Expériences d'inversion basées sur un modèle articulatoire

Blaise Potard, Yves Laprie, Slim Ouni

Équipe 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 : {Blaise.Potard, Yves.Laprie, Slim.Ouni}@loria.fr - http ://www.loria.fr/equipes/parole/

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 base on Maeda's articulatory model. After an overall description of the inversion method the paper presents how inversion can be used to investigate acoustic properties of the articulatory model, which helps us to formulate the way of incorporating effective constraints to obtain phonetically realistic inverse solutions. First, we improved the articulatory codebook construction by achieving a better approximation of the articulatory space boundaries. Then, we adapted geometrical parameters of the articulatory model to guaranty a better acoustical faithfulness with respect to the original acoustical data measured. Finally, we show some inversion results obtained on vowels uttered by the female subject whose images were used to build the articulatory model.

1. INTRODUCTION

La perspective de réaliser des systèmes de codage articulatoire de la parole est à l'origine d'une part importante des travaux consacrés à l'inversion acoustique articulatoire. De ce point de vue le système développé chez AT&T [2] constitue l'un des premiers systèmes complets même s'il souffrait de nombreuses limitations. En particulier, le seul critère d'évaluation consistait à assurer que le spectrogramme de parole synthétisée n'était pas trop éloigné du spectrogramme du signal de parole d'origine. L'utilisation du modèle articulatoire de Mermelstein [4] permettait à la fois de garantir une certaine vraisemblance des formes articulatoires produites et de contraindre l'espace des solutions. En revanche, une validation réellement articulatoire est difficile à réaliser compte tenu de la petite quantité de données disponibles. En effet, même s'il existe un grand nombre de bases de données articulatoires, il est rare d'en trouver une disposant de suffisamment de données d'une qualité suffisante, soit que les images ne couvrent pas tout le conduit vocal, soit que la qualité des images ou du signal sonore soit insuffisante pour évaluer l'inversion acoustique articulatoire.

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 qu'elles soient quantitatives ou qualitatives sous la forme de connaissances phonétiques. Nos travaux reposent sur le modèle de Maeda[3] et c'est pour cette raison que nous avons récemment commencé à utiliser les données articulatoires qui ont servi à construire ce modèle. Ce papier décrit l'adaptation du modèle articulatoire de manière à améliorer sa fidélité acoustique par rapport au signal enregistré avec les images, l'amélioration de la construction de la table articulatoire de façon à la rendre plus compacte et plus précise aux frontières de l'espace articulatoire que peut réaliser la locutrice et la récupération des lieux d'articulation des voyelles qu'elle a produites.

En testant la méthode d'inversion sur les données qui ont servi à la construction du modèle articulatoire utilisé pour l'inversion nous nous sommes placés a priori dans des conditions très favorables. En fait, les images sont celles de la coupe sagittale du conduit vocal et la troisième dimension est récupérée à l'aide d'une heuristique relativement imprécise. Par ailleurs, les images d'origine ne couvrent que très partiellement la région du larynx. Enfin, la simulation acoustique fait appel à un plusieurs constantes physiques qui ne sont pas connues avec précision. Par conséquent, même dans cette situation très favorable, il n'est pas possible de re-synthétiser fidèlement la parole produite par la locutrice (cf. § 3.1). Le travail de validation porte donc sur l'influence des erreurs citées au-dessus sur la pertinence articulatoire des résultats de l'inversion.

2. DESCRIPTION DE NOTRE MÉTHODE D'INVERSION

Notre méthode d'inversion comporte trois étapes. La première étape consiste à générer un grand nombre de solutions potentielles : pour cela, nous utilisons 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, le triplet des fréquences des 3 premiers formants). Un vecteur acoustique étant donné, il existe *a priori* une infinité de vecteurs articulatoires permettant de l'obtenir, nous n'avons donc pas l'ambition de générer toutes les solutions inverses possibles. Cependant, il est nécessaire, pour avoir une inversion de qualité, que les échantillons retenus soient suffisamment représentatifs pour contenir des solutions proches de la solution réelle.

La deuxième étape de notre méthode consiste en la reconstruction d'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 en l'amélioration de la fidélité acoustique et de la régularité articulatoire de la solution obtenue à l'étape précédente en utilisant un algorithme de régularisation variationnelle.

2.1. Construction du codebook articulatoire

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 récursivement en évaluant à chaque étape la linéarité locale de la relation articulatoire acoustique. Si la relation n'est pas suffisamment linéaire, on subdivise l'espace.

Plus précisément, les paramètres articulatoires du modèle de Maeda variant entre -3σ et $+3\sigma$, où σ est l'écart type, l'espace articulatoire peut-être vu comme un hypercube à 7 dimensions (de rayon 6σ). L'échantillonnage de l'espace articulatoire se fait en cherchant des points qui délimitent les zones linéaires ; un hypercube étant donné, on évalue sa linéarité en considérant tous les segments reliant deux sommets de l'hypercube : les valeurs des vecteurs acoustiques obtenues par synthèse au niveau des sommets sont linéairement interpolées au milieu du segment, et le vecteur correspondant est comparé à la valeur du vecteur obtenu par synthèse au milieu du segment. Si la différence entre les deux est inférieure à un certain seuil prédéfini, alors la relation est considérée comme linéaire pour le segment. Si la relation est linéaire pour tous les segments, alors l'hypercube est considéré linéaire, et on sauvegarde l'hypercube dans le codebook. Sinon, on subdivise l'hypercube, et on applique récursivement les tests de linéarité dans tous les sous-hypercubes. Comme nous ne testons que la linéarité des segments reliant deux sommets, nous n'avons aucune garantie sur le comportement de la relation à l'intérieur de l'hypercube ; mais la simplicité du test est dictée par le temps de calcul important que prend la synthèse de l'image d'un vecteur articulatoire. Expérimentalement, nous avons montré[5] qu'en utilisant un seuil de 0.3 bark par formant lors du test, l'erreur moyenne sur les formants pour les points générés par la méthode d'inversion était inférieure à 10Hz.

En pratique, pour des raisons d'explosion combinatoire, on est obligé de limiter le niveau de subdivisions (en dimension 7, faire une subdivision oblige à explorer 2^7 nouveaux hypercubes, donc augmenter d'un niveau la subdivision multiplie *a priori* le temps de calcul par 128). On arrête de subdiviser quand le test de linéarité est satisfait, ou quand la taille d'un côté de l'hypercube passe au-dessous d'un certain seuil. Dans les deux cas, on sauve les cubes dans le codebook. Par ailleurs, on subdivise également si un ou plusieurs points d'un cube ne donnent pas une fonction d'aire de voyelle réaliste, c'est-à-dire si l'aire à la constriction est nulle ou trop faible. Ces cubes sont situés à la frontière de l'espace articulatoire ; comme pour le test de linéarité, on arrête la subdivision quand la taille du cube est inférieure à un certain seuil (différent du précédent), et dans ce cas on rejette le cube.

2.2. Amélioration de la construction

La principale difficulté, dans notre cas, est de couvrir l'espace articulatoire des voyelles de façon aussi complète que possible, en supprimant les parties non pertinentes, tout en limitant au maximum le temps de calcul nécessaire (le test de linéarité prend un temps de l'ordre d'une seconde sur une machine récente). Le principal problème de la version originale était lié au seuil de subdivision pour les cubes à la frontière de l'espace articulatoire : dans un premier temps, on avait essayé de mettre une valeur petite pour ce seuil, mais cela conduisait malheureusement à un nombre trop important de petits cubes, dans des zones qui semblent *a priori* moins utiles et plus difficiles à représenter (la relation articulatoire acoustique est particulièrement chaotique le long de la frontière). Pour que le temps de calcul reste raisonnable, on avait alors limité la subdivision à 2 niveaux. Malheureusement, des zones importantes de l'espace articulatoire étaient totalement supprimées, même si elles n'avaient qu'un point invalide, et en particulier l'inversion du [u] pour une locutrice féminine ne donnait plus aucune solution. En augmentant la subdivision d'un niveau, la précision est meilleure sans être vraiment satisfaisante, et les temps de calcul bien plus importants. Nous avons alors décidé d'arrêter la subdivision des cubes qui ont des sommets invalides seulement si le nombre de points invalides dans l'hypercube dépasse un certain seuil, (qui peut dépendre du niveau de subdivision), ou si le jacobien n'est pas calculable au centre.

Un autre problème préoccupant était l'espace disque occupé par le codebook. Une modification de la façon de sauvegarder les hypercubes nous a permis, pour une même précision, de diviser par 20 la place occupée par le codebook.

Nous avons également commencé à intégrer des contraintes phonétiques supplémentaires. Nous avions remarqué[7] que nous trouvions un grand nombre de solutions avec des aires à la constriction beaucoup trop importantes par rapport à celles observées par Wood[9]. Désormais, on supprime les hypercubes contenant des solutions avec une aire à la constriction trop importante.

Dans le codebook actuel, après suppression des hypercubes indésirables (constriction trop étroite ou trop importante), 29% de l'espace articulatoire potentiel, c'est-à-dire l'hypercube de dimension 7 $[-3\sigma, 3\sigma]$ est conservé. Dans la version précédente, seul 9.8% de l'espace était conservé.

2.3. Exploration de l'espace nul de la relation articulatoire acoustique

Pour chaque vecteur acoustique représentée par les trois premières fréquences formantiques, le processus d'inversion consiste en la recherche de tous les hypercubes qui peuvent générer le triplet de formants observé. Il faut ensuite trouver un ensemble de solutions dans chacun de ces cubes. Comme l'inversion consiste à trouver 7 paramètres à partir de 3, l'espace des solutions a *a priori* 4 degrés de liberté. La relation articulatoire acoustique (notée R) est supposée être localement linéaire au niveau du centre P_0 de l'hypercube (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 4 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 dimen-

sions, ce que l'on ne sait pas faire de manière formelle. Une première approximation de l'intersection est obtenue par programmation linéaire. Puis l'espace nul est échantillonné, et l'appartenance à l'intersection de chacun des points est testée[6].

3. EXPÉRIENCES AUTOUR DE L'INVERSION

3.1. *Améliorer la fidélité acoustique du modèle de Maeda*

Le modèle articulatoire de Maeda a été construit en appliquant des méthodes d'analyse statistique à des cinéradiographies d'une locutrice prononçant 10 phrases courtes. Comme le signal acoustique a été enregistrée pendant les prises d'images, ces données sont très intéressantes pour vérifier la pertinence de la méthode d'inversion. Cependant, bien que les formants synthétisés aient des trajectoires assez similaires à celles extraites du signal de parole, on observe un net écart entre les deux. Il se trouve que la précision du modèle géométrique dépendait de deux facteurs d'échelle qui avaient été choisis arbitrairement, car les réglages de la machine à rayon X n'étaient pas connus avec précision. L'ajustement de ces facteurs d'échelle n'était pas possible en 1979 quand le modèle a été construit, car cela aurait demandé un temps de calcul trop important. Nous avons réalisé l'ajustement de ces facteurs en échantillonnant un domaine de valeurs raisonnables pour ces deux facteurs, en comparant les formants synthétisés en utilisant les nouvelles valeurs avec les formants du signal d'origine, et en prenant le couple de valeurs qui donne le meilleur[1] résultat. Il apparaît alors que le facteur ad-hoc d'augmentation de l'aire, fixé à 40% dans le modèle original, peut être supprimé, si on prend un facteur d'échelle (`TEKvt`) pour les radiographies de 196 (au lieu de 187). L'erreur moyenne pour le premier formant passe alors de 114Hz à 54Hz, ce qui est encore important. Cette erreur est essentiellement due à l'absence du larynx sur les radiographies, qui fait que le paramètre correspondant dans le modèle n'est pas connu et donc fixé arbitrairement à 0.

3.2. *Récupération du lieu d'articulation de voyelles*

L'utilisation de données réelles - des images au rayons X ou par IRM par exemple - est la meilleure façon d'étudier les lieux d'articulation des voyelles. Mais cette approche a plusieurs défauts, liés soit à la dangerosité des techniques d'imagerie (dans le cas des rayons X), soit aux conditions anormales d'enregistrement (voir en particulier l'étude de Engwall[1] sur les effets de la position allongée dans le cas des images IRM). Ces techniques présentent de plus un coût prohibitif qui limite considérablement la quantité de données que l'on peut espérer enregistrer. Wood[9], pour son travail, considéré pourtant comme l'une des études les plus complètes des lieux d'articulation des voyelles en utilisant des données réelles, ne disposait que d'une quantité de données assez limitée qui ne permettait pas de couvrir l'ensemble des configurations possibles.

Une solution pour étudier les lieux d'articulation des voyelles est de reconstituer l'information articulatoire directement à partir du signal de parole. Cela présente l'avantage de pouvoir analyser simplement un nombre importants de voyelles différentes. Notre méthode garantit que toutes les configurations articulatoires possibles (c'est-à-dire les solutions inverses pour une voyelle donnée, représentée par son triplet de formants) sont trouvées. Nous avons ainsi utilisé notre méthode pour étudier les lieux d'articulation pour des voyelles du Français prononcées par un locuteur[7], et la locutrice ayant servi à construire le modèle de Maeda[3]. Ces expériences nous permettent également d'étudier le comportement acoustique de ce modèle articulatoire.

Nous présentons ici résultats pour la locutrice pour les voyelles /u/ et /a/.

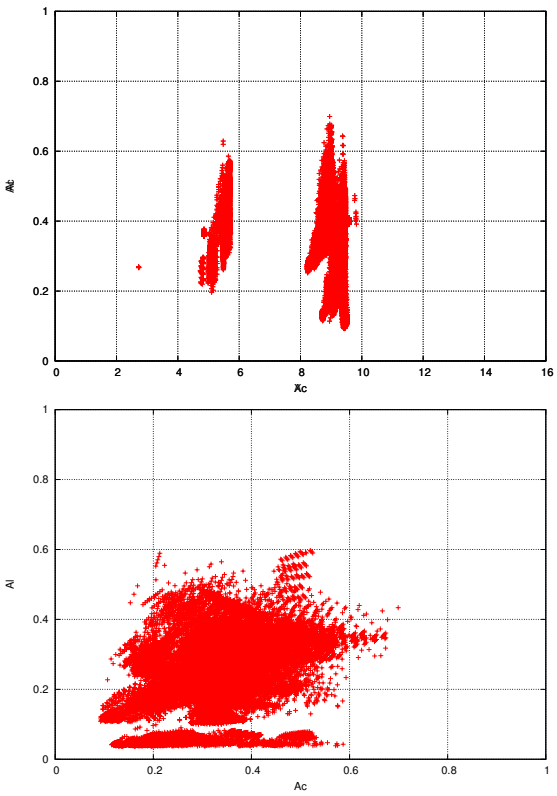

FIG. 1: Solutions inverses pour la voyelle [u] dans les plans Xc/Ac et Ac/Al, où Xc est la position de la constriction principale mesurée comme la distance à la glotte le long du conduit, en cm, Ac l'aire à la constriction en cm^2 et Al l'aire aux lèvres en cm^2.

La Fig. 1 présente le résultats de l'inversion de la voyelle [u] dans deux représentations différentes : l'aire aux lèvres (Al) en fonction de l'aire à la constriction (Ac), et l'aire à la constriction en fonction de la position de la constriction (Xc). Xc varie entre 0 (au niveau de la glotte) et 16 cm (au niveau des lèvres). Chaque croix représente une solution particulière. Deux lieux d'articulation bien définis peuvent être observés pour le [u] : une vers 9 cm (constriction palatale), et une vers 5 cm (constriction vélaire). On remarque également quelques points indiquant une constriction au niveau du pharynx. En Français, seul le [u] vélaire est normalement observé, mais le [u] palatal est observé dans d'autres langages. Le [u] pharyngal n'a

[1] Nous utilisons comme critère de discrimination la somme des différences (en bark) pour les trois premiers formants.

jamais été observé naturellement, mais il a par contre été observé comme stratégie compensatoire dans une expérience de l'ICP où le sujet devait parler avec un tube entre les lèvres[8].

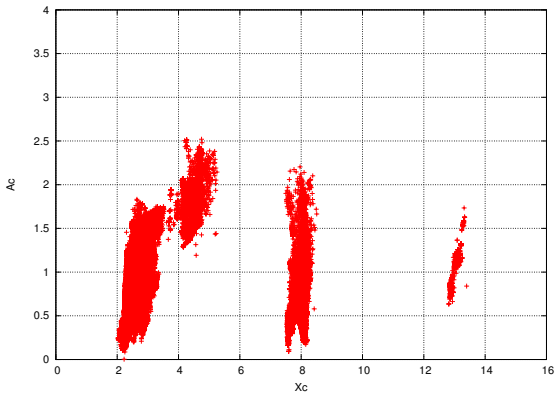

FIG. 2: Solutions inverses pour la voyelle [a] dans le plan Xc/Ac.

La figure 2 présente le résultats de l'inversion de la voyelle [a] dans le plan Xc/Ac. On constate que pour cette voyelle, les lieux d'articulations possibles sont beaucoup plus étalées. Le nombre de solutions trouvées est également nettement plus important : pour les même réglages, on trouve un peu moins de 4000 solutions pour le [u] et plus de 100000 pour le [a].

4. CONCLUSION ET PERSPECTIVES

La force de notre méthode d'inversion est de garantir, à l'aide de la décomposition en hypercubes quasi linéaires couplée à l'exploration de l'espace nul de la relation articulatoire acoustique, que pratiquement toutes les solutions possibles peuvent être parcourues, si l'on utilise un quadrillage de l'espace articulatoire suffisamment fin.

Les expériences d'inversion montrent qu'un grand nombre de trajectoires articulatoires peuvent être obtenues. Certaines d'entres elles sont dues à des mouvements compensatoires qui pourraient réellement être observés chez des locuteurs humains, mais d'autres sont dues au manque de contrôle du modèle articulatoire, qui accepte des formes de conduits irréalistes. Notre objectif est de contrôler de façon précise l'intégration de contraintes pour étudier leurs mérites. Notre méthode d'inversion permet cela, car elle est très neutre, contrairement à d'autres qui favorisent certaines solutions au détriment d'autres qui pourraient être réalistes.

Nous travaillons également à l'incorporation de contraintes statiques et dynamiques au sein du processus d'inversion. Des contraintes statiques peuvent être construites à partir de connaissances phonétiques sur certains sons du langage, pour pénaliser les entrées du codebook qui ne satisfont pas certaines caractéristiques attendues pour un certain triplet de formants : par exemple, pour la voyelle [y], on peut pénaliser les solutions pour lesquelles la protrusion des lèvres est faible. Nous avons commencé par pénaliser les solutions dont l'aire à la constriction est trop importante (par rapport aux aires relevées par Wood[9]). Ce type de contraintes peut aussi être intégré dynamiquement, mais l'intégration statique permet d'accélérer l'inversion en supprimant ou en pénalisant un grand nombre de solutions ; il faut cependant s'assurer de ne pas supprimer de solutions importantes. Comme contraintes dynamiques, nous étudions en ce moment l'utilisation de contraintes sur les paramètres visuels (lèvres, position de la mâchoire), acquis automatiquement par stéréovision.

Remerciements : Nous remercions Shinji Maeda et Jean Schoentgen pour les discussions fructueuses.

RÉFÉRENCES

[1] O. Engwall. Are statical mri data representative of dynamic speech ? results from a comparative study using mri, ema and epg. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pages 17–20, Beijing, Chine, Octobre, 2000.

[2] 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.

[3] 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.

[4] P. Mermelstein. Articulatory model for the study of speech production. *J. Acoust. Soc. Am.*, 53 :1070–1082, 1973.

[5] S. Ouni and Y. Laprie. Improving acoustic-to-articulatory inversion by using hypercube codebooks. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, Chine*, volume II, pages 178–181, October 2000.

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

[7] S. Ouni and Y. Laprie. A study of the french vowels through the main constriction of the vocal tract using an acoustic-to-articulatory inversion method. In *15th International Congress of Phonetic Sciences 2003 - ICPhS'2003, Barcelone, Espagne*, Aug 2003.

[8] P. Savariaux, C. Perrier and J.-P. Orliaguet. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube : A study of the control space in speech production. *Journal of the Acoustical Society of America*, 98 :2428–2442, 1995.

[9] S. Wood. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7 :25–43, 1979.