Inversion acoustique-articulatoire dynamique par codebook hypercuboïque : premiers résultats

Blaise Potard

Equipe Parole - LORIA
Campus scientifique - BP 239 - 54506 Vandoeuvre-lès-Nancy Cedex, France
potard@loria.fr
http ://parole.loria.fr/

ABSTRACT

Our goal is to recover articulatory information from the speech signal by acoustic-to-articulatory inversion. Like most inversion methods proposed in the literature, our method relies on the analysis-by-synthesis paradigm, here based on Maeda's articulatory model. After an overall description of the inversion method the paper presents a few inversions of formants frequencies trajectories obtained from synthesizing articulatory data and compare the obtained articulatory trajectories to the original.
Keywords: acoustic-to-articulatory, inversion, codebook, constraints

1. Introduction

L'inversion acoustique-articulatoire, c'est-à-dire l'obtention d'informations sur la position des articulateurs (position des lèvres, langue...) à partir du signal de parole, est l'un des domaine clé de l'étude de la parole. Les nombreuses études consacrées au sujet restent malheureusement assez peu concluantes, notamment concernant la possibilité d'utiliser un modèle donné pour un locuteur quelconque.

Cet article a pour but d'évaluer une évolution de la méthode à codebooks hypercubiques de Ouni[9]. La validation d'une méthode d'inversion acoustique-articulatoire est l'un des problèmes cruciaux dans ce domaine de recherche et nous organisons donc nos travaux sur l'inversion de manière à pouvoir exploiter facilement les données disponibles. Nos travaux reposent sur le modèle de Maeda[6] et c'est pour cette raison que nous utilisons les données articulatoires qui ont servi à construire ce modèle.

Nous présentons dans cet article les premières experiences d'inversions de phrases effectuées à l'aide d'un codebook hypercuboïque. Les paramètres acoustiques que nous utilisons sont les fréquences des premiers formants.

2. Processus d'inversion

Notre méthode d'inversion est fondée, comme beaucoup d'autres, sur l'analyse par synthèse, et le processus d'inversion comporte trois étapes.

Une étape préalable au processus d'inversion est la construction d'une table articulatoire, ou codebook, qui associe des vecteurs articulatoires (à 7 dimensions, correspondant aux 7 paramètres du modèle de Maeda) à leurs correspondants acoustiques, dans notre cas les fréquences des premiers formants. La force de notre méthode d'inversion réside dans la résolution acoustique quasi uniforme du codebook. Cette propriété est garantie par la façon dont est construite la table : on explore l'espace articulatoire récursivement en évaluant à chaque étape la linéarité locale de la relation articulatoire-acoustique[9]. Cette table est organisée de manière à retrouver facilement tous les vecteurs articulatoires qui permettent de générer un tuple de fréquences formantiques donné. D'importantes innovations ont recemment été apportées à notre méthode de construction de codebook[11].

La première étape du processus d'inversion proprement dit consiste à générer un grand nombre de solutions potentielles à partir du codebook. Comme il existe *a priori* une infinité de vecteurs articulatoires permettant d'obtenir un vecteur acoustique il est nécessaire d'échantillonner l'espace des solutions de façon suffisamment concise mais précise pour trouver des solutions proches de la solution réelle.

La deuxième étape de notre méthode consiste à reconstruire une trajectoire articulatoire qui soit suffisamment régulière au cours du temps. Nous utilisons pour cela un algorithme de programmation non-linéaire, qui minimise une fonction de coût quantifiant la « régularité » des trajectoires articulatoires.

La dernière étape consiste à améliorer la fidélité acoustique et la régularité articulatoire de la solution obtenue à l'étape précédente en utilisant un algorithme de régularisation variationnelle[5].

La régularité des trajectoires articulatoires peut être envisagée à plusieurs niveaux. Ici il s'agit d'une combinaison linéaire de l'écart à la position neutre des articulateurs (i.e. $\int \alpha^2(t)dt$) et du mouvement des articulateurs (i.e. $\int \left(\frac{d\alpha(t)}{dt}\right)^2 dt$).

2.1. Codebook hypercuboïque

Le codebook hypercuboïque est une évolution du codebook hypercubique de Ouni[9]. En résumé, il s'agit d'une représentation arborescente de la linéarisation par morceaux de la relation articulatoire ⇒ acoustique. Comparé aux codebooks présentés dans les autres études consacrées à l'inversion acoustique-articulatoire[1, 12, 2, 13], celui-ci présente plusieurs

particularités appréciables : une exploration intégrale de l'espace articulatoire, et une précision acoustique quasi-homogène garantie sur l'ensemble de la table.

La principale différence de notre méthode avec celle de Ouni réside dans la structure élémentaire utilisée pour modéliser la relation articulatoire ⇒ acoustique locale : dans le cas de Ouni il s'agissait d'hypercubes (c'est-à-dire la généralisation du carré dans un espace de dimension strictement supérieure à 3), et dans notre cas il s'agit d'hypercuboïdes, c'est-à-dire la généralisation du rectangle dans un espace de dimension strictement supérieure à 3.

Cette modification très simple de la structure permet d'avoir un codebook beaucoup plus souple ; mais pour pouvoir exploiter efficacement cette structure, il est nécessaire de complexifier l'algorithme de construction. Ce dernier peut se résumer en une exploration récursive de l'espace articulatoire, arrêtant l'exploration dès que la relation articulatoire vers acoustique est « suffisamment » linéaire, c'est-à-dire lorsque l'erreur commise est inférieure à un certain seuil.

Dans la méthode de Ouni, la technique d'exploration était élémentaire : lorsque la linéarité locale n'était pas assurée dans un hypercube en cours d'exploration, celui-ci été subdivisé en sous-hypercubes de côté moitié qu'il explorait alors récursivement ; dans un espace de dimension 2, cela ne fait que $2^2 = 4$ sous-carrés à explorer... mais dans un espace de dimension 7, comme celui correspondant aux commandes articulatoires du modèle de Maeda[8], cela fait $2^7 = 128$ sous-hypercubes. Chaque niveau de subdivision supplémentaire multipliait ainsi la taille du codebook par un facteur très important.

Dans cette nouvelle méthode, lorsqu'un hypercuboïde ne permet pas d'assurer la linéarité locale, nous ne divisons par 2 que la taille d'un seul des côtés, ou en d'autre termes, nous ne subdivisons que dans une seule direction ; chaque subdivision n'entraîne ainsi que l'exploration de 2 sous-hypercuboïdes. Le choix de la direction dans laquelle on effectue la subdivision a bien entendu son importance, mais un simple choix aléatoire permet déjà d'économiser une place considérable par rapport à la méthode utilisant les hypercubes, pour une même précision. Des expériences effectuées précédemment[11] montrent qu'une subdivision dans une direction aléatoire permet de gagner un facteur 4 par rapport à la subdivision hypercubique, tandis qu'une heuristique relativement simple subdivisant dans la direction qui maximise l'erreur d'interpolation à partir du polynôme de Taylor calculé au centre de l'hypercuboïde permet de gagner un facteur 16.

Cette structure permet ainsi d'obtenir des codebooks nettement plus concis pour une même précision. D'autres améliorations décrites dans [11] permettent d'améliorer encore la concision du codebook.

2.2. Inversion statique

Pour chaque vecteur acoustique représentée par les n premières (entre 3 et 5) fréquences formantiques, le processus d'inversion consiste en la recherche de tous les hypercuboïdes qui peuvent générer le tuple de formants observé. Il faut ensuite trouver un ensemble de solutions dans chacun de ces cuboïdes. Comme l'inversion consiste à trouver 7 paramètres à partir de n, l'espace des solutions a *a priori* $7 - n$ degrés de liberté. La relation articulatoire acoustique (notée R) est supposée être localement linéaire au niveau du centre P_0 de l'hypercuboïde (c'est-à-dire que l'application $P - P_0 \mapsto R(P) - R(P_0)$ est supposée être une application linéaire). Trouver l'ensemble des solutions n'est pas un problème trivial car il s'agit de trouver l'intersection d'un espace à $7-n$ dimensions (l'espace nul de la relation précédente, c'est-à-dire l'ensemble des antécédents de 0 pour l'application linéaire) et d'un hypercube à 7 dimensions, ce que l'on ne sait en général pas faire de manière formelle. Une première approximation de l'intersection peut être obtenue par programmation linéaire. Puis l'espace nul est échantillonné aléatoirement, et l'appartenance à l'intersection de chacun des points est testée[10].

3. Expérimentation

Les premières experiences que nous effectuons ici ont pour objectif de valider le processus d'inversion, en particulier sa capacité à retrouver une articulation correcte à partir de bons vecteurs acoustiques. Pour éliminer toutes les sources d'erreurs externes, notamment celles liées au traitement du signal acoustique, à l'adaptation du modèle articulatoire au locuteur, ou aux imperfections du synthétiseur articulatoire, nous avons pratiqué l'inversion sur un signal acoustique *synthétique*, c'est-à-dire généré grâce au synthétiseur sur des données articulatoires.

Le corpus de données sur lequel nous travaillons, issu de l'étude de Bothorel et al.[3] à l'institut de Phonétique de Strabourg, est constitué d'une dizaine de phrases prononcées par la locutrice PB en cinéradiographie. Le signal de parole enregistré simultanément est également disponible, mais est très bruité. Ce sont ces mêmes données qui ont servi à établir le modèle articulatoire[8] que nous utilisons ; S. Maeda nous a fourni les données radiographiques correspondantes transposées dans le modèle sous la forme de vecteurs articulatoires. Nous travaillons directement sur les vecteurs articulatoires, nous n'avons pas cherché à les recalculer à partir des données radiographiques originales. À partir de ces données, il est aisé à l'aide du synthétiseur articulatoire intégré au modèle[7] de générer un signal de parole synthétique.

Les données articulatoires ne permettent cependant pas d'obtenir le signal acoustique original ; deux explications peuvent être avancées :
- le synthétiseur acoustique intégré au modèle fait des hypothèse erronées et par conséquent produit un signal acoustique erronées,
- les données articulatoires sont erronées.

En l'occurrence, les deux explications sont valables : le synthétiseur utilise une approximation, certes courante[4], mais néanmoins une approximation, pour retrouver la troisième dimension du conduit vocal et ainsi en déduire la fonction d'aire ; ensuite les hypothèses formulées pour calculer le spectre du signal acoustique ne sont a priori valables que pour les fré-

quences inférieures à 4kHz, et ne sont donc a priori pas toujours respectées pour les 4^{e}et 5^{e}formants. Les données articulatoires sont également loin d'être parfaites : un détourage manuel des images radiographiques a d'abord été effectué, par des personnes différentes (ce qui induit un manque de cohérence dans les frontières), et les croquis ont ensuite été numérisés avec parfois des défauts d'orientation qui n'ont pas toujours été corrigés.

Par conséquent, il n'est guère étonnant de ne pas retrouver le signal acoustique original à partir des données articulatoires; de même, il n'est guère raisonnable d'espérer retrouver les trajectoires articulatoires originales à partir du signal acoustique réel. Il est cependant tout à fait envisageable de pouvoir les retrouver à partir d'un signal acoustique synthétique, ou de retrouver des trajectoires articulatoires *proches* de l'originale à partir du signal réel.

Protocole 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, 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.

Résultats Nous présentons ici les résultats de l'inversion effectuée sur la première phrase du corpus, « Ma chemise est roussie » (/maʃmizɛʁusi/).

La figure 1 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 rouges), les solutions de l'inversion statique (points noirs), la trajectoire obtenue à l'issue du lissage non-linéaire (étoiles vertes), et enfin la trajectoire obtenue après régulation variationnelle (carrés bleus). On voit que dès la première étape la trajectoire articulatoire se devine déjà, la contrainte acoustique est ici très forte pour ce paramètre articulatoire. On constate également sur cette figure que la similitude entre les courbes inverses et originale 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 deux décrochements non négligeables (une erreur d'une demi-unité) autour de 500ms et autour de 1100ms. On observe ici également certaines limitations du codebook : pour certains vecteurs acoustiques, on ne trouve ici aucun antécédant (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. Cependant, ces vecteurs acoustiques correspondant à des fricatives, il n'est pas étonnant que ceux-ci ne soient pas présents dans le codebook utilisé, destiné à l'inversion des voyelles.

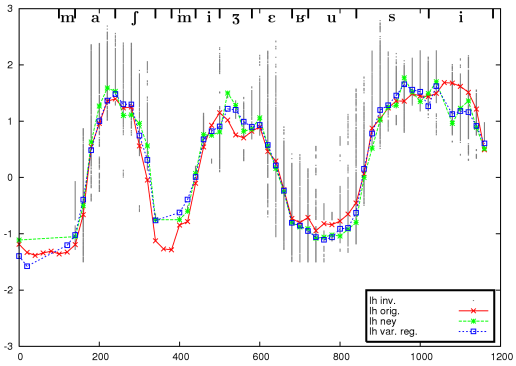

Fig. 1: Paramètre d'ouverture des lèvres original et retrouvé par inversion. Les points noirs éparpillés correspondent aux solutions de l'inversion statique, la trajectoire articulatoire originale est representée pqr des croix rouges, la trajectoire articulatoire trouvée par lissage non-linéaire par des étoiles vertes, celle trouvée après régulation variationnelle par des carrés bleus. Les ordonnées sont en unités de paramètre articulatoire, l'abscisse en ms.

La figure 2 présente les résultats de l'inversion pour le paramètre articulatoire correspondant à la protrusion des lèvres. On constate ici que l'inversion ne retrouve pas la bonne trajectoire articulatoire. On constate également lors de la première étape de l'inversion que la contrainte acoustique liée à ce paramètre est très faible : les solutions possibles couvrent pratiquement intégralement l'intervalle de variation pour ce paramètre. Par ailleurs, la courbe articulatoire à retrouver est ici particulièrement chaotique et les contraintes dynamiques mises en place ne peuvent pas permettre de retrouver une telle courbe.

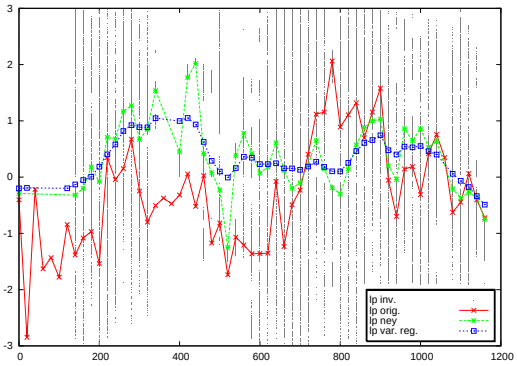

Fig. 2: Paramètre de protrusion des lèvres original et retrouvé par inversion. Les courbes suivent les mêmes conventions que pour la figure 1.

Le tableau 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

Err.	d_1	d_2	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. 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.

(*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 régulation variationnelle augmente de façon importante la fidélité à l'originale, 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. Conclusion et perspective

Les premières experiences confirment la qualité de la méthode d'inversion dans des conditions idéales : en effet, nous parvenons à retrouver certaines trajectoires articulatoires pratiquement identiques aux données originales, en utilisant les 5 premières fréquences formantiques. Les principales sources d'erreurs ici sont liées au faits que les trajectoires articulatoires ne sont pas tout à fait conformes aux hypothèses habituelles sur la régularité des trajectoires articulatoires. En recalculant les trajectoires à partir des cinéradiographies originales en imposant une condition de régularité, il est probable que l'on puisse obtenir des trajectoires plus facilement atteignables.

Il serait intéressant de voir si avec des paramètres acoustiques plus réduits (4, voire 3 fréquences formantiques) on parvient toujours à retrouver d'aussi bons résultats. La prochaine étape est bien entendu de pratiquer des expériences d'inversion sur des signaux acoustiques réels pour vérifier que le système parvient à compenser les erreurs liées au modèle et au synthétiseur articulatoires.

Références

[1] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63(5) :1535–1555, May 1978.

[2] L.-J. Boë, P. Perrier, and G. Bailly. The geometric vocal tract variables controlled for vowel production : proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20 :27–38, 1992.

[3] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling. *Cinéradiographies des voyelles et consonnes du Français*. Travaux de l'institut de Phonétique de Strasbourg, 1986.

[4] J. M. Heinz and K. N. Stevens. On the relations between lateral cineradiographs, area functions and acoustic spectra of speech. In *Proceedings of the 5th International Congress on Acoustics*, page A44., 1965.

[5] Y. Laprie and 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, May 1998.

[6] 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.

[7] S. Maeda. A digital simulation of the vocal tract system. *Speech Communication*, 1 :199–229, 1982.

[8] S. Maeda. Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

[9] Slim Ouni and Yves Laprie. Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire. In *Actes des Journées d'Étude sur la parole, Aussois*, June 2000.

[10] Slim Ouni and Yves Laprie. Studying articulatory effects through hypercube sampling of the articulatory space. In *17th International Congress on Acoustics , Rome, Italy*, volume 4, September 2001.

[11] B. Potard and Y. Laprie. Compact representations of the articulatory-to-acoustic mapping. In *Interspeech, Anvers*, August 2007.

[12] J. Schroeter and M. M. Sondhi. Speech coding based on physiological models of speech production. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 231–267. Dekker, New York, 1992.

[13] V.N. Sorokin and A.V. Trushkin. Articulatory-to-acoustic mapping for inverse problem. *Speech Communication*, 19 :105–118, 1996.