Inversion acoustique-articulatoire en utilisant des contraintes phonétiques

Blaise Potard, Yves Laprie

Équipe Parole
LORIA, Campus Scientifique - BP 239
54506 VANDŒUVRE-lès-NANCY CEDEX, France
Tél.: +33 (0)3 83 59 30 00 - Fax: +33 (0)3 83 27 83 19
Mél: {Blaise.Potard, Yves.Laprie}@loria.fr - http://parole.loria.fr/

Inversion acoustique-articulatoire en utilisant des contraintes phonétiques

Blaise Potard, Yves Laprie

Résumé: Le but de l'inversion acoustique articulatoire est d'obtenir la position des articulateurs à partir du signal de parole. L'une des difficultés majeures de l'inversion est qu'une infinité de formes de conduits peut donner un même spectre de parole. Une façon de réduire cette difficulté est de contraindre davantage le problème, en utilisant par exemple des contraintes d'ordre visuel (on suppose connaître en plus du signal de parole, la position des articulateurs visibles ; ce qui peut se faire en utilisant une ou plusieurs caméras), ou d'ordre phonétique (les caractéristiques phonétiques des voyelles du Français sont connues, par exemple). Mais cette difficulté peut aussi se transformer en avantage : en permettant d'obtenir toutes les configurations du conduit vocal correspondant à un son donné, l'inversion fournit potentiellement un moyen d'étudier des stratégies compensatoires préservant l'acoustique. Nous montrerons comment l'utilisation de contraintes d'origine phonétique permet de réduire considérablement l'espace des solutions et d'améliorer la pertinence des solutions.

1 Introduction

Le but de l'inversion acoustique articulatoire est d'obtenir une description du conduit vocal à partir d'un signal de parole. Historiquement, le conduit était simplement représenté par sa fonction d'aire, et le signal de parole par les fréquences des formants. Atal et ses collègues[*1*] ont montré qu'une infinité de fonctions d'aires peut générer exactement le même triplet de formants.

L'une des principales difficultés de l'inversion acoustico-articulatoire est ainsi d'ajouter des contraintes permettant de restreindre l'espace des solutions tout en gardant les solutions pertinentes. Une approche courante est l'utilisation de modèles articulatoires, pilotés par un petit nombre de paramètres, et imitant la morphologie du conduit vocal ([2, 3]). Ces modèles sont généralement issus de l'étude statistique d'images d'origine médicale (Rayons X, IRM, ...).

Même si ces modèles réduisent considérablement le domaine des formes de conduits envisageables, l'indétermination reste encore trop importante, et il reste un très grand nombre de solutions pour chaque triplet de formants. Cela n'a rien d'étonnant : la variabilité articulatoire est une caractéristique essentielle de la parole. Les articulateurs de la parole ont de très grandes capacités compensatoires, qui permettent d'enchaîner des sons ayant des caractéristiques articulatoires intrinsèques très différentes. Malgré cette grande variabilité, il existe un certain nombre d'invariants articulatoires. Le but de ce travail est d'exploiter ces invariants articulatoires, issus de connaissances phonétiques traditionnelles sur les lieux d'articulation, sous la formes de contraintes imposées aux paramètres articulatoires.

Nous commencerons par décrire les contraintes phonétiques et leur implémentation dans notre système d'inversion[4], qui utilise une table articulatoire (ou codebook), générée en utilisant le modèle articulatoire de Maeda[3]. Nous évaluerons ensuite la validité de ces contraintes, d'abord dans le cas de voyelles isolées pour étudier leur effet sur le lieu et l'importance de la constriction, puis dans le cas de séquences de parole où les paramètres articulatoires sont connus.

2 Exprimer les caractéristiques phonétiques sous forme de contraintes articulatoires

L'idée principale qui sous-tend l'utilisation de contraintes d'origine phonétique est l'hypothèse que chaque phonème a des caractéristiques articulatoires invariantes, comme par exemple une forte protrusion pour le /y/. Dans le cas des voyelles, qui présentent des structures acoustiques relativement stables dans le temps, comparées à d'autres phonèmes comme les plosives, ces caractéristiques peuvent être aisément traduites en contraintes sur les paramètres articulatoires.

Dans la suite, ces contraintes articulatoires liées aux caractéristiques phonétiques des phonèmes seront simplement désignées sous le terme de «contraintes phonétiques».

2.1 Contraintes phonétiques pour les voyelles non-nasales

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. L'importance relative de chaque caractéristique dépend de la voyelle considérée. Comme nous l'avons mentionné dans l'introduction, la variabilité inter-locuteurs est importante ; par conséquent, nous avons défini des contraintes numériques (plutôt que booléennes) de façon à associer à un vecteur articulatoire un score de pertinence phonétique en fonction de l'adéquation de ses caractéristiques phonétiques à celles de la voyelle associée (elle-même identifiée par la valeur des formants correspondants).

Le tableau 1 résume notre classification pour les 10 voyelles non-nasales du français. Dans ce tableau, *D* correspond à la position du dos de la langue, *O* correspond à l'ouverture de la bouche, *E* à l'étirement des lèvres, et *P* à la protrusion. La convention retenue est intuitive : plus le nombre est grand, et plus la valeur associée à la contrainte correspondante sera élevée. Par exemple, une contrainte de O_1 signifie que la bouche a une ouverture faible, et O_4 correspond à une ouverture de la bouche très importante. Ces données sont des valeurs moyennes de la façon dont un locuteur Français va articuler les voyelles. 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/). Ceci explique pourquoi *D* ne varie qu'entre 6 et 8 pour les voyelles.

Voyelle	D	O	E	P
i	D6	01	E4	P1
e	D6	02	E3	P1
ɛ	D6	03	E2	P1
a	D7	04	E1	P1
y	D6	01	E1	P4
ø	D6	02	E1	P3
œ	D6	03	E1	P2
u	D8	01	E1	P4
o	D8	02	E1	P3
ɔ	D8	03	E1	P2

Tableau 1: *Classification des voyelles du français.*

2.2 Transposer les 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 notre cas, nous utilisons le modèle de Maeda[3], dans lequel les paramètres peuvent facilement s'interpréter comme des articulateurs au sens phonétique. Par conséquent, l'expression des contraintes phonétiques sous forme de paramètres articulatoires est très simple (sauf pour l'étirement des lèvres, que l'on ne peut pas exprimer) : 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 [5] : 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

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

où *Tp* est la position du dos de la langue, *Jw* est la position de la mâchoire et α 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 cinéradiographiques. 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 PB, le coefficient que Maeda avait trouvé expérimentalement (approximativement égal à 0.66). Cet effet compensatoire permettait à Maeda d'expliquer la majeure partie de la variabilité articulatoire intra-locuteur de /a/ et /i/.

2.3 Partitionnement de l'espace acoustique

Pour chaque phonème, nous devons définir un domaine acoustique où les contraintes phonétiques vont être considérées comme valides, c'est-à-dire, un domaine où il est probable d'observer des configurations articulatoires qui respectent les contraintes données. Pour le moment, nous utilisons des modèles simples centrés sur les fréquences formantiques moyennes de locuteurs français (les valeurs de [6]).

Notre modèle d'inversion travaille actuellement sur l'espace tridimensionnel des trois premières fréquences formantiques. Nous avons partitionné l'espace acoustique en utilisant différents modèles : diagramme de Voronoï sur les centres des voyelles (cf. Fig. 1); diagramme de Voronoï pondéré par l'écart type de chaque fréquence formantique (cf. Fig. 2).

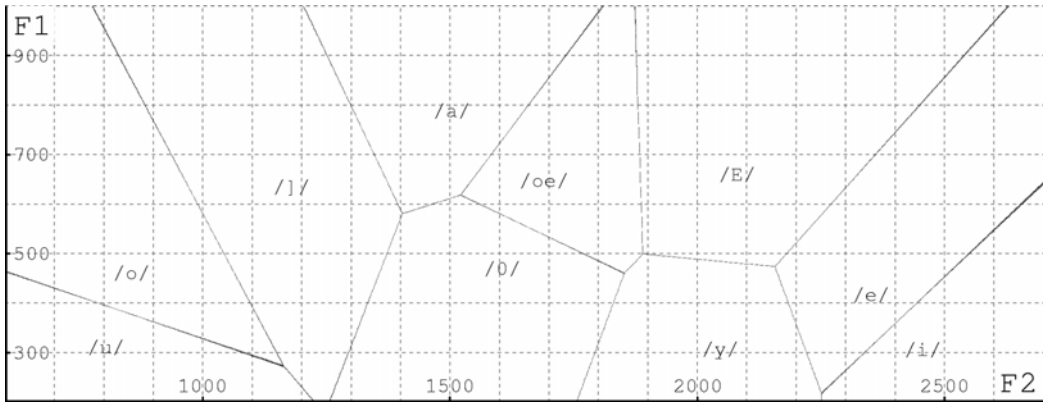

Figure 1: *Diagramme de Voronoï.*

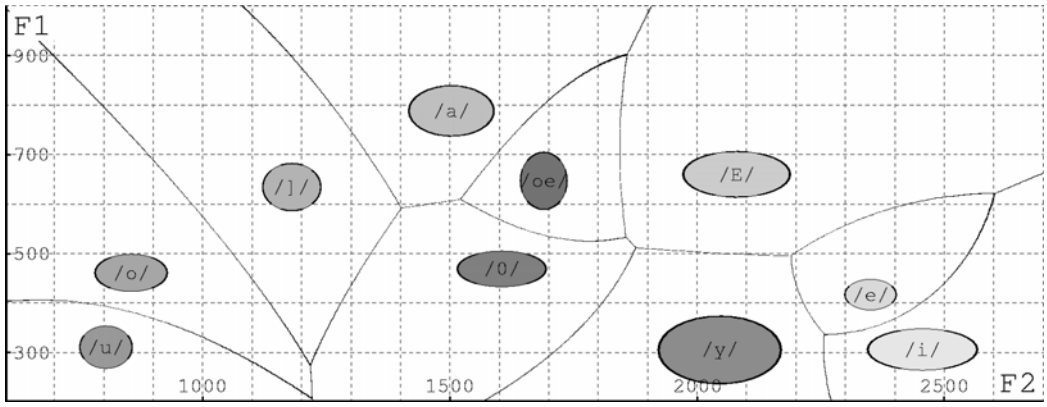

Figure 2: *Diagramme de Voronoï pondéré.*

2.4 Score Phonétique

Après avoir choisi un modèle de partitionnement de l'espace acoustique, nous devons encore expliquer comment un «score phonétique» (c'est-à-dire, une évaluation de la pertinence phonétique) peut être associé à chaque solution de l'inversion. En résumé, chaque vecteur acoustique est rattaché à un «domaine articulatoire idéal» (défini par les valeurs des contraintes du tableau 1), en fonction de la région de l'espace acoustique à laquelle il appartient. À chaque vecteur articulatoire *V* généré par l'inversion (dont l'image par le synthétiseur articulatoire est très proche de ce vecteur acoustique, donc), on peut ainsi associer un «score phonétique», en fonction de la distance de *V* au «domaine idéal». Une façon de faire cela consisterait en le calcul de 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), mais cela n'est pas forcément simple à calculer. Nous préférons calculer un score relatif à chaque type de contrainte (position du dos de la langue, ouverture de la bouche, protrusion des lèvres), qui est plus simple à calculer et plus flexible (on peut facilement choisir de privilégier certaines 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 $\sigma(v,t)$, qui va définir un intervalle de validité $I(v,t) = [\theta(v,t) - \sigma(v,t); \theta(v,t) + \sigma(v,t)]$. Si le calcul de la contrainte de type t pour V (qui est une simple application affine d'une de ses composantes pour $t=D$ et $t=P$, un peu plus complexe pour $t=O$, où interviennent 3 composantes) conduit à une valeur dans l'intervalle $I(v,t)$, alors on lui attribue un score parfait (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, sauf l'étirement des lèvres qui a un poids nul : le modèle de Maeda, qui a été développé à partir d'images aux rayons X de coupes sagittales du conduit vocal, ne peut en effet pas du tout prendre en compte l'étirement des lèvres.

3 Expériences

Nous avons mené des expériences d'inversion sur les données utilisées par Maeda pour concevoir son modèle. Il s'agissait d'un corpus de 10 phrases, d'une durée totale d'environ 20 secondes de cinéradiographies, accompagné d'un enregistrement sonore (assez bruité). Les voyelles cardinales et quelques séquences VV ont été sélectionnées dans le signal de parole, et les fréquences des trois premiers formants ont été extraites manuellement. Un codebook haute précision adapté à la locutrice de Maeda a également été construit. Bien que nous étudions la locutrice ayant permis l'élaboration du modèle articulatoire, nous avons cependant du adapter le modèle de façon à améliorer la fidélité acoustique[4], principalement parce que la calibration géométrique de l'acquisition des rayons X n'est pas connue avec précision.

Il est important de noter que, malgré cette adaptation, le modèle articulatoire et sa simulation acoustique restent incapables de générer les fréquences formantiques mesurées sur le signal sonore, à partir des paramètres articulatoires, eux-mêmes mesurés sur les radiographies. Même avec la meilleure adaptation géométrique, l'erreur moyenne sur F1 reste de 54 Hz. Cette divergence non-négligeable s'explique par l'approximation commise lors du passage de la représentation 2D (correspondant à la coupe sagittale du conduit vocal) à la représentation 3D du conduit (la fonction d'aire). Cette approximation, basée sur la méthode proposée par Heinz et Stevens[7], n'est pas capable de décrire l'aire de la glotte jusqu'au lèvres de façon précise.

Pour ces raisons, malgré la situation apparemment favorable étudiée (le signal de parole à inverser a été prononcé par la locutrice dont les données ont servi à l'élaboration du modèle articulatoire), l'inversion est non triviale et n'a aucune chance de conduire aux trajectoires articulatoires originales.

3.1 Vérification de la cohérence du modèle

Comme les contraintes phonétiques, ainsi que le partitionnement de l'espace acoustique, sont indépendants du locuteur dans notre modèle actuel, nous avons vérifié que le domaine acoustique de chaque voyelle correspondait bien à l'image du domaine idéal des contraintes phonétiques correspondantes : nous avons calculé des images de vecteur articulatoires ayant des scores phonétiques parfaits, et avons pu observer, pour chaque voyelle, que l'image générale du domaine articulatoire englobait bien le domaine acoustique associé. Nous avons également calculé une nouvelle partition de l'espace acoustique en attribuant chaque zone de l'espace acoustique à la voyelle qui avait la densité maximale dans cette zone, c'est-à-dire le plus d'images de vecteurs du domaine articulatoire idéal (chaque voyelle avait le même nombre de points synthétisés, choisis de façon aléatoire dans le domaine idéal). La figure 3 représente le graphe F1/F2 correspondant (F1 en ordonnée décroissante, F2 en abscisse croissante), qui est assez proche de nos modèles acoustiques (sauf pour la voyelle /e/, qui a une très petite zone de dominance).

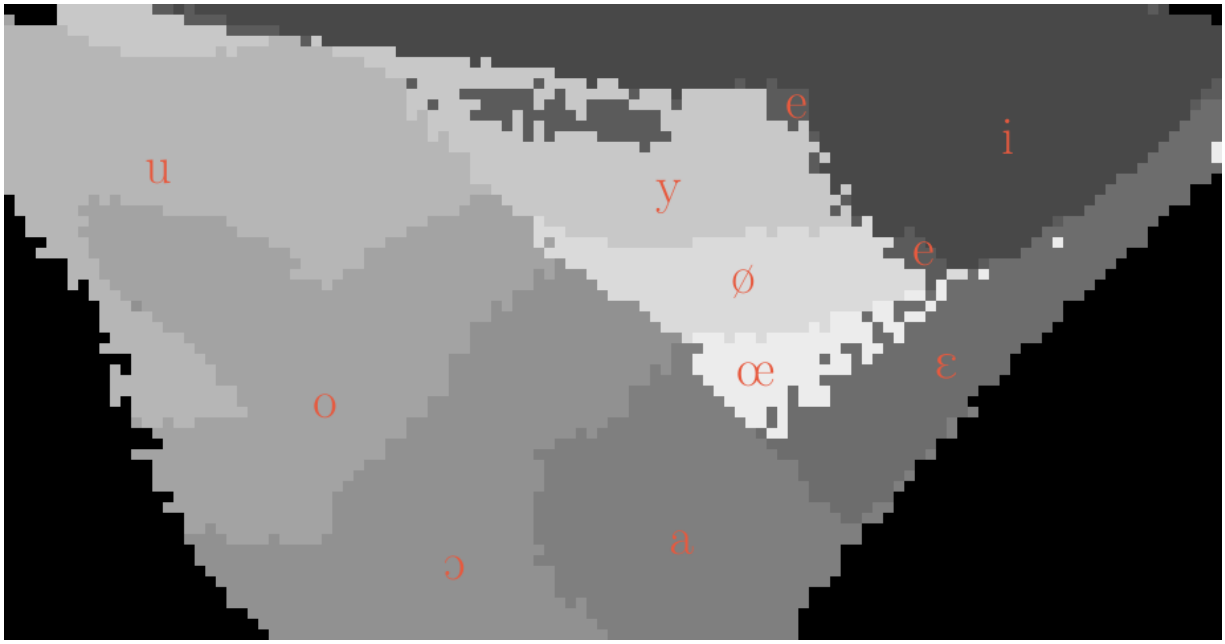

Figure 3: *Diagramme de dominance des images des domaines articulatoires.*

3.2 Inversion de voyelles isolées

Les voyelles /a,i,u,e,o/ ont été inversées en utilisant les contraintes phonétiques. Chaque solution de l'inversion se voit attribuer un score phonétique en fonction de sa distance au «domaine idéal» de la voyelle. La figure 4 représente l'aire (en cm^2) au point de constriction maximale en fonction de sa position (en cm, en partant de la glotte) pour chaque solution de l'inversion trouvée. Le niveau de gris de chaque point est une fonction de son score phonétique ; les points sombres ont un score plus élevé. Bien que les contraintes soient appliquées au niveau des paramètres articulatoires, elle ont un effet global cohérent, i.e. elles favorisent l'émergence de régions bien localisées dans le plan défini par la position et l'aire de la constriction maximale, et affaiblit certains lieux d'articulation secondaires. Les régions retenues sont également plus cohérentes avec les données articulatoires de Wood[8]. On peut également observer que ces contraintes phonétiques pénalisent les formes du conduit vocal ayant une aire importante à la constriction. Cet aspect est important car les propriétés acoustiques des formes du conduit ne sont pas très sensibles à une augmentation uniforme de l'aire du conduit ; de cette façon ce type de formes irréalistes est pénalisé.

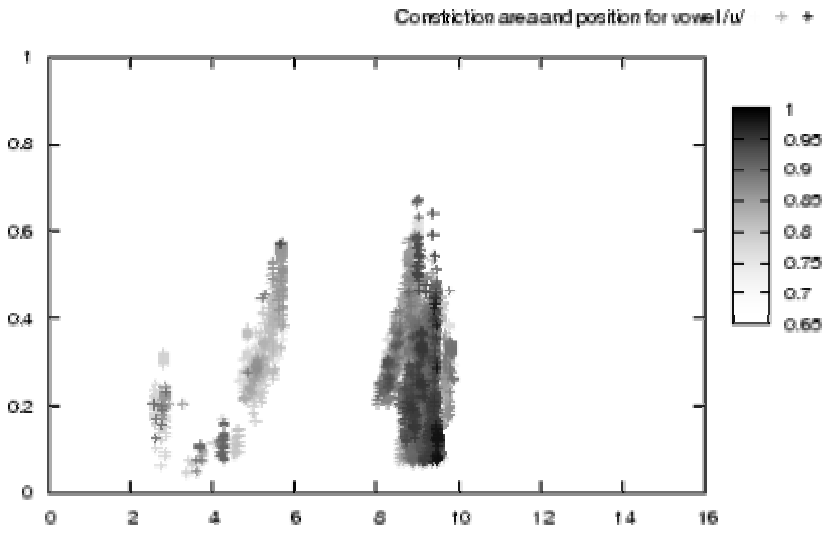

Figure 4: *Score phonétique des solutions de l'inversion pour /u/*

3.3 Inversion de séquences voyelle-voyelle

Nous avons pu extraire plusieurs séquences VV parmi les phrases prononcées par PB: /ui/, /yi/, /ie/.

Comme le signal sonore était bruité, nous avons du extraire les formants de façon manuelle. Après cela, les séquences ont été inversées en utilisant différents types de contraintes. Ici nous présentons les résultats pour la séquence /yi/. Pour chacune des figures, l'unité de temps est la *ms* et les paramètres articulatoires sont donnés en écart-types par rapport à la position neutre.

La Fig. 5 représente les valeurs des trois paramètres principaux (mâchoire, position de la langue, protrusion des lèvres) telles que mesurées sur les images rayons X. Il faut noter que ces trajectoires ont une fréquence d'échantillonnage plus faible (50 Hz) que les trajectoires articulatoires obtenues par inversion.

La Fig. 6 est la trajectoire inverse obtenue en utilisant uniquement des contraintes biodynamiques sur les paramètres articulatoires ; c'est-à-dire que le mouvement global des articulateurs est

minimisé. Bien que la solution inverse ait une très bonne précision acoustique, elle est très différente de la solution observée, et elle n'est pas phonétiquement réaliste. Comme on pouvait s'y attendre, la minimisation du mouvement donne des trajectoires pratiquement droites.

La Fig. 7 est la séquence inversée, en utilisant à la fois les contraintes biodynamiques et phonétiques, à poids égaux. Cette fois, la solution est bien plus réaliste. Le mouvement global des articulateurs correspond aux données de la figure 5 (sauf pour la mâchoire), même si les valeurs absolues des paramètres articulatoires ne sont pas toutes égales aux valeurs originales.

Cette expérience montre que des contraintes très générales, issues de connaissances phonétiques standard, permettent (dans certains cas au moins) d'obtenir des trajectoires articulatoires réalistes. L'impact de ces contraintes phonétique est d'autant plus important que notre méthode d'inversion exploite une description quasi-exhaustive de l'espace articulatoire.

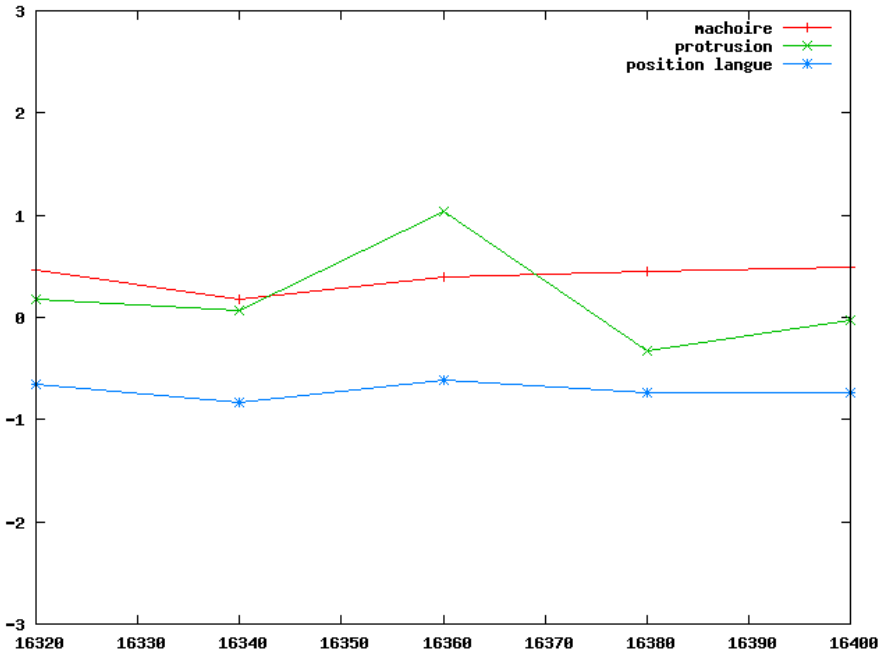

Figure 5: *Paramètres articulatoires mesurés*

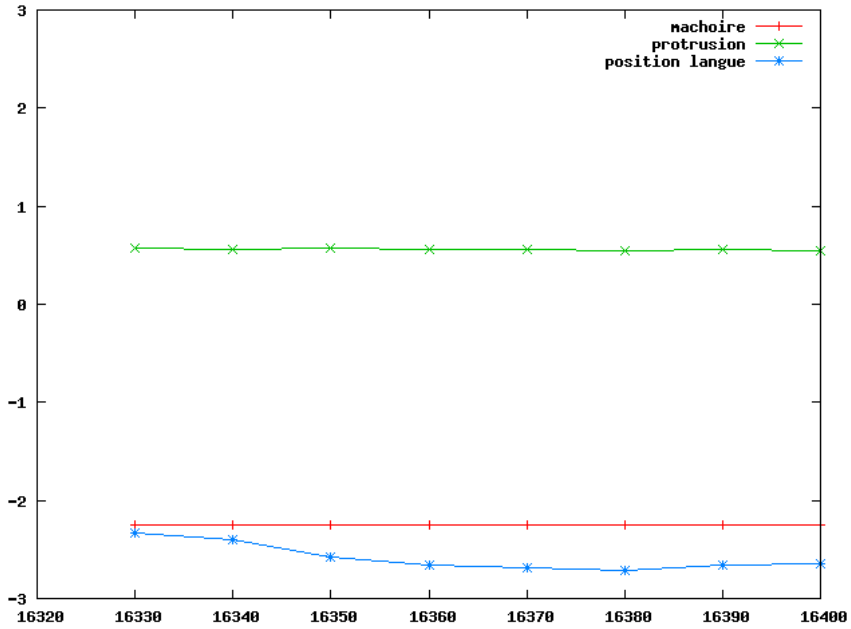

Figure 6: *Inversion avec uniquement des contraintes biodynamiques*

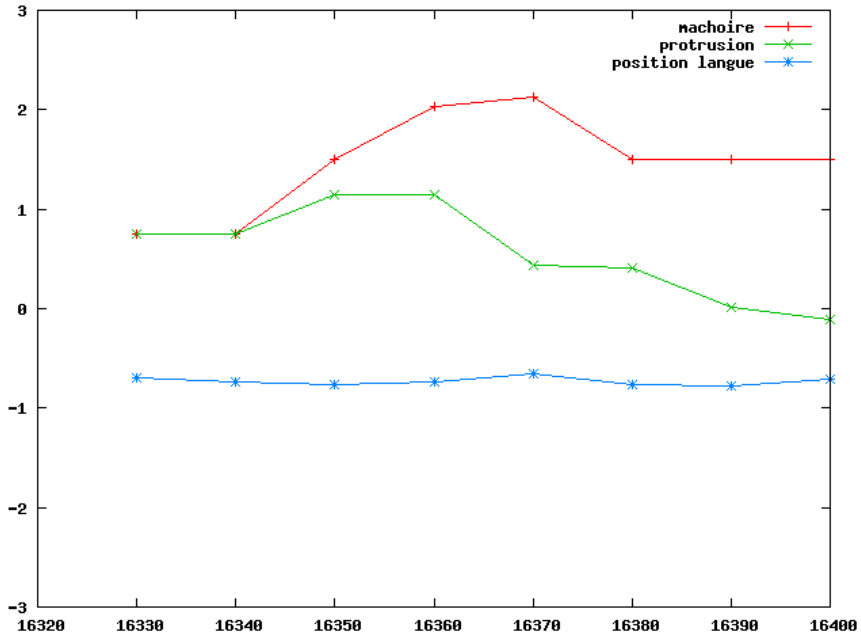

Figure 7: *Inversion avec des contraintes phonétiques et biodynamiques*

4 Conclusion et perspectives

La sous-détermination du problème d'inversion acoustique-articulatoire a donné naissance à de nombreuses pistes de recherche visant à élaborer des contraintes permettant de compenser le manque de données. Cependant, la plupart des contraintes envisagées (voir par exemple [9]) nécessite la connaissance de constantes numériques difficiles à estimer, ou très difficilement applicable à un autre locuteur que celui de l'étude. Ces contraintes phonétiques présentent ainsi deux avantages par rapport à ces contraintes : d'une part elles ne nécessitent pas d'utiliser de nombreux paramètres numériques ; d'autre part, elles sont générales et indépendantes du locuteur et ont été validées (puisqu'elles sont issues de connaissances phonétiques standard). De plus, ces contraintes phonétiques pourraient facilement être couplées à d'autres, en particulier des contraintes dérivées de l'observation du visage du locuteur.

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,” *JASA*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[2] P. Mermelstein, “Articulatory model for the study of speech production,” *JASA*, vol. 53, pp. 1070–1082, 1973.

[3] S. Maeda, “Un modèle articulatoire de la langue avec des composantes linéaires,” in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, Mai 1979, pp. 152–162.

[4] Y. Laprie, S. Ouni, B. Potard, and S. Maeda, “Inversion experiments based on a descriptive articulatory model,” in *6th International Seminar on Speech Production*, Sydney, Autralia, Dec. 2003.

[5] S. Maeda, “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,” in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marschal, Eds.1em plus 0.5em minus 0.4emKluwer Academic Publishers, 1990.

[6] F. Lonchamp, “Les sons du Français — Analyse acoustique descriptive,” Institut de Phonétique, Université de Nancy II,” Cours de Phonétique, 1984.

[7] 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*, 1965, p. A44.

[8] S. Wood, “A radiographic analysis of constriction locations for vowels,” *Journal of Phonetics*, vol. 7, pp. 25–43, 1979.

[9] V. Sorokin, A. Leonov, and A. Trushkin, “Estimation of stability and accuracy of inverse problem solution for the vocal tract,” *Speech Communication*, vol. 30, pp. 55–74, 2000.