Inversion des fricatives par codebook hypercuboïque

Farid Feiz, Blaise Potard, Yves Laprie

Equipe PAROLE
LORIA, Campus Scientifique - BP 239, 54506 VANDOEUVRE-lès-NANCY CEDEX, France
Tél. : +33 (0)3 83 59 30 00 - Fax : +33 (0)3 83 27 83 19
Mél : laprie@loria.fr - http ://parole.loria.fr

ABSTRACT

The objective is to recover the vocal tract shape dynamics from the speech signal of vowels and fricatives. The method relies on the analysis-by-synthesis paradigm and is an extension of the method proposed by Ouni and Laprie which exploits a hypercubic articulatory table to represent the synthesis facet, i.e. Maeda's articulatory model. The first major modification is the use of parallelepiped instead of cubes. The new construction strategy only subdivides the articulatory space in the articulatory direction which gives rise to the strongest non-linearities. This enables a substantial reduction of the table size. The second major modification is the inversion of fricative sounds. In addition to the articulatory parameters the relative location of the noise source downstream the constriction is taken into account. This gives rise to three different articulatory codebooks, each corresponding to the relative position of the source with respect to the main constriction. This new inversion method has been evaluated on VCV sequences.

Keywords: voiceless fricative, acoustic-to-articulatory mapping, articulatory inversion

1. Introduction

L'objectif de l'inversion est de récupérer l'évolution temporelle du conduit vocal à l'origine de la production du signal de parole. Les outils mis en oeuvre pour y parvenir, reposent souvent sur une méthode d'analyse par synthèse. Cela signifie que le spectre de parole à inverser est comparé à un spectre calculé par synthèse articulatoire. L'étape de synthèse fait appel à un modèle articulatoire qui calcule la forme du conduit vocal, à partir de la position de la mâchoire, la position et la forme de la langue, du larynx et des lèvres. Pour éviter de faire appel trop souvent à l'étape de synthèse coûteuse en temps de calcul, la synthèse est représentée sous la forme d'une table (ou codebook) formée de couples associant les paramètres articulatoires aux paramètres acoustiques.

Notre méthode d'inversion repose sur le modèle articulatoire de Maeda [8] et sa méthode de simulation du conduit vocal dans le domaine fréquentiel [7]. Nous présentons dans cet article les premières expériences sur l'inversion de séquences VFV à l'aide d'un codebook hypercuboïque.

2. Notre méthode d'inversion

Notre méthode d'inversion comporte trois étapes. La première consiste à trouver toutes (à un échantillonnage près) les solutions potentielles pouvant donner le spectre à inverser. Les vecteurs de paramètres articulatoires retenus sont donc ceux dont l'image par synthèse est proche des données acoustiques.

A partir de l'ensemble des solutions retenues, la seconde étape consiste à reconstruire une trajectoire articulatoire qui soit suffisamment régulière au cours du temps. Nous utilisons pour cela un algorithme de programmation dynamique qui minimise une fonction de coût représentant la «distance» couverte par les articulateurs.

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.

2.1. Construction du codebook hypercuboïque

Il existe a priori une infinité de solutions articulatoires permettant d'obtenir le même spectre de parole. Nous utilisons une méthode de tabulation afin d'obtenir tous les candidats potentiels, à un échantillonnage près, dont l'image acoustique est proche du spectre à inverser. Il est ensuite possible de déterminer parmi ces candidats, ceux qui minimisent la «distance» couverte par les articulateurs au cours du temps.

Pour construire la table articulatoire, il faut réaliser le pavage de l'espace articulatoire en petits éléments, où la relation de l'articulatoire vers l'acoustique peut être évaluée localement très rapidement. L'algorithme de construction du codebook peut se résumer en une exploration récursive de l'espace articulatoire, arrêtant l'exploration dès que la relation articulatoire $\Rightarrow$ acoustique est «suffisamment» linéaire, c'est-à-dire lorsque l'erreur commise est inférieure à un certain seuil.

Nous avons choisi d'utiliser une structure hypercuboïdale. Cette structure est une généralisation de la structure hypercubique créée par Slim Ouni [9]. Comparé aux codebooks présentés dans les autres études consacrées à l'inversion acoustique-articulatoire [1, 14, 5, 4, 15, 2, 3, 12], le codebook hypercubique de Ouni présente plusieurs particularités appréciables : une exploration intégrale de l'espace articulatoire, et

une précision acoustique 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.

Pour la méthode de Ouni, la subdivision des cubes avait lieu dès que la linéarité locale n'était pas assurée. La subdivision consistait à diviser la taille de l'arête par deux, ce qui donnait donc $2^7 = 128$ sous-hypercubes puisque l'espace articulatoire est de dimension 7. 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 tables articulatoires nettement plus concis pour une même précision acoustique.

2.2. Modèle d'articulation des voyelles et fricatives

Compte tenu des différences acoustiques et articulatoires importantes entre les voyelles et les fricatives, notre étude a consisté à développer une nouvelle approche de l'inversion. Notre travail a porté en premier lieu sur l'adaptation de la simulation acoustique utilisée dans le synthétiseur articulatoire de Maeda et la construction en conséquence d'une table articulatoire adaptée aux fricatives.

Sur un plan articulatoire, une fricative est produite lorsque la constriction devient suffisamment étroite pour que l'écoulement de l'air devienne turbulent. Contrairement aux voyelles pour lesquelles il n'y a qu'une seule source, la vibration des cordes vocales, il existe souvent plusieurs sources d'excitation pour les fricatives. Une fricative se caractérise en effet, par la combinaison d'une source de bruit de friction située au niveau de la constriction, avec éventuellement plusieurs sources turbulentes de pression en aval de la constriction, au niveau des dents par exemple. Notre étude se limite aux fricatives sourdes, donc en l'absence de vibration de la glotte.

Pour réaliser les labio-dentales [f], les incisives supérieures viennent partiellement au contact de la lèvre inférieure, provoquant à cet endroit une zone de turbulence. Dans le cas des alvéolaires [s], la constriction se place juste derrière les incisives supérieures ; l'aire en aval de la constriction subit une pression au contact des dents. Pour la production des post-alvéolaires [ʃ], le conduit subit un resserrement entre les alvéoles et le début du palais ; la présence des incisives renforce la turbulence du flot d'air dans la cavité avant.

De façon comparable à [13], aux 7 paramètres du modèle de Maeda décrivant la configuration du conduit vocal, on adjoint un paramètre supplémentaire qui décrit la position relative de la source de bruit éventuelle par rapport à la constriction. Il est important de noter qu'il n'est pas possible de considérer ce paramètre comme les autres, puisqu'il dépend de la forme que donne les sept premiers paramètres. Le paramètre décrivant cette position relative, est exprimé en nombre de sections en aval de la constriction ; il vaut soit 0 (source en sortie de la constriction), soit 1, 2, etc. Dans notre étude, nous avons considéré trois positions relatives de la source de bruit par rapport à la constriction, et avons généré un codebook hypercuboïque pour chaque position relative. Cette couverture de la position relative permet donc de retrouver une grande variété de solutions inverses.

Pour les voyelles, seuls les 7 premiers paramètres du modèle contrôlent le spectre des voyelles, et un seul codebook hypercuboïque leur est donc dédié.

2.3. Analyse de la séquence à inverser

Le corpus acoustique est constitué de fricatives sourdes [ʃ s f] enregistrées par le locuteur YL dans des contextes vocaliques [i a u y]. Le modèle articulatoire a été adapté au locuteur YL, notamment grâce aux paramètres décrivant l'élongation des sections orale et pharyngale. Ces deux paramètres, obtenus à l'aide d'images IRM, représentent la longueur du conduit vocal du sujet YL par rapport à celle du modèle articulatoire.

La fréquence fondamentale est utilisée comme critère de segmentation entre la fricative et les voyelles contextuelles. Une fois le signal de départ segmenté en une suite V-F-V, nous avons choisi les 4 premiers formants comme vecteur de paramétrisation des voyelles, estimés tous les 4 ms par les racines du polynôme de prédiction LPC, et réajustées manuellement sur le spectrogramme. Pour la fricative, nous avons décidé d'évaluer son enveloppe cepstrale tous les 4 ms. La présence des creux entre les harmoniques, capturés pour un nombre suffisamment élevé de coefficients cepstraux (de l'ordre de 30) et les pics de fréquence, peuvent servir à accentuer le contraste acoustique entre les fricatives et renseigner sur une position éventuelle de la source de pression. Nous avons en définitive fait le choix de 4 pics de fréquence pour en avoir autant que dans le cas des voyelles.

2.4. *Utilisation du codebook hypercuboïque pour l'inversion*

Comme nous l'avons vu précédemment, le signal est décomposé en segments de 4 ms chacun. Chaque segment constitue une entrée acoustique s à inverser. On cherche à déterminer l'ensemble des vecteurs articulatoires v dont l'image par le synthétiseur articulatoire F est proche de s. Nous utilisons pour cela l'approximation du codebook hypercuboïque, approprié à l'entrée acoustique à inverser (voyelle/fricative) :

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

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,

$$P(X) - s = 0$$

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. Dans notre cas, P est un polynôme de degré 1 : le système devient linéaire, et la résolution peut se faire simplement grâce aux méthodes classiques d'algèbre linéaire. Si nous cherchons à inverser des quadruplets de fréquences, $N = 7$ et $M = 4$. Nous obtenons ainsi un système sous-déterminé ; le sous-espace vectoriel des solutions est a priori de dimension $N - M = 3$, 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 3 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.

La procédure précédente a permis de rechercher l'ensemble des points articulatoires possibles à un instant. Pour trouver les trajectoires articulatoires, il faut choisir à chaque instant du segment de parole, un point articulatoire parmi ceux qui viennent d'être trouvés. La recherche d'une trajectoire articulatoire s'effectue à l'aide d'un algorithme de programmation dynamique qui minimise la «distance» couverte par les articulateurs [6].

Une fois la meilleure trajectoire trouvée, elle est régularisée à l'aide d'un algorithme [6] qui améliore la régularité tout en assurant que les trajectoires acoustiques obtenues par synthèse à partir des trajectoires articulatoires inverses, sont proches des trajectoires mesurées dans la séquence de parole.

La fidelité des trajectoires articulatoires inverses a par ailleurs été évaluée [10] sur les données articulatoires qui ont servi à construire le modèle.

3. Expériences

Nous présentons nos expériences d'inversion sur quelques séquences /aʃV/ de manière à mettre en évidence l'influence du contexte phonétique sur la nature des trajectoires articulatoires. Bien souvent, les gestes consonantiques effectués pour la production de la fricative démarrent durant les configurations associées à la production de la voyelle précédente. De la même manière, les gestes vocaliques sont souvent anticipés durant l'apparition de la fricative qui précède la voyelle. Nous avons représenté sur les figures 2 et 3, l'évolution des articulateurs critiques au cours du temps.

Pour faciliter la prononciation du /ʃ/, l'apex de la langue (P4) reste en position relevée tout au long de la séquence /aʃa/ (Fig. 2). Le paramètre (P2) décrit la variation de la position du corps de la langue. Il décroît, et donc, le corps de la langue s'avance pour réaliser la constriction ; il croît ensuite pour permettre la production de la voyelle /a/. La prononciation du /ʃ/ s'accompagne d'un mouvement de protrusion des lèvres (P6).

Pour réaliser la séquence /aʃy/, la langue se masse en arrière de la cavité orale, tandis que le corps de la langue (P3) s'abaisse progressivement dans la transition /aʃ/ (Fig. 3) pour s'élever légèrement en fin de séquence durant la transition /ʃy/. La protrusion (P6) qui caractérise la voyelle /y/, est relativement précoce et anticipée durant l'apparition de la fricative.

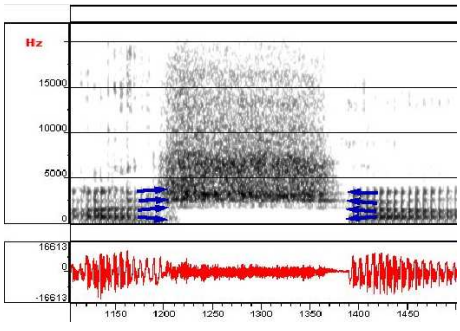

Fig. 1: *Direction des transitions formantiques de la voyelle* /a/ *vers la fricative* /ʃ/ *dans la séquence* /aʃa/

4. Conclusion et perspectives

La force de notre méthode d'inversion repose sur le pavage fin et étendue de l'espace articulatoire, notamment grâce au contrôle de la position relative de la source de friction dans le conduit. Notre méthode garantit non seulement l'obtention d'une grande variété de solutions inverses, mais une régularité dans les

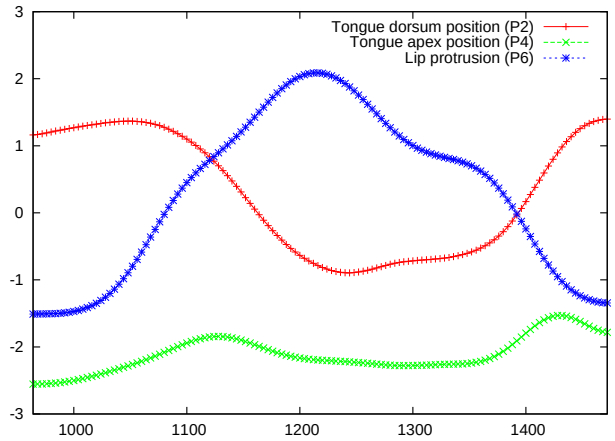

Fig. 2: *Evolution temporelle des paramètres articulatoires (position et pointe de la langue) (pour la transition /aʃa/) obtenue après inversion*

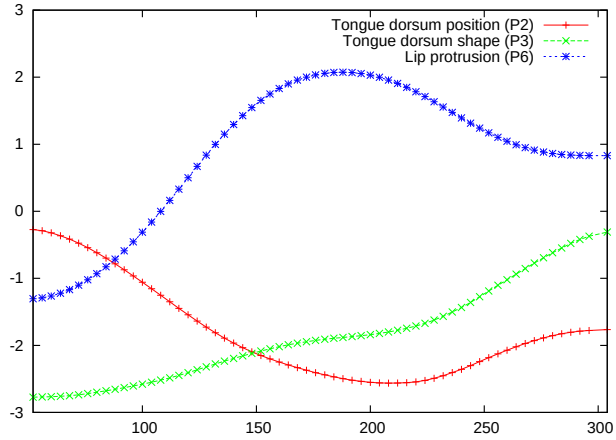

Fig. 3: *Evolution temporelle des paramètres articulatoires (position de la langue, forme de la langue et protrusion) (pour la transition /aʃy/) obtenue après inversion*

trajectoires articulatoires de fricatives en contexte vocalique. Cette première étude doit bien entendu être complétée par l'étude des autres consonnes fricatives, en occurence des fricatives voisées.

La source d'excitation était jusqu'ici placée à la jonction entre une section et une autre dans la grille semi-polaire de Maeda. Nous souhaitons en perspective adapter le modèle articulatoire pour gérer de façon plus fine la position de la source de friction dans le conduit.

Remerciements Nous remercions vivement Shinji Maeda et Martine Toda pour leur disponibilité et les discussions fructueuses.

Références

[1] M. V. Mathews B. S. Atal, J. J. Chang 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, 1978.

[2] F. Charpentier. Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic non-linearities. *Speech Communication*, 3 :291–308, 1984.

[3] V. Gracco I. Zlokarnik P. Rubin J. Hogden, A. Lofqvist and E. Saltzman. Accurate recovery of articulator positions from acoustics : new conclusions based on human data. *Journal of the Acoustical Society of America*, 100(12) :1819–1834, 1996.

[4] J. Schroeter J. N. Larar and M. M. Sondhi. Vector quantization of the articulatory space. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(12) :1812–1818, 1988.

[5] P. Perrier L.-J. Boë 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.

[6] Y. Laprie and B. Mathieu. A variational approach for estimating vocal tract shapes from the speech signal. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 929–932, 1998.

[7] S. Maeda. A digital simulation method 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] S. Ouni and Y. Laprie. Improving acoustic-to-articulatory inversion by using hypercube codebooks. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, China*, volume 2, pages 178–181, 2000.

[10] B. Potard. Inversion acoustique-articulatoire dynamique par codebook hypercuboïque : premiers résultats. In *Journées d'Etudes sur la Parole - JEP'08 , Avignon*, 2008.

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

[12] 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.

[13] E. L. Riegelsberger. *The acoustic-to-articulatory mapping of voiced and fricated speech.* PhD thesis, The Ohio State University, 1997.

[14] 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.

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