Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire

Slim Ouni, Yves Laprie

Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire

Slim OUNI et Yves LAPRIE

LORIA - UMR 7503
BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex - France
Mél: ouni,laprie@loria.fr

ABSTRACT

This paper presents an articulatory codebook construction method which gives rise to a good articulatory space coverage with a limited number of points. It is a new representation of the articulatory space by hypercubes. A hypercube is a region represented by a few number of points. We also present an interpolation method to retrieve points from a hypercube and an inversion method based on the same interpolation method. The advantage of the codebook and the inversion method is its strength against articulatory-to-acoustic mapping non-linearity problems.

1 INTRODUCTION

L'inversion acoustico-articulatoire consiste à récupérer les paramètres articulatoires décrivant la forme du conduit vocal à partir du signal de parole. L'une des familles de méthodes d'inversion est celles des méthodes par tabulation qui exploitent un dictionnaire de formes articulatoires indexées par les paramètres acoustiques. Ce dictionnaire est construit en utilisant un synthétiseur articulatoire qui à partir des paramètres articulatoires calcule les paramètres acoustiques.

Notre méthode d'inversion repose sur le modèle articulatoire de Maeda [MAE79] et consiste à régulariser les trajectoires articulatoires initiales obtenues à l'aide d'une méthode par tabulation. Le dictionnaire de formes nous permet de construire les trajectoires initiales, et conditionne donc fortement la qualité finale de l'inversion. Il faut donc construire le dictionnaire de formes articulatoires avec une grande attention. Pour ce faire, plusieurs méthodes existent:

- échantillonnage aléatoire des paramètres articulatoires [SCH90] ;
- échantillonnage autour des trajectoires liant les formes de base correspondant aux voyelles [LAR88] ;
- échantillonnage implicite de l'espace articulatoire pour entraîner un système neuromimétique comme le modèle "forward" [LAB95].

Pour la construction de notre dictionnaire, nous utilisons un système d'hypercubes. Comme nous le verrons, notre méthode est destinée à réduire au minimum l'espace et le temps nécessaires à une exploration systématique de l'espace articulatoire. Le dictionnaire final est certes encore volumineux, mais permet d'être assuré que l'échantillonnage articulatoire n'a plus d'influence sur l'inversion, ce qui est loin d'être le cas pour les autres méthodes. A titre d'exemple nous montrons dans notre étude expérimentale la comparaison entre la trajectoire récupérée avec le dictionnaire hypercubique et celle récupérée avec un dictionnaire de 600000 formes choisies aléatoirement. En effet, une trajectoire linéaire dans l'espace articulatoire ne s'accompagne pas forcément d'une trajectoire linéaire dans l'espace acoustique [OUN99]. Par conséquent, une région qui présente ce genre de non-linéarité peut être omise s'il n'y a pas suffisament de points couvrant cette région. En fait, si nous voulions faire un échantillonnage régulier et fin des sept paramètres du modèle articulatoire de Maeda dans l'intervalle $[-3\sigma, 3\sigma]$ (σ étant l'écart type) avec un pas d'échantillonnage relativement grossier de $1/3\ \sigma$, nous obtiendrions $(19^7) \cong 900$ millions de points, ce qui est très coûteux pour les machines actuelles en espace de stockage comme en temps d'accès.

L'idée est d'avoir un échantillonnage moins coûteux, mais précis. Il doit être précis, pour pouvoir être sûr de récupérer toutes les solutions possibles, pour étudier l'influence articulatoire des contraintes ajoutées à l'inversion et pour pouvoir trouver la trajectoire articulatoire qui est à l'origine du signal de la parole à inverser. En effet, les méthodes d'inversion existantes exploitent, voire abusent, de l'effet compensatoire du conduit vocal ce qui peut fausser l'interprétation des résultats. Dans les paragraphes qui suivent nous présentons notre méthode de construction du dictionnaire avec une meilleure couverture de l'espace articulatoire et ensuite les outils d'inversion pour un tel dictionnaire.

1. LA CONSTRUCTION DU DICTIONNAIRE DE FORMES

1.1 Le dictionnaire hypercubique

L'idée s'inspire du fait que la relation articulatoire-acoustique (qu'on notera $\mathcal{M}$) est non-linéaire. C'est un problème qui doit être pris en compte si nous voulons obtenir une couverture efficace de l'espace articulatoire. Nous rappelons que la non-linéarité de la relation $\mathcal{M}$ est inévitable vu qu'elle est liée à la nature physique et géométrique du conduit vocal [CHA84]. Pour cela, nous décomposons l'espace articulatoire d'une manière fine dans les régions où la relation $\mathcal{M}$ est fortement non-linéaire. Dans ce but, nous utilisons les hypercubes. Un hypercube d'ordre N est une région d'un espace de dimension N délimitée par des hyperplans. L'espace articulatoire sera représenté par une arborescence d'hypercubes. Chaque hypercube représente une région de l'espace articulatoire où la relation $\mathcal{M}$ peut être considérée comme linéaire. Le lecteur peut trouver dans [OUN99] les détails de la construction de l'hypercube. Rappelons ici seulement le principe de la méthode.

1.2 Le principe de la méthode

Nous supposons que tout l'espace articulatoire est contenu dans un hypercube. Si la relation $\mathcal{M}$ est non-linéaire à l'intérieur de cet hypercube, ce dernier est décomposé en sous-hypercubes. Pour chaque sous-hypercube, nous testons de nouveau la linéarité. Si la relation est quasi-linéaire, nous gardons cet hypercube sinon, nous le décomposons à nouveau. Cette procédure est répétée récursivement jusqu'à l'obtention d'un hypercube de taille suffisamment petite pour pouvoir considérer que le comportement de la relation $\mathcal{M}$ dans cet hypercube est linéaire.

1.3 Le test de linéarité

Le test proposé par Charpentier [CHA84] consiste à calculer la courbure acoustique le long d'un chemin articulatoire à l'intérieur de la région à explorer. Cette méthode acceptable dans le cas d'un modèle de fonction d'aire qui utilise peu de paramètres conduirait à des calculs trop longs dans notre cas. C'est pourquoi nous utilisons le test suivant. Pour tous les segments qui relient les sommets d'un hypercube, nous considérons les milieux de ces segments et nous interpolons linéairement les valeurs acoustiques correspondantes. Ensuite, nous comparons ces valeurs avec celles calculées directement avec le synthétiseur articulatoire. Si la différence entre les valeurs acoustiques synthétisées et les valeurs acoustiques interpolées est inférieure à un seuil prédéfini $\Delta\varepsilon$, la relation $\mathcal{M}$ dans cet hypercube est considérée comme linéaire. Nous disons que $\mathcal{M}$ est linéaire avec une marge d'erreur de $\Delta\varepsilon$ dans le domaine acoustique. Pour un hypercube de dimension 7, nous avons 128 sommets (2^7) et le nombre de segments possibles entre ces sommets est 8128, ce qui correspond au nombre de tests. Nous supposons que ce test de linéarité est suffisant.

1.4 La description du dictionnaire hypercubique.

En résumé, un hypercube est défini par ses sommets qui sont des vecteurs de l'espace articulatoire. Dans un dictionnaire hypercubique, un hypercube est représenté par un sommet origine, la longueur d'un des cotés (avec ces deux informations seulement, nous pouvons construire l'hypercube) et les valeurs acoustiques des sommets. En conséquence, le dictionnaire est composé par des hypercubes plus ou moins fins, selon la linéarité de la relation $\mathcal{M}$ dans la région représentée par l'hypercube. Plus l'hypercube est grand, plus la relation $\mathcal{M}$ est linéaire.

1.5 L'interpolation dans un hypercube

Nous pouvons récupérer toutes les informations dont nous avons besoin à partir des sommets de l'hypercube. En effet, la seule connaissance des vecteurs articulatoires et des paramètres acoustiques leurs correspondant, nous permet de retrouver toutes les informations concernant les vecteurs articulatoires se trouvant à l'intérieur de cet hypercube par une interpolation à partir des sommets. Pour rendre l'interpolation plus robuste et plus précise, nous interpolons par rapport au sommet le plus proche du vecteur dont nous cherchons les paramètres acoustiques en calculant le gradient en ce sommet. Considérer le sommet le plus proche permet, en effet, de renforcer l'hypothèse de linéarité. Le sommet le plus proche est retrouvé à partir des 128 sommets de l'hypercube.

Soit $\vec{P}$ le vecteur articulatoire (ses composantes sont les paramètres articulatoires du modèle de Maeda) dont nous cherchons son correspondant acoustique $\vec{F}$ (ses composantes sont les trois premiers formants) par interpolation dans l'hypercube H_c.

$$\vec{F} = \mathcal{M}(\vec{P}) \quad (1)$$

L'interpolation au sens du gradient par rapport au sommet le plus proche P_0 est donnée par l'équation suivante :

$$\vec{F} = \vec{F}_0 + \nabla\vec{F} \cdot (\vec{P} - \vec{P}_0) \quad (2)$$

Où $\vec{F}_0 = \mathcal{M}(\vec{P}_0)$ et $\nabla\vec{F}$ est le gradient de $\vec{F}$.

Connaissant $\vec{P}$, P_0 et F_0 , nous calculons $\vec{F}$. Grâce à l'équation (2), pour tout vecteur articulatoire se trouvant à l'interieur de l'hypercube, nous pouvons retrouver toute l'information acoustique qui lui est relative par interpolation.

1.6 Vérification expérimentale de l'interpolation

Afin de tester cette méthode d'interpolation, nous avons généré une trajectoire acoustique qui a été synthétisée avec le synthétiseur articulatoire à partir d'une trajectoire articulatoire. Par ailleurs, cette trajectoire est interpolée à partir du dictionnaire hypercubique. Nous comparons la proximité du signal acoustique correspondant à la trajectoire synthétisée et le signal correspondant à la trajectoire interpolée.

Nous obtenons de bons résultats du point de vue de la proximité acoustique, comme le montre la figure 1. En effet, pour la construction du dictionnaire nous avons fixé une marge d'erreur assez grande pour le test de linéarité (50Hz pour le premier formant, 75Hz pour le deuxième formant et 100Hz pour le troisième formant). Nous avons généré 37 trajectoires articulatoires. L'erreur moyenne ne dépasse pas 10Hz pour les deux premiers formants et 20Hz pour le troisième formant. Ceci constitue une bonne approximation formantique et est rassurant à propos de la qualité de l'interpolation.

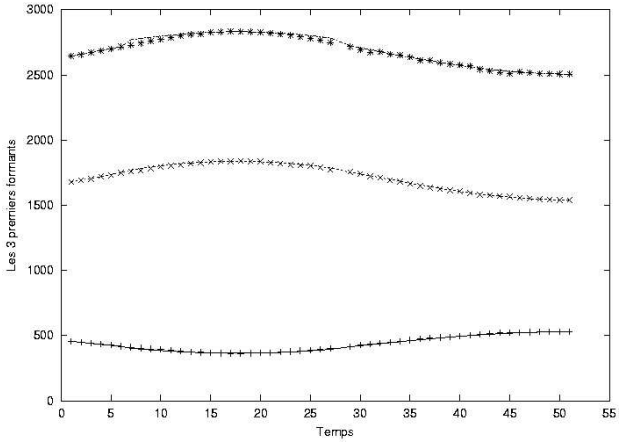

Figure 1 - Le paramètre correspondant à la mâchoire varie linéairement (les autres paramètres restent constants). Nous représentons la trajectoire synthétisée (trait fin) et interpolée (points) dans l'espace acoustique des trois premiers formants.

2 UTILISATION DU DICTIONNAIRE HYPERCUBIQUE POUR L'INVERSION

Etant donné un signal acoustique, nous voulons maintenant récupérer les paramètres articulatoires qui sont à l'origine de ce signal. Les solutions obtenues doivent vérifier les deux critères suivants:

- une bonne proximité acoustique avec les données de départ;
- la régularité des trajectoires articulatoires. Si cela est possible, nous voulons retrouver les trajectoires articulatoires qui sont à l'origine du signal acoustique mesuré ou, au moins, une solution proche de l'originale.

Nous rappelons qu'un hypercube représente une région où la relation $\mathcal{M}$ est quasi-linéaire. L'image d'un hypercube articulatoire est donc un polygone inscrit dans un hyperplan acoustique.

Le signal est décomposé en segments de quelques millisecondes chacun. Chaque segment constitue une entrée acoustique à inverser. Nous cherchons les hypercubes du dictionnaire

hypercubique dont l'image par $\mathcal{M}$ (donc dans l'hyperplan acoustique) contient cette entrée acoustique. Nous augmentons la taille des hyperplans acoustiques de quelques Hertz afin d'éviter les problèmes aux limites. A partir du dictionnaire de formes nous récupérons tous les hypercubes répondant à cette requête. En effet, une entrée acoustique peut appartenir à plusieurs hypercubes. Mais, chaque hypercube ne fournit qu'une seule solution (hypothèse de linéarité dans un hypercube).

Soit $\vec{F}$ le vecteur acoustique (représenté par les trois premiers formants) à inverser. Soit H_c un hypercube tel que $\vec{F} \in \mathcal{M}(H_c)$. Soit $\vec{P}$ le vecteur articulatoire (représenté par les sept paramètres du modèle articulatoire de Maeda) cherché associé à $\vec{F}$.

Nous utilisons le même principe pour l'inversion que celui de l'interpolation. Nous faisons l'hypothèse que le point inversé est proche d'un sommet P_0 . Ce dernier est choisi comme étant le sommet ayant son vecteur acoustique F_0 le plus proche de $\vec{F}$. L'inversion en utilisant l'interpolation au sens du gradient se fait en résolvant l'équation suivante (qui découle de l'équation (2)):

$$\vec{F} - \vec{F}_0 = \nabla\vec{F} \cdot (\vec{P} - \vec{P}_0) \qquad (3)$$

L'équation (3) est un système d'équations linéaires qui peut être écrit de la manière suivante:

$$\begin{pmatrix} F^1 - F_0^1 \\ F^2 - F_0^2 \\ F^3 - F_0^3 \end{pmatrix} = \begin{pmatrix} \frac{\partial F^1}{\partial \alpha_1} & \frac{\partial F^1}{\partial \alpha_2} & \dots & \frac{\partial F^1}{\partial \alpha_1} \\ \frac{\partial F^2}{\partial \alpha_1} & \frac{\partial F^2}{\partial \alpha_2} & \dots & \frac{\partial F^2}{\partial \alpha_7} \\ \frac{\partial F^3}{\partial \alpha_1} & \frac{\partial F^3}{\partial \alpha_2} & \dots & \frac{\partial F^3}{\partial \alpha_7} \end{pmatrix} \cdot \begin{pmatrix} (P^1 - P_0^1) \\ (P^2 - P_0^2) \\ (P^3 - P_0^3) \\ (P^4 - P_0^4) \\ (P^5 - P_0^5) \\ (P^6 - P_0^6) \\ (P^7 - P_0^7) \end{pmatrix}$$

Où F^i, F_0^i représentent les ièmes composantes des vecteurs $\vec{F}$ et F_0 et P^i, P_0^i sont les ièmes composantes des vecteurs $\vec{P}$ et P_0

Pour résoudre un tel système, dont le nombre d'équations est inférieur au nombre des inconnues, nous avons eu recours à un alogrithme se basant sur la méthode SVD (décomposition en valeurs singulières) [GOL89]. Cette méthode permet d'obtenir toutes les solutions du système d'équations. Dans notre cas, nous voulons une solution dans le voisinage immédiat du sommet P_0 pour respecter l'hypothèse de calcul du jacobien qui est la dérivée calculée au sommet P_0 . Nous choisissons donc le point de l'espace solution le plus proche de $\vec{P}$, c'est à dire le point de l'hypercube pour lequel la distance ($\vec{P}$ - P_0) est minimale[1]. Grâce à cet algorithme nous obtenons le point inverse. A ce stade nous vérifions l'hypothèse de départ concernant la proximité du vecteur inverse $\vec{P}$ par rapport au sommet P_0 . Pratiquement il faut donc tester l'hypothèse de proximité pour tous les sommets (2^7=128 sommets) ou s'arrêter dès que cette hypothèse est bien vérifiée. Ce processus est appliqué à tous les hypercubes H_i tels que $\vec{F} \in \mathcal{M}(H_i)$, et de même, il est appliqué à l'ensemble des entrées acoustiques.

[1] Plus précisément, SVD permet d'obtenir l'espace des solutions sous la forme d'un point Q et des vecteurs de base qui correspondent au noyau du système. Le point Q est le point de l'espace solution le plus proche de l'origine [GOL89] qui est dans notre cas le sommet de l'hypercube considéré P_0. Comme nous nous sommes placés dans l'hypothèse de trouver la solution la plus proche du sommet P_0, Q est donc le point cherché à condition qu'il appartienne à l'hypercube.

2.1 Contraindre l'inversion

Pour que l'inversion réussisse, trois conditions doivent être vérifiées:

- $F_0 = \mathcal{M}(P_0)$ est l'image acoustique du sommet P_0 , la plus proche de $\vec{F}$ (vecteur acoustique à inverser): $F_0 = \min_i d(F, F_i)$, F_i étant les paramètres acoustiques correspondant au sommet i (i=1..128),
- P_0 reste toujours le sommet le plus proche de $\vec{P}$ (vecteur articulatoire), résultat de l'inversion qui est calculé par interpolation: $P_0 = \min d(P, P_i)$, avec P_i le sommet i de l'hypercube (i=1.. 128)
- $\vec{P}$, le vecteur articulatoire trouvé par inversion, doit être à l'intérieur de l'hypercube qui a servi pour l'inversion.

La première condition est la définition même du sommet le plus proche. La deuxième condition traite le cas où la relation $\mathcal{M}$ n'est pas suffisament linéaire dans H_c (à cause des erreurs dans le test de linéarité). Ces deux contraintes sont concurrentes: elles assurent la proximité du sommet le plus proche dans les deux espaces à la fois. Ce qui améliore la linéarité de la relation $\mathcal{M}$ dans l'hypercube. La troisième condition élimine les erreurs dues au fait que l'inversion donne des résultats en dehors de l'hypercube articulatoire considéré. Si l'une des trois premières conditions n'est pas vérifiée, nous rejettons la solution. En combinant l'interpolation par rapport au sommet le plus proche au sens du gradient et ces trois conditions, nous augmentons la précision de l'inversion.

3 ETUDE EXPÉRIMENTALE ET DISCUSSION

Afin de tester la qualité de l'inversion, nous avons construit un dictionnaire de formes de dimensions réduites avec seulement les 5 premiers paramètres. Cela ne change en rien notre méthode d'inversion, car seul le dictionnaire change et faire les tests avec un dictionnaire de dimensions 5 permet d'examiner les solutions plus facilement. Ce dictionnaire contient 5526 hypercubes de dimensions 5. Le nombre de sommets total est donc 176.832. Nous avons pris comme marge d'erreur pour le test de linéarité le triplet (50Hz,75Hz,100Hz) pour les trois premiers formants. Des trajectoires tests ont été générées avec le synthétiseur articulatoire en faisant varier un ou plusieurs paramètres articulatoires. Nous utilisons des trajectoires simulées pour pouvoir faire une comparaison pertinente puisque l'on compare les résultats à une trajectoire connue. Nous passons à notre procédure d'inversion le signal acoustique synthétisé, et obtenons en sortie toutes les solutions obtenues à partir du dictionnaire hypercubique. Nous vérifions deux critères:

- la proximité acoustique par rapport au signal à inverser.
- le fait que la trajectoire articulatoire initiale soit parmi les solutions obtenues ou non.

Pour évaluer la proximité acoustique, nous générons le signal acoustique à partir des résultats de l'inversion et nous le comparons au signal acoustique original. Le deuxième critère est destiné à assurer que nous récupérons bien toutes les solutions, et donc nous avons une bonne couverture de l'espace articulatoire. Si l'inversion fonctionne correctement nous devons retrouver parmi ces solutions la trajectoire originale.

Dans la figure 2, nous présentons les solutions de l'inversion d'un signal acoustique obtenu en faisant varier sinusoïdalement le paramètre articulatoire de la mâchoire et en laissant les autres constants. Pour évaluer acoustiquement l'inversion nous avons resynthétisé le signal acoustique à partir des paramètres articulatoires obtenus par inversion. Pour chaque trame de parole, nous obtenons plusieurs solutions proches de la

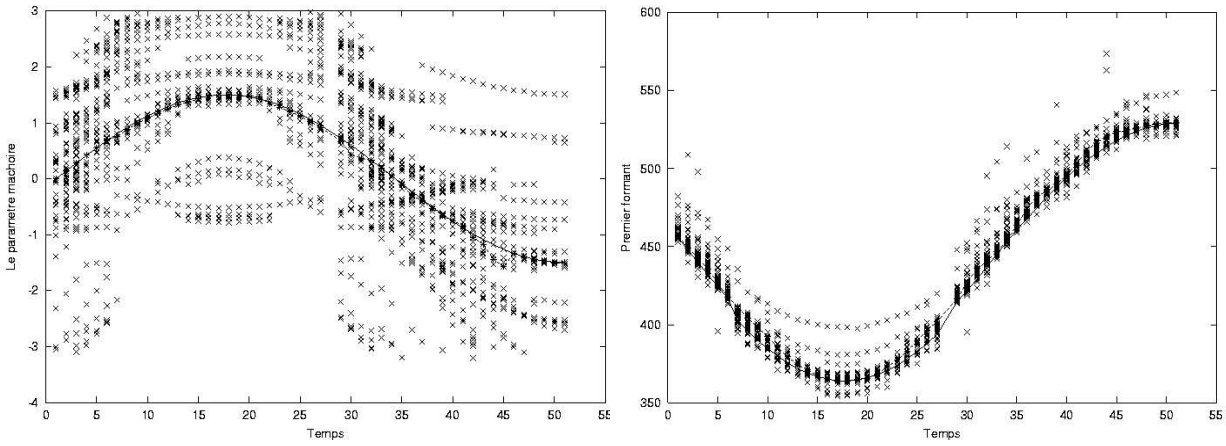

Figure 2 - Représentation des solutions de l'inversion dans l'espace articulatoire (le 1er graphique) et dans l'espace acoustique, premier formant (le 2ème graphique). Les trajectoires initiales sont représentées en trait continu, et les solutions par des points (X).

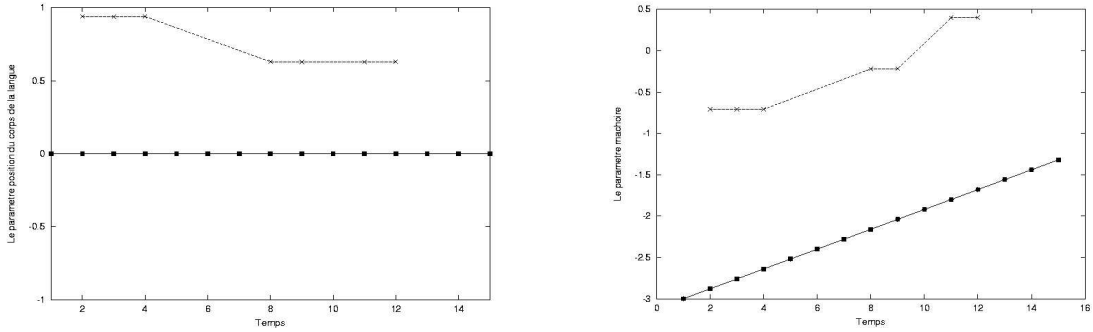

Figure 3 - Comparaison entre l'inversion par un dictionnaire à échantillonnage aléatoire et l'inversion par le dictionnaire hypercubique. Nous présentons la trajectoire dans le plan paramètre position du corps de la langue (1er graphique) et dans le plan du paramètre mâchoire (2ème graphique). La trajectoire de départ (—) et la solution de l'inversion par le dictionnaire hypercubique (▪) se superposent. La solution de l'inversion par le dictionnaire à échantillonnage aléatoire (×) est loin de la trajectoire de départ.

trajectoire acoustique initiale. Dans l'espace articulatoire, nous avons plusieurs solutions possibles. Parmi ces solutions, nous retrouvons bien la trajectoire de départ, qui correspond à la sinusoïde du départ, grâce à une méthode de lissage non-linéaire [LAP98]. Le point fort de cette méthode d'inversion est qu'elle ne contraint pas implicitement le processus d'inversion. Il est donc possible d'étudier très précisement comment l'introduction de contraintes d'origine physiologiques ou acoustiques influence l'inversion de manière à récupérer les trajectoires articulatoires proches des trajectoires réalisées par le locuteur. Pour cela, nous envisageons l'introduction des masses différentes suivant les articulateurs et faire un apprentissage à partir des données réelles .

Nous terminons notre étude expérimentale par une comparaison entre l'inversion avec le dictionnaire hypercubique et l'inversion avec un dictionnaire à échantillonnage aléatoire. Comme pour les autres tests, nous effectuons l'inversion avec les deux dictionnaires de formes, et nous appliquons le même algorithme de lissage pour obtenir la solution. Comme cela apparaît sur la figure 3, pour le paramètre mâchoire et le paramètre position du corps de la langue, la trajectoire obtenue en utilisant le dictionnaire hypercubique coincide avec la trajectoire de départ, alors que la deuxième, obtenue en utilisant le dictionnaire de formes aléatoires de 600000 formes, ne donne pas l'inversion de tous les points et le résultat est même totalement différent de la trajectoire articulatoire initiale. La deuxième solution est obtenue en abusant de l'effet de compensation.

Comme la taille du dictionnaire complet, du moins avec la précision que nous nous sommes imposés, est trop importante pour une machine traditionnelle et que notre algorithme se prête bien au parallélisme, nous sommes en train d'implanter l'algorithme sur une machine parallèle.

BIBLIOGRAPHIE

[CHA 84] Charpentier F. (1984), "Determination of the vocal tract shape from the formants by analysis of the artiuclatory -to-acoustic non-linearities", Speech Com. vol. 3, pp.291-308.

[GOL 89] Golub G. H. & Van Loan C. F. (1989), "Matrix computations", 2ème edition, JHU Press, §8.3, chap. 12.

[LAB 95] Laboissière R. & Galvan A. (1995), "Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences", ICPhS'95, vol. 1, pp. 358-361.

[LAP 98] Laprie Y. & Mathieu B. (1998) "A variational approach for etimating vocal tract shapes from speech signals", ICASSP98, vol. 2, p929-932.

[LAR 88] Larar J. N. & Sondhi M. M. "Vector quantization of the articulatory space", IEEE Trans. Acou., Speech, Signal Processing, vol. 36 n°12, pp. 1812-1818.

[MAE 79] Maeda S. (1979) "Un modèle articulatoire de la langue avec des composantes linéaires", JEP 79, p152-162.

[OUN 99] Ouni S. & Laprie Y. (1999), "Design of hypercube codebooks for the acoustic-to-articulatory inversion respecting the non-linearities of the articulatory-to-acoustic mapping", Eurospeech'99, vol. 1, pp. 141-144.

[SCH 90] Schroeter J., Meyer P. et Parthasarathy S. (1990), "Evaluation of improved articulatory codebooks and codebook access distance measures", ICASSP'90, vol. 1, pp. 393-396.