3	0.5 ± 0.2
ɔ	1.5 ± 0.3	-0.5 ± 0.2	0.5 ± 0.2

TAB. 5.4: *Contraintes pour les voyelles du français. D correspond à la position du dos de la langue, O à l'ouverture de la bouche, et P à la protrusion des lèvres.*

La formule exacte utilisée pour le calcul du score phonétique est la suivante :

$$s(\alpha, v) = 1 - \sum_{t \in \{D,O,P\}} \lambda_t g(p(\alpha, v, t)),$$

où λ_t est le poids de la contrainte de type t, g est une fonction de transition, et $p(\alpha, v, t)$ la distance par rapport à la contrainte idéale, i.e. :

$$p(\alpha, v, t) = \max\left(|f(t, \alpha) - \theta(v, t)| - \sigma(v, t), 0\right)$$

où $f(t, \alpha)$ est la valeur de la contrainte de type t pour le vecteur articulatoire α, soit :

$$f(D, \alpha) = Tp = \alpha_2 \tag{5.1}$$
$$f(P, \alpha) = Lp = \alpha_6 \tag{5.2}$$
$$f(O, \alpha) = \min(Tp + \gamma Jw, Lh) = \min(\alpha_2 + \gamma\alpha_1, \alpha_5) \tag{5.3}$$

La fonction de transition pour calculer le score en dehors des intervalles que nous avons retenue est la fonction :

$$g(x) = e^{\frac{-1}{x^2}}.$$

Il est important de noter que le score phonétique d'une forme de conduit n'est toujours calculé que relativement à l'une des classes de contraintes. Ce score est en général calculé relativement à l'image acoustique du vecteur articulatoire, en partitionnant l'espace des vecteurs acoustiques pour que toute zone de l'espace acoustique soit associée à l'une de ces classes, ce qu'on peut écrire :

$$s(\alpha) = s(\alpha, v_m(F(\alpha))),$$

où $v_m(F(\alpha))$ désigne le phonème associé à l'image acoustique de α par le modèle acoustique ; mais on pourrait calculer un score phonétique global autrement : on pourrait, par exemple, attribuer à un vecteur articulatoire le score phonétique maximal parmi toutes les classes de contraintes, i.e. :

$$s_1(\alpha) = \max_{v \in V} s(\alpha, v).$$

Un tel score a notamment l'avantage d'être continu dans l'espace articulatoire, et de ne pas dépendre du modèle acoustique – ce qui permet d'éviter d'avoir à calculer l'image acoustique du vecteur avant de pouvoir déterminer son score phonétique[13]. En revanche, l'indication apportée par cette mesure n'est a priori pertinente que si l'on a une cohérence des classes de contraintes et des images acoustiques attendues pour les phonèmes correspondants : on n'a, par exemple, aucune envie d'obtenir un bon score phonétique pour un /u/ que l'on articule comme un /i/. La cohérence articulatoire-acoustique des classes de contraintes a donc été étudiée (cf. section 6.1).

5.5 Inversion avec contraintes phonétiques

Les contraintes phonétiques se traduisent ainsi comme un score attribué à un vecteur articulatoire en fonction de sa proximité à un domaine articulatoire de référence que l'on peut choisir de façon arbitraire, mais qui est en général déterminé en fonction de l'image acoustique du vecteur articulatoire.

Ce score peut être intégré au processus d'inversion de multiples façons : au niveau de la construction de codebooks, au niveau de l'inversion statique, ou au niveau de l'inversion dynamique.

5.5.1 Construction de codebooks

L'une des premières façons dont il est envisageable d'utiliser les contraintes phonétiques est au sein de la construction même de codebooks. En restreignant l'exploration de l'espace articulatoire aux zones permettant d'obtenir un score phonétique supérieur à un seuil fixé, il est possible d'obtenir des codebooks beaucoup plus compacts et limitant considérablement les solutions trouvées et accélérant considérablement l'inversion. Mais cela suppose que les réalisations du locuteur à inverser vont toujours obtenir un score phonétique supérieur à un certain seuil, ce qui semble une hypothèse très forte, puisque de nombreuses formes de la parole naturelle ne sont pas du tout modélisées par ces contraintes : les consonnes, et les transitions entre voyelles. Notre synthétiseur articulatoire étant de toute manière restreint à la modélisation des voyelles, la limitation relative aux consonnes n'est a priori pas trop gênante dans notre cas ; mais il serait en revanche particulièrement problématique que les transitions entre voyelles ne soient pas représentées au sein du codebook.

[13] À noter que, dans le cadre de l'inversion, si le modèle acoustique utilise les mêmes vecteurs acoustiques que ceux mesurés en entrée, cela n'est de toute façon pas nécessaire.

Les contraintes phonétiques ont été intégrées au processus de construction de codebook, mais les performances n'ont guère été encourageantes, et l'intérêt s'est avéré somme toute très limité. Le plus gros inconvénient à effectuer le calcul des scores phonétiques à ce niveau est que cela implique un modèle de sélection acoustique statique, ne permettant donc plus de modifier le modèle acoustique de façon à appliquer des contraintes différentes, par exemple liées à l'intention articulatoire. Cela ne permet pas non plus d'étudier finement les contraintes, par exemple pour ajuster les poids à donner à chacun des types de contraintes, ou ajuster les classes de contraintes. Par ailleurs, il s'est avéré que le seuil sur le score phonétique que l'on devait imposer pour conserver toutes les formes de conduits observées dans les séquences articulatoires de la locutrice de référence était trop bas pour que les gains de place soient réellement intéressants. Enfin, suite aux diverses optimisations apportées au processus, l'inversion statique est devenue une opération très rapide ; les éventuels gains de temps à ce niveau n'ont donc guère d'intérêt.

Il apparaît donc qu'il n'est pas intéressant d'utiliser les contraintes lors de la construction du codebook ; les contraintes phonétiques ne sont donc pas employées à ce niveau.

5.5.2 Inversion statique

Une autre façon d'utiliser les contraintes, celle que nous utiliserons le plus souvent en pratique, est de calculer le score phonétique de chacune des solutions de l'inversion, pour ne converser que celles ayant les meilleurs scores. Lorsque le modèle acoustique prend en entrée le même vecteur acoustique que pour l'inversion, il n'est pas nécessaire de calculer l'image acoustique des vecteurs articulatoires solutions, et le temps nécessaire au calcul des scores phonétiques devient négligeable par rapport à celui nécessité par l'inversion. Il est à noter que, le codebook n'ayant pas une précision acoustique parfaite, l'image acoustique d'un vecteur articulatoire solution de l'inversion est généralement un peu différente du vecteur acoustique inversé. Une légère erreur est donc commise lorsque l'on identifie l'image d'un vecteur articulatoire à son antécédent lors de l'inversion par codebook. Mais il faut bien observer que l'erreur est commise au niveau du vecteur articulatoire, pas au niveau de l'image acoustique, puisque, par définition, on souhaiterait que l'image acoustique d'un vecteur articulatoire solution de l'inversion soit égale au vecteur acoustique inversé. Il est donc plus logique de calculer le score phonétique en fonction de l'*image théorique* – c'est-à-dire l'antécédent – d'un vecteur solution plutôt qu'en fonction de son image réelle.

On se retrouve ainsi avec des vecteurs articulatoires ayant une composante supplémentaire : leur score phonétique. En général, les vecteurs articulatoires ayant un score phonétique faible ne sont pas éliminés à cette étape.

5.5.3 Inversion dynamique

Le score attribué aux vecteurs articulatoires peut servir à l'inversion dynamique sous deux formes et dans deux optiques assez différentes, mais non exclusives :

- limiter l'espace articulatoire à explorer, et ainsi accélérer le processus (en particulier le lissage non-linéaire),
- améliorer le réalisme des solutions en intégrant un terme supplémentaire basé sur le score phonétique dans les fonctions de coûts des procédures de lissage.

La première correspond essentiellement à une utilisation statique des contraintes : seules les formes ayant un score phonétique assez important seront considérées. En pratique, plutôt que de ne garder que les formes ayant un score phonétique au-dessus d'un seuil fixe, ce qui risquerait de pénaliser inutilement les transitions entre voyelles, on garde un certain pourcentage des solutions

à chaque instant. Une diminution de moitié du nombre global de solutions à explorer permet déjà d'accélérer de manière conséquente le lissage non-linéaire (cf. section 4.1) : le gain est de 2^{D+1} – où D est le niveau maximal de dérivation utilisée pour le terme articulatoire dans la fonction de coût – soit un facteur 16 si $D = 3$, et un facteur 4 pour le cas – plus courant en pratique dans nos expériences – où $D = 1$.

La deuxième utilisation consiste à ajouter un terme à la fonction de coût visant à maximiser le score phonétique global sur la séquence. En pratique, nous utilisons un terme de la forme suivante :

$$C_p = \beta_p(S_t - s(\alpha)),$$

où S_t correspond, en quelque sorte, à un score cible minimal : les paramètres articulatoires ayant un score supérieur à cette valeur bénéficient d'un bonus dans le lissage non-linéaire. Ils ont donc moins de chance d'être éliminés.

Cela conduit à la fonction de coût global suivante :

$$C(j,\alpha) = \sum_{k=1}^{K}(t_{j(k)} - t_{j(k-1)}) \left(\sum_{d=0}^{D} \beta_d \left| d! * [\alpha(j(k-d)), \ldots, \alpha(j(k))] \right|_{\mathrm{Ar}} + \beta_p(S_t - s(\alpha(j(k)))) - B \right)$$

En pratique, nous fixons le plus souvent $S_t = 1$. Cette deuxième forme n'accélère pas le processus d'inversion, mais conditionne en partie la sélection des solutions à leur score phonétique.

5.6 Exemples

Dans cette section, nous présentons quelques exemples d'utilisation des contraintes phonétiques pour l'inversion : nous inversons une voyelle de notre locutrice de référence et présentons différentes statistiques relatives au score phonétique de ces solutions.

Il s'agit de la voyelle /i/ présentée à la section 3.2.1.

La figure 5.2a représente les mêmes données qu'à la figure 3.2c, c'est-à-dire les solutions de l'inversion pour une voyelle synthétique de notre locutrice de référence représentées en fonction de l'aire et de la position de la constriction minimale, à la différence près que les solutions sont désormais colorées en fonction de leur score phonétique. Plus un point est sombre, plus le score de la solution correspondante est élevé.

La figure 5.2b représente la répartition des scores de l'inversion : plus un score donné a une fréquence d'apparition élevée, plus sa valeur sera grande. On peut constater que les scores qui apparaissent le plus souvent sont le score parfait, et un score de 0,66 : c'est-à-dire des solutions qui remplissent parfaitement deux des contraintes, mais ne satisfont pas du tout la troisième. On peut aussi étudier indépendamment chacune des contraintes (cf. figure 5.2c) : on peut alors constater que la contrainte d'ouverture est remplie par la majorité des solutions, celle sur la position de la langue est à peu près répartie uniformément, avec toutefois, en dehors du pic important autour de 1, deux pics assez importants autour de 0,28 et 0,15, et un creux important entre les deux. Enfin, la contrainte sur la protrusion est certainement la plus intéressante : une majorité des solutions a un score très proche de 0 pour cette contrainte ; on peut ainsi en déduire qu'elle est responsable du pic visible sur la figure 5.2b.

On peut étudier plus précisément la distribution des solutions suivant les paramètres articulatoires associés à chacune des contraintes. La figure 5.3a représente la densité des solutions en fonction de la protrusion (trait continu), ainsi que le score associé (trait pointillé). On constate que la densité des solutions décroît linéairement lorsque la protrusion augmente, mais que le

modèle articulatoire parvient tout de même à générer un grand nombre de solutions pour le /i/ dans tout le domaine articulatoire du paramètre. On retrouve également le fait qu'un très grand nombre de solutions ont un faible score.

La figure 5.3b représente la densité des solutions en fonction de la position du dos de la langue. On constate qu'ici la contrainte phonétique est pratiquement inutile, puisque l'intervalle ayant des solutions est très réduit, et la courbe de densité correspond pratiquement parfaitement avec la courbe du score phonétique. Cela indique au passage que les critères articulatoire retenus pour élaborer les contraintes correspondent ici à la réalité acoustique du synthétiseur articulatoire. On peut également expliquer grâce à cette figure les irrégularités des scores correspondant à la position de la langue : le creux important au-dessous de 0,28 est dû à la frontière inférieure de l'espace articulatoire.

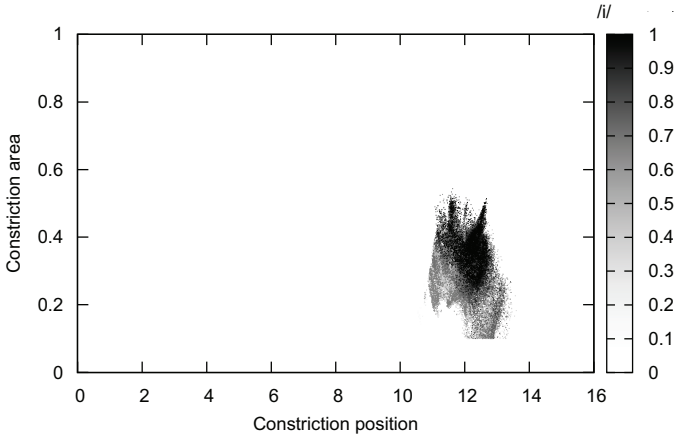
Enfin, la figure 5.3c représente la densité des solutions en fonction de l'ouverture de la bouche, telle que définie dans l'équation 5.3, ainsi que le score phonétique associé. À nouveau, la courbe de densité correspond de façon assez fidèle à la courbe des scores.

On peut conclure de ces expériences que, concernant l'inversion du /i/, la contrainte sur la protrusion est probablement celle qui supprimera le plus de solutions, l'ouverture et la position de la langue étant déjà considérablement contraintes par l'acoustique.

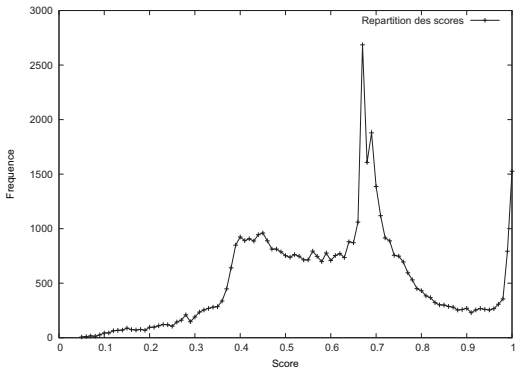
Conclusion

Nous avons introduit dans ce chapitre des contraintes génériques, basées sur des connaissances phonétiques empiriques, dans l'optique d'améliorer la qualité des solutions de l'inversion. Les premières expériences réalisées montrent que ces contraintes semblent correspondre à la réalité acoustique du modèle articulatoire, c'est-à-dire que les formes privilégiées par les contraintes sont a priori aussi les plus nombreuses parmi les solutions de l'inversion.

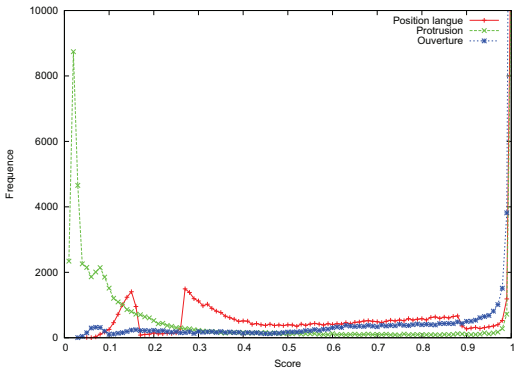
Il reste cependant à effectuer une évaluation à grande échelle de ces contraintes en comparant les formes privilégiées par les contraintes à des données réelles, et à mesurer les gains apportés par l'utilisation de ces contraintes, tant pour l'inversion statique que pour l'inversion dynamique, par rapport à l'inversion classique. Cette évaluation est l'objet du chapitre suivant.



(a) Représentations des solutions de l'inversion : aire à la constriction minimale (en cm^2 en fonction de la position dans le conduit (en cm, depuis la glotte) et du score phonétique (niveaux de gris : plus un point est sombre, plus son score est élevé).

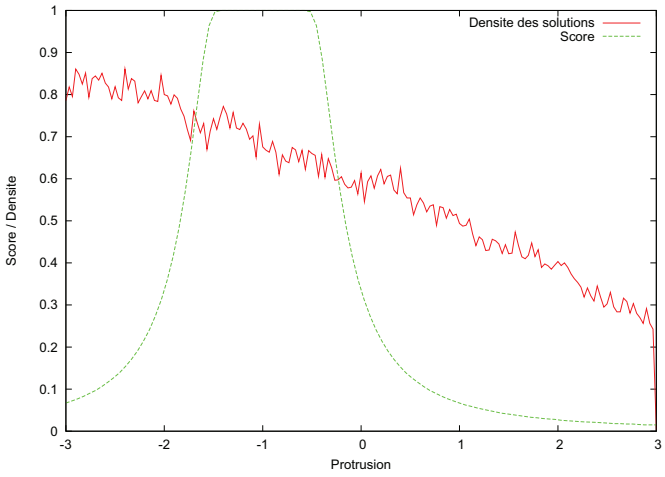


(b) Densité des solutions en fonction du score phonétique.

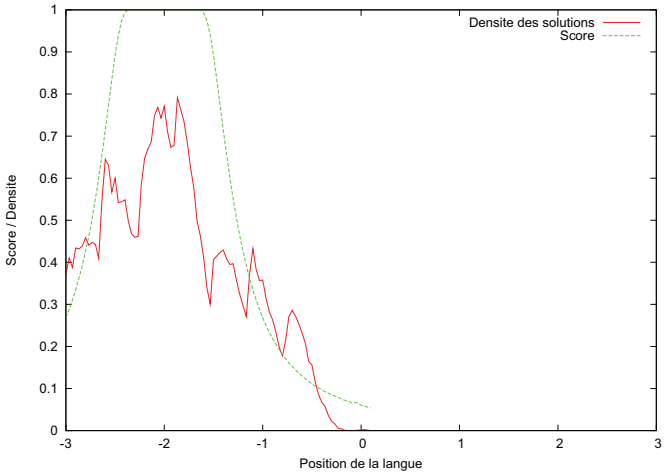


(c) Densité des solutions en fonction du score phonétique, par contrainte

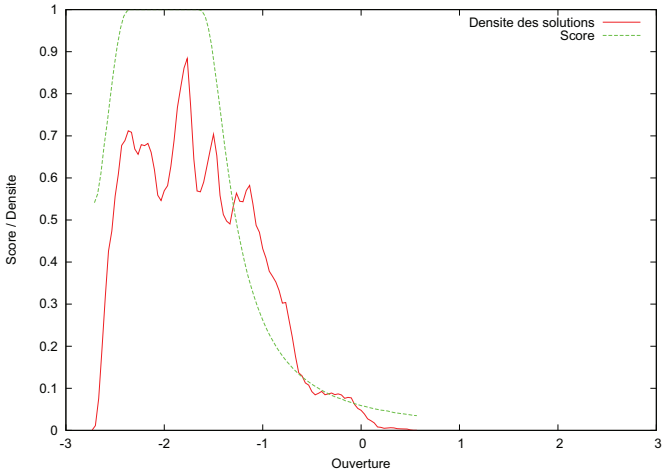
FIG. 5.2: *Solutions de l'inversion et scores phonétiques pour la voyelle /i/.*



(a) Protrusion des solutions du /i/ et score associé



(b) Position de la langue pour les solutions du /i/ et score associé



(c) Ouverture pour les solutions du /i/ et score associé

6

Évaluation des contraintes phonétiques

Introduction

Dans ce chapitre sont rassemblées différentes expériences réalisées pour évaluer les contraintes phonétique introduites au chapitre 5.

Les premières expériences portent sur la correspondance entre les domaines articulatoires associées aux phonèmes et leurs images acoustiques.

Dans un deuxième temps nous évaluons l'inversion statique, en comparant les formes de conduit ayant des scores phonétiques élevés à des formes de conduit réelles.

Enfin, nous emploierons les contraintes phonétiques pour réaliser de l'inversion dynamique.

6.1 Correspondance articulatoire-acoustique

Les classes de contraintes que nous avons définies – à partir de critères phonétiques puis transcrites sous forme de domaines articulatoires – sont supposées être spécifiques à l'articulation d'un phonème donné. Ces domaines articulatoires correspondant à des phonèmes, il est important que l'image acoustique d'un domaine articulatoire donné corresponde au domaine acoustique du phonème.

Une façon simple de vérifier que les classes de contraintes ont été définies correctement est ainsi de générer l'image acoustique du domaine articulatoire de la classe de contraintes, et de comparer celle-ci au domaine acoustique du phonème correspondant.

Dans cette section, nous cherchons à vérifier une hypothèse encore plus forte : nous étudions si l'image acoustique d'un domaine articulatoire contient bien le domaine acoustique correspondant, et réciproquement, nous vérifions que la zone acoustique est bien spécifique à la contrainte considérée, c'est-à-dire que, si nous inversons un vecteur acoustique d'un phonème particulier, ce sont les contraintes de ce phonème qui lui permettent d'obtenir les meilleurs scores phonétiques. Imaginons que l'on inverse, par exemple, un vecteur acoustique correspondant à la voyelle /i/ et que, par erreur, nous appliquons les contraintes correspondant au /a/. On s'attend, dans ce cas là, à ce que les solutions de l'inversion n'obtiennent pas d'aussi bons scores que si l'on avait appliqué les contraintes correctes, ce qui du reste pourrait être un moyen de se rendre compte de l'erreur commise.

En d'autres termes, nous cherchons à étudier s'il y a une correspondance bijective entre les domaines articulatoire et acoustique.

Deux expériences ont été réalisées pour répondre à cette question : la première (cf. figure 6.1a) réalise un échantillonnage aléatoire homogène de chacun des domaines articulatoires idéaux (donc

des vecteurs articulatoires ayant un score phonétique de 1 pour les contraintes du phonème), calcule l'image acoustique de chacun des points, et les projette dans l'espace acoustique F1-F2. Ensuite, chacun des points de l'espace acoustique se voit attribuer une couleur en fonction du phonème dominant parmi les images acoustiques situées au voisinage[14] de ce point.

Ce processus comporte plusieurs inconvénients. D'une part, il n'est pas clair que l'échantillonnage retenu pour l'espace articulatoire soit le plus adéquat ; en effet, nous générons le même nombre de vecteurs articulatoires pour chacun des phonèmes ; les phonèmes ayant des domaines articulatoires plus grands (tel que le /a/) risquent d'avoir une image acoustique plus éparpillée, avoir une densité acoustique moyenne plus faible, et ainsi être moins en position de dominance qu'ils ne devraient l'être.

La deuxième expérience consiste à inverser l'intégralité de l'espace acoustique, et à colorier chacun des points en fonction du « phonème dominant » pour un critère particulier, par exemple le phonème qui maximise le score phonétique moyen, ou celui ayant le plus grand nombre de vecteurs articulatoires ayant un score supérieur à un certain seuil. Il n'est cependant pas simple de définir un critère réellement pertinent pour qualifier la notion de « phonème dominant ». Pour le graphique présenté à la figure 6.1b, nous avons inversé l'espace acoustique avec un pas très fin (10Hz pour chaque fréquence formantique), et en utilisant un critère de prépondérance simple : on « colorie » la case du phonème dont le score phonétique moyen est le plus grand dans la zone considérée, en imposant également que le nombre de solutions de l'inversion soit supérieur à un seuil minimum (pour ne pas prendre en compte inutilement les points situés dans des zones acoustiques limites). En pratique, l'inversion effectuée porte sur les trois premiers formants. Pour réaliser la figure 6.1b, nous avons intégré les résultats sur la dimension correspondant au formant F3 et appliqué un lissage identique à celui de la figure 6.1a. Le nombre de points générés lors de l'inversion, notamment dans les zones centrales, ne permet pas de garantir des résultats statistiquement significatifs pour toutes les normes ; à titre indicatif, 11 millions de points environ ont été générés au total pour cette figure, soit environ 1000 en moyenne par couple F1-F2.

L'étude de ces figures nous permet de formuler plusieurs remarques : d'une part, les zones de dominance des images acoustiques des domaines articulatoires idéaux dessinent bien des « clusters » distincts dans l'espace acoustique, et c'est également le cas sur le graphique de dominance des phonèmes de score phonétique moyen maximal. Les emplacements précis des zones distinctes (les espaces acoustiques des phonèmes) varient cependant de manière notable entre les deux figures, et on peut également constater que les zones acoustiques des phonèmes diffèrent sensiblement des zones correspondantes dans le modèle acoustique que nous utilisons pour déterminer quel classe de contraintes appliquer, notamment pour le /y/ sur la figure 6.1b et le /e/ sur les deux figures. Le triangle vocalique reste cependant respecté, et les valeurs des centres vocaliques restent dans la norme. La principale modification par rapport aux données acoustiques standard porte sur F2 : tous les phonèmes ont un F2 significativement plus élevé que sur la figure 5.1 ; ce qui, du reste, s'observe également – mais de façon moins marquée – dans les réalisations de notre locutrice de référence.

Il est intéressant de constater que les domaines acoustiques « intrinsèques » déterminés lors de ces expériences sont finalement assez proches de ceux que nous utilisons réellement. Cela prouve d'une part que la correspondance articulatoire-acoustique est assez bonne, et aussi qu'en définitive le module d'identification acoustique des phonèmes n'est pas totalement indispensable : les scores phonétiques peuvent suffire dans une certaine mesure à identifier de façon assez pertinente une

[14] En pratique, le voisinage est un cercle intégrateur à règle exponentielle, dont le rayon dépend de la densité des points générés : celui-ci est déterminé de façon à ce que le nombre moyen d'échantillons dans chaque cercle soit d'environ 1000 points, pour que les résultats soient statistiquement significatifs. Sur la figure 6.1a, 1000000 de points ont été générés, le rayon est ainsi d'environ 20Hz.

classe de contraintes articulatoires à appliquer.

Le module d'identification acoustique reste cependant utile : l'utilisation des contraintes phonétiques pour identifier la classe de contraintes à appliquer nécessite des calculs importants pour obtenir un résultat statistiquement significatif, et il peut également être utile d'utiliser un modèle d'identification pour appliquer des contraintes différentes de celles que l'on pourrait vouloir appliquer en se basant uniquement sur l'acoustique : par exemple pour appliquer des contraintes liées à l'intention articulatoire.

6.2 Inversion statique

6.2.1 Expériences d'inversion

Boë et al. (Boë *et al.* 1992) ont utilisé le modèle articulatoire de Maeda pour étudier les lieux d'articulation des voyelles du français. Mais leur travail ne prenait pas en considération la totalité des solutions, car au lieu d'utiliser une méthode d'inversion, ils n'ont pris en compte qu'un nombre limité de formes articulatoires tirées au hasard. Les 60000 configurations articulatoires utilisées dans leur étude correspondent approximativement au choix de 4 valeurs différentes pour chacun des paramètres articulatoires. Le codebook que nous avons utilisé pour cette étude était construit grâce à notre méthode hypercuboïdale et comprenait environ 60000 hypercuboïdes décrivant localement l'approximation à l'ordre 1 de la relation articulatoire vers acoustique, pour une erreur acoustique tolérée de 1 Bark sur les trois premières fréquences formantiques. Lors de la construction de ce codebook, 30 millions de vecteurs articulatoires différents ont été synthétisés, à comparer aux 300000 de Boë et al. Il faut toutefois remarquer que le nombre de formes utilisées lors de la construction de notre codebook ne correspond qu'au choix d'environ 10 valeurs différentes pour chaque paramètre.

Nous avons réalisé des expériences d'inversion en utilisant des données radiographiques d'une locutrice française, PB. Ce sont les données que Maeda a utilisées pour construire son modèle articulatoire. Le signal acoustique enregistré en même temps que les radiographies est de mauvaise qualité, mais reste exploitable. Les formants ont été extraits à partir d'un spectre calculé par l'algorithme d'« enveloppe vraie » (Halle 1983), qui est une procédure de lissage itératif prenant seulement en compte les pics du spectre, c'est-à-dire principalement les harmoniques, puis vérifiés et au besoin corrigés manuellement. Toutes les voyelles inversées ont été vérifiées auditivement, pour s'assurer qu'elles soient correctes perceptivement. Les formants F2 et F3 du /u/ ont été particulièrement difficiles à extraire, parce que l'énergie de cette voyelle est faible, et ces formants étaient par conséquent dominés par le bruit de la machine à rayons-X. L'exemplaire retenu correspond à un /u/ légèrement plus marqué que les autres.

Pour chaque voyelle, un échantillonnage des solutions possibles a été effectuée en appliquant la procédure d'inversion sur les fréquences des trois premiers formants. Pour vérifier la précision des résultats de l'inversion, les spectres acoustiques des images des vecteurs articulatoires ont été calculés, les fréquences des formants en ont été extraites et comparées aux valeurs originales. Tous les résultats de l'inversion ayant des images acoustiques trop éloignées ont été supprimées : nous avons imposé une précision de 30Hz sur F1, 50Hz pour F2 et 75Hz pour F3. Le tableau 6.1 présente les 5 voyelles inversées : les trois fréquences formantiques, le nombre de solutions retenues, et l'erreur moyenne des images acoustiques des solutions resynthétisées (les solutions supprimées ont également été prises en compte dans ce calcul).

Les solutions de l'inversion sont présentées ici sous la forme du diagramme suivant : aire à la constriction (A_c, cm^2)/ position de la constriction dans le conduit vocal (X_c, cm). La position de la constriction est exprimée en centimètres à partir du larynx, en abscisse pseudo-curviligne le

long de la « médiane » du conduit. Cela signifie que nous calculons la longueur du segment milieu de chaque section, cf. figure 1.4) dans la représentation du conduit. La position de la constriction est déterminée en calculant la position de la section d'aire minimale le long du conduit vocal. Nous ne prenons pas en compte la constriction dans la partie basse du pharynx, ni celle formée au niveau des lèvres : nous mesurons la constriction linguale, c'est-à-dire celle formée entre la langue et la paroi fixe du conduit vocal.

Pour chaque voyelle, nous présentons deux types de résultats : le diagramme représentant l'aire à la constriction en fonction de sa position (les figures 6.1b, 6.2b, 6.3b, 6.4b et 6.5b), et des coupes medio-sagittales de formes de conduit caractéristiques (les figures 6.1a, 6.2a, 6.3a, 6.4a et 6.5a).

La position de la constriction varie entre 0 cm au niveau de la glotte, et environ 16 cm au niveau des lèvres. De façon à garder les aires à la constriction cohérentes avec le modèle de production des voyelles, nous avons éliminé les formes présentant une aire à la constriction inférieure à 0.2 cm^2. Nous n'avons pas éliminé d'autre solution sur les diagrammes des figures 6.1b, 6.2b, 6.3b, 6.4b et 6.5b. Ces diagrammes présentent également le score phonétique de chacune des solutions sous la forme d'un niveau de gris : plus le point est sombre, plus son score est élevé.

Pour chacune des voyelles, deux ou trois tracés de conduits vocaux caractéristiques pour notre locutrice, et issue du livre de Bothorel et al. (Bothorel *et al.* 1986), sont présentés (les figures 6.1c, 6.2c, 6.3c, 6.4c et 6.5c). À l'exception du /e/, ces tracés caractéristiques sont issus de contextes différents de celui des voyelles inversées. En effet, on on retrouve pas les contextes phonétiques des tracés manuels dans le fichier audio que nous avons inversé.

Enfin, la distance géométrique moyenne des solutions de l'inversion par rapport à la forme de conduit vocal observée (elle aussi issue du livre de Bothorel) est également calculée, en fonction du score phonétique des formes. Les résultats de cette expérience sont présentés pour les voyelles /a/ et /i/ sur les figures 6.7 et 6.8. La distance géométrique est évaluée en projetant les formes de conduit sur la grille de Maeda, et en calculant la norme euclidienne des projetés par rapport à la projection de la forme de référence. Plus précisément, si l'on désigne par $\{P_{0,j}\}_{0\leq\ j\leq\ M_g}$ les coordonnées de la projection de la forme de référence sur la grille de Maeda, et $\{P_{i,j}\}_{0\leq\ j\leq\ M_g}$ les coordonnées d'une forme de conduit particulière, nous pouvons calculer une distance géométrique en utilisant la formule suivante :

$$d(0,i) = \sqrt{\frac{\sum_{j=1}^{M_g} |P_{i,j} - P_{0,j}|^2}{M_g}}, \tag{6.1}$$

dans laquelle M_g est le nombre total de points sur la grille de Maeda. Les figures 6.7 et 6.8 ont été obtenues en calculant la distance géométrique moyenne de toutes les formes de conduits ayant un score phonétique compris dans une fenêtre de taille réduite (en l'occurrence 0.001) : par exemple, l'erreur pour un score de 0.5 est obtenue en calculant la moyenne des distances géométriques pour toutes les formes ayant un score phonétique compris entre 0.499 et 0.501.

6.2.2 Analyse des résultats

Commençons par observer le nombre de solutions trouvées. La première étape de la procédure d'inversion est conçue pour échantillonner l'espace articulatoire de façon uniforme, ce qui signifie que le nombre de solutions trouvées est fortement lié à l'étendue des régions articulatoires correspondant aux voyelles. Si le modèle articulatoire représentait fidèlement le comportement

Voyelle	contexte	F1	F2	F3	Nb. de sol.	ΔF1	ΔF2	ΔF3
a	t**a**bac	749	1701	2785	103578	19.1	25.0	24.4
e	t**e**s beaux	458	2341	3070	208502	10.7	27.8	35.8
i	rouss**i**es	349	2305	3345	52799	15.8	19.4	54.1
u	b**ou**gies	367	1050	2495	5147	22.6	49.8	10.7
y	d**u** guet	341	1956	2523	21748	11.1	60.3	27.4

TAB. 6.1: *Une sélection de voyelles du Français pour la locutrice PB, leur contexte phonétique, les trois premières fréquences formantiques, le nombre de solutions trouvées, et l'erreur moyenne entre les formants originaux et ceux obtenus par la resynthèse des vecteurs articulatoires trouvés par inversion.*

acoustique du conduit humain, ces chiffres correspondraient au degré de précision nécessaire pour articuler une voyelle (ou tout du moins, un son ayant les trois mêmes premières fréquences formantiques). Dans notre cas, malgré les conditions favorables (le modèle articulatoire a été construit à partir d'images radiographiques de la locutrice que nous inversons, et nous nous sommes efforcé d'adapter le modèle de façon à avoir la meilleure correspondance acoustique possible), il subsiste tout de même une disparité importante entre les images acoustiques synthétisées à partir des données articulatoires obtenues sur les radiographies, et les vecteurs acoustiques mesurés sur les enregistrements sonores. Ceci étant, les résultats trouvés pour les vecteurs acoustiques mesurés (cf. le tableau 6.1) montrent clairement que l'articulation des voyelles /i,y,u/, surtout /u/ d'ailleurs, nécessitent une plus grande précision articulatoire que pour /e/ et /a/. Le petit nombre de solutions trouvées pour /u/ est probablement lié au fait que les régions articulatoires correspondant à cette voyelle sont proches des frontières de l'espace articulatoire. Il semble également que la configuration articulatoire particulière du /u/ rend sa synthèse difficile par les modèles de synthèse « classiques » (c'est-à-dire à fonction d'aire). Ce problème n'est en effet pas spécifique au synthétiseur intégré au modèle articulatoire de Maeda : on peut observer que le /u/ est difficile à obtenir avec d'autres modèles de synthèse articulatoire, tels que ASY (Rubin *et al.* 1981) ou TractSyn (Birkholz & Jackèl 2003) par exemple.

6.2.2.1 Lieu d'articulation

L'examen des figures 6.1b, 6.2b, 6.3b, 6.4b et 6.5b nous permet d'observer des propriétés intéressantes sur les lieux de constriction.

Premièrement, la discrétisation du conduit vocal, et par conséquent de la fonction d'aire, occasionne souvent des positions de constriction discrètes, ce qui correspond aux lignes presque verticales sur les figures 6.1b, 6.2b, 6.3b, 6.4b et 6.5b. Cependant, malgré cette répartition localisée, on observe aussi que les lieux de constriction s'organisent en un petit nombre de régions connexes, toujours inférieur à 3. Dans certains cas, ces régions fusionnent quand l'aire à la constriction augmente, ce qui est particulièrement visible dans le cas du /e/ (cf. figure 6.2b).

Deuxièmement, le calcul du lieu d'articulation, c'est-à-dire le point du conduit vocal où la fonction d'aire est minimale, dépend à la fois de la forme de la paroi fixe du conduit vocal et de la langue. Certains lieux d'articulation qui pourraient sembler extrêmement différents au premier abord, peuvent en réalité être dus à un petit déplacement de la langue et correspondre à des formes de conduit très similaires, notamment dans le cas du /a/. Les zones distinctes d'articulation du /a/ correspondent toutes à la partie pharyngale du conduit vocal, comme le montrent les coupes medio-sagittales (figure 6.1a).

Troisièmement, les données sont en accord avec les données de Wood (Wood 1979), à la fois pour les lieux d'articulation et pour les aires à la constriction. L'aire à la constriction pour le /e/ est en moyenne plus importante que celle du /i/, tout comme sur les données de Wood. Ces données confirment également que le lieu d'articulation du /a/ peut être réparti sur une grande partie du pharynx.

Enfin, les formes de conduits phonétiquement pertinentes ainsi que les formes irréalistes partagent les mêmes lieux d'articulation. Cela vient du fait que les propriétés acoustiques des voyelles imposent de très fortes contraintes sur le lieu d'articulation ; par conséquent, la connaissance du lieu d'articulation ne peut pas être utilisée seule comme critère pour distinguer les formes de conduits irréalistes.

6.2.2.2 Coupes medio-sagittales

L'examen des figures 6.1a, 6.2a, 6.3a, 6.4a et 6.5a (c'est-à-dire des exemples de formes de conduits vocaux obtenues par inversion) et leur comparaison avec des coupes medio-sagittales des mêmes voyelles (les figures 6.1c, 6.2c, 6.3c, 6.4c et 6.5c) obtenues sur les radiographies originales de Bothorel et al. (Bothorel *et al.* 1986), nous permet d'analyser plus finement les formes de conduits.

Les lieux d'articulation sont conformes aux connaissances phonétiques ainsi qu'aux résultats des modèles du conduit vocal à deux ou trois tubes proposés par Fant (Fant 1960). Malgré cette bonne conformité avec l'approximation à deux ou trois tubes, il s'avère que le modèle permet une très grande variabilité articulatoire, comme on peut le constater sur les coupes medio-sagittales des figures 6.1a, 6.2a, 6.3a, 6.4a et 6.5a. Une petite partie seulement de cette variabilité correspond à des formes de conduit réalistes. Pour chacune des voyelles étudiées, les premières coupes présentées sont les moins réalistes au sens des contraintes présentées précédemment.

Nous donnons un exemple de « bonne » et de « mauvaise » forme (au sens du score phonétique) pour chacune des trois zones d'articulation du /a/, c'est-à-dire respectivement à environ 3cm, 4.7cm, et 8cm de la glotte. On peut constater que les coupes les moins réalistes correspondent à des positions extrêmes des articulateurs. La coupe située en haut à gauche de la figure 6.1a, par exemple, présente une forte ouverture des lèvres et une faible ouverture de la mâchoire, et une position très basse de la langue, ce qui crée une constriction très proche de la glotte. Clairement, cette forme de conduit a extrêmement peu de chances d'être réalisée par un locuteur humain.

Dans le cas du /e/, les trois formes les moins réalistes (la ligne supérieure de la figure 6.2a) présentent une assez forte protrusion de lèvres. De plus, sur les deux premiers exemples on observe une forte ouverture des lèvres en conjonction avec une langue en position haute, ce qui semble difficile à réaliser pour un locuteur humain.

De manière similaire, les formes du conduit vocal avec les plus faibles scores phonétiques du /i/ et du /u/ correspondent à des configurations très improbables. Pour le /y/ la configuration avec le score phonétique le plus faible correspond à une protrusion et une ouverture des lèvres très faibles, ainsi que l'apex en position basse et une langue ramassée.

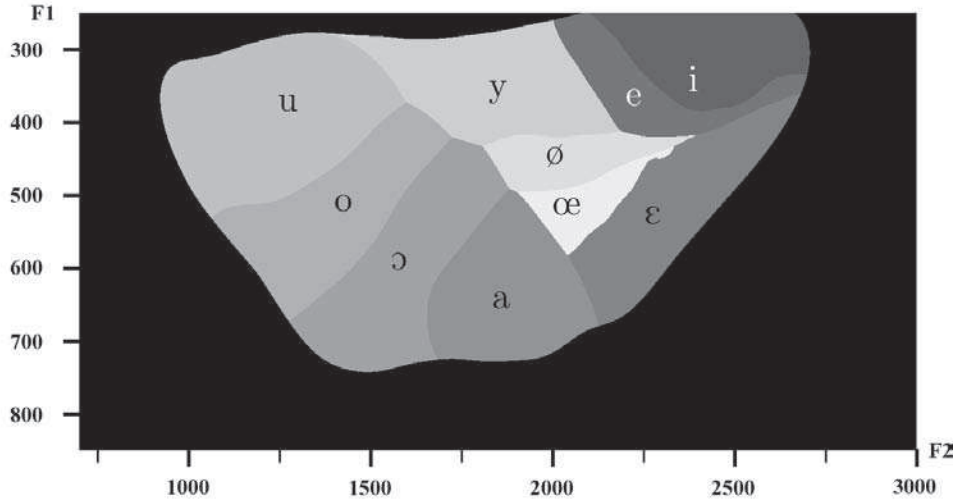
Deux lieux d'articulations existent pour le /u/ ; le second, représentée par la troisième coupe medio-sagittale de la figure 6.4a, est situé dans la partie basse du pharynx. Cependant, la figure 6.6, qui présente la fonction d'aire, montre que le pharynx entier est très étroit. On peut noter que le nombre de solutions correspondant à ce lieu d'articulation est nettement plus faible que pour le premier lieu d'articulation. Cela signifie que ce genre de formes de conduit ne peut pas être atteint aussi facilement que celles du premier lieu, d'un point de vue articulatoire.

On constate également que les formes de conduits ayant un score phonétique élevé (ligne

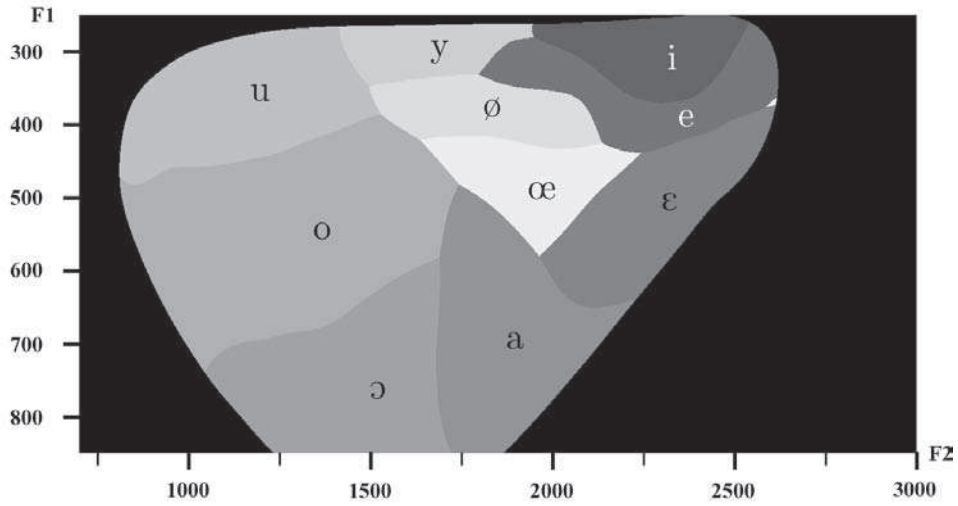
inférieure des figures 6.1a, 6.2a, 6.3a, 6.4a et 6.5a) correspondent très bien aux coupes radiographiques originales. Cela est d'autant plus remarquable que la simulation acoustique sur les coupes originales génère des formants sensiblement différents de ceux que l'on peut mesurer sur l'enregistrement sonore simultané. Malgré cette inadéquation acoustique, la procédure d'inversion est parvenue à retrouver les spécificités de la locutrice. Cela est particulièrement clair pour l'inversion de la voyelle /i/ : comme on peut le voir sur la figure 6.3a, la deuxième forme de conduit trouvée a une ouverture des lèvres plutôt supérieure à ce que l'on pourrait attendre pour cette voyelle. Cependant, il apparaît que cette locutrice réalise parfois le /i/ avec une ouverture importante des lèvres (contour pointillé de la figure 6.3c), comparé aux autres locuteurs de l'étude de Bothorel (Bothorel *et al.* 1986). Par conséquent, même si la deuxième coupe sagittale présente une ouverture des lèvres légèrement plus importante que sur la radiographie, cela n'a rien d'incompatible avec l'articulation de la locutrice PB.

Enfin, nous pouvons observer sur les figures 6.7 et 6.8, où l'on a représenté l'erreur géométrique moyenne en fonction du score phonétique pour les voyelles /a/ et /i/, que de manière générale, la moyenne de l'erreur géométrique entre les solutions de l'inversion et la forme de conduit vocal original diminue lorsque que le score phonétique augmente. Cela est particulièrement clair pour le /i/ pour lequel l'erreur géométrique passe de 0.7 cm pour un score phonétique minimal à 0.2 cm pour un score proche de 1. Cela est moins clair pour le /a/ : la distance moyenne passe de 0.47 cm pour un score de 0 à 0.32 cm pour les scores proches de 1. La même tendance est observée pour toutes les voyelles inversées dans cette étude, avec les contrastes les plus forts pour les voyelles ayant des lieux d'articulation bien localisés, tels que /i/, et plus faiblement pour les voyelles ayant de très larges zones d'articulation, tels que /a/. L'erreur résiduelle est d'autant plus importante que les lieux d'articulation sont éparpillés le long du conduit vocal.

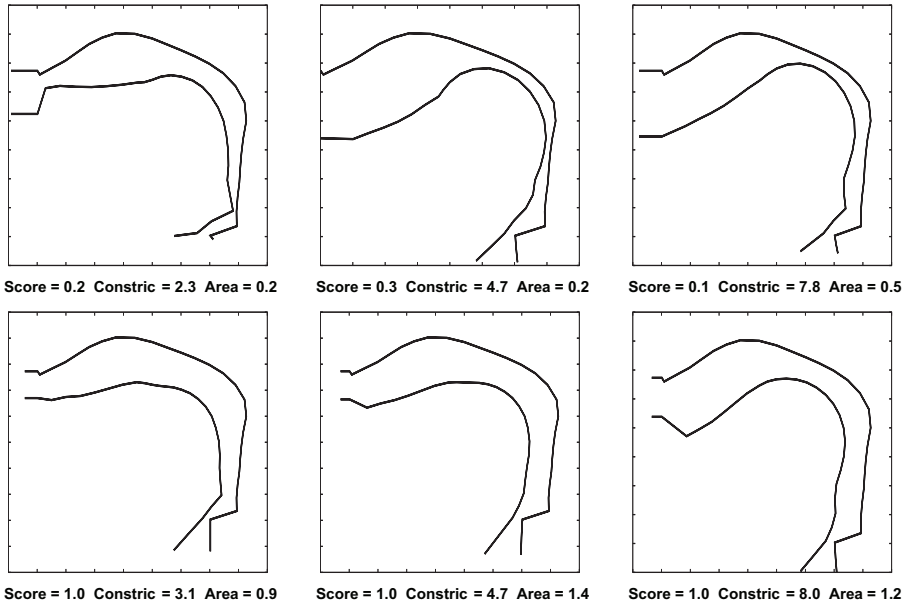
Ces expériences montrent que les contraintes phonétiques proposées pénalisent des formes du conduit vocales irréalistes et donnent un score élevé aux formes correctes. Elle permettent également de clarifier les rôles respectifs du modèle et des contraintes phonétiques au sein du processus d'inversion : le modèle articulatoire induit des contraintes sur la position géométrique de la constriction par le biais de la simulation acoustique, alors que les contraintes phonétiques portent plus particulièrement sur les interdépendances des articulateurs.



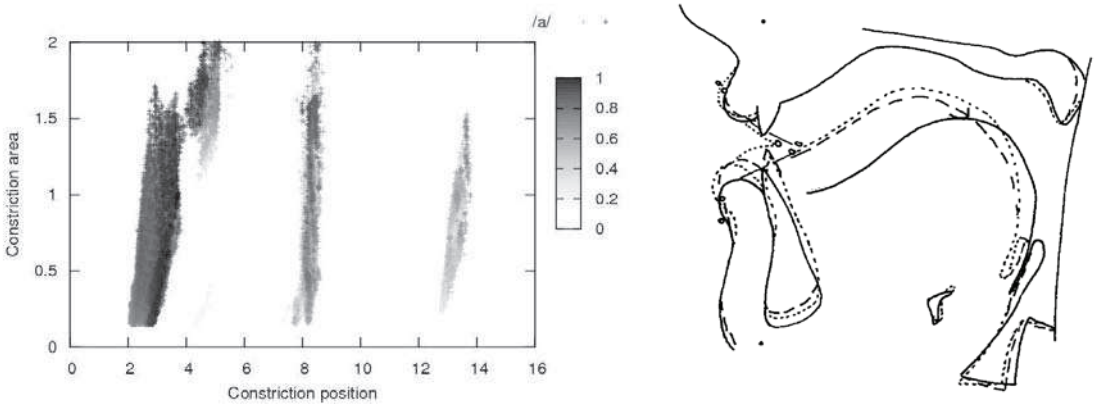
(a) Classes de contraintes dont l'image acoustique a une densité locale maximale dans l'espace acoustique F1-F2.



(b) Classes de contraintes assurant un score phonétique moyen maximal dans l'espace F1-F2.

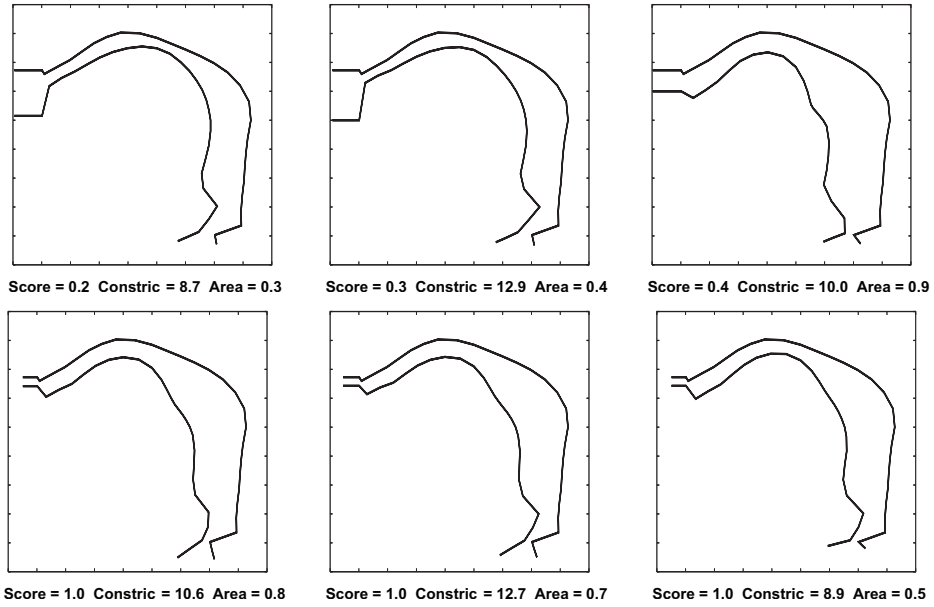


(a)

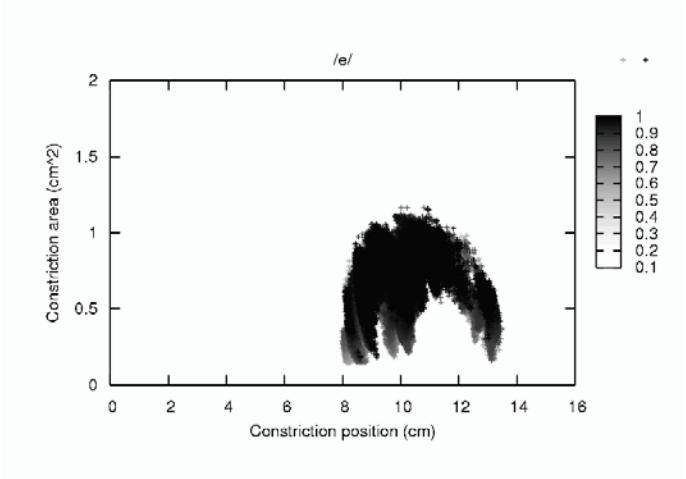


(b) (c)

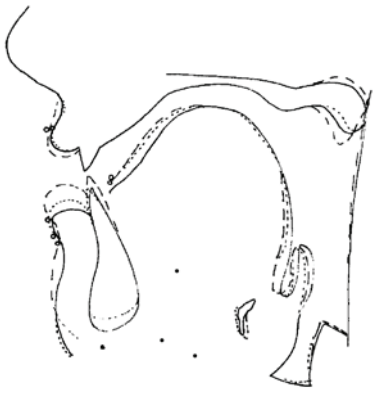
FIG. 6.1: **6.1a** *: Coupes medio-sagittales du conduit vocal pour la voyelle /a/. Pour chacun des coupes, on donne le score phonétique, le lieu de constriction maximale (mesuré à partir de la glotte), et l'aire au niveau de cette constriction en* cm^2*.* **6.1b** *: Résultats de l'inversion pour la voyelle /a/ représenté sur un diagramme position de la constriction / aire à la constriction. La position de la constriction est en cm, comptée à partir de la glotte, et l'aire à la constriction est en* cm^2*. La couleur d'un point indique son score phonétique (plus le point est sombre, meilleur est son score).* **6.1c** *: Radiographies de coupes medio-sagittale pour la voyelle /a/, et leur contexte : /ab***a***/ (trait plein) /m***a***ʃ/ (trait discontinu) /vw***a***l/ (trait pointillé).*



(a)

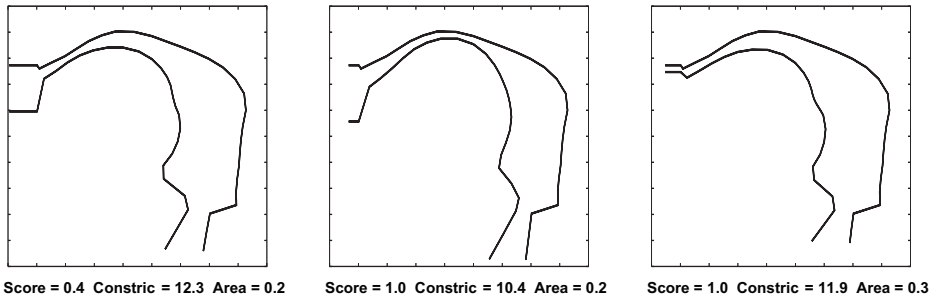


(b)

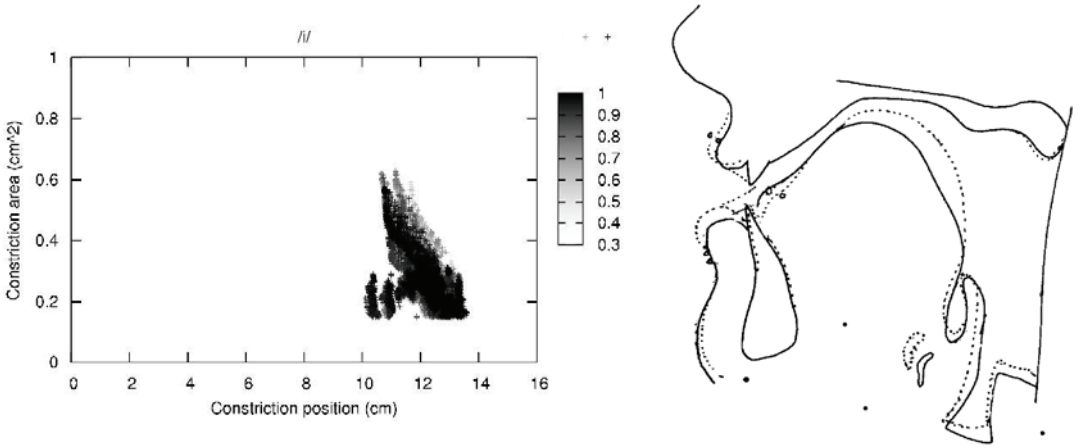


(c)

FIG. 6.2: **6.2a** *: Coupes medio-sagittales du conduit vocal pour la voyelle /e/. Pour chacun des coupes, on donne le score phonétique, le lieu de constriction maximale (mesuré à partir de la glotte), et l'aire au niveau de cette constriction en* cm^2*.* **6.2b** *: Résultats de l'inversion pour la voyelle /e/ représenté sur un diagramme position de la constriction / aire à la constriction. La position de la constriction est en cm, comptée à partir de la glotte, et l'aire à la constriction est en* cm^2*. La couleur d'un point indique son score phonétique (plus le point est sombre, meilleur est son score).* **6.2c** *: Radiographies de coupes medio-sagittale pour la voyelle /e/, et leur contexte : /dyg***e***/ (trait plein), /d***e***bu/ (trait discontinu), /t***e***bo/ (trait pointillé).*

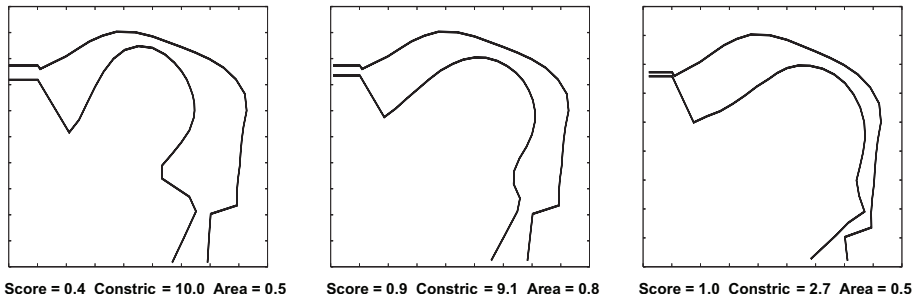


(a)

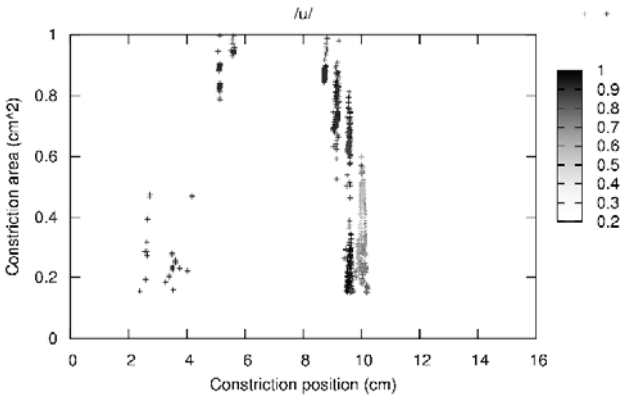


(b) (c)

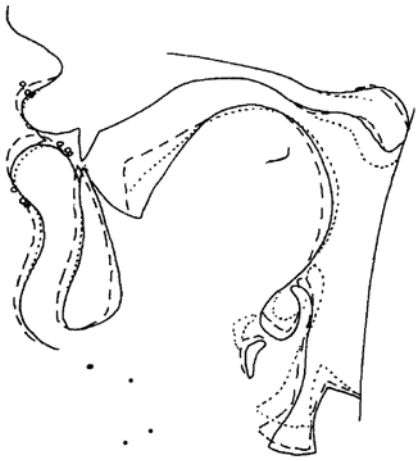
FIG. 6.3: **6.3a** *: Coupes medio-sagittales du conduit vocal pour la voyelle /i/. Pour chacun des coupes, on donne le score phonétique, le lieu de constriction maximale (mesuré à partir de la glotte), et l'aire au niveau de cette constriction en* cm^2. **6.3b** *: Résultats de l'inversion pour la voyelle /i/ représentés sur un diagramme position de la constriction / aire à la constriction. La position de la constriction est en cm, comptée à partir de la glotte, et l'aire à la constriction est en* cm^2*. La couleur d'un point indique son score phonétique (plus le point est sombre, meilleur est son score).* **6.3c** *: Radiographies de coupes medio-sagittale pour la voyelle /i/, et leur contexte : /ab***i***/ (trait plein), /lw***i***pã/ (trait pointillé).*



(a)

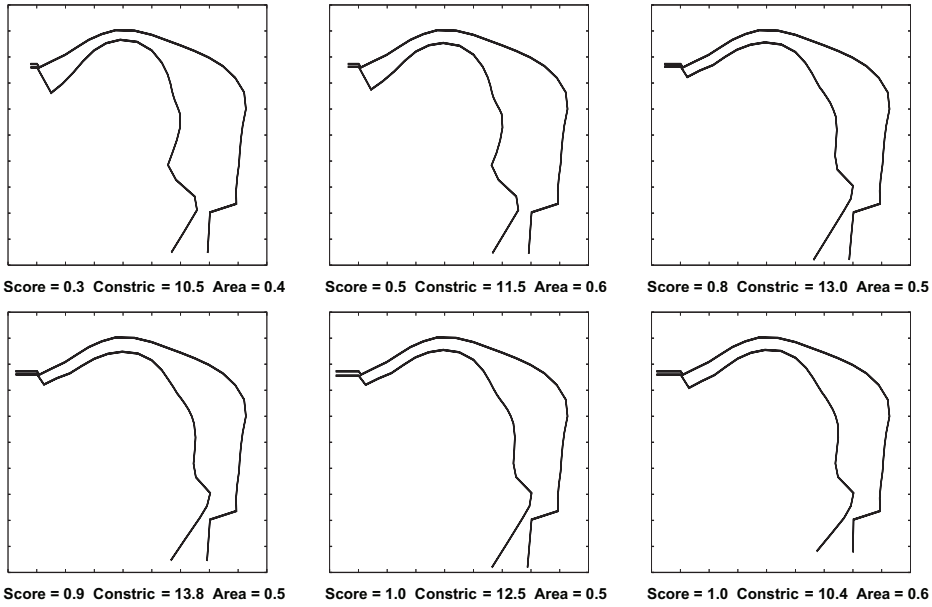


(b)

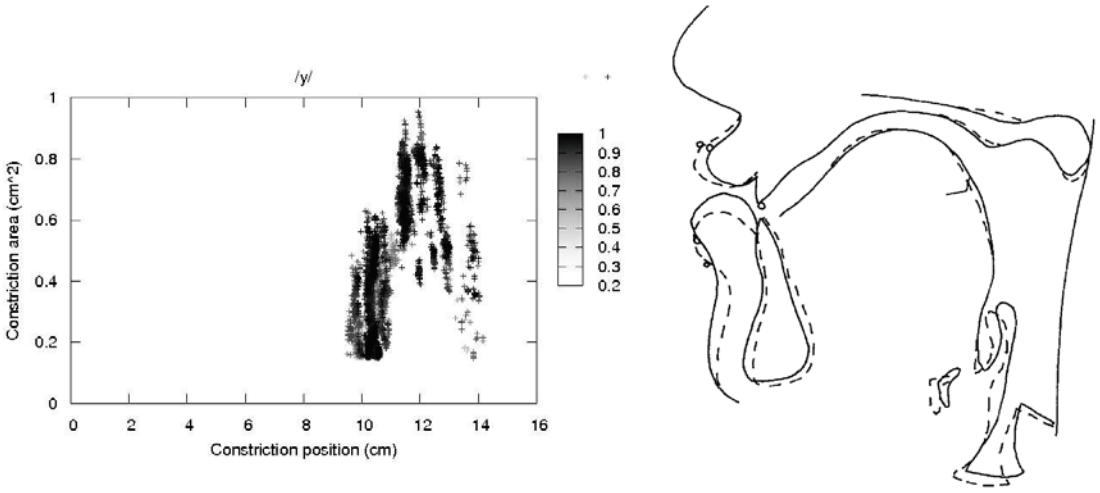


(c)

FIG. 6.4: **6.4a** *: Coupes medio-sagittales du conduit vocal pour la voyelle /u/. Pour chacun des coupes, on donne le score phonétique, le lieu de constriction maximale (mesuré à partir de la glotte), et l'aire au niveau de cette constriction en* cm^2. **6.4b** *: Résultats de l'inversion pour la voyelle /u/ représentés sur un diagramme position de la constriction / aire à la constriction. La position de la constriction est en cm, comptée à partir de la glotte, et l'aire à la constriction est en* cm^2*. La couleur d'un point indique son score phonétique (plus le point est sombre, meilleur est son score).***6.4c** *: Radiographies de coupes medio-sagittale pour la voyelle /u/, et leur contexte : /ik***u***/ (trait plein), /ʃ***u***/ (trait discontinu), /ɾ***u***si/ (trait pointillé).*



(a)



(b) (c)

FIG. 6.5: **6.5a** *: Coupes medio-sagittales du conduit vocal pour la voyelle /y/. Pour chacun des coupes, on donne le score phonétique, le lieu de constriction maximale (mesuré à partir de la glotte), et l'aire au niveau de cette constriction en* cm^2. **6.5b** *: Résultats de l'inversion pour la voyelle /y/ représentés sur un diagramme position de la constriction / aire à la constriction. La position de la constriction est en cm, comptée à partir de la glotte, et l'aire à la constriction est en* cm^2*. La couleur d'un point indique son score phonétique (plus le point est sombre, meilleur est son score).* **6.5c** *: Radiographies de coupes medio-sagittale pour la voyelle /y/, et leur contexte : /ig***y***/ (trait plein), /***y***n/ (trait discontinu).*

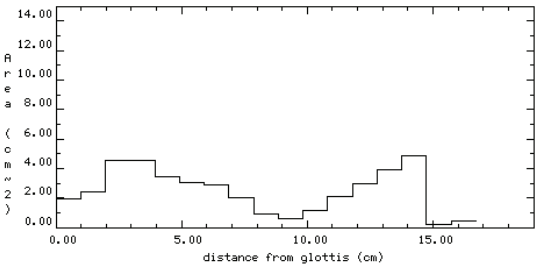


FIG. 6.6: *Fonction d'aire pour un /u/ avec une constriction pharyngale.*

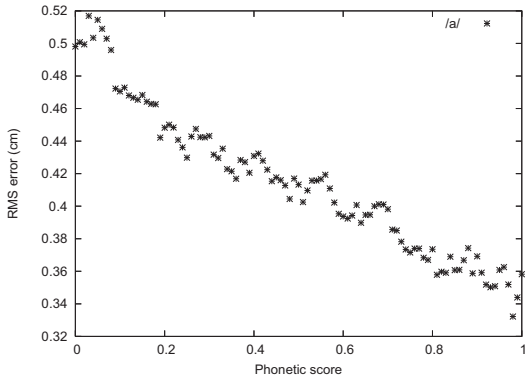


FIG. 6.7: *Distance RMS moyenne (en cm) de la forme de conduit réelle en fonction du score phonétique, pour les solutions de l'inversion du /a/.*

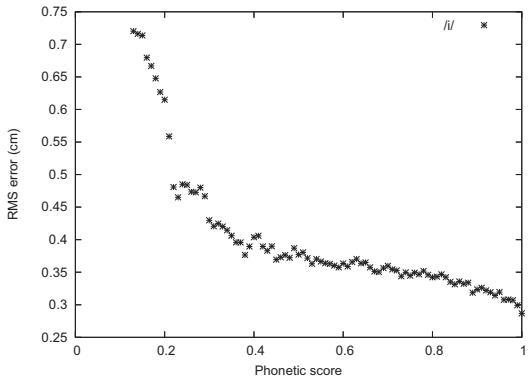


FIG. 6.8: *Distance RMS moyenne (en cm) de la forme de conduit réelle en fonction du score phonétique, pour les solutions de l'inversion du /i/.*

6.3 Inversion dynamique

L'inversion dynamique est la composante la plus intéressante de l'inversion. La principale difficulté de notre méthode (comme de la majorité des méthodes d'inversion) est que l'on manque de références articulatoires auxquelles comparer les résultats. Dans de nombreux cas, cette difficulté est ignorée en ne mesurant pas l'écart par rapport à une référence articulatoire, mais tout simplement en ne prenant en compte que l'écart du signal acoustique resynthétisé par rapport au signal acoustique de départ. Cela permet notamment d'éliminer les problèmes liés à l'inexactitude du synthétiseur articulatoire (la synthèse des vecteurs articulatoires mesurés sur les radiographies ne permet pas d'obtenir exactement les vecteurs acoustiques mesurés), ainsi que le fait que l'inversion acoustique articulatoire est un problème mal posé : il existe une infinité de configurations du conduit vocal permettant d'obtenir un vecteur acoustique donné. Nous désignerons cette façon d'inverser la version « faible » de l'inversion. Un autre problème, légèrement plus simple que le problème général, est l'inversion d'un signal de parole artificiel produit par un synthétiseur à partir de trajectoires articulatoires. Cela permet de ne pas souffrir des insuffisances du système de synthèse puisque cela implique une parfaite correspondance. Nous appellerons ce problème l'inversion « moyenne ». Nous nous intéressons ici à la version « forte » du problème : nous tentons de retrouver les configurations articulatoires initiales à partir du signal sonore réel.

Dans cette étude, nous disposons d'une référence assez fiable : nous disposons des trajectoires des vecteurs articulatoires pour la locutrice PB sur une dizaine de phrases françaises. Malheureusement, notre méthode n'était pas encore adaptée, au moment de la rédaction de cette thèse, à l'inversion d'autres sons que les voyelles. L'adaptation à tous les types de sons (par le biais de l'utilisation de paramètres cepstraux) permettra dans l'avenir d'inverser toutes les séquences. Nous avons cependant pu inverser des séquences VV (Voyelle-Voyelle) pour la locutrice PB.

La première difficulté concerne l'extraction de formants elle-même. Le signal est généralement trop bruité pour pouvoir extraire une information précise sur les formants, même manuellement (ce qui limite encore d'avantage les échantillons de parole sur lesquels on peut travailler). Par ailleurs, les séquences VV sont très peu nombreuses.

6.3.1 Inversion de transitions Voyelle-Voyelle

Du corpus de PB, qui comporte 10 phrases (cf. l'annexe A), nous n'avons pu extraire aucune véritable séquences VV, mais uniquement des séquences d'une semi-voyelle suivie d'une voyelle : /wa/ dans « Voilà », /wi/ dans « Louis », et /ɥi/ dans « Lui ». Ce n'est de toute évidence pas une situation très favorable pour notre étude, les trajectoires formantiques des semi-voyelles ayant été particulièrement difficiles à extraire ; les séquences ont toutefois pu être inversées. À noter que le suivi de formants a été effectué de façon semi-automatique, et non plus manuellement comme pour l'inversion statique. Différentes inversions ont été réalisées :

- inversion des formants synthétiques (générés à partir des données articulatoires),
- inversion des formants mesurés,
- inversion des formants synthétiques avec contraintes phonétiques,
- inversion des formants mesurés avec contraintes phonétiques.

Pour l'inversion avec contraintes phonétiques, différentes valeurs pour le poids des contraintes phonétiques ont été testées. En parallèle, différents poids pour la contrainte d'écart à la position neutre (cf. section 4.1.1) ont été testés. Le poids des contraintes phonétiques sera noté dans la suite Rp, celui pour la contrainte d'écart au neutre Rk.

Signal inversé	d_1 (cm)			d_2		
	inv. stat.		dyn.	inv. stat.		dyn.
	min.	moy.	moy.	*min.*	moy.	moy.
synthétique	*0.05*	0.48	0.12	*0.21*	1.52	0.62
naturel	*0.20*	0.67	0.22	*0.60*	1.52	0.71

TAB. 6.2: *Erreurs géométriques et articulatoires des trajectoires trouvées par inversion des formants synthétiques et naturels pour la séquence /*wa*/ (sans contraintes phonétiques) : erreur minimale parmi l'ensemble des solutions statiques, erreur moyenne pour le même ensemble de solutions, et enfin erreur de la solution trouvée après régularisation variationnelle.*

6.3.1.1 Séquence /wa/

La figure 6.9 présente les trajectoires des formants que nous avons inversées : celles du signal synthétique généré à partir des paramètres articulatoires déduits des radiographies (en pointillés), et celles mesurées sur le signal acoustique réel (en trait continu). On constate une certaine disparité entre les formants synthétiques et réels, qui est due à plusieurs facteurs. D'une part, le signal acoustique original étant particulièrement bruité, il y a une erreur conséquente pour les formants mesurés, en particulier pour F2 : il y a en effet un renforcement spectral important autour de 1500Hz dû à la machine aux rayons-X (cf. figure 3.1), ce qui perturbe la mesure de F2. D'autre part, les données articulatoires ne correspondent pas exactement aux radiographies, mais aux vecteurs articulatoires dont la représentation géométrique est la plus proche de l'original. Enfin, la synthèse articulatoire n'est pas parfaite. La différence entre les formants originaux et synthétiques n'est donc guère étonnante ; on constate cependant que leurs trajectoires ont des tendances similaires, il est donc envisageable que les trajectoires des vecteurs articulatoires trouvés par inversion aient également des tendances similaires.

Le tableau 6.2 présente une mesure de l'erreur – au sens des distances d_1 et d_2 présentées à la section 4.3 – des solutions de l'inversion, par rapport aux trajectoires articulatoires originales, sans utiliser les contraintes phonétiques. Pour l'inversion statique, on présente l'erreur moyenne, ainsi que l'erreur minimale moyenne ; pour l'inversion dynamique, on présente l'erreur moyenne pour la solution finale (i.e. après lissage non-linéaire et régularisation variationnelle).

On constate qu'au sens de la norme géométrique d_1, la solution issue de l'inversion des formants synthétiques est nettement meilleure que celle issue de l'inversion des formants naturels. Les performances des deux solutions sont en revanche très comparables concernant la distance articulatoire d_2.

On constate également que la solution finale de l'inversion des formants naturels a une erreur proche de l'erreur minimale, pour les deux normes, alors que la solution finale de l'inversion des formants synthétiques reste assez éloignée de l'erreur minimale. Cela indique que les contraintes dynamiques sont assez pertinentes : elles permettent de retrouver des trajectoires articulatoires ayant des tendances correctes, même sur une entrée erronée. Il est particulièrement remarquable de constater que la solution finale trouvée est pratiquement la solution acoustiquement correcte la plus proche possible de l'originale. On remarquera au passage que par rapport à l'expérience réalisée à la section 4.3, l'erreur minimale est nettement plus faible : les séquences étant plus courtes, on s'est permis ici de réaliser l'inversion dans des conditions plus favorables, en utilisant un codebook plus précis et en générant un nombre de solutions plus important lors de l'inversion statique.

Nous présentons également les trajectoires (originale, obtenue par inversion des formants

synthétique, obtenue par inversion des formants naturels) pour les paramètres articulatoires principaux sur la figure 6.10. On constate que les trajectoires inverses ont généralement les bonnes tendances (voir en particulier l'ouverture des lèvres et la forme de la langue), et que les erreurs articulatoires sont comparables pour l'inversion des deux signaux.

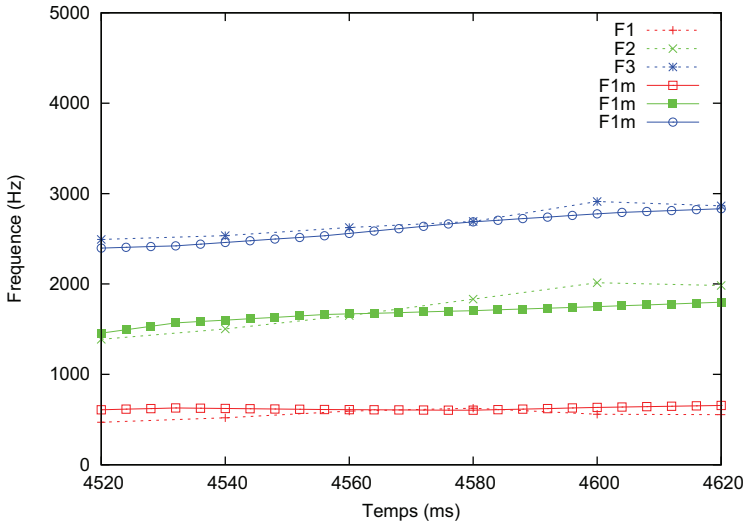


FIG. 6.9: *Trajectoires acoustiques à inverser : les trois premiers formants du signal synthétique sont en pointillés, les formants mesurés sur le signal acoustique original sont représentés en trait continu. L'abscisse représente le temps, en millisecondes, et l'ordonnée la fréquence, en Hertz.*

Enfin, à la figure 6.11 on présente les résultats de l'inversion dynamique pour différentes valeurs des paramètres Rk et Rp, sur les signaux acoustiques synthétiques et naturels, et ce pour les deux distances d_1 et d_2. Il s'agit des résultats à l'issue du lissage non-linéaire.

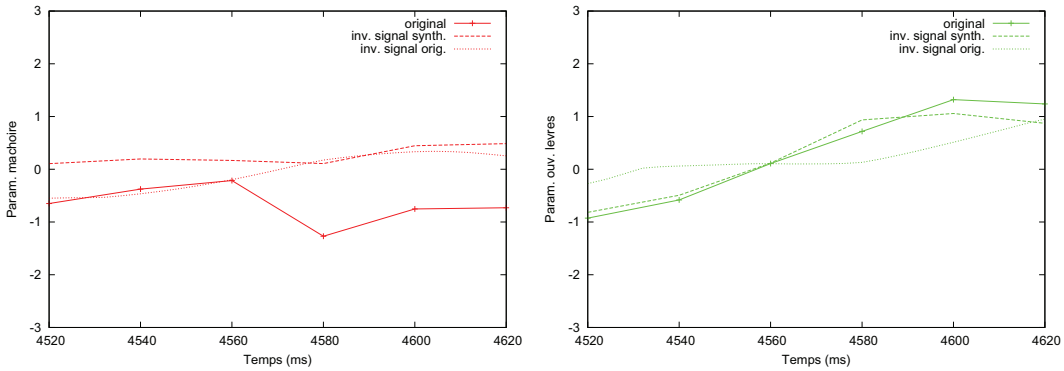
On constate que pour l'inversion du signal synthétique, l'application des contraintes phonétiques n'améliore pas les résultats, que ce soit pour la distance géométrique (Figure 6.11c) ou pour la distance articulatoire (Figure 6.11d), l'erreur minimale étant obtenue pour $Rp \sim 0$ et $Rk \sim 1$. On remarque par ailleurs sur ces figures que la contrainte d'écart au neutre (valeurs élevées de Rk) améliore significativement les résultats par rapport à la contrainte purement dynamique, ce qui confirme le bien fondée de cette contrainte.

Pour le signal naturel, en revanche, l'utilisation des contraintes phonétiques couplées à la contrainte d'écart au neutre améliore significativement la trajectoire pour la distance géométrique d_1 (voir figure 6.11c), le minimum global de l'erreur étant obtenu autour de ($Rp \sim 2$, $Rk \sim 15$).

Pour la distance articulatoire d_2 (figure 6.11d), l'effet des contraintes phonétiques est moins significatif. L'erreur atteint un minimum local autour du point précédent, mais le minimum global est atteint pour les valeurs élevées de R_k.

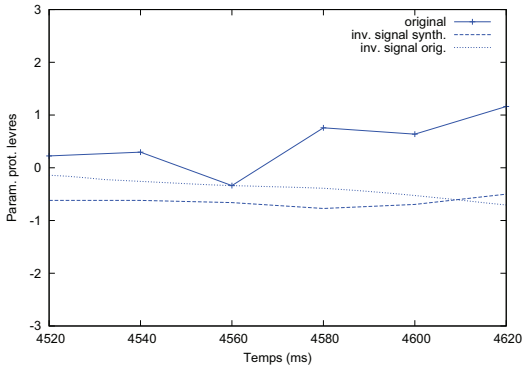
L'effet des contraintes phonétiques est ici loin d'être convainquant : elles semblent ne pas améliorer la trajectoire obtenue lors de l'inversion du signal synthétique, et elles ne semblent améliorer que très légèrement la solution lors de l'inversion du signal naturel, et ce seulement pour la distance géométrique.

Ces mauvais résultats doivent cependant être relativisés : concernant le signal synthétique, les résultats de l'inversion sont de toute façon exceptionnellement bons donc difficilement améliorables, et il difficile d'obtenir des améliorations pour la distance d_2, du fait de la forme particulière des trajectoires articulatoires originales (cf. l'exemple de la section 4.3).

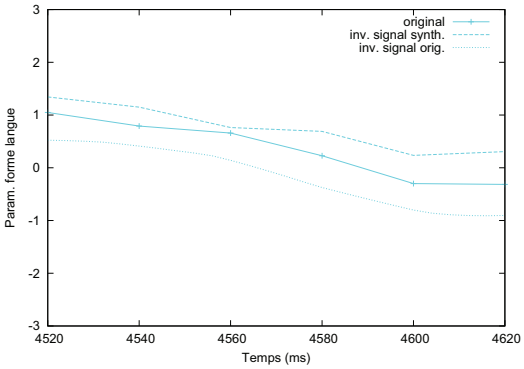


(a) Mâchoire

(b) Ouverture des lèvres



(c) Protrusion des lèvres



(d) Forme de la langue

FIG. 6.10: *Trajectoires articulatoires principales pour /wa/. Pour chacun des paramètres présentés, la trajectoire originale (trait continu), la trajectoire issue de l'inversion du signal synthétique (tirets), et la trajectoires issue de l'inversion du signal naturel (pointillés) sont présentées.*

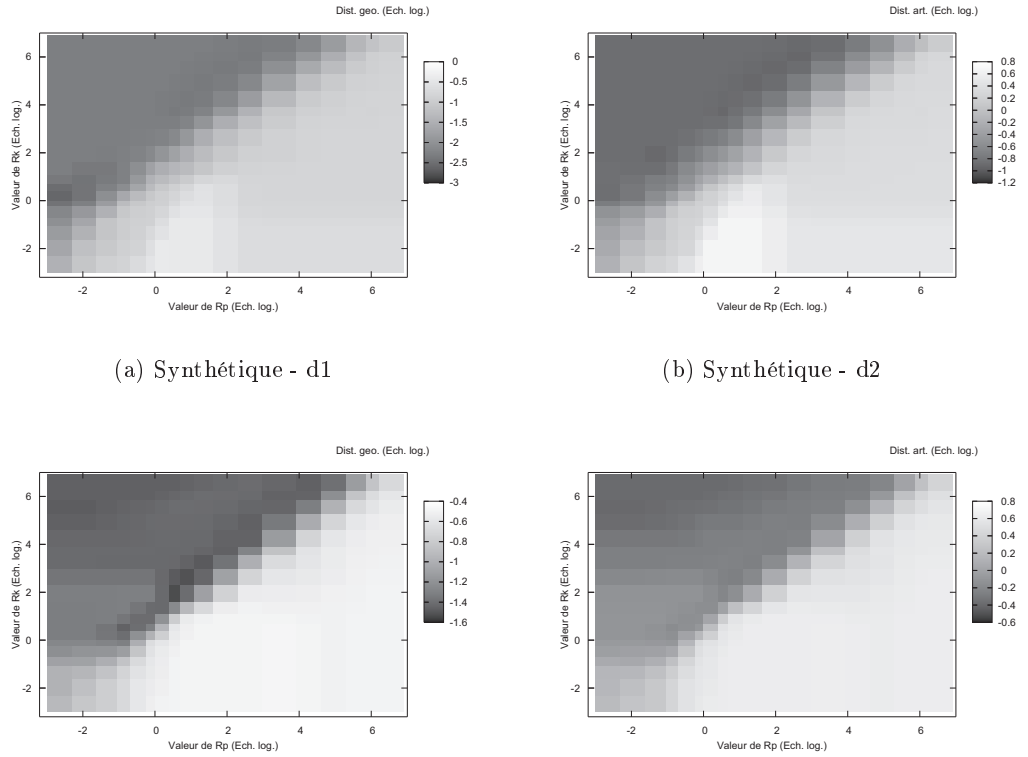
Par ailleurs, il est également possible que le modèle de sélection acoustique ne parvienne pas à déterminer les classes de contraintes appropriées à partir du signal acoustique ; ou encore simplement parce que la semi-voyelle /w/ n'a pas de classe de contraintes appropriée dans le modèle actuel, uniquement adapté aux voyelles.

6.3.1.2 Séquence /wi/

La figure 6.12 présente les trajectoires des formants pour la séquence /wi/. On observe à nouveau une importante disparité entre les formants synthétiques et naturels : au début de la séquence, il y a visiblement une importante erreur sur le formant F2 mesuré. À l'instant 12580, il y a probablement une erreur dans les données articulatoires.

Le tableau 6.3 présente une mesure de l'erreur des solutions de l'inversion, par rapport aux trajectoires articulatoires originales. On constate que les résultats sont très différents de ceux obtenus pour la séquence précédente : l'inversion de la séquence acoustique synthétique obtient d'assez mauvais résultats, et l'inversion de la séquence acoustique naturelle obtient des résultats catastrophiques : la séquence obtenue fait à peine mieux que la moyenne de l'erreur des solutions statiques.

On présente sur la figure 6.13 les résultats de l'inversion dynamique pour différentes valeurs



(a) Synthétique - d1

(b) Synthétique - d2

(c) Naturel - d1

(d) Naturel - d2

FIG. 6.11: *Erreur de la trajectoire articulatoire trouvée par lissage non linéaire, en fonction de Rp et Rk, pour la séquence /*wa*/. Échelles logarithmiques. Le dégradé de gris indique la valeur de l'erreur : plus la couleur est sombre, plus l'erreur est faible.*

Signal inversé	d_1 (cm)			d_2		
	inv. stat.		dyn.	inv. stat.		dyn.
	min.	moy.	moy.	*min.*	moy.	moy.
synthétique	*0.05*	0.52	0.28	*0.22*	1.60	0.87
naturel	*0.35*	0.66	0.60	*0.88*	1.77	1.59

TAB. 6.3: *Erreurs géométriques et articulatoires des trajectoires trouvées par inversion des formants synthétiques et naturels pour la séquence /*wi*/ (sans contraintes phonétiques).*

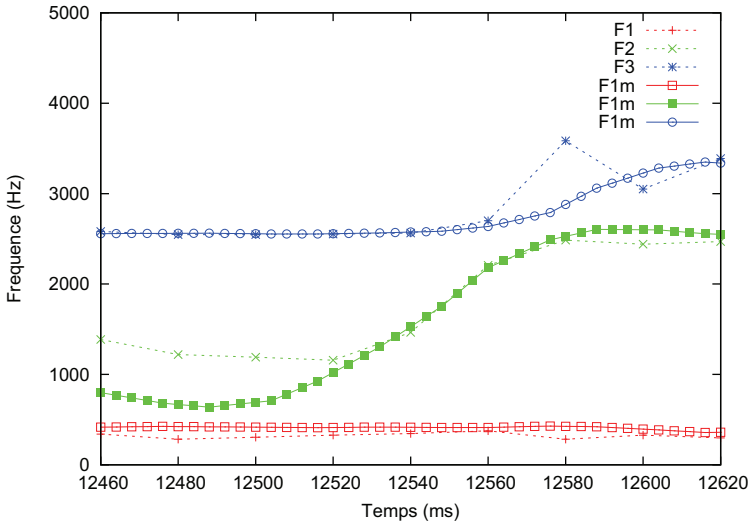


FIG. 6.12: *Trajectoires acoustiques à inverser : les trois premiers formants du signal synthétique sont en pointillés, les formants mesurés sur le signal acoustique original sont représentés en trait continu. L'abscisse représente le temps, en millisecondes, et l'ordonnée la fréquence, en Hertz.*

des paramètres Rk et Rp, sur les signaux acoustiques synthétiques et naturels, et ce pour les deux distances d_1 et d_2.

On constate ici que les contraintes phonétiques améliorent les trajectoires obtenues, que ce soit pour l'inversion du signal synthétique ou celle du signal naturel, et ce pour les distances géométriques et articulatoires (avec une amélioration nettement plus sensible concernant la distance géométrique).

À nouveau, on constate que le minimum global de l'erreur est atteint pour une combinaison de Rk et Rp, ce qui indique l'utilité de ces deux types de contraintes. En revanche, les couples de valeurs permettant de minimiser l'erreur semblent ne pas être identiques pour tous les cas, même si on retrouve plusieurs fois le couple ($Rp \sim 2$, $Rk \sim 15$).

On remarque également que les gains sont à nouveau moins significatifs pour la distance articulatoire que pour la distance géométrique.

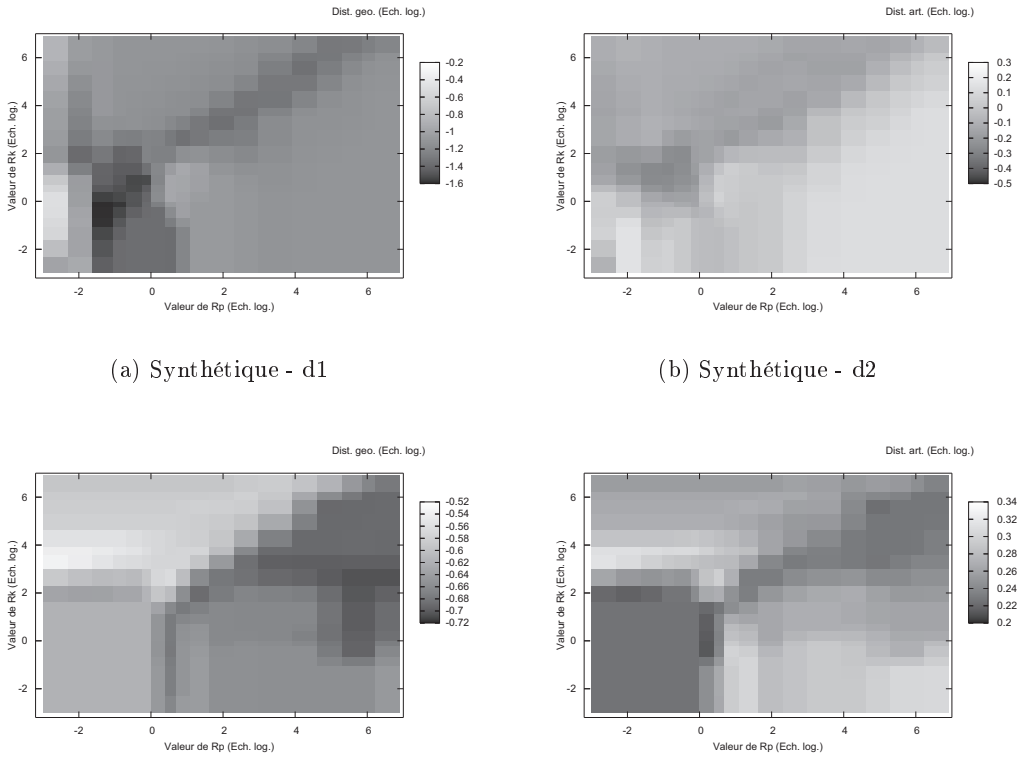
Enfin, comparé aux résultats de la séquence /wa/ vue précédemment, on obtient des erreurs nettement plus importantes, tant pour le signal naturel que pour le signal synthétique.

6.3.1.3 Séquence /ɥi/

La figure 6.14 présente les trajectoires des formants pour la séquence /ɥi/. Cette fois-ci, les formants synthétiques et mesurés sont assez proches, avec toutefois un écart relativement important pour F2.

On présente à la figure 6.15 la distance à l'original de la trajectoire trouvée par inversion dynamique pour différentes valeurs des paramètres Rk et Rp, sur les signaux acoustiques synthétiques et naturels, pour les deux distances d_1 et d_2.

L'effet des contraintes phonétiques est ici assez mitigé : elles semblent améliorer ponctuellement les solutions pour l'inversion du signal naturel (ligne sombre autour de $\log(Rp) = -1$), et les minimaux globaux sont obtenus pour des valeurs de Rp différentes de 0, mais le gain apporté n'est guère significatif. Les gains les plus importants sont à nouveau apportés par la contrainte



(a) Synthétique - d1

(b) Synthétique - d2

(c) Naturel - d1

(d) Naturel - d2

FIG. 6.13: *Erreur de la trajectoire articulatoire trouvée par lissage non linéaire, en fonction de Rp et Rk, pour la séquence /*wi*/. Échelles logarithmiques. Le dégradé de gris indique la valeur de l'erreur : plus la couleur est sombre, plus l'erreur est faible.*

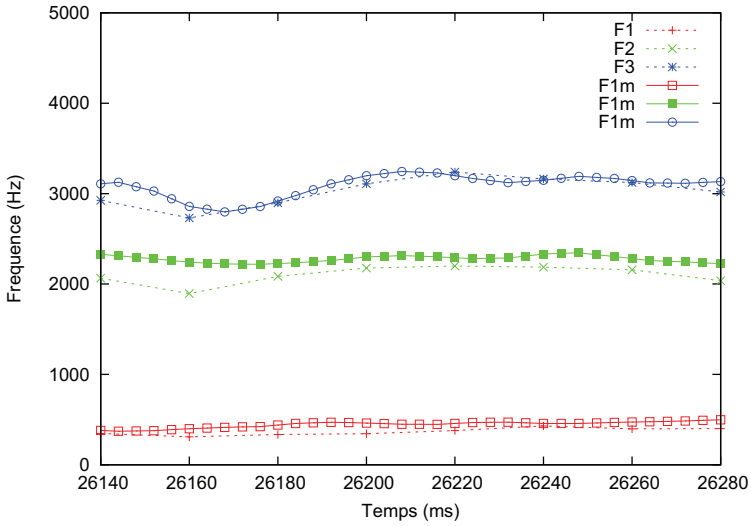


FIG. 6.14: *Trajectoires acoustiques à inverser : les trois premiers formants du signal synthétique sont en pointillés, les formants mesurés sur le signal acoustique original sont représentés en trait continu. L'abscisse représente le temps, en millisecondes, et l'ordonnée la fréquence, en Hertz.*

d'écart au neutre.

Il est à nouveau probable que les mauvais résultats pour cette séquence puisse être imputés au modèle de sélection acoustique, ou à l'inexistence d'une classe de contraintes spécifique pour la semi-voyelle /ɥ/.

6.3.2 Récapitulatif

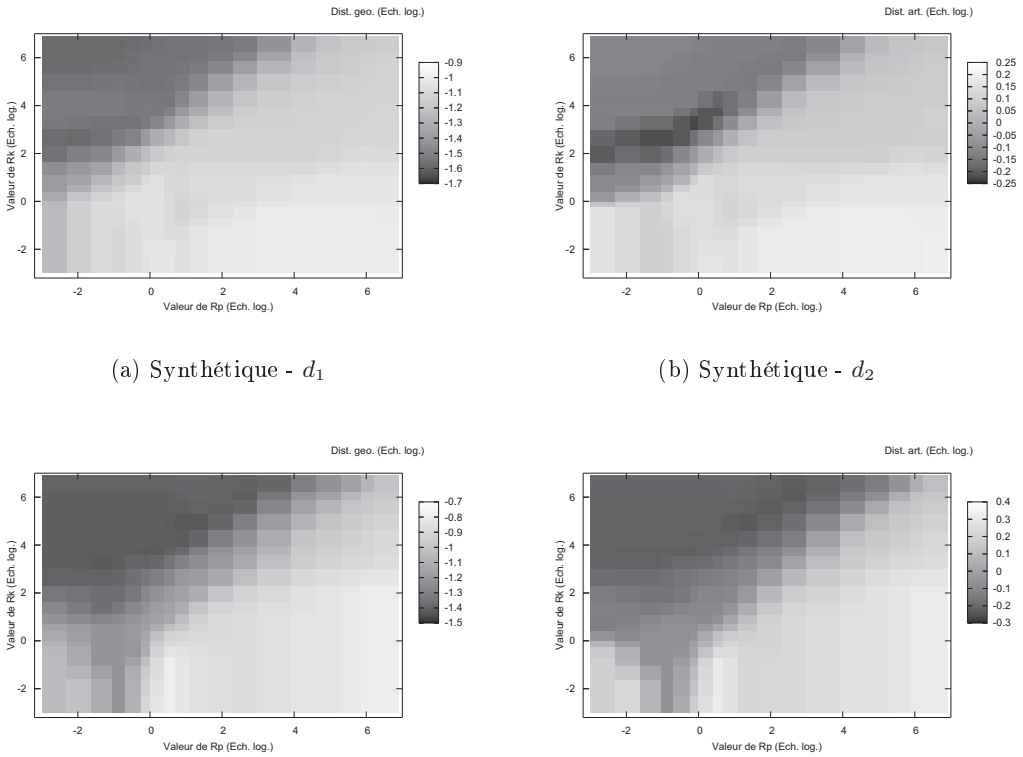
Nous avons vu dans les paragraphes précédents des expériences préliminaires, dans l'optique d'utiliser les contraintes phonétique de façon générique pour l'inversion dynamique.

Le principal constat est que l'utilisation des contraintes phonétiques seules ne diminue l'erreur que de façon médiocre. Un second constat est que la contrainte d'écart au neutre utilisée de façon adéquate améliore considérablement les performances. Un troisième constat est que malheureusement il ne semble pas exister de paramètres miracles donnant systématiquement les meilleurs résultat ; cependant le couple ($Rp \sim 2$, $Rk \sim 15$) semble donner d'assez bons résultats dans pratiquement tous les cas.

Le tableau 6.4 récapitule la distance moyenne par rapport à la trajectoire articulatoire originale au sens des normes articulatoire et géométrique après lissage non-linéaire, sur les trois séquences, et pour différentes modalité : contraintes dynamiques seules, solution optimale en utilisant une combinaison des contraintes dynamiques et contraintes d'écart au neutre, solution optimale en utilisant une combinaison des contraintes dynamiques et contraintes phonétiques, solution optimale en utilisant une combinaison de toutes ces contraintes, et enfin solution « générique » – c'est-à-dire utilisant le couple de valeurs évoqué précédemment.

L'étude de ce tableau révèle plusieurs éléments.

Premièrement, par rapport aux contraintes dynamiques seules, la combinaison avec la contrainte d'écart au neutre « optimale » peut améliorer la solution trouvée de façon spectaculaire, dépassant souvent la précision de la solution trouvée après régularisation variationnelle. Cela est cependant à relativiser : les conditions d'obtention ne sont pas réalistes, puisqu'elles supposent



(a) Synthétique - d_1

(b) Synthétique - d_2

(c) Naturel - d_1

(d) Naturel - d_2

FIG. 6.15: *Erreur de la trajectoire articulatoire trouvée par lissage non linéaire, en fonction de Rp et Rk, pour la séquence /ɥi/. Échelles logarithmiques. Le dégradé de gris indique la valeur de l'erreur : plus la couleur est sombre, plus l'erreur est faible.*

Signal inversé		d_1 (cm)					d_2				
		A	*B*	*C*	*D*	E	A	*B*	*C*	*D*	E
/wa/	synth.	0.46	0.10	0.28	0.075	**0.095**	1.48	0.42	0.89	0.37	**0.39**
	nat.	0.37	0.23	0.37	0.21	**0.21**	1.19	**0.67**	1.18	0.67	0.77
/wi/	synth.	0.57	0.41	**0.21**	0.20	0.26	1.02	0.69	**0.65**	0.65	0.81
	nat.	0.54	0.54	0.51	0.49	**0.51**	1.25	1.25	1.25	1.22	**1.24**
/ɥi/	synth.	0.29	**0.20**	0.29	0.20	0.26	1.17	**0.79**	1.09	0.79	0.80
	nat.	0.35	**0.24**	0.30	0.23	0.25	1.16	0.81	0.94	0.79	**0.80**
Moyenne		0.43	0.29	0.33	0.23	**0.26**	1.21	**0.77**	1.00	0.75	0.80

TAB. 6.4: *Erreurs géométriques et articulatoires des trajectoires trouvées par inversion des formants synthétiques et naturels pour les séquence /wa/,/wi/,/ɥi/ avec différents type de contraintes : contraintes purement dynamique (A), dynamique + écart au neutre minimal (B), dynamique + score phonétique minimal (C), dynamique + écart au neutre + score phonétique minimal (D), dynamique + écart au neutre + score phonétique générique (E). Pour chaque expérience, le type de contraintes donnant le meilleur résultat (exception faite de la condition D) est mis en valeur.*

de disposer de la trajectoire articulatoire originale.

Deuxièmement, si les performances des contraintes phonétiques seules ne sont pas extraordinaires, elles permettent tout de même – dans des conditions là aussi artificielles – d'obtenir de bien meilleurs résultats qu'avec les contraintes dynamiques seules.

Enfin, et c'est certainement le plus intéressant, la combinaison des contraintes phonétiques et d'écart au neutre « génériques » permet une diminution importante de l'erreur (-40% pour d_1, -34% pour d_2) par rapport aux contraintes dynamiques seules.

6.4 Conclusions et perspectives

Les contraintes phonétiques constituent une manière relativement simple d'améliorer les résultats de l'inversion statique comme nous l'avons vérifié pour les voyelles isolées. Leur intérêt pour l'inversion dynamique n'a pour le moment pas été vraiment démontré : pour certaines séquences, la fidélité des trajectoires articulatoires trouvées est effectivement nettement meilleure, pour d'autres le gain est négligeable. Il convient de souligner que ces derniers tests ne portent que sur un nombre très limité d'exemples, qui plus est des exemples sans doute assez peu pertinents. Les données actuellement en cours d'acquisition dans le cadre du projet ASPI permettront d'obtenir une évaluation plus pertinente.

Ce qui fait l'intérêt de ces contraintes est leur généralité puisqu'elles ont été définies à partir de connaissances absolument indépendantes du locuteur. Cela est sans doute aussi une faiblesse puisqu'elles sont donc assez imprécises et peut-être trop laxistes puisque les domaines articulatoires pour lesquels le score phonétique vaut 1 sont très étendus, et surtout ont été définis manuellement. Il serait utile de les déterminer à partir de l'analyse statistique de données articulatoires réelles sur un nombre important de locuteurs. Il serait également intéressant d'utiliser les contraintes phonétiques avec un modèle non plus générique, mais adapté au locuteur, et d'étudier les gains de performance résultants.

Une autre perspective porte sur la définition de contraintes phonétiques dynamiques, i.e. qui

s'appliquent à des séquences (diphone ou triphone), avec l'avantage de capturer des effets de réorganisation et de planification articulatoires dus à la coarticulation. Mais pour cela, il sera nécessaire de disposer de données articulatoires en quantité très importante pour établir des statistiques pertinentes.

7

Contraintes visuelles

Introduction

Une des voies prometteuses pour contourner les difficultés de l'inversion acoustique-articulatoire est de ne plus se contenter d'utiliser le signal acoustique comme entrée, mais d'utiliser des données supplémentaires provenant d'autres sources, soit, en d'autres termes, faire de l'inversion multimodale. Plusieurs sources peuvent être envisagées, mais il s'agit généralement de données articulatoires partielles relativement faciles à obtenir, et cela à un coût modéré : données EMA, échographie, ou tout simplement vidéo.

L'approche présentée dans ce chapitre repose sur un système d'acquisition à deux caméras permettant de suivre la position tridimensionnelle de marqueurs peints sur le visage d'un locuteur. Ce système permet de déterminer de façon précise, peu coûteuse et avec une bonne résolution temporelle la position des marqueurs sur le visage. Les positions de ces marqueurs peuvent être alors être exploitées comme information supplémentaire pour l'inversion.

Dans le cadre de cette thèse, nous avons développé un système permettant de déterminer automatiquement, à partir des coordonnées de ces marqueurs, les valeurs des paramètres du modèle articulatoire correspondant aux articulateurs visibles : position de la mâchoire, ouverture et protrusion des lèvres, ce qui permet de restreindre considérablement l'espace des solutions de l'inversion. Cette étude reste cependant préliminaire : nous n'avons malheureusement pas été en mesure de quantifier les gains de performance apportés par cette méthode par rapport à l'inversion effectuée sur les données acoustiques seules.

7.1 Origine

La nature multimodale de la parole est connue et étudiée depuis longtemps. La parole est en effet un signal bimodal, comportant une composante acoustique, et une composante visuelle : la vue du locuteur. Ces deux modalités sont fortement corrélées et redondantes. Il a été observé à de nombreuses reprises que les locuteurs et auditeurs humains exploitent la nature multimodale de la parole, et plus particulièrement les indices articulatoires : l'intelligibilité de la parole augmente dans des conditions d'écoute difficiles (déficiences auditives, environnement bruyant...) lorsque l'auditeur voit le locuteur (Sumby & Pollack 1954; Benoît *et al.* 1994; Robert-Ribes *et al.* 1994; Le Goff 1997). La bimodalité de la parole se manifeste aussi de façon particulièrement spectaculaire avec l'effet McGurk (McGurk & MacDonald 1976), où la perception d'une vidéo truquée modifie la perception acoustique.

L'inversion audiovisuelle, ou l'utilisation de données visibles pour faciliter l'inversion, est une

idée relativement récente. Plusieurs études ont déjà été réalisées dans ce cadre, notamment par Engwall (Engwall 2005) ou Katsamanis (Katsamanis *et al.* 2007). L'originalité de notre approche réside dans l'origine des données visibles, en particulier notre système d'acquisition vidéo utilisant deux caméras tout en permettant d'acquérir des données précises et en quantité importante. De plus, nous sommes – à notre connaissance – les seuls à effectuer de l'inversion multimodale à l'aide d'une méthode d'analyse par synthèse.

Le développement du module permettant d'exploiter les données d'origine visible pour l'inversion est paradoxalement le premier à avoir été réalisé, puisqu'il s'agissait de l'objet de mon stage de DEA, et aussi celui qui a mis le plus de temps à parvenir à maturité, car l'acquisition et le traitement des données – en particulier le développement des différents modules permettant l'acquisition – ont pris beaucoup de temps, pour diverses raisons. En premier lieu, il n'y avait pas de logiciel d'acquisition permettant d'acquérir le flux des deux caméras de façon synchronisée, il a donc été nécessaire de l'écrire. Ce logiciel a presque entièrement été réalisé par mes soins. Il a ensuite été nécessaire de réaliser les différents logiciels permettant de retrouver la position tridimensionnelle des marqueurs à partir des images. Il a également fallu traiter ces données de façon à les rendre utilisables pour l'inversion, ce qui s'est avéré être un travail long et difficile. Enfin, les données articulatoires qui permettraient de valider la méthode n'ont pu être disponibles qu'à la fin de ma thèse. Ce travail sera donc complété dans un avenir proche en exploitant les données en cours d'acquisition dans le cadre du projet ASPI.

7.2 Données multimodales

Différents corpus de données multimodales ont été enregistrés, dans le cadre de plusieurs projets ; les premiers corpus, qui sont ceux sur lesquels j'ai principalement travaillé, ont été enregistrés dans le but de réaliser une tête parlante qui soit compréhensible par lecture labiale par des malentendants. Les données visuelles ont été acquises par deux caméras et traitées par un système de stéréovision développé par l'équipe Magrit du LORIA (Wrobel-Dautcourt *et al.* 2005). Les données visuelles sont des vidéos en stéréovision du visage des locuteurs sur lequel des marqueurs ont été peints. La méthode développée permet de retrouver la position tridimensionnelle à chaque instant pour chacun de ces marqueurs. Différents marqueurs placés dans des zones immobiles du visage permettent de transposer chaque image dans un repère relatif au visage ; la méthode permet ainsi une relative liberté de mouvements et il n'est par conséquent pas nécessaire de contraindre la position du visage des locuteurs.

Le système d'acquisition est plus flexible que les systèmes de *motion-capture* qui utilisent généralement des caméras infrarouges et des marqueurs collés sur la peau. Il utilise simplement deux caméras, un PC, et des marqueurs peints qui ne perturbent pas l'articulation ; il permet une acquisition suffisamment rapide pour reconstituer de façon précise les trajectoires des points 3D.

Pour faire une reconstruction des mouvements des articulateurs en stéréovision, il est nécessaire d'être capable de suivre les même points physiques au cours du temps. Comme la peau naturelle n'est pas assez contrastée, nous avons choisi de peindre des marqueurs sur le visage du locuteur. Cette méthode permet de contrôler la taille, la densité et la position des points intéressants. Ainsi 210 marqueurs ont été peints sur le visage (46 sur les lèvres) afin d'obtenir une information précise sur la déformation de la forme des lèvres (fig. 7.1) dans l'optique de construire une tête parlante de bonne qualité.

Dans le cas du corpus utilisé pour l'étude de la variabilité interlocuteur de la coarticulation labiale 15 marqueurs seulement ont été peints sur le visage (dont seulement 4 marqueurs sur



FIG. 7.1: *210 marqueurs blancs sont peints sur le visage du locuteur.*



FIG. 7.2: *Images en stéreovision de deux locuteurs, 15 marqueurs sont peints sur le visage de chaque locuteur.*

les lèvres, fig. 7.2) de façon à conserver un temps de préparation raisonnable pour les sujets de l'étude. En plus des marqueurs utilisés pour étudier les mouvements des lèvres, nous avons placé 6 marqueurs dans la partie supérieure du visage de façon à compenser le mouvement global de la tête. Deux caméras monochromes sont utilisées, car leur vitesse d'acquisition (environ 120 images par seconde) étant plus rapide que celles des caméras couleurs dont nous disposions, elles permettent de suivre des mouvements très rapides des articulateurs, par exemple lors de l'articulation des occlusives.

Le matériel utilisé est le suivant : deux caméras N&B JAI A33 équipées d'optiques Tamron, reliées à une carte d'acquisition X64-CL Dual à deux entrées IEEE-1394, sur un PC classique. Deux projecteurs à lumière tamisée sont utilisés pour éviter les ombres sur le visage des locuteurs.

7.2.1 Minicorpus

Il s'agit du corpus que nous avons le plus étudié. Il comporte 10 locuteurs différents (5 masculins et 5 féminins), et permet d'étudier la variabilité interlocuteurs de l'articulation. Le nombre de marqueurs dans cette étude est particulièrement réduit : il y avait 16 marqueurs sur le visage, dont 8 étaient exploitables pour mon travail. Il n'y avait en particulier que 4 marqueurs sur les lèvres, ce qui un peu insuffisant pour déterminer l'ouverture de la bouche. Nous sommes cependant parvenus à développer une méthode qui nous permet d'exploiter ces données de façon relativement satisfaisante.

Notons que notre objectif était d'utiliser des données portant sur les articulateurs visibles, et non pas de développer des algorithmes permettant de récupérer la position de ces articula-

teurs. Deux familles d'algorithmes sont en cours d'étude dans le cadre du projet européen ASPI, la première est celle des modèles d'apparence actifs (Larsen *et al.* 2007). La seconde est inspirée d'une méthode de suivi par indexation d'une image inconnue à l'aide d'une base d'images clés (Fontecave-Jallon & Berthommier 2008) et donne d'ores et déjà des résultats tout à fait intéressants.

7.2.2 Corpus AL

Il s'agit du corpus le plus conséquent de notre étude, réalisé pour une locutrice. Le nombre de marqueurs est très élevé : 190 marqueurs, principalement situés sur et autour des lèvres. De nouveau, les marqueurs ne nous permettent pas de déterminer fidèlement l'ouverture de la bouche, en particulier lors d'une protrusion importante, car ces marqueurs ne sont situés que sur l'extérieur des lèvres.

7.2.3 Données du projet ASPI

Dans le cadre du projet européen ASPI (Audiovisual-to-articulatory SPeech Inversion)de nombreuses données articulatoires ont été acquises (données IRM, échographe, articulographe, stéréovision), permettant de valider la méthode d'inversion développée ici, car comportant des données visuelles, et des informations sur la position des articulateurs internes. Malheureusement, ces données n'ont pu être exploitées dans le cadre de cette thèse car elles ont été acquises après la réalisation de ce travail.

7.3 Traitement des données

Dans cette section nous expliquons comment la position tridimensionnelle des marqueurs est déterminée à partir des images stéréo.

Une étape de prétraitement est appliquée sur les images pour détecter les marqueurs. Pour cela on exploite les caractéristiques des marqueurs peints : il s'agit de points blancs de forme circulaire et de rayon inférieur à 3 pixels. Le logiciel utilisé est capable de détecter la plupart des marqueurs, à l'exception de certains points absents dans l'une ou l'autre des images stéréo – par exemple les points des tempes qui disparaissent parfois lorsque le locuteur tourne la tête, ou certains marqueurs des lèvres masqués lors de la protrusion ou la fermeture de la bouche. Il arrive également que ce processus détecte des points erronés qui ne correspondent pas à des marqueurs, mais ont des caractéristiques similaires, comme la lumière réfléchie sur les yeux ou les dents.

Un algorithme de stéréovision classique est alors appliqué pour apparier et calculer les coordonnées 3D des marqueurs présents sur chaque paire d'images de la séquence. L'algorithme utilisé respecte les contraintes classiques d'unicité et de géométrie épipolaire (les points homologues doivent appartenir aux droites épipolaires, cf. figure 7.3).

Il est ensuite nécessaire de construire les trajectoires temporelles des marqueurs, c'est-à-dire identifier les marqueurs en correspondance parmi les ensembles trouvés sur chacune des paires d'images. Cela se fait généralement en utilisant un critère de proximité géométrique : un point à l'instant t sera connecté au point le plus proche parmi l'ensemble des points calculés à l'instant $t+1$. En raison des déformations importantes du visage occasionnées par le processus de production de la parole, des ambiguïtés peuvent survenir pour le choix des points en correspondance : les trajectoires des points des lèvres supérieures et inférieures peuvent en effet se croiser lors des fermetures et ouvertures de la bouche. Des trajectoires temporelles fiables, parfois incomplètes,

ne sont construites – en utilisant des critères de similitude de position et vitesse – que lorsque qu'une correspondance non ambiguë est possible.

Enfin, une stratégie globale utilisant un maillage déformable du visage est utilisé pour combler les trous et interpoler les données manquantes. Un maillage initial est construit automatiquement à l'instant 0 à partir de l'ensemble des marqueurs 3D, puis corrigé manuellement. Ce maillage est évolutif au cours du temps, tout en imposant une topologie invariante. Les points manquants peuvent ainsi être retrouvés à partir de la connaissance de leur voisins en utilisant un schéma d'interpolation classique. En moyenne, 7% des marqueurs sont estimés par ce processus, les autres étant directement déterminés par stéréovision.

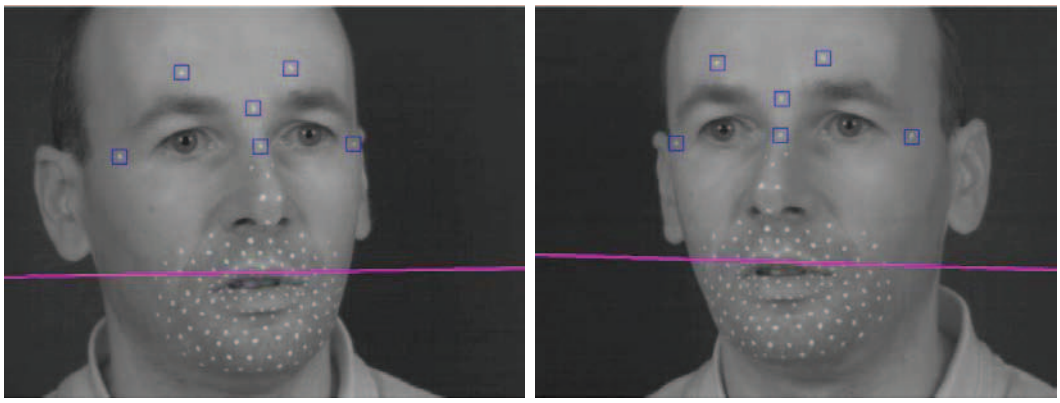


FIG. 7.3: *Une paire d'images en stéréovision du locuteur VR avec les marqueurs blancs peints sur le visage ; 6 marqueurs (encadrés en bleu) sont utilisés pour calculer le mouvement global de la tête. La géométrie épipolaire, calculée à l'aide d'une mire de calibration, décrit les relations qui existent entre les deux images : les points homologues doivent appartenir aux droites épipolaires telles celles représentées sur les images.*

7.4 Implémentation des contraintes visuelles

Pour l'implémentation des contraintes visuelles en elles-mêmes, il n'y avait en définitive que deux choses à faire : en premier lieu, établir une correspondance entre les positions observées des marqueurs et une grandeur utilisable dans le modèle articulatoire, dans notre cas le modèle de Maeda ; en deuxième lieu étudier la façon la plus intéressante d'imposer les contraintes visuelles lors de l'inversion.

Ces deux éléments se sont avérés plus problématiques que prévu. Pour établir la correspondance entre la position des marqueurs et des grandeurs utilisables en pratique dans le modèle (en l'occurrence, les paramètres de commande du modèle correspondant aux articulateurs visibles), il faut pouvoir tenir compte des caractéristiques des marqueurs : positionnement mal défini par rapport au contour de la bouche pour les marqueurs situés sur les lèvres, déformation et déplacement relativement à la mandibule pour les marqueurs sur la peau du menton. Ces caractéristiques, où beaucoup de paramètres inconnus entrent en jeu, ne permettent pas aisément de retrouver les grandeurs appropriées, et par conséquent les paramètres articulatoires visibles trouvés manquent de précision. Pour utiliser ensuite les valeurs obtenues pour les paramètres visibles lors de l'inversion, il faut prendre en compte, d'une part le fait que nos mesures ne sont pas très précises, et d'autre part le fait qu'il n'est pas forcément possible – du fait des défauts

inhérents à la méthode d'inversion acoustique, discutés précédemment – de trouver des solutions qui permettent de satisfaire ces contraintes.

7.4.1 Correspondance entre les marqueurs et le modèle articulatoire

Le modèle articulatoire de Maeda (Maeda 1979) a été établi à partir d'images radiographiques de coupes sagittales du conduit vocal, en appliquant une analyse factorielle permettant de choisir explicitement les composantes linéaires pertinentes. Les mouvements de la mâchoire, en particulier, peuvent être facilement déterminés en mesurant la position des incisives qui apparaissent très clairement sur les radiographies.

Trois paramètres articulatoires correspondent aux articulateurs « visibles » du visage : jw, qui est un paramètre contrôlant l'ouverture de la mâchoire, lh qui est le paramètre d'ouverture intrinsèque de la bouche, et enfin lp qui est un paramètre correspondant à la protrusion des lèvres[15]. À partir de ces 3 paramètres, toutes les grandeurs géométriques utiles peuvent être déterminées : l'ouverture de la mâchoire (entièrement déterminée par jw), l'ouverture effective de la bouche (entièrement déterminée par jw et lh), la protrusion et l'étirement des lèvres (déterminés par jw, lh et lp).

Les positions tridimensionnelles des marqueurs situés sur le visage du locuteur permettent de mesurer directement l'étirement et l'ouverture des lèvres à partir de la position des marqueurs situés sur les lèvres (voir Fig. 7.2). La protrusion peut aussi être estimée à partir de ces points, mais comme il s'agit d'un mouvement complexe qui implique un « dépliement » des lèvres, les mouvements de marqueurs peints sur les lèvres dans le plan sagittal ne peuvent rendre compte que partiellement de ce mouvement complexe. Par conséquent, la protrusion est probablement légèrement erronée.

Contrairement aux images radiographiques, les positions des marqueurs du visage du locuteur ne permettent pas de mesurer précisement les mouvements de la mâchoire. En effet, la position des marqueurs peints sur le menton (utilisés pour évaluer les mouvements de la mâchoire) est liée à la mandibule, mais aussi à celui de la lèvre inférieure qui déplace ces marqueurs quand elle bouge. Par conséquent, le mouvement de la mâchoire n'est pas non plus connu avec précision.

À partir des données visuelles acquises, nous calculons 3 grandeurs : l'ouverture de la bouche, la position de la mâchoire – ces paramètres se calculant facilement à partir de la position des marqueurs – et la protrusion des lèvres – dont l'évaluation est plus complexe.

- L'ouverture de la bouche est donnée par la distance entre les deux marqueurs des lèvres situés dans le plan medio-sagittal.
- La position de la mâchoire est la distance entre les points du menton et un point fixe. Nous prenons la moyenne des positions des 4 points du menton. En faisant cela, nous négligeons de façon implicite l'influence des mouvements des lèvres sur la position de ces marqueurs.
- La protrusion des lèvres est plus complexe à calculer. Le paramètre est déterminé en projetant le centre de gravité des marqueurs des lèvres sur un plan de référence construit à partir de la position moyenne des 4 marqueurs des lèvres.

Comme notre objectif est d'utiliser les données 3D obtenues avec le système de stéréovision comme des contraintes sur les paramètres régissant les articulateurs visibles du modèle de Maeda, nous devons établir une correspondance entre les paramètres observés que nous venons de définir et les paramètres articulatoires du modèle.

Deux approches très différentes ont été développées à cette fin : une première méthode qui consiste à employer la même méthode d'analyse statistique que Maeda pour retrouver les valeurs

[15] En réalité, il s'agit de la première composante de l'ACP entre les contributions de l'étirement et de la protrusion des lèvres.

des paramètres articulatoires à partir des grandeurs géométriques, et une deuxième méthode plus complexe qui applique la correspondance existante entre les grandeurs géométriques et les paramètres « visibles » du modèle de Maeda, tout en tentant de faire en sorte de supprimer les erreurs de mesure en utilisant un critère de régularité des trajectoires.

7.4.1.1 Méthode analytique

Cette première méthode tente de retrouver les paramètres articulatoires en utilisant la méthode analytique décrite par Maeda (Maeda 1990) pour obtenir la correspondance entre les données géométriques et les paramètres du modèle. L'utilisation de cette méthode est discutable : même en supposant que les mesures géométriques sont fiables, il n'y a pas de raison que l'analyse statistique nous permette de retrouver des paramètres articulatoires qui correspondent fidèlement à ce qu'ils devraient être réellement, puisque pour que cela fonctionne, il faudrait d'une part étudier des corpus de parole comparables à celui utilisé par Maeda – ce qui n'est pas le cas pour notre étude – et d'autre part une similarité de l'articulation entre les locuteurs – ce qui n'est a priori pas le cas.

Les grandeurs géométriques mesurées par Maeda étaient centrées et normalisées avant d'être traitées par analyse factorielle. Nous appliquons la même transformation aux paramètres construits à partir des données tridimensionnelles : chacun des paramètres est centré autour de sa position moyenne et réduit.

L'étape de normalisation précédente permet d'obtenir des paramètres observés qui ont les mêmes caractéristiques que les paramètres articulatoires. Cependant, l'analyse factorielle de Maeda permettait de soustraire l'effet de la mâchoire des autres paramètres, de façon à obtenir des paramètres indépendants. Nous devons donc retirer l'effet des mouvements de la mâchoire des autres paramètres. De la même façon que Maeda nous calculons la corrélation entre la mâchoire et chacun des deux autres paramètres (l'ouverture et la protrusion des lèvres puisque nous n'utilisons pas l'étirement) et soustrayons la corrélation des mesures normalisées.

Le principal problème de cette méthode est que, contrairement aux radiographies où le mouvement de la mâchoire est mesurable avec précision, la position de la mandibule inférieure n'est ici connue que de manière approchée. Cette étape de décorrélation ne permet d'obtenir par conséquent que des approximations de chacun des paramètres.

Nous obtenons ainsi trois paramètres compatibles avec le modèle de Maeda, mais malheureusement imprécis. Cette imprécision doit par conséquent être compensée, ce que nous pouvons faire en relâchant la contrainte relative aux paramètres visibles pour les solutions de l'inversion, c'est-à-dire en acceptant des solutions ayant des valeurs relativement distantes des valeurs cibles pour les paramètres articulatoires visibles.

7.4.1.2 Méthode géométrique régulée

Cette deuxième méthode exploite simplement la relation exacte existant entre les grandeurs géométriques visibles et les paramètres du modèles. Cependant, les grandeurs visibles mesurées sur les visages des locuteurs ne correspondant pas vraiment à celles mesurées par Maeda, il est nécessaire de faire une adaptation.

La relation qui permet d'obtenir les grandeurs géométriques à partir des paramètres articulatoires du modèle est simplement une application affine (ou, en d'autres termes, la composition d'une application linéaire et d'une translation), il est ainsi assez simple d'inverser la relation et ainsi retrouver les paramètres articulatoires à partir des grandeurs géométriques que nous permettent de trouver le modèle. Le modèle de Maeda nous permet d'obtenir à partir des paramètres

de mâchoire, d'aperture et de protrusion, la position de la mâchoire, les ouvertures verticale et longitudinale de la bouche, la protrusion des lèvres. Le problème principal est que les mesures que nous permettent d'obtenir les positions des marqueurs ne correspondent pas exactement aux mesures géométriques obtenues par cette transformation.

Un problème récurrent concernant les grandeurs obtenues à partir des marqueurs est qu'elle correspondent en première approximation à la bonne valeur géométrique, mais à une translation près. Nous pouvons déterminer cette translation (tout du moins, une bonne approximation), en faisant en sorte que les moyennes pour chacun des ensembles de données mesurées coïncident avec les moyennes des grandeurs correspondantes pour le modèle articulatoire. Cet ajustement peut être biaisé, vu que les valeurs moyennes n'ont pas forcément de raison de coïncider (d'une part parce que les corpus ne sont pas identiques, d'autre part parce que l'articulation peut varier fortement d'une personne à une autre).

Par ailleurs, il est également nécessaire de corriger certaines des mesures pour soustraire l'influence parasite des autres articulateurs. En particulier, pour mesurer l'ouverture verticale de la bouche, il est nécessaire d'enlever l'influence de la protrusion des lèvres sur la distance entre les deux points des lèvres situés sur le plan medio-sagittal ; en effet, quand les lèvres sont arrondies, la position de ces deux points relativement au contour de l'ouverture effective de la bouche varie. De la même façon, il est nécessaire de soustraire l'influence de l'ouverture de la bouche sur la position des marqueurs du menton – qui nous permettent de déterminer la position de la mâchoire. Les autres sources de perturbation sont négligées, notamment l'influence de la protrusion sur les marqueurs du menton, ainsi que le facteur multiplicatif que l'on devrait utiliser pour rendre compte du fait que la mesure de l'ouverture de la mâchoire n'est pas réalisée au même endroit (la distance entre les deux incisives dans le modèle de Maeda, la distance entre le menton et un point fixe du front pour les données visuelles).

Ce deuxième problème est résolu en exprimant l'ouverture de la bouche comme une combinaison linéaire de la distance entre les points medio-sagittaux des lèvres et du paramètre de protrusion, et la position de la machoire comme une combinaison linéaire de la valeur mesurée et du paramètre intrinsèque d'ouverture verticale de la bouche.

$$\text{Ouverture de la bouche} = \text{Distance entre les marqueurs medio-sagittaux des lèvres} + \alpha * lp \tag{7.1}$$

$$\text{Position dela mâchoire} = \text{Distance menton-point fixe} + \beta * lh \tag{7.2}$$

Dans la formule précédente, lp représente le paramètre de la protrusion des lèvres, et lh est le paramètre d'ouverture intrinsèque de la bouche. Les paramètres α et β sont déterminés en minimisant une fonction de coût basée sur la régularité des trajectoires des paramètres articulatoires et la distance à la position neutre.

Le critère exactement utilisé est de la forme suivante :

$$\underbrace{\int_t \sum_i d^2\alpha_i(t)/dt^2}_{\text{Terme de régularité}} + \underbrace{\int_t \sum_i \alpha_i^2(t)}_{\text{Terme d'écart au neutre}}$$

Le terme de régularité permet de pénaliser les changements de direction des trajectoires, le terme d'écart au neutre permet de pénaliser les trajectoires qui atteignent des valeurs trop importantes pour les paramètres articulatoires.

En pratique, on utilise une variante d'un algorithme génétique pour déterminer la solution optimale.

Préalablement à ce calcul, une première étape consiste à adapter le modèle articulatoire au locuteur étudié, dans notre cas en utilisant la méthode de Galvan (Galván-Rodrigez 1997). Une vérification manuelle est effectuée à partir des paramètres finals obtenus en construisant les formes du contour intérieur de la bouche à partir des paramètres et en les comparant à celles de la vidéo.

7.4.2 Inversion avec contraintes visuelles

Ces deux méthodes nous permettent d'obtenir des valeurs pour les paramètres articulatoires visibles, mais comme nous l'avons signalé à plusieurs reprises, ces paramètres sont intrinsèquement erronés ; pour utiliser ces valeurs comme contraintes lors de l'inversion, il est donc nécessaire de laisser une certaine marge d'erreur.

En pratique, les paramètres articulatoires cibles sont employées à deux niveaux : lors de la génération de solutions statiques pour chacun des vecteurs acoustiques (i.e. l'inversion à l'aide du codebook), et lors du lissage non-linéaire à l'aide de l'algorithme de Ney.

Pour l'inversion par codebook, on ne retiendra que les solutions dont les paramètres articulatoires visibles ne sont pas trop éloignés – au sens de la norme euclidienne – des valeurs cibles. Ceci à l'avantage de supprimer un grand nombre de solutions *a priori* inutiles et donc d'accélérer le processus d'inversion.

Pour le lissage non linéaire, on rajoute un terme supplémentaire à la fonction de coût :

$$C_v = \beta_v \times d_v(\alpha, \alpha_v),$$

où β_v est un coefficient de pondération, α_v représente le triplet de valeurs cibles pour les paramètres visibles, et d_v est la distance euclidienne restreinte au triplet des paramètres visibles. En d'autres termes, on pénalise les solutions proportionnellement à leur distance euclidienne à la valeur cible.

7.5 Expériences d'inversion

Plusieurs expériences ont été effectuées pour évaluer les paramètres articulatoires obtenus à l'aide de chacune de ces deux méthodes. Ces expériences ont été effectuées sur trois locuteurs français natifs : deux hommes (BP et EK) et une femme (AB). En premier lieu, les paramètres articulatoires visibles obtenus à l'aide de chacune des méthodes sont étudiés. Ensuite, des séquences de parole ont été inversées en utilisant ces paramètres comme contraintes supplémentaires.

7.5.1 Comparaison des deux méthodes

En comparant les trajectoires articulatoires obtenues à l'aide des deux méthodes, il est possible de découvrir des éléments intéressants sur l'articulation. Pour la première méthode, nous supposons implicitement que les mouvements articulatoires de chacun des locuteurs présentent les mêmes caractéristiques statistiques que le locuteur ayant servi à élaborer le modèle. Dans la deuxième méthode, les paramètres articulatoires sont calculés de façon à avoir des dimensions géométriques cohérentes ainsi que des trajectoires lisses. En comparant les données obtenues dans chacun des cas, nous pouvons vérifier la validité des hypothèses formulées.

Sur la figure 7.4, l'étendue de l'intervalle de valeurs possibles – exprimées en multiple de l'écart-type relatif à l'ensemble de données étudié pour la première méthode, en unité de paramètre articulatoire dans le deuxième cas – pour le paramètre de la mâchoire obtenu par la deuxième méthode, dépasse parfois largement l'étendue que nous permettons généralement lors de

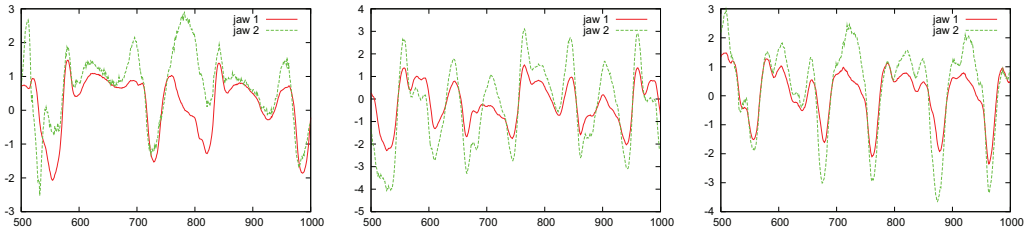


FIG. 7.4: *Paramètres de la mâchoire obtenus respectivement pour les locuteurs AB, BP et EK, en utilisant respectivement la première méthode (trait plein) ou la seconde (tirets). L'axe des abscisses correspond au numéro de trame, chaque trame durant environ 8ms ; l'axe des ordonnées est gradué en multiple d'écart-type relatif à l'ensemble de données étudié dans le premier cas, en unité de paramètre articulatoire dans le deuxième cas.*

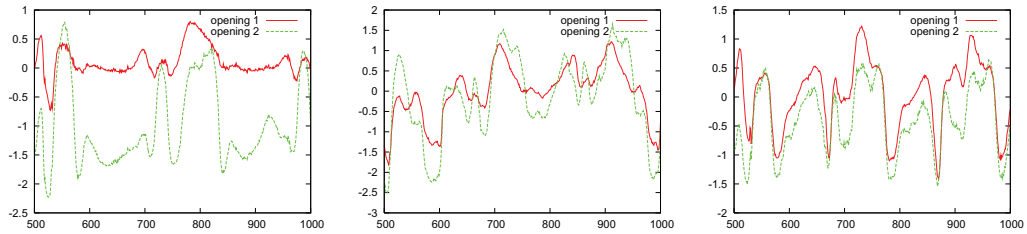


FIG. 7.5: *Paramètres d'ouverture des lèvres obtenus pour les locuteurs AB, BP et EK, en utilisant respectivement la première méthode (trait plein), ou la deuxième méthode (tirets).*

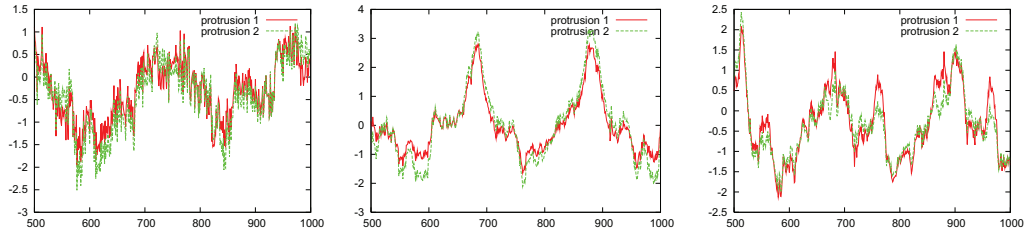


FIG. 7.6: *Paramètres de protrusion obtenus pour les locuteurs AB, BP et EK, en utilisant respectivement la première méthode (trait plein), ou la deuxième méthode (tirets).*

nos expériences d'inversion : les valeurs varient entre -5.5 (au lieu de -3) à 4.5 (au lieu de 3). Le domaine de variation peut sembler excessif, mais après vérification, cela n'est en définitive pas le cas. Les contours de bouche générés à partir des paramètres obtenus par la deuxième méthode sont très semblables à ceux que l'on peut observer sur les vidéos. Par ailleurs, Maeda (Maeda 1990) a également confirmé que comparé à d'autres locuteurs, celui retenu pour l'élaboration de son modèle utilisait une stratégie articulatoire qui sollicitait assez peu la mâchoire ; il n'est ainsi guère surprenant que d'autres locuteurs puissent manifester une plus grande amplitude de variation pour ce paramètre. Cela indique également que pour ce paramètre, l'une des hypothèses sur lesquelles repose la première méthode – des amplitudes de variation semblables d'un locuteur à l'autre – n'est pas respectée.

Le paramètre d'ouverture des lèvres (figure 7.5) a un domaine de variation légèrement plus faible pour le locuteur EK que pour le locuteur de référence utilisé par Maeda, mais légèrement plus important pour les locuteurs AB et BP. On observe également des différences importantes

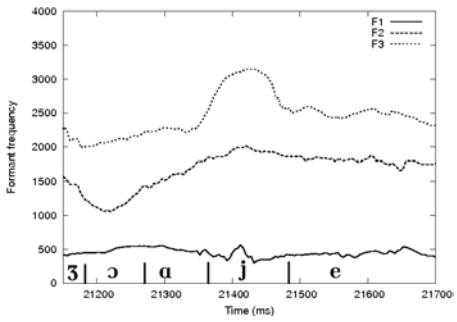


FIG. 7.7: *Trajectoires des formants de la séquence inversée ; l'axe des abscisses correspond au temps, en ms., l'axe des ordonnées à la fréquence, en Hz.*

entre les valeurs obtenues à l'aide des deux méthodes pour la locutrice AB.

Enfin, le paramètre de protrusion des lèvres (figure 7.6) est probablement le plus intéressant. Pour les trois locuteurs étudiés, on obtient des trajectoires pratiquement identiques à l'aide de deux méthodes. Ceci indique que l'étendue de variation pour le paramètre de protrusion est relativement constant parmi tous nos locuteurs, ce qui tendrait à indiquer qu'il est constant parmi tous les locuteurs du Français. Il est également assez étonnant de parvenir à trouver des résultats identiques à l'aide des deux méthodes, sachant que les calculs effectués sont très différents.

7.5.2 Expériences d'inversion

En utilisant ces paramètres comme entrées supplémentaires, nous avons effectué des expériences d'inversion sur une phrase de notre corpus, « Le joaillier a broyé les cailloux de la voyageuse. » Cette phrase est particulièrement appropriée pour évaluer l'inversion à l'aide de notre méthode puisque la plupart des sons sont des voyelles, des semi-voyelles ou d'autres sons voisés.

Dans cette expérience, l'entrée de l'inversion est constituée des trois premières fréquences formantiques obtenues automatiquement, auxquelles ont été adjointes les valeurs de paramètres obtenues respectivement à l'aide de la première ou de la deuxième méthode. À noter que les résultats présentées ici sont les solutions trouvées à l'issue de la deuxième étape de l'inversion, c'est-à-dire le lissage non-linéaire effectué sur les solutions statiques obtenues par exploration du codebook. Comme les paramètres visuels sont par nature peu fiables, leur utilisation en tant que contraintes est relâchée : on a retenu ici toutes les solutions dont les paramètres articulatoires visibles sont à une distance euclidienne inférieure à 1 des paramètres visibles cibles. À titre de comparaison, le seuil d'erreur admis pour les paramètres acoustiques lors de l'inversion par codebook est inférieure à 3%.

Les trajectoires formantiques de la séquence à inverser sont présentées à la figure 7.7.

La figure 7.8 présente les résultats de l'inversion pour la séquence /ʒɔɑje/ en utilisant les contraintes issues de la première méthode. Nous y présentons les trajectoires articulatoires trouvées pour les 4 principaux articulateurs (mâchoire, ouverture de la bouche, protrusion des lèvres, position du corps de la langue). Les contraintes visuelles respectives sont également présentées en traits pointillées.

Il apparaît clairement sur la figure correspondant au paramètre de la mâchoire que le système d'inversion rencontre des difficultés au milieu de la séquence, lors de la transition /ɑj/, puisqu'aucune solution n'y a été trouvée. On constate que les solutions trouvées respectent assez bien les deux autres contraintes visuelles. On peut également observer que la trajectoire du paramètre

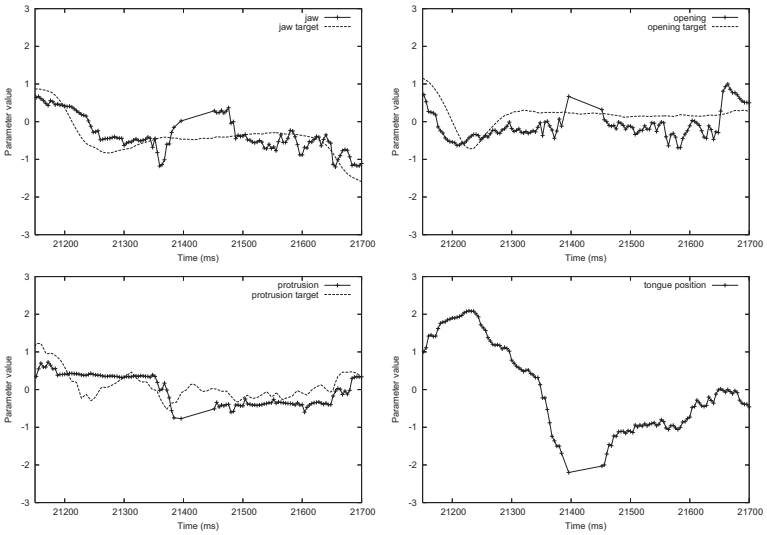


FIG. 7.8: *Résultats de l'inversion pour « joaillier » en utilisant la première méthode. Quatre paramètres articulatoires sont présentés : la mâchoire, la protrusion et l'ouverture des lèvres, et la position du dos de la langue. Les trajectoires des paramètres articulatoires cibles sont présentés en pointillés pour les paramètres articulatoires visibles, les trajectoires trouvées par inversion par des points reliés par des segments. Chaque paramètre peut varier dans l'intervalle [-3 ;3], l'axe des abscisses correspond au temps, en ms.*

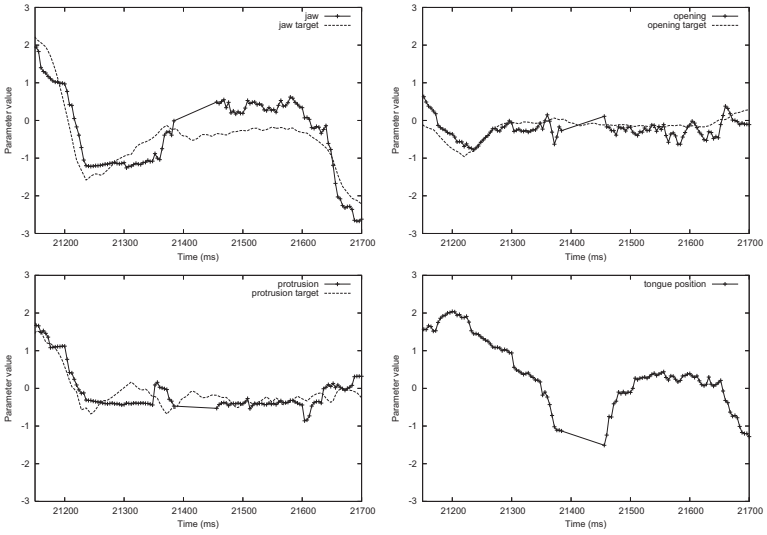


FIG. 7.9: *Résultats de l'inversion pour « joaillier » en utilisant comme contraintes les paramètres articulatoires obtenus à l'aide de la seconde méthode. Les conventions sont identiques à celles de la figure 7.8.*

de position de la langue correspond à ce à quoi on pourrait s'attendre – lorsque la valeur du paramètres augmente, cela signifie que la langue recule – : la langue commence à reculer pour prononcer /ɔɑ/, puis avance pour le /j/, puis recule à nouveau pour /e/.

Les résultats de l'inversion en utilisant les contraintes visuelles issues de la seconde méthode sont assez similaires, bien que les contraintes visuelles – et donc les trajectoires articulatoire

correspondant aux paramètres visuels – soient assez différentes. On observe les mêmes phénomènes : la transition /ɑj/ est difficile à inverser, les trajectoires articulatoires des paramètres visibles suivent bien leur cible, et le paramètre de position de la langue a un comportement correct.

Cette expérience démontre une nouvelle fois les remarquables capacités de compensation du modèle articulatoire de Maeda, puisque les contraintes visuelles ont beau être assez différentes, il est toujours possible de trouver des solutions respectant les contraintes et phonétiquement plausibles. Les deux modèles d'adaptation de données visuelles apparaissent équivalents dans cette expérience particulière.

7.6 Conclusion

Notre objectif était de concevoir une stratégie d'utilisation de données tridimensionnelles du visage et en particulier la fusion de ces données articulatoires visibles avec les paramètres du modèle articulatoire. La difficulté est de rendre les données du visages compatibles avec les paramètres du modèle. De ce point de vue il convient de souligner que cette question reste la même que les paramètres articulatoires visibles résultent d'une mesure directe comme c'est le cas dans notre travail, ou de l'utilisation de techniques de suivi comme les modèles d'apparence.

Bien que les contraintes visuelles soient opérationnelles et semblent donner des résultats prometteurs – les solutions que l'on trouve semblent phonétiquement réalistes – nous n'avons pas pu pour le moment les valider expérimentalement sur des bases de données plus vastes. Nous compléterons donc cette première série de tests à l'aide des données acquises dans le cadre du projet européen ASPI.

Cette lacune devrait cependant être corrigée très prochainement.

Conclusions et perspectives

Conclusions et perspectives

Avec l'augmentation de la puissance des ordinateurs, ainsi que l'amélioration des techniques d'imagerie médicale non invasives et sans danger, tous les sujets liés à la modélisation articulatoire, et en particulier l'inversion acoustique-articulatoire, connaissent un regain d'intérêt depuis quelques années, après être restés longtemps un peu confidentiels. Ce regain d'intérêt se manifeste sous diverses formes, et pour divers objectifs.

L'un des objectifs essentiels, et qui a constitué pendant longtemps le terme espéré de ma thèse, est l'animation de têtes parlantes à partir de l'inversion. Les têtes parlantes ont de nombreuses applications : elle peuvent jouer le rôle d'agent interactif « humain », de répétiteur pour l'apprentissage des langues ou pour les personnes malentendantes ne pouvant pas percevoir le visage du locuteur, par exemple dans une salle de classe.

L'animation d'une tête parlante pour une association de parents d'enfants malentendants a constitué pendant longtemps un objectif prioritaire pour cette thèse, et les recherches ont été organisées dans cet objectif. Nous avions initialement choisi de reprendre la technique d'inversion déjà développée dans le laboratoire pour avoir une base de travail éprouvée, et ainsi travailler plus précisément à l'amélioration de la méthode à l'aide de contraintes.

Il s'est avéré que la méthode d'inversion souffrait d'un certain nombre de défauts, les petits défauts ayant rapidement été corrigés, mais les défauts plus complexes étant restés longtemps en suspens et n'ayant véritablement été corrigés que lors de la dernière année. Ce travail a aussi donné lieu à un vaste effort d'un point de vue purement informatique. Nous avons ainsi sérieusement amélioré la robustesse des algorithmes et des programmes, ajouté de nombreux outils (des interfaces graphiques d'une part et des scripts facilitant grandement l'utilisation des outils d'inversion d'autre part), rendu disponible cet environnement à l'aide du logiciel de partage de fichier `cvs`, assuré une bien meilleure portabilité à l'aide du logiciel `configure`, contribué au développement des outils d'acquisition d'images de stéréovision qui nécessitent de synchroniser deux flux de données...

Cette thèse présente ainsi une nouvelle technique de construction de codebook articulatoire, échantillonnant l'espace articulatoire sous forme d'hypercuboïdes, et garantissant une précision acoustique homogène lors de la resynthèse à l'aide du codebook. Par rapport à la méthode d'Ouni, notre technique est nettement plus performante ; nous avons en effet travaillé à tous les niveaux (échantillonnage de l'espace, tests de régularité acoustique, direction de subdivision...), et en utilisant des techniques de calcul numérique assez avancées, pour réduire le temps de construction et la taille du codebook. Pour une précision acoustique donnée, le nombre d'hypercuboïdes est considérablement plus faible que le nombre d'hypercubes nécessaire avec la méthode d'Ouni, l'homogénéité acoustique est mieux respectée, et les temps de construction sont également réduits. Par ailleurs, une extension avec des polynômes multivariés de la modélisation articulatoire $\Rightarrow$ acoustique nous permet de réaliser des synthétiseurs articulatoires à codebook avec une précision acoustique pratiquement parfaite.

Le codebook hypercuboïdale reste assez volumineux, mais permet de contrôler l'influence de

l'erreur acoustique sur l'inversion. Il permet également une couverture de l'espace articulatoire largement suffisante pour représenter toutes les voyelles. La robustesse de l'inversion a également été améliorée grâce à diverses extensions permettant de générer un nombre de points respectant la densité des solutions dans l'espace articulatoire.

Les contraintes phonétiques introduites ont prouvé leur efficacité pour l'inversion de voyelles isolées, et en particulier leur capacité à conserver les solutions réalistes, tout en éliminant un grand nombre de solutions irréalistes. Il resterait cependant à valider ces contraintes avec davantage de données articulatoires réelles, et étudier davantage leur utilisation dans le cadre de l'inversion dynamique.

Les contraintes visuelles permettent de faire de l'inversion multimodale assez convaincante. Il reste cependant nécessaire de valider la méthode, en enregistrant des données articulatoires internes simultanément à l'enregistrement vidéo, ce qui est fait dans le cadre du projet ASPI. Il serait également intéressant de réaliser des tests perceptifs en animant une tête parlante à partir des résultats de l'inversion.

Par ailleurs, il reste un défaut important : il s'agit de l'incapacité du synthétiseur articulatoire à reconstituer de façon fidèle les formants originaux à partir de données articulatoires réelles, même pour les formes de conduits ayant permis son élaboration. Nous envisageons de modifier les valeurs de la fonction $\alpha - \beta$ permettant de trouver la fonction d'aire à partir de la largeur du conduit mesurée sur la coupe medio-sagittale à l'aide d'IRM en 3D du conduit vocal acquises dans le cadre du projet européen ASPI.

Il est également nécessaire de valider davantage toutes ces expériences : dans le cadre du projet européen ASPI, tout un dispositif d'acquisition a été mis en place pour valider l'inversion, et les données commencent à être disponibles. Nous serons bientôt en mesure de comparer de façon rigoureuse des trajectoires articulatoires obtenues par inversion avec des données réelles.

Enfin, malgré leurs défauts, les outils développés au cours de cette thèse donnent tout de même des résultats suffisamment encourageants pour être utilisés dans des projets d'envergure internationale. Ils sont notamment l'un des composants essentiels du projet européen ASPI, et ont également constitué l'un des éléments clé d'un projet portant sur l'étude du vieillisement vocal lors du Workshop de l'été 2008 du *Centre for Speech and Language Processing*, à la Johns Hopkins University.

A

Corpus PB

Le corpus que nous avons utilisé comprend 10 courtes phrases en langue française prononcées par une locutrice native. Il est issu de l'étude de Bothorel et al. de l'Institut de Phonétique de Strasbourg. Les données dont nous disposions sont le livre de Bothorel (Bothorel *et al.* 1986), le signal acoustique original, ainsi que les séquences articulatoires (sous forme de paramètres du modèle articulatoire de Maeda) pour les 10 phrases prononcées par PB. Ces données nous ont été gracieusement fournies par Shinji Maeda.

Le signal acoustique a été enregistré simultanément au film aux rayons-X ; sa qualité est assez médiocre du fait du bruit de la caméra. Le bruit est non stationnaire, mais néanmoins pratiquement périodique, il a donc été possible de l'atténuer légèrement.

1. Ma chemise est roussie
2. Voilà des bougies
3. Donne un petit coup
4. Une réponse ambiguë
5. Louis pense à ça
6. Mets tes beaux habits
7. Une pâte à choux
8. Prête-lui seize écus
9. Chevalier du gué
10. Il fume son tabac

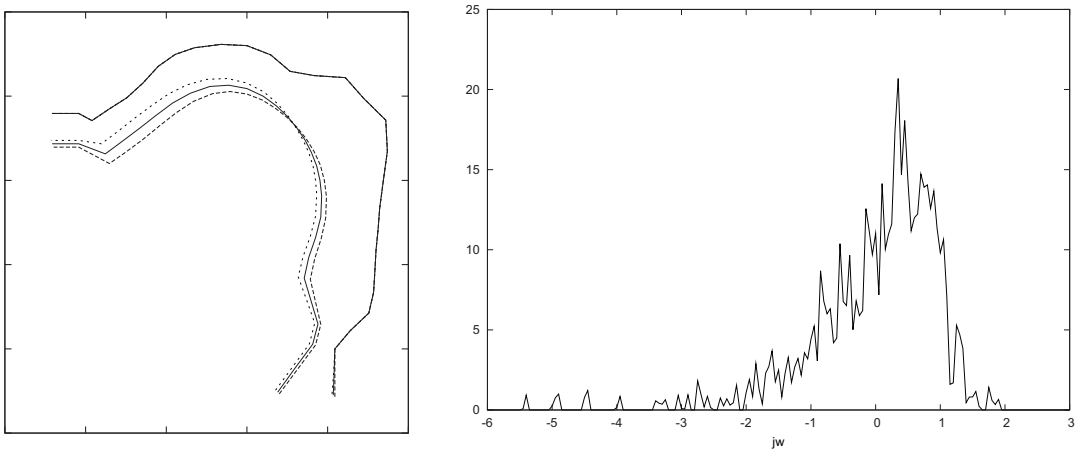
A.1 Paramètres articulatoires

Nous présentons dans cette section les déformations du conduit auxquelles correspondent chacun des paramètres, ainsi que la répartition de leurs valeurs sur les 10 phrases du corpus.

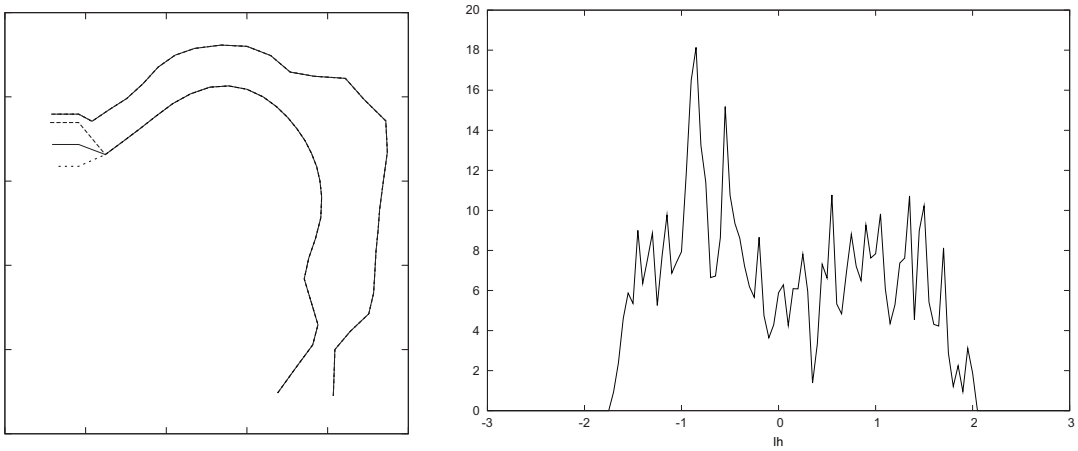
Une observation rapide des répartitions montrent que pour pratiquement tous les paramètres articulatoires, les valeurs suivent une loi gaussienne simple. Il y a cependant une exception notable : le paramètre d'ouverture des lèvres, dont la répartition des valeurs suit une loi gaussienne à deux centres.

On constate également que pour pratiquement tous les paramètres, les valeurs sont incluses dans l'intervalle [-3 ; 3]. Il y a cependant deux exceptions notables : le paramètres contrôlant la pointe de la langue, et le paramètre de mâchoire.

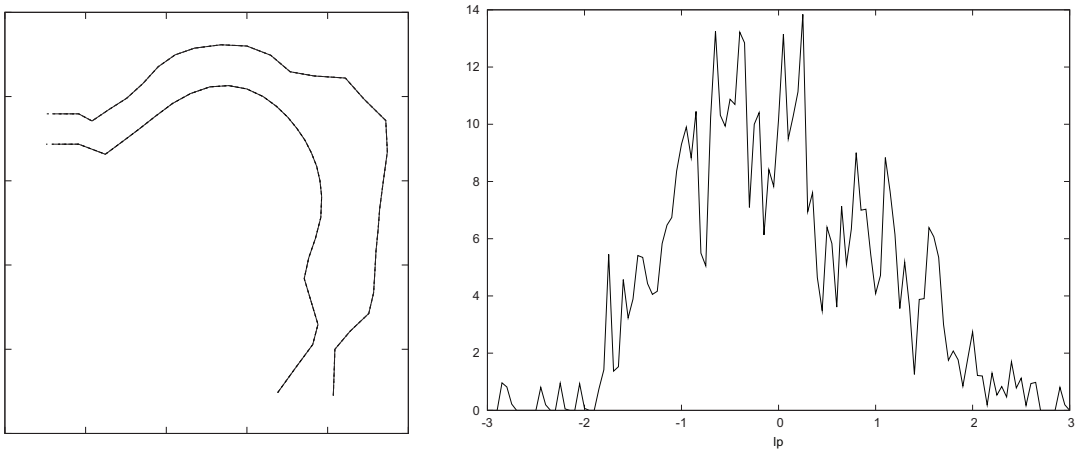
L'observation des modes de déformation fait également apparaître une compensation géométrique relative entre les paramètres de mâchoire et de pointe de la langue. Il est donc probable que lors de l'inversion ces paramètres articulatoires seront déterminés avec une moins bonne précision. L'ouverture et la protrusion des lèvres, ainsi que la forme et la position de la langue, présente des déformations géométriques uniques, il est ainsi probable que l'on parvienne à les retrouver plus facilement.



(a) Mâchoire.

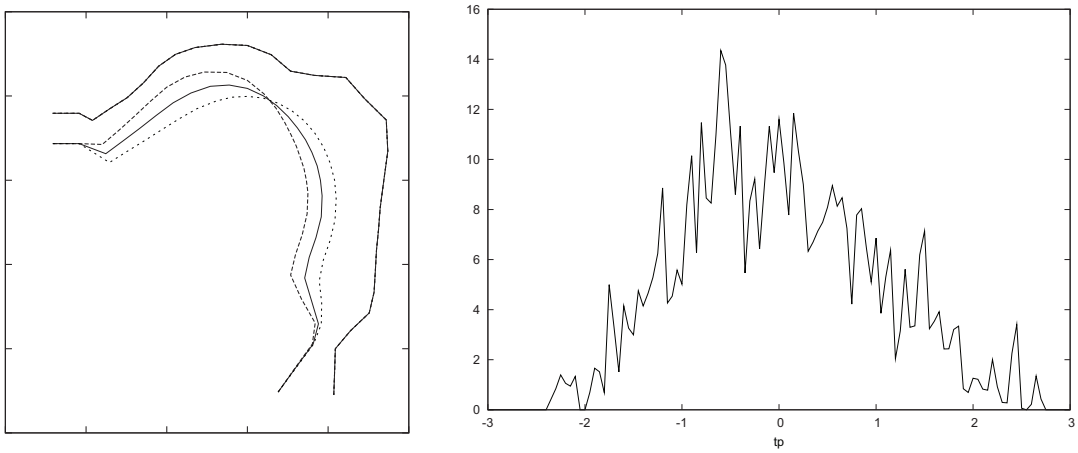


(b) Ouverture des lèvres.

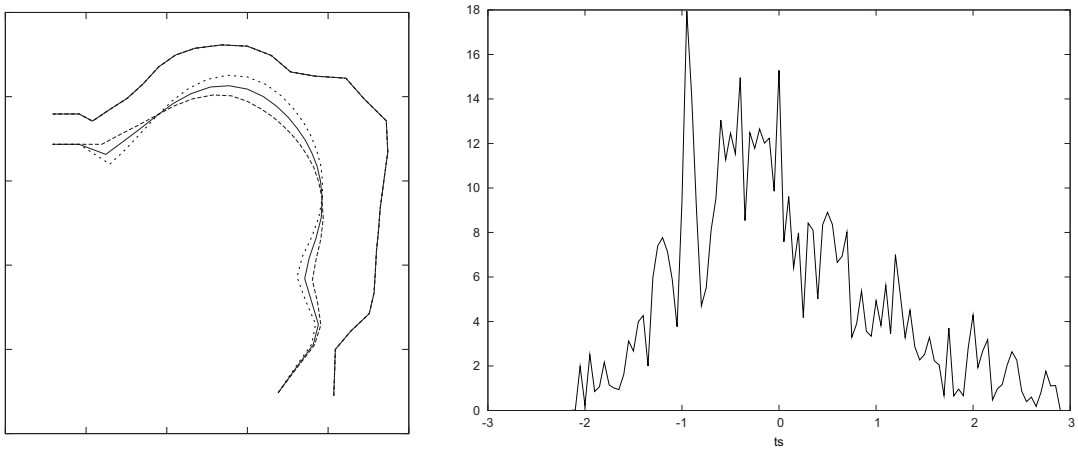


(c) Protrusion des lèvres.

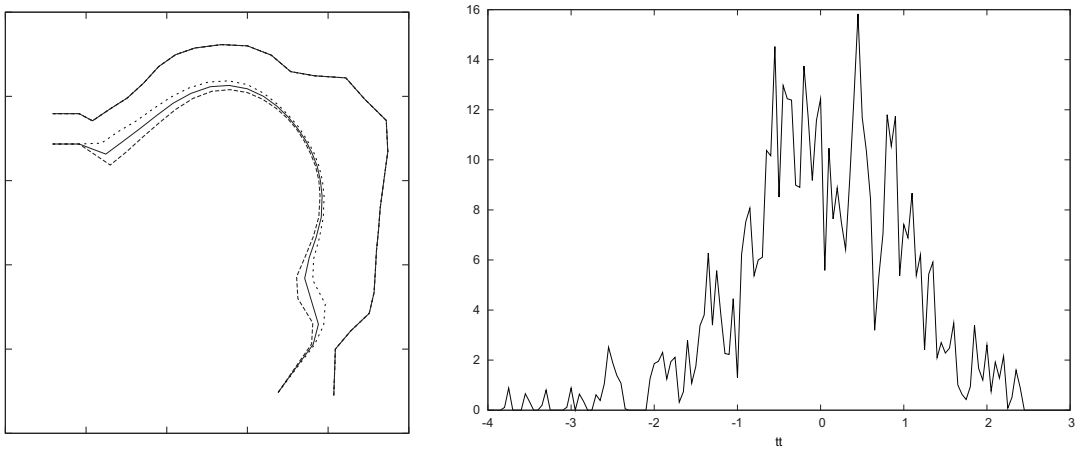
FIG. A.1



(d) Position de la langue.



(e) Forme de la langue.



(f) Pointe de la langue.

FIG. A.1: *Les 6 premiers paramètres du modèle articulatoire de Maeda : modification de la forme neutre, et pour une variation d'une unité positive (pointillés) ou négative (tirets). La distribution des valeurs sur le corpus de PB est également présentée.*

Bibliographie

[Atal & Rioul 1989] B. S. Atal et O. Rioul. *Neural Networks for Estimating Articulatory Positions from Speech.* J. Acoust. Soc. Am. Suppl. 1, vol. 86, Novembre 1989.

[Atal *et al.* 1978] B. S. Atal, J. J. Chang, M. V. Mathews et J. W. Tukey. *Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer-Sorting Technique.* Journal of the Acoustical Society of America, vol. 63, no. 5, pages 1535–1555, Mai 1978.

[Aurenhammer & Klein 1999] F. Aurenhammer et R. Klein. *Voronoi Diagrams.* In J. R. Sack et J. Urrutia, editeurs, Handbook of computational geometry. Elsevier, 1999.

[Badin *et al.* 2002] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth et C. Savariaux. *Three-Dimensional Linear Articulatory Modelling of Tongue, Lips and Face Based on MRI and Video Images.* Journal of Phonetics, vol. 30, no. 3, pages 533–553, 2002.

[Badin *et al.* 2005] P. Badin, I. S. Makarov et V. N. Sorokin. *Algorithm for Calculating the Cross-Section Areas of the Vocal Tract.* Acoustical Physics, vol. 51, pages 38–43, 2005.

[Benoît *et al.* 1994] C. Benoît, T. Mohamadi et S. Kandel. *Effect of Phonetic Context on Audio-Visual Intelligibility of French.* Journal of Speech, Language and Hearing Research, vol. 37, pages 1195–1203, Octobre 1994.

[Beutnagel *et al.* 1999] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou et A. Syrdal. *The AT&T Next-Gen TTS System.* In Proceedings of the Joint Meeting of ASA, EAA, and DAGA (AED), pages 18–24, Berlin, Germany, 1999.

[Birkhoff 1979] G. Birkhoff. The Algebra of Multivariate Interpolation. Academic Press Inc., 1979.

[Birkholz & Jackèl 2003] P. Birkholz et D. Jackèl. *A Three-Dimensional Model of the Vocal Tract for Speech Synthesis.* In 15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain, pages 2597–2600, Août 2003.

[Birkholz 2007] P. Birkholz. *Control of an Articulatory Speech Synthesizer Based on Dynamic Approximation of Spatial Articulatory Targets.* In Proc. INTERSPEECH, Août 2007.

[Black & Campbell 1995] A. W. Black et N. Campbell. *Optimising Selection of Units from Speech Databases for Concatenative Synthesis.* In Proceedings of the 4th European Conference on Speech Communication and Technology, pages 581–584, Madrid, Espagne, 1995.

[Boë *et al.* 1992] L.-J. Boë, P. Perrier et G. Bailly. *The Geometric Vocal Tract Variables Controlled for Vowel Production : Proposals for Constraining Acoustic-to-Articulatory Inversion.* Journal of Phonetics, vol. 20, pages 27–38, 1992.

[Bothorel *et al.* 1986] A. Bothorel, P. Simon, F. Wioland et J.-P. Zerling. Cinéradiographies des Voyelles et Consonnes du Français. Travaux de l'institut de Phonétique de Strasbourg, 1986.

[Calliope 1989] Calliope. *Description Acoustique.* In La Parole et son Traitement Automatique, chapitre 3. Masson, Paris, 1989.

[Charpentier 1984] F. Charpentier. *Determination of the Vocal Tract Shape from the Formants by Analysis of the Articulatory-to-Acoustic Non-Linearities.* Speech Communication, vol. 3, pages 291–308, 1984.

[Ciocea 1997] S. Ciocea. *Semi-Analytic Formant-to-Area Mapping.* PhD thesis, Université Libre de Bruxelles, Bruxelles, Belgique, 1997.

[Coker 1973] C. H. Coker. *Synthesis by Rule from Articulatory Parameters.* In J. L. Flanagan et L. R. Rabiner, editeurs, Speech Synthesis, pages 396–397. Dowden, Hutchinson & Ross, 1973.

[Deng *et al.* 2006] L. Deng, A. Acero et I. Bazzi. *Tracking Vocal Tract Resonances Using a Quantized Nonlinear Function Embedded in a Temporal Constraint.* IEEE Trans. on Speech, and Audio Processing, vol. 14, no. 2, pages 425–434, 2006.

[Engelbrecht 2001] S. Engelbrecht. *Minimum Principles in Motor Control.* The Journal of Mathematical Psychology, vol. 45, pages 497–542, 2001.

[Engwall 2005] O. Engwall. *Introducing Visual Cues in Acoustic-to-Articulatory Inversion.* In Proceedings of Interspeech 2005., Lisbonne, Portugal, 2005.

[Ericsdotter 2007] C. Ericsdotter. *Detail in Vowel Area Functions.* In Proc. of the International Congress of Phonetic Sciences, pages 513–516, Sarrebruck, Allemagne, Août 2007.

[Fant 1960] G. Fant. Acoustic Theory of Speech Production. The Hague : Mouton & Co., 1960.

[Fant 1970] 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.

[Flanagan *et al.* 1980] J. Flanagan, K. Ishizaka et K. Shipley. *Signal Models for Low Bit-Rate Coding Speech.* Journal of the Acoustical Society of America, vol. 68, pages 780–791, Mars 1980.

[Flanagan 1972] J. L. Flanagan. Speech Analysis, Synthesis and Perception. Springer-Verlag, 2nd ed, New York, 1972.

[Flash & Hogan 1985] T. Flash et N. Hogan. *The Coordination of Arm Movements : An Experimentally Confirmed Mathematical Model.* The Journal of Neuroscience, vol. 5, no. 7, pages 1688–1703, 1985.

[Fontecave-Jallon & Berthommier 2008] J. Fontecave-Jallon et Frédéric Berthommier. *A Semi-Automatic Method for Extracting Vocal Tract Movements from X-Ray Films.* Speech Communication, vol. 51 (à paraître), 2008.

[Galván-Rodrigez 1997] A. Galván-Rodrigez. 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.

[Gérard *et al.* 2003] J.-M. Gérard, R. Wilhelms-Tricarico, P. Perrier et Y. Payan. *A 3D Dynamical Biomechanical Tongue Model to Study Speech Motor Control.* Research Developments in Biomechanics, pages 49–64, 2003.

[Goldstein 1980] U. G. Goldstein. *An Articulatory Model for the Vocal-Tracts of Growing Children.* PhD thesis, MIT, Cambridge, 1980.

[Golub & Loan 1989] G.H. Golub et C.F. Van Loan. Matrix Computations. Johns Hopkins University Press, Baltimore, 1989.

[Halle 1983] P. Halle. *Techniques Cepstrales Améliorées pour l'Extraction d'Enveloppe Spectrale et la Détection du Pitch.* In Actes du séminaire "Traitement du signal de parole", pages 83–93, Paris, 1983.

[Heinz & Stevens 1965] J. M. Heinz et 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.

[Hélie 2002] T. Hélie. *Modélisation Physique d'Instruments de Musique en Systèmes Dynamiques et Inversion.* PhD thesis, Université Paris XI, Ircam, Paris, 2002.

[Hiroya & Honda 2004] S. Hiroya et M. Honda. *Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model.* IEEE Trans. on Speech, and Audio Processing, vol. 12(2), pages 175–185, 2004.

[Hogden *et al.* 1996] J. Hogden, A. Löfqvist, V. Gracco, I. Zlokarnik, P. Rubin et E. Saltzman. *Accurate Recovery of Articulator Positions from Acoustics : New Conclusions Based on Human Data.* Journal of the Acoustical Society of America, vol. 100, pages 1819–1834, Septembre 1996.

[Ishizaka & Flanagan 1972] K. Ishizaka et J. L. Flanagan. *Acoustic Properties of a Two-Mass Model of the Vocal Cords.* Bell Syst. Technol. J., vol. 51, pages 1233–1268, 1972.

[Katsamanis *et al.* 2007] A. Katsamanis, G. Papandreou et P. Maragos. *Audiovisual-to-Articulatory Inversion Using Hidden Markov Models.* In Proc. IEEE Workshop on Multimedia Signal Processing (MMSP-2007), pages 457–460, Chania, Greece, 2007.

[King & Richmond 2005] S. King et K. Richmond. *Projet COUGAR (Concatenation Of Units Guided by ARticulation)*, 2005.

[Laboissière 1992] R. Laboissière. *Préliminaires pour une Robotique de la Communication Parlée : Inversion et Contrôle d'un Modèle Articulatoire du Conduit Vocal.* PhD thesis, Institut National Polytechnique de Grenoble, ICP, Grenoble, 1992.

[Ladefoged 2005] P. Ladefoged. A Course in Phonetics, 5th edition. Heinle, 2005.

[Laprie & Mathieu 1998a] Y. Laprie et B. Mathieu. *Inversion Acoustique Articulatoire par une Méthode Variationnelle.* In Actes des 22èmes Journées d'Étude sur la Parole, pages 295–298, Martigny, Switzerland, Juin 1998.

[Laprie & Mathieu 1998b] Y. Laprie et B. Mathieu. *A Variational Approach for Estimating Vocal Tract Shapes from the Speech Signal.* In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, volume 2, pages 929–932, Seattle, USA, Mai 1998.

[Laprie 2004] Y. Laprie. *A Concurrent Curve Strategy for Formant Tracking.* In Proc. ICSLP, Jegu, Korea, Octobre 2004.

[Larar *et al.* 1988] J. N. Larar, J. Schroeter et M. M. Sondhi. *Vector Quantization of the Articulatory Space.* IEEE Trans. ASSP, vol. 36, no. 12, pages 1812–1818, Décembre 1988.

[Larsen *et al.* 2007] R. Larsen, M. B. Stegmann, S. Darkner, S. Forchhammer, T. F. Cootes et B. K. Ersbøll. *Texture Enhanced Appearance Models.* Computer Vision and Image Understanding, vol. 106, pages 20–30, 2007.

[Le Goff 1997] B. Le Goff. *Automatic Modeling of Coarticulation in Text-to-Visual Speech Synthesis.* In Eurospeech'97 Proceedings, volume 3, pages 1667–1670, Rhodes, Greece, 1997. European Speech Communication Association.

[Lefèvre & Zimmermann 2004] V. Lefèvre et P. Zimmermann. *Arithmétique Flottante.* Rapport technique RR-5105, INRIA, Février 2004.

[Lindblom *et al.* 1979] B. Lindblom, J. Lubker et T. Gay. *Formant Frequencies of Some Fixed-Mandible Vowels and a Model of Speech Motor Programming by Predictive Simulation.* J. Phonetics, vol. 7, pages 147–161, 1979.

[Lonchamp 1984] F. Lonchamp. *Les Sons du Français — Analyse Acoustique Descriptive.* Cours de phonétique, Institut de Phonétique, Université de Nancy II, 1984.

[Maeda *et al.* 2006] S. Maeda, M.-O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard et J. Schoentgen. *Acoustic-to-Articulatoy Inversion : Methods and Acquisition of Articulatory Data.* Rapport technique, ASPI Consortium, Novembre 2006.

[Maeda 1972] S. Maeda. *Conversion of Midsagittal Dimensions to Vocal Tact Area Function.* Journal of the Acoustical Society of America, 1972.

[Maeda 1979] 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.

[Maeda 1990] S. Maeda. *Compensatory Articulation During Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model.* In W.J. Hardcastle et A. Marchal, editeurs, Speech Production and Speech Modelling, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[Marchal 1980] A. Marchal. Les Sons et la Parole. Guérin, Montréal, 1980.

[Mathieu 1999] B. Mathieu. Modèles de Production de Parole et Reconnaissance à Partir d'Automates. Thèse de L'Université Henri Poincaré, Décembre 1999.

[Mawass *et al.* 2000] K. Mawass, P. Badin et G. Bailly. *Synthesis of French Fricatives by Audio-Video to Articulatory Inversion.* Acta Acustica, vol. 86, no. 1, pages 136–146, 2000.

[McGurk & MacDonald 1976] H. McGurk et J. MacDonald. *Hearing Lips and Seeing Voices.* Nature, vol. 246, pages 745–746, 1976.

[Mermelstein 1967] P. Mermelstein. *Determination of the Vocal-Tract Shape from Measured Formant Frequencies.* Journal of the Acoustical Society of America, vol. 41, pages 1283–1294, 1967.

[Mermelstein 1973] P. Mermelstein. *Articulatory Model for the Study of Speech Production.* Journal of the Acoustical Society of America, vol. 53, pages 1070–1082, 1973.

[Ney 1983] H. Ney. *A Dynamic Programmation Algorithm For Nonlinear Smoothing.* Signal Processing, vol. 5, no. 2, pages 163–173, Mars 1983.

[Ouni & Laprie 2000] S. Ouni et Y. Laprie. *Utilisation d'un Dictionnaire Hypercubique pour l'Inversion Acoustico-Articulatoire.* In Actes des Journées d'Étude sur la parole, Aussois, Juin 2000.

[Ouni & Laprie 2001] S. Ouni et Y. Laprie. *Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook.* In Proc. EUROSPEECH, volume 1, pages 277–280, Aalborg, Septembre 2001.

[Ouni & Laprie 2005] S. Ouni et Y. Laprie. *Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion.* Journal of the Acoustical Society of America, vol. 118, no. 1, pages 444–460, 2005.

[Ouni 2001] S. Ouni. Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico-articulatoire. Thèse de L'Université Henri Poincaré, Décembre 2001.

[Overall 1962] J. E. Overall. *Orthogonal Factors and Uncorrelated Factor Scores.* Psychological Reports, pages 651–662, 1962.

[Papcun *et al.* 1992] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks et S. Levy. *Inferring Articulation and Recognizing gestures from Acoustics with a Neural Network Trained on X-ray Microbeam Data.* Journal of the Acoustical Society of America, vol. 92, pages 688–700, 1992.

[Perkell 1974] J.S Perkell. *A Physiologically-Oriented Model of Tongue Activity in Speech Production.* PhD thesis, MIT, Boston, 1974.

[Perrier *et al.* 1992] P. Perrier, L.J. Boë et R. Sock. *Vocal Tract Area Function Estimation from Midsagittal Dimensions with CT Scans and a Vocal Tract Cast.* Journal of Speech and Hearing Research, pages 53–67, 1992.

[Pfitzinger 2005] H.R. Pfitzinger. *Concatenative Speech Synthesis with Articulatory Kinematics obtained via Three-Dimensional Electro-Magnetic Articulography.* In Proc. Deutsche Jahrestagung für Akustik, DAGA, pages 769–770, Mars 2005.

[Potard & Laprie 2007] B. Potard et Y. Laprie. *Compact Representation of the Articulatory-to-Acoustic Mapping.* In Interspeech, Anvers, Août 2007.

[Potard *et al.* 2004] B. Potard, Y. Laprie et S. Ouni. *Expériences d'Inversion Basées sur un Modèle Articulatoire.* In Journées d'Études sur la Parole - JEP'04 , Fès, Maroc, Avril 2004.

[Qin & Carreira-Perpiñán 2007] C. Qin et M. Á. Carreira-Perpiñán. *An Empirical Investigation of the Nonuniqueness in the Acoustic-to-Articulatory Mapping.* In Interspeech, Anvers, Août 2007.

[Ramsay & Shadle 2006] Gordon Ramsay et Christine Shadle. *The Influence of Geometry on the Initiation of Turbulence in the Vocal Tract During the Production of Fricatives.* In ISSP 2006, Ubatuba, Brésil, Décembre 2006.

[Remez 1934] E. Remez. *Sur un Procédé Convergent d'Approximations Successives pour Déterminer les Polynômes d'Approximation.* In Comptes-rendus de l'Académie des Sciences, volume 198, Paris, 1934.

[Richmond 2001] 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.

[Richmond 2006] K. Richmond. *A Trajectory Mixture Density Network for the Acoustic-Articulatory Inversion Mapping.* In Proc. INTERSPEECH, Pittsburgh, USA, Septembre 2006.

[Robert-Ribes *et al.* 1994] J. Robert-Ribes, J-L. Schwartz et P. Escudier. *A Comparison of Models for Fusion of the Auditory and Visual Sensors in Speech Perception.* Artificial Intelligence Review, vol. 9, pages 323–346, 1994.

[Robert *et al.* 2005] V. Robert, B. Wrobel-Dautcourt, Y. Laprie et A. Bonneau. *Strategies of Labial Coarticulation.* In Interspeech, Lisboa, Septembre 2005.

[Rubin *et al.* 1981] P. Rubin, T. Baer et P. Mermelstein. *An Articulatory Synthesizer for Perceptual Research.* Journal of the Acoustical Society of America, vol. 70, no. 2, pages 321–328, 1981.

[Sagisaka 1988] Y. Sagisaka. *Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units.* In Proceedings of the 13th International Conference on Acoustics, Speech, and Signal Processing, pages 679–682, New York, USA, Mai 1988.

[Savariaux & Orliaguet 1995] P. Savariaux C. Perrier et J.-P. Orliaguet. *Compensation Strategies for the Perturbation of the Rounded Vowel [u] Using a Lip-Tube : A Study of the Control Space in Speech Production.* Journal of the Acoustical Society of America, vol. 98, pages 2428–2442, 1995.

[Schoentgen & Ciocea 1997] J. Schoentgen et S. Ciocea. *Kinematic Formant-to-Area Mapping.* Speech Communication, vol. 21, pages 227–244, 1997.

[Schroeder 1967] M. R. Schroeder. *Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements.* Journal of the Acoustical Society of America, vol. 41, pages 1002–1010, 1967.

[Schroeter & Sondhi 1992] J. Schroeter et M. M. Sondhi. *Speech Coding Based on Physiological Models of Speech Production.* In S. Furui et M. M. Sondhi, editeurs, Advances in Speech Signal Processing, pages 231–267. Dekker, New York, 1992.

[Schroeter & Sondhi 1994] J. Schroeter et M. M. Sondhi. *Techniques for Estimating Vocal-Tract Shapes from the Speech Signal.* IEEE Trans. on Speech and Audio Processing, vol. 2, no. 1, Part. II, pages 133–150, Janvier 1994.

[Sondhi 1986] M.M. Sondhi. *Resonances of a Bent Vocal Tract.* Journal of the Acoustical Society of America, vol. 79, pages 1113–1116, Avril 1986.

[Soquet *et al.* 1990] A. Soquet, M. Saerens et P. Jospa. *Acoustic-Articulatory Inversion Based on a Neural Controller of a Vocal Tract Model.* In Proceedings of the ESCA workshop on speech synthesis, Autrans, France, Septembre 1990.

[Soquet *et al.* 1991] A. Soquet, M. Saerens et P. Jospa. *Acoustic-Articulatory Inversion Based on a Neural Controller of a Vocal Tract Model : Further Results.* In O. Simula T. Kohonen K. Mokisara et J. Kangas, editeurs, Artificial Neural Networks, pages 371–376. North Holland : Elsevier, 1991.

[Sorokin & Trushkin 1996] V.N. Sorokin et A.V. Trushkin. *Articulatory-to-Acoustic Mapping for Inverse Problem.* Speech Communication, vol. 19, pages 105–118, 1996.

[Sorokin *et al.* 2000] V.N. Sorokin, A.S. Leonov et A.V. Trushkin. *Estimation of Stability and Accuracy of Inverse Problem Solution for the Vocal Tract.* Speech Communication, vol. 30, pages 55–74, 2000.

[Sorokin 1992] V. N. Sorokin. *Determination of Vocal Tract Shape for Vowels.* Speech Communication, vol. 11, pages 71–85, 1992.

[Stevens & House 1955] K. N. Stevens et A. S. House. *Development of a Quantitative Description of Vowel Articulation.* Journal of the Acoustical Society of America, vol. 27, pages 484–493, 1955.

[Sumby & Pollack 1954] W. H. Sumby et I. Pollack. *Visual Contribution to Speech Intelligibility in Noise.* Journal of the Acoustical Society of America, vol. 26, no. 2, pages 212–215, Mars 1954.

[Toda *et al.* 2004] T. Toda, A. W. Black et K. Tokuda. *Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model.* In Proc. ICSLP, Jegu, Korea, Octobre 2004.

[von Helmholtz 1867] H. von Helmholtz. Handbuch der physiologischen Optik. L. Voss, Leipzig, 1867.

[Wood 1979] S. Wood. *A Radiographic Analysis of Constriction Locations for Vowels.* Journal of Phonetics, vol. 7, pages 25–43, 1979.

[Wrobel-Dautcourt *et al.* 2005] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie et S. Ouni. *A Low Cost Stereovision Based System for Acquisition of Visible Articulatory Data.* In Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05), pages 145–150, Vancouver, 2005.

[Yehia & Itakura 1996] H. Yehia et F. Itakura. *A Method to Combine Acoustic and Morphological Constraints in the Speech Production Inverse Problem.* Speech Communication, vol. 18, no. 2, pages 151–174, 1996.

[Zwicker & Feldtkeller 1981] E. Zwicker et R. Feldtkeller. Psychoacoustique : l'Oreille, Récepteur d'Information. Masson, 1981.

Résumé

Cette thèse porte sur l'inversion acoustique-articulatoire, c'est-à-dire la récupération des mouvements des articulateurs de la parole à partir du signal sonore. Nous présentons dans ce mémoire une évolution importante des méthodes de tabulation à *codebooks* utilisant une table de correspondants acoustique-articulatoire précalculée à l'aide d'un modèle de synthèse acoustique. En dehors de la méthode d'inversion proprement dite, nous présentons également l'introduction de deux types de contraintes : des contraintes phonétiques génériques, issues de l'analyse par des experts humains de l'invariance articulatoire des voyelles, et des contraintes visuelles, c'est-à-dire des contraintes obtenues automatiquement à partir de l'enregistrement et l'analyse d'images en stéréovision du locuteur.

Mots-clés: Inversion, acoustique, articulatoire, analyse par synthèse, contraintes, phonétique, stéréovision

Abstract

This thesis investigates acoustic-to-articulatory inversion, i.e. recovering articulatory movements from the speech signal. In this work, we present an important evolution of *codebooks* methods, i.e. methods using acoustic-articulatory tuples precomputed using an acoustic synthesis model. Apart from the inversion method, we present the introduction of two types of constraints : generic phonetic constraints, derived from the analysis by human experts of articulatory invariance for vowels, and visual constraints, i.e. constraints derived automatically from a video signal, in our case a stereo video signal, thus allowing us to perform multimodal inversion.