Introduction de contraintes pour l'inversion acoustico-articulatoire utilisant une table hypercubique

Yves Laprie, Slim Ouni

Introduction de contraintes pour l'inversion acoustico-articulatoire utilisant une table hypercubique

Slim Ouni et Yves Laprie

LORIA/INRIA
615 rue du jardin botanique 54602 Villers-lès-Nancy FRANCE

Mél : {Slim.Ouni, Yves.Laprie}@loria.fr

RÉSUMÉ

Our acoustic to articulatory inversion method exploits an original codebook representing the articulatory space by hypercubes. The articulatory space is decomposed into regions where the articulatory-to-acoustic mapping is linear. Each region is represented by a hypercube. The inversion procedure retrieves articulatory vectors corresponding to an acoustic entry from the hypercube codebook. As the dimension of the articulatory space is greater than the dimension of the acoustic space, the corresponding null space is sampled by linear programming to retrieve all the possible solutions. A dynamic procedure is used to recover the best articulatory trajectory according to a minimum articulatory rate criterion. The addition of constraints allows the inversion process to be focused on realistic inverse articulatory trajectories.

1. INTRODUCTION

La plupart des méthodes d'inversion acoustico-articulatoire repose sur une approche d'analyse par synthèse. Cela signifie que la transformation articulatoire acoustique doit être représentée, soit explicitement sous la forme d'une table donnant les paramètres acoustiques (en général les formants) pour un certain nombre de points échantillonnant l'espace articulatoire [4], soit implicitement sous la forme d'un réseau neuromimétique par exemple. La qualité de la représentation influence fortement les solutions inverses récupérées puisque ces trajectoires doivent s'appuyer, au moins en partie, sur les points de la table articulatoire. Pour cette raison nous avons développé un algorithme d'échantillonnage adaptatif qui assure que la résolution acoustique est relativement indépendante de la région de l'espace articulatoire considérée [8]. L'échantillonnage adaptatif conduit à une table structurée sous la forme d'une arborescence d'hypercubes à l'intérieur desquels la transformation articulatoire acoustique peut être considérée comme presque linéaire.

Dans notre cas nous avons choisi le modèle articulatoire de Maeda [6] qui décrit la forme du conduit vocal à l'aide de sept paramètres exprimés en écarts type par rapport à la moyenne des mesures articulatoires. L'inversion consiste donc à retrouver les trajectoires des sept paramètres articulatoires à partir de la donnée des trois premiers formants pour un signal de parole.

Lors de l'inversion d'un signal il faut récupérer à chaque instant tous les points articulatoires donnant les formants mesurés, et ensuite, il faut construire les meilleures trajectoires articulatoires, ce qui revient à trouver les meilleurs chemins à partir des points articulatoires récupérés.

2. RÉCUPÉRATION DE TOUS LES POINTS INVERSES DANS UN HYPERCUBE

Pour trouver tous les points articulatoires qui peuvent donner un triplet de formants mesurés sur le signal de parole à inverser, on recherche tous les hypercubes compatibles avec ces formants. Il faut ensuite trouver tous les points articulatoires possibles à l'intérieur de chaque hypercube. Soit F le vecteur des trois premiers formants, F_0 le vecteur des formants au centre de l'hypercube et ∇F la matrice jacobienne de la transformation articulatoire acoustique calculée au centre de l'hypercube P_0. On cherche donc tous les points P tels que :

$$F = F_0 + \nabla F(P - P_0)$$

La décomposition en valeurs singulières [3] de la matrice jacobienne fournit un point solution et la base de l'espace nul. Comme la dimension du vecteur F est 3 et celle de $P - P_0$ est 7, la dimension de l'espace nul est en général 4. Pour connaître les points de l'hypercube qui peuvent donner les formants mesurés, il faut échantillonner l'intersection de l'hypercube avec l'espace formé par le point trouvé par la méthode SVD et la base de l'espace nul. Soit P_{svd}, le point trouvé par SVD, et $\{v_j\}_{j=1..4}$ la base de l'espace nul, les points P_s solution sont donc :

$$P_s = P_{svd} + \sum_{j=1}^{4} \beta_j v_j \qquad (1)$$

Les coordonnées $\beta_{j=1..4}$ doivent être choisies pour assurer que $P_s \in H_c$, c'est-à-dire :

$$\alpha^i_{inf} \leq P^i_{svd} + \sum_{j=1}^{4} \beta_j v^i_j \leq \alpha^i_{sup} \qquad i = 1..7 \qquad (2)$$

où α^i_{inf} et α^i_{sup} définissent la plus petite et la plus grande valeur du $i^{ème}$ paramètre articulatoire dans l'hypercube étudié. Ce problème est très simple à résoudre en dimension 2 mais il n'existe pas de solution connue dans le cas général. Par conséquent, nous l'avons résolu de manière approchée en considérant deux ensembles de programmes linéaires, l'un pour trouver les plus petites valeurs de β_j, et l'autre pour trouver les plus grandes valeurs β_j. Comme généralement la dimension de l'espace nul est quatre on résout donc quatre programmes linéaires par la méthode du simplexe. Les valeurs extrêmes des β_j définissent un sur-ensemble des solutions recherchées, il faut donc échantillonner cet ensemble pour trouver les points

qui appartiennent à l'hypercube. En pratique, nous avons adapté la finesse de l'échantillonnage pour conserver un nombre raisonnable de points (moins d'une centaine).

La résolution fréquentielle imposée lors de la construction de la table articulatoire permet d'atteindre une précision moyenne de l'ordre de 10 Hz entre les formants du signal de départ et ceux synthétisés à partir des points articulatoires inverses.

3. CONSTRUCTION DES TRAJECTOIRES ARTICULATOIRES INVERSES

La procédure précédente donne pour chaque triplet de formants l'ensemble des points articulatoires possibles à un instant. Pour retrouver les trajectoires articulatoires il faut choisir à chaque instant du segment de parole à inverser 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 l'effort articulatoire du locuteur. Il faut noter que la minimisation ne porte que sur le critère articulatoire car les points articulatoires qui ont été récupérés à l'étape précédente produisent des formants très proches (moins de 10 Hz en moyenne) de ceux qui ont été mesurés dans le signal de parole.

Nous notons $s(i)$ l'ensemble des points articulatoires récupérés à l'instant t_i. La suite de ces ensembles est : $S = (s(0) \dots s(i) \dots s(N))$ où N est le nombre d'instants pour lesquels l'inversion a été effectuée. La construction d'une trajectoire donne lieu à une double sélection : les instants auxquels la trajectoire est définie et le point articulatoire choisi à l'intérieur de l'ensemble $s(i)$. Cette double sélection permet d'éliminer éventuellement tous les points articulatoires de l'un des instants du signal pour lequel l'inversion a échoué. Le choix des instants se fait par l'intermédiaire d'une fonction de sélection strictement croissante j parmi les instants i. La séquence résultat est de la forme : $\overline{S} = (s(j(0)) \dots s(j(k)) \dots s(j(K)))$ où $K < N$.

Le choix de l'un des points parmi les ensembles sélectionnés donne la trajectoire articulatoire résultat qui est de la forme : $\overline{A} = (\alpha(j(0)) \dots \alpha(j(k)) \dots \alpha(j(K)))$. Le coût de choisir $\alpha(j(k))$ après $\alpha(j(k-1))$ est :

$$C(\alpha(j(k)), \alpha(j(k-1))) = \lambda \sum_{i=1}^{7} m_i (\alpha_i(j(k)) - \alpha_i(j(k-1)))^2 - Bonus(\alpha(j(k)))$$

où m_i est la pondération donnée au $i^{ème}$ paramètre articulatoire et $Bonus(\alpha(j(k))$ est un bonus qui représente l'intérêt de conserver le point $\alpha(j(k))$ dans la trajectoire articulatoire finale. Le bonus doit être supérieur en valeur absolue au premier terme de $C(\alpha(j(k)), \alpha(j(k-1)))$ pour éviter de trouver une trajectoire articulatoire vide et il permet aussi d'exprimer des contraintes sur les trajectoires que l'on souhaite récupérer. Le critère minimisé par programmation dynamique est donc : $D = \sum C(\alpha(j(k)), \alpha(j(k-1)))$.

Une fois la meilleure trajectoire trouvée elle est régularisée à l'aide d'un algorithme [5] qui améliore la régularité tout en assurant que les trajectoires des formants obtenues par synthèse à partir des trajectoires articulatoires inverses sont proches des trajectoires mesurées dans le signal de parole. L'intérêt de cet algorithme est de prendre en compte les trajectoires articulatoires globalement et d'intégrer directement le comportement acoustique du modèle articulatoire.

4. INTRODUCTION DE CONTRAINTES SUR LES TRAJECTOIRES INVERSES

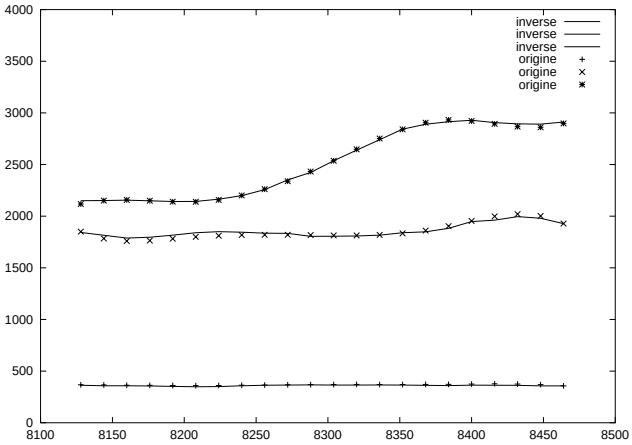

FIG. 1: *Trajectoires des trois premiers formants (pour la transition /yi/) obtenus par synthèse à partir des paramètres articulatoires inverses sans imposer de contraintes. Les trajectoires synthétiques sont représentées sous la forme de lignes. Les valeurs originales des formants sont représentées par des points.*

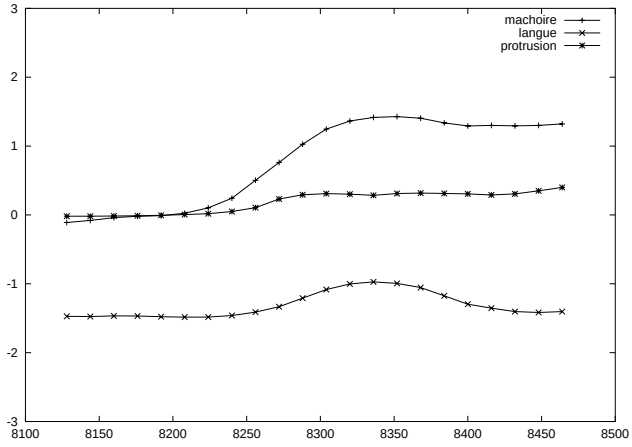

FIG. 2: *Évolution temporelle des trois paramètres articulatoires (mâchoire, position de langue et protrusion) sans imposer aucune contrainte.*

La thèse de Slim Ouni [8] contient les résultats d'inversion pour un grand nombre de transitions entre voyelles et pour des séquences /VCV/. Ces résultats ont été obtenus sur un sujet pour lequel le modèle de Maeda a été adapté à l'aide de la méthode proposée par Galvan [2]. La table articulatoire a été calculée pour les facteurs d'échelle de la bouche et du pharynx de ce locuteur. Elle ne peut donc pas servir à inverser la parole d'un autre locuteur.

Nous présentons ici un résultat d'inversion sur la transition /yi/ pour lequel il faut introduire des contraintes sur le point initial afin de trouver la solution inverse attendue. La figure Fig. 1 donne les trajectoires des trois premiers formants de la transition /yi/ pour notre sujet.

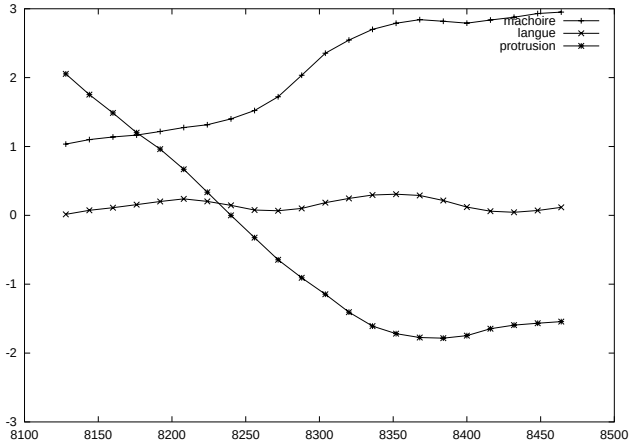

FIG. 3: *Évolution temporelle des trois paramètres articulatoires (mâchoire, position de langue et protrusion) en imposant la protrusion à 2,7σ et l'ouverture de la mâchoire à 1,5σ pour le premier point.*

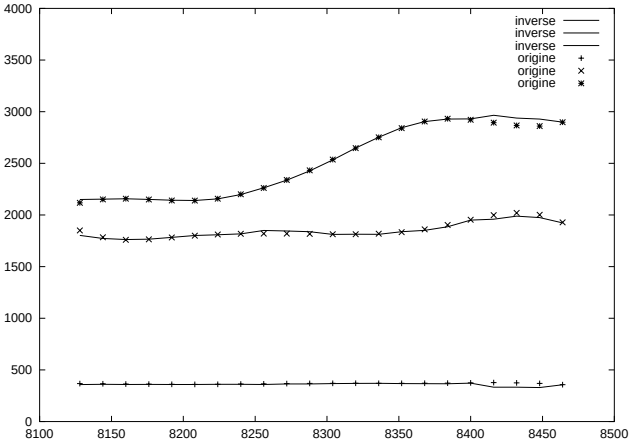

FIG. 4: *Trajectoires des trois premiers formants (pour la transition /yi/) obtenus par synthèse à partir des paramètres articulatoires inverses en imposant la protrusion à 2,7σ et l'ouverture de la mâchoire à 1,5σ pour le premier point. Les trajectoires synthétiques sont représentées sous la forme de lignes. Les valeurs originales des formants sont représentées par des points.*

L'application de la procédure d'inversion décrite audessus conduit aux trajectoires articulatoires de la figure Fig. 2 (seules les trajectoires de la mâchoire, de la position de la langue et de la protrusion sont représentées pour ne pas trop charger la figure).

La caractéristique la plus notable de ces résultats est la faible protrusion de /y/ alors que c'est l'une des caractéristiques articulatoires essentielles de /y/, même si les trajectoires des formants sont reproduites avec une grande fidélité par l'inversion (cf. Fig. 1). Cela signifie que le critère adopté pour rechercher le meilleur chemin - la minimisation de la vitesse articulatoire - est insuffisant à lui seul. Plutôt que de le modifier, nous avons donc ajouté une contrainte sur le point de départ de la protrusion en la fixant à $2{,}7 \pm 0{,}1\,\sigma$. L'adjonction de contraintes se réduit à donner un bonus très fort aux points de départ dont la protrusion est $2{,}7 \pm 0{,}1\,\sigma$ lors de l'application de la programmation dynamique. Il est d'ailleurs très simple d'ajouter des contraintes à d'autres instants de l'inversion en spécifiant un bonus très élevé pour les points vérifiant les contraintes à ces instants.

La nouvelle trajectoire articulatoire inverse pour la protrusion est cette fois plus correcte (cf. Fig. 5) mais elle reste insuffisante et surtout, la mâchoire inférieure part d'une position relativement ouverte.

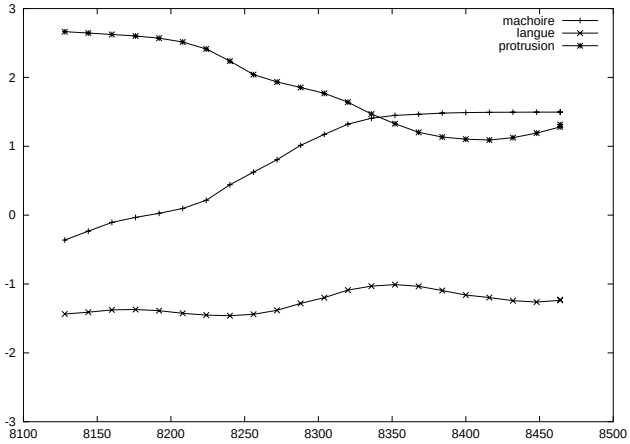

FIG. 5: *Évolution temporelle des trois paramètres articulatoires (mâchoire, position de langue et protrusion) en imposant la protrusion à 2,7σ pour le premier point.*

Nous avons donc complété la contrainte imposée au point de départ en fixant la position de la mâchoire à une valeur assez forte $1{,}5 \pm 0{,}5\sigma$. Cette fois les résultats deviennent plus corrects puisque la protrusion décroît fortement de /y/ à /i/ et la fermeture augmente légèrement sans que la position du corps de la langue n'évolue notablement (cf. Fig. 3). Il est très important de noter que toutes ces trajectoires articulatoires permettent toujours de reproduire avec une grande fidélité les formants de la transition réelle (cf. Fig. 4).

Cet exemple montre que l'adjonction de contraintes très simples permet de guider l'inversion vers les solutions attendues. Malgré tout, il faut noter que le nombre de solutions satisfaisant les contraintes décroît très fortement avec leur niveau d'exigence. Ainsi, nous avons dû donner une marge d'erreur de $0{,}5\,\sigma$ pour qu'il existe une solution. Cela signifie donc que nous avons atteint les limites de la compensation acoustique du modèle articulatoire adapté à notre sujet, ou encore que l'adaptation du modèle articulatoire n'est pas suffisante pour reproduire fidèlement la parole de notre sujet. Lors des expériences préliminaires sur l'adaptation du modèle de Maeda nous avions utilisé des images IRM de notre sujet [7]. Le modèle adapté était donc vraisemblablement plus fidèle à la géométrie du conduit vocal du sujet étudié, que dans le cas de cette expérience pour laquelle le modèle était adapté à l'aide de la procédure proposée par Galvan [2]. Malgré tout, l'erreur moyenne sur les fréquences des trois premiers formants de 10 voyelles orales du français restait assez élevée (49 Hz sur F1, 125 Hz sur F2 et 170 Hz sur F3). Ces erreurs sont comparables à celles relevées par Story et al. [9], par exemple. Elles proviennent d'erreurs sur la géométrie du conduit vocal et d'erreurs sur les paramètres physiques utilisés pour le calcul des pertes sans qu'il soit possible d'en connaître les contributions respectives.

L'un des avantages de notre méthode d'inversion est

qu'elle assure que toutes les solutions d'inversion possibles, compte tenu de la précision fréquentielle fixée pour la récupération des formants et du modèle articulatoire, ont été explorées. À notre connaissance il s'agit de la seule méthode d'inversion qui puisse assurer que toutes les trajectoires articulatoires autorisées par le modèle articulatoire ont été explorées. Pour l'inversion de la suite /yi/ présentée au-dessus, on constate ainsi qu'il n'existe qu'un petit nombre de solutions vraisemblables vérifiant une contrainte imposée à un seul point des trajectoires articulatoires. Cela confirme la pertinence du modèle articulatoire qui, à partir d'une contrainte très faible, permet de générer des déformations du conduit vocal rendant compte des observations réelles.

La quasi exhaustivité de l'exploration des trajectoires articulatoires est coûteuse en complexité puiqu'il existe en moyenne 10000 points inverses à chaque instant. La complexité de l'algorithme de programmation dynamique standard nécessiterait approximativement $N \times 10^8$ calculs de coût partiel, N étant le nombre d'instants auxquels l'inversion est effectuée. Comme nous tolérons des chemins éventuellement incomplets avec des sauts de longueur limitée (au plus 3 points) cela conduit à une complexité approximative de $N \times 4 \times 10^8$. Cette complexité importante est due en grande partie à l'exploration de l'espace nul. Les vecteurs choisis dans cet espace n'influencent pas les paramètres acoustiques. Au contraire, les vecteurs de base du complémentaire du noyau définissent les commandes articulatoires « efficaces » acoustiquement. Nous projetons de réduire fortement la complexité de l'inversion en retardant l'exploration de l'espace nul après la détermination de la partie des trajectoires articulatoires efficaces acoustiquement. Le critère à minimiser lors de la première étape sera la cohérence de la stratégie de commande mesurée par les produits scalaires entre les vecteurs de base efficaces acoustiquement aux instants t et $t + 1$.

5. CONCLUSION

Le point faible des méthodes d'inversion utilisant une table articulatoire construite à partir d'un échantillonnage partiel de l'espace articulatoire est d'orienter l'inversion vers des trajectoires articulatoires ne s'appuyant que sur les points de cette table. Par conséquent, certaines trajectoires articulatoires acceptables d'un point de vue articulatoire ne sont pas trouvées. Au contraire, notre méthode d'inversion ne favorise aucune trajectoire articulatoire, et l'exemple d'inversion précédent montre que certaines solutions inverses ne respectent pas les caractéristiques articulatoires observées chez l'être humain. Nous avons donc le projet de compléter la table articulatoire en donnant pour chaque hypercube la probabilité qu'il soit utilisé par un locuteur. L'apprentissage de ces probabilités peut se faire de deux façons. La première est d'exploiter les connaissances articulatoires classiques sous la forme de contraintes pour récupérer les trajectoires articulatoires prévisibles à partir de la suite de sons prononcée. L'exemple d'inversion de /yi/ donne une idée de ce processus. En imposant la protrusion et la position de la mâchoire on récupère des trajectoires acceptables d'un point de vue phonétique, et il est possible de renforcer la probabilité d'émission des points articulatoires correspondants. Cette solution assez semblable à celle proposée par Bailly [1] nécessite la construction d'un nombre suffisant de séquences de sons pour lesquelles il est possible de prédire correctement les gestes articulatoires. La seconde solution est d'utiliser de vrais locuteurs pour réaliser cet apprentissage, en utilisant des données radiographiques traditionnelles ou obtenues par microfaisceaux, ou encore, des données électromagnétographiques.

RÉFÉRENCES

[1] G. Bailly. Learning to speak. sensori-motor control of speech movements. *Speech Communication*, 22 :251–267, 1997.

[2] A. Galván-Rdz. *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.

[3] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1989.

[4] S. K. Gupta and J. Schroeter. Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *Journal of Acoustical Society of America*, 94(5) :2517–2530, Nov 1993.

[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] B. Mathieu and Y. Laprie. Adaptation of Maeda's model for acoustic to articulatory inversion. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2015–2018, Rhodes, Greece, September, 1997.

[8] S. Ouni. *Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico-articulatoire*. Thèse de L'Université Henri Poincaré, Dec 2001.

[9] B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *Journal of Acoustical Society of America*, 100(1) :537–553, July 1996.