Suivi de contours d'articulateurs orofaciaux à partir d'IRM dynamique

Mathieu Labrunie, Pierre Badin, Laurent Lamalle, Coriandre Emmanuel Vilain, Louis-Jean Boë, Jens Frahm, Peter Birkholz

Suivi de contours d'articulateurs orofaciaux à partir d'IRM dynamique

Mathieu Labrunie [1, 2] Pierre Badin [1, 2] Laurent Lamalle [3] Coriandre Vilain [1, 2]
Louis-Jean Boë [1, 2] Jens Frahm [4] Peter Birkholz [5]
(1) Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
(2) CNRS, GIPSA-Lab, F-38000 Grenoble, France
(3) Inserm US 17 — CNRS UMS 3552 — Univ. Grenoble Alpes & CHU de Grenoble UMS IRMaGE, France
(4) Biomedizinische NMR Forschungs GmbH am Max-Planck-Institut für biophysikalische Chemie, Göttingen, Germany
(5) Institute of Acoustics and Speech Communication, TU Dresden, Germany
(mathieu.labrunie, pierre.badin)@gipsa-lab.grenoble-inp.fr

RESUME

Nous présentons une méthode de prédiction de contours médiosagittaux des organes orofaciaux de la parole et la déglutition à partir d'images IRM dynamiques. Pour chaque locuteur, un ensemble de 60 images représentatives pour lesquelles les contours ont été tracés manuellement permet d'entraîner des modèles ACP d'images et de contours articulatoires, ainsi qu'un modèle multilinéaire qui prédit les paramètres des contours à partir des paramètres des images. Les contours obtenus sont ensuite corrigés par des modèles de forme actifs (ASM) modifiés utilisant les informations locales de profils d'intensité de pixels le long des normales aux contours. Les performances de cette méthode (erreurs moyennes « points à contour » entre 0,57 et 0,70 mm) sont insensibles au type de séquence IRM (écho de gradient avec échantillonnage synchronisé ou écho de gradient radial hautement sous-échantillonné), sont meilleures que celles de la littérature, et rendent possible le traitement de volumineux corpus d'images IRM dynamiques.

ABSTRACT

Orofacial articulators tracking from dynamic MRI.

We introduce a method for predicting midsagittal contours of orofacial organs during speech and swallowing from dynamic MRI images. For each speaker, a set of 60 representative images for which contours have been manually traced allows for training PCA images. The data serve to derive articulatory contour PCA models and a multilinear model that predicts contour parameters from image parameters. The obtained contours are then corrected by a modified Active Shape Model (ASM) using the local information of the pixel intensity profiles along the contour normals. The performance of this method (mean "points to contour" errors between 0.57 and 0.70 mm) is insensitive to the type of MRI sequence (conventional gradient echoes with synchronized sampling or highly undersampled radial gradient echoes), better than those in the literature, and make it possible to process large corpora of dynamic MRI images.

MOTS-CLES : IRM dynamique ; articulateurs orofaciaux de la parole ; suivi automatique de contours ; régression linéaire multiple ; modèles de forme actifs.

KEYWORDS: Dynamic MRI; speech orofacial articulators; automatic contour tracking; multiple linear regression; Active Shape Models.

1 Introduction

Les progrès considérables accomplis en IRM dynamique temps réel dans la dernière décennie (*cf.* Uecker *et al.* (2010) ou Niebergall *et al.* (2013)) ont rendu cette technique d'imagerie médicale extrêmement intéressante pour l'étude des mouvements des articulateurs orofaciaux dans les tâches de parole (Silva & Teixeira (2015)) ou de déglutition (Olthoff *et al.* (2014)), en permettant l'acquisition de volumineux corpus d'images médiosagittales à 30-60 images par seconde avec une résolution de l'ordre de 1,5 mm/pixel. Afin de pouvoir caractériser et modéliser ces données, il est donc nécessaire de développer des méthodes automatiques de suivi des contours des articulateurs à partir de ces images fournissant des résultats aussi précis et fiables que les méthodes (semi-) manuelles traditionnelles (voir p. ex. Serrurier & Badin (2008)).

Cet article décrit notre approche pour développer une telle méthode pour tous les organes articulatoires orofaciaux impliqués dans la production de la parole et la déglutition. Deux types de structures ont été considérés : les structures osseuses rigides, et les organes déformables. Les structures rigides du crâne doivent être suivies pour contrôler les mouvements de tête involontaires des sujets. Les autres structures rigides intéressantes sont la mâchoire et l'os hyoïde qui ont des mouvements spécifiques aussi bien en parole qu'en déglutition. Les organes déformables sont les lèvres supérieure et inférieure, la langue, l'épiglotte, le voile du palais, et l'ensemble de la paroi naso- et oro-pharyngée postérieure.

2 Travaux précédents

Avant de décrire notre approche, nous présentons une revue de la littérature sur le suivi d'articulateurs à partir d'IRM basée sur le travail très détaillé de Silva & Teixeira (2015).

Bresch & Narayanan (2009) proposent une méthode d'ajustement de contours dans l'espace de Fourier des images. Un modèle de contours est constitué des trois principales régions des organes articulatoires (au-dessus du palais dur, au-dessous de la langue et en arrière de la paroi pharyngée) délimitées par des frontières polygonales. Une descente de gradient minimise la distance entre l'image et les contours dans leurs espaces de Fourier. Bien qu'aucune évaluation quantitative ne soit donnée, les exemples de contours déterminés attestent d'une qualité raisonnable de la méthode.

Proctor *et al.* (2010) identifient les points de la ligne centrale du conduit vocal sur les images IRM en cherchant le chemin optimal entre les différentes positions de minima d'intensité de pixels le long de lignes d'un système de grilles perpendiculaires au conduit positionnées manuellement. Les frontières des tissus sont ensuite déterminées comme les positions de plus fort gradient d'intensité des pixels de chaque côté de la ligne centrale. Cette approche non supervisée fournit des contours de conduit indifférenciés, avec une erreur RMS de reconstruction de la distance sagittale variant de 0,82 à 1,61 mm. Kim *et al.* (2014) ont tenté d'améliorer cette méthode en optimisant la qualité d'image à l'aide d'une carte de correction de sensibilité de pixels et d'une réduction de bruit de grain par traitement d'image local, et un lissage des contours estimés. Ils obtiennent une erreur RMS sur la distance sagittale entre 2,13 et 2,79 mm avec des images de résolution de 3 mm/px.

Eryildirim & Berger (2011) ont développé un algorithme de segmentation de langue à partir d'images IRM statiques semblable aux algorithmes de type modèles de forme actifs (Active Shape

Models, ASM). Cet algorithme est basé sur un modèle de forme construit à partir des contours édités manuellement pour 38 images par analyse en composantes principales (ACP). La détection des points terminaux des contours de langue est améliorée par l'utilisation d'une méthode de recalage non rigide. Ils obtiennent une erreur moyenne de reconstruction des distances entre contours de 1,6 mm avec des images de résolution de 0,625 mm/px.

Finalement, Silva & Teixeira (2015) ont récemment proposé un modèle actif d'apparence modifié (Active Appearance Model, AAM) pour le suivi de contours articulatoires à partir d'images IRM dynamiques. Ils utilisent deux modèles AAM, l'un construit à partir des contours édités manuellement sur 30 images d'articulations non-nasales, et l'autre à partir de 21 articulations nasales. Ils trouvent que leur approche est plus rapide et converge mieux que les AAM traditionnels. Notons que chaque articulateur est clairement identifié à la fois lors de la segmentation manuelle et dans le modèle de forme (lèvres, corps et pointe de la langue, voile du palais, palais dur et pharynx). Ils mesurent les erreurs en termes du coefficient de similarité de Dice qui reflète la différence du nombre de pixels de part et d'autre des contours ; cette erreur n'est pas directement comparable à une distance RMS entre contours.

Nous introduisons dans cet article une méthode proche des ASM initialisée à l'aide d'une procédure de prédiction des contours à partir des intensités des pixels de l'image par modèle linéaire multiple. Nous décrivons l'implémentation de cette méthode et présentons des résultats d'évaluation sur des corpus d'images IRM de qualités différentes, ainsi que pour des méthodes basées sur le recalage d'images.

3 Segmentation basée sur le recalage d'image

Le recalage d'image consiste à déterminer la transformation spatiale – rigide ou élastique – qui permet de mettre en correspondance une image source avec une image cible. L'algorithme de recalage détermine la transformation optimale minimisant la distance, selon la mesure de similarité ou de dissimilarité choisie, entre les intensités des pixels de l'image source transformée et celles des pixels de l'image cible. Appliquer cette transformation de recalage à des contours de référence tracés sur l'image source permet ensuite de prédire les contours de l'image cible.

3.1 Recalage des structures rigides par comparaison à un motif standard

Pendant l'acquisition des données, la tête du sujet est stabilisée au mieux à l'aide de coussins en mousse, mais il est impossible d'empêcher complètement les mouvements parallèles au plan médiosagittal. Il est donc nécessaire de suivre les mouvements du crâne afin de les compenser. Par ailleurs, les mouvements d'autres structures rigides importantes pour l'articulation – la mâchoire et l'os hyoïde – devront être déterminés. La rigidité de ces structures n'autorise que des mouvements de translation et de rotation dans le plan médiosagittal. Une méthode de segmentation adaptée à ces propriétés rigides de l'objet d'intérêt est la comparaison à des motifs standards (*template matching*). La première étape de cette méthode consiste à choisir un motif contenant l'objet d'intérêt (p. ex. le palais) dont le contour est connu sur une image de référence. Ce motif est délimité par un masque excluant les tissus voisins non caractéristiques ou variables. Pour une image à traiter, l'objectif est ensuite de déterminer les paramètres de rotation et translation 2D pour lesquels le motif source de l'image de référence se superpose au motif cible correspondant sur l'image à traiter.

3.2 Recalage des organes déformables par démons

Pour les organes déformables tels la langue, les lèvres, ou le voile du palais il faut calculer un champ de transformation potentiellement non-linéaire pour pouvoir transformer l'image source en image cible. Nous avons testé le recalage par démons de Kroon & Slump (2009). Le champ de transformation associé à cette méthode est influencé par deux forces : une force interne dirigée par le gradient en chaque point de l'image, et une force externe dirigée par la différence entre intensités correspondantes de l'image source transformée et de l'image cible. Un facteur α permet de pondérer ces deux forces. Dans la procédure itérative de minimisation de la distance entre les intensités des images source et cible, il est possible de régulariser certains champs de déformation. A chaque itération, un champ de mise à jour du champ de transformation rajoute des déplacements au champ de transformation préalablement obtenu. Nous avons employé une régularisation fluide sur ce champ de mise à jour (filtrage gaussien d'écart-type σ_{fluide}) et un effet de diffusion en réalisant un filtrage gaussien du champ de transformation (écart-type : σ_{diff}). Les paramètres optimaux que nous avons trouvés pour l'ensemble de nos corpus sont $\alpha = 12$, $\sigma_{diff} = 1$, et $\sigma_{fluide} = 4$.

4 Segmentation basée sur des méthodes d'apprentissage

Dans la méthode précédente, seules les informations d'une image et d'un contour de référence sont prises en compte pour obtenir le contour associé à une nouvelle image. Les méthodes de recalage d'image, rigide ou élastique, ne prennent en compte que les propriétés de l'image, mais ignorent les propriétés des contours recherchés. Disposer de tracés manuels experts des contours d'intérêt sur un corpus représentatif de l'ensemble des données offre la possibilité d'introduire une information pertinente sur les contours recherchés qui permet d'améliorer considérablement les résultats en contraignant l'espace de recherche des contours. Nous décrivons ci-dessous trois méthodes utilisant l'entraînement de modèles à partir d'une base d'images et de contours associés, après avoir indiqué comment nous sélectionnons les images du corpus d'apprentissage.

4.1 Sélection du corpus d'apprentissage

Pour construire un modèle suffisamment général pour représenter toutes les articulations d'intérêt, le corpus d'apprentissage doit couvrir aussi exhaustivement que possible la diversité des articulations que peut produire le locuteur, tout en minimisant le nombre d'images dont les contours devront être édités manuellement. Pour construire cet ensemble, nous avons supposé que la distance euclidienne entre l'intensité des pixels des images était une métrique corrélée avec la distance euclidienne entre contours associés (a posteriori nous avons trouvé des corrélations supérieures à 0,85, ce qui valide cette hypothèse). Nous avons donc réparti toutes les images en n_{cl} classes par classification ascendante hiérarchique en utilisant cette métrique : différents tests ont montré que $n_{cl} = 60$ constitue un bon compromis entre le nombre d'images à tracer et un niveau d'erreur de l'ordre du millimètre, et qu'en outre cette métrique produit un dendrogramme cohérent au sens du coefficient de corrélation cophénétique. Le représentant de chaque classe est ensuite choisi comme l'élément de la classe le plus éloigné des éléments des autres classes, de façon à assurer que la périphérie de l'espace soit également bien représentée.

4.2 Régression linéaire multiple (Multiple Linear Regression, MLR)

Le modèle de prédiction des contours en fonction des images le plus simple est celui qui prédit chacune des coordonnées de chacun des contours comme combinaison linéaire des intensités des

pixels de la zone d'intérêt. Deux types de zones ont été utilisés : une zone cadrée globalement (*cf.* le cadre jaune en Fig. 1a), ou des zones cadrées sur chaque articulateur (*cf.* les autres cadres de Fig. 1a). Nous avons réduit la dimensionnalité de l'espace des intensités des zones par ACP, en retenant n_{int_gbl} composantes pour le cadrage global, et $n_{int\_orgs}(1{:}n_{org})$ pour les cadrages des n_{org} organes. De même, les coordonnées des contours des organes sont modélisées soit par un seul ensemble de n_{cnt_gbl} composantes, soit séparément par $n_{cnt\_orgs}(1{:}n_{org})$ composantes pour chaque organe. Les nombres de composantes sont choisis pour minimiser les erreurs de prédiction pour chaque méthode. La deuxième méthode apporte une flexibilité supplémentaire qui permet de mieux approcher chaque contour d'organe ; les composantes des organes peuvent alors être partiellement corrélées entre organes, comme par exemple pour la lèvre inférieure et la mâchoire. Le modèle d'association entre les contours et les images s'obtient finalement par régression linéaire multiple des prédicteurs des contours en fonction des prédicteurs des intensités sur l'ensemble des n_{cl} données d'apprentissage, soit de manière globale, soit organe par organe.

4.3 Modèles de forme actifs (Active Shape Models, ASM)

La méthode générale des ASM (Cootes *et al.* (1995)) vise à ajuster les points d'un contour aux limites d'un objet dans une image en les déplaçant de manière itérative afin de minimiser la distance entre l'apparence mesurée au voisinage de ces points (un profil d'intensité par exemple) et celle prédite par un *modèle d'apparence*, tout en contraignant le contour par un modèle (*modèle de forme*). Ces modèles sont établis lors d'une phase d'apprentissage à partir des images et des contours tracés. Dans notre implémentation[1], que nous appellerons ASM modifié (ASMM) nous utilisons les modèles de forme décrits en 4.2. Des modèles d'apparence sont développés pour chaque point de chaque organe, pour trois niveaux d'échantillonnage des images (échelles de 2, 1, et 0.5), de la manière suivante. En chaque point du contour considéré, un profil d'intensité est échantillonné par interpolation sur $n_{pfl} = 13$ points distribués le long d'un segment normal centré sur le point par pas d'un pixel (voir Fig. 1b, haut). Au lieu de modéliser l'apparence de ces profils d'intensité par ACP comme dans les ASM traditionnels, nous associons ces profils à des classes. Deux classes principales sont déterminées. La classe « non-contact » regroupe les profils pour lesquels la distance du point du contour aux organes voisins le long de la normale est supérieure à un seuil de 2 pixels. La classe « contact » regroupe tous les autres cas, y compris donc ceux pour lesquels la distance entre contours est inférieure au seuil de 2 pixels. Cependant une classification plus fine est nécessaire, car plusieurs sortes de profils peuvent être obtenues pour une même classe, du fait de la variabilité de l'orientation des normales et des niveaux de gris pour les tissus. Ces deux classes ont donc été divisées par un algorithme des k-moyennes en sous-classes dont le nombre a été optimisé (jusqu'à 10 en pratique). Chaque sous-classe est finalement représentée par son profil moyen (voir deux exemples à la Fig. 1c, haut).

La procédure de segmentation débute par une initialisation du contour de l'organe considéré. Pour chaque point de chaque contour, on explore l'apparence dans le voisinage (voir Fig. 1b, bas) en déterminant les profils d'intensité à n_{pfl} points le long de la normale en faisant varier la position i_{ctr} du centre par pas de 1 pixel sur une plage de ±4 pixels (voir Fig. 1c, bas). On calcule ensuite les distances de tous ces profils aux profils moyens de toutes les sous-classes. Si la distance minimale correspond à la classe « contact » il est impossible de déterminer avec précision la position du point de contour qui est alors ignoré dans l'étape suivante. Sinon, le point d'indice i_{ctr} est considéré comme le point de contour corrigé. L'étape suivante consiste à ajuster le modèle de forme aux

[1] Notre implémentation est basée sur le script fourni par Dirk J Kroon dans http://www.mathworks.com/matlabcentral/fileexchange/26706-active-shape-model—asm--and- activeappearance- model--aam-

points déterminés à l'étape précédente. Notons que les points non déterminés dans les zones de contact sont reconstruits par le modèle de forme lors de ce processus. Cette procédure est exécutée pour chacune des trois résolutions, de la plus grossière à la plus fine, le résultat de l'une servant d'initialisation à la suivante.

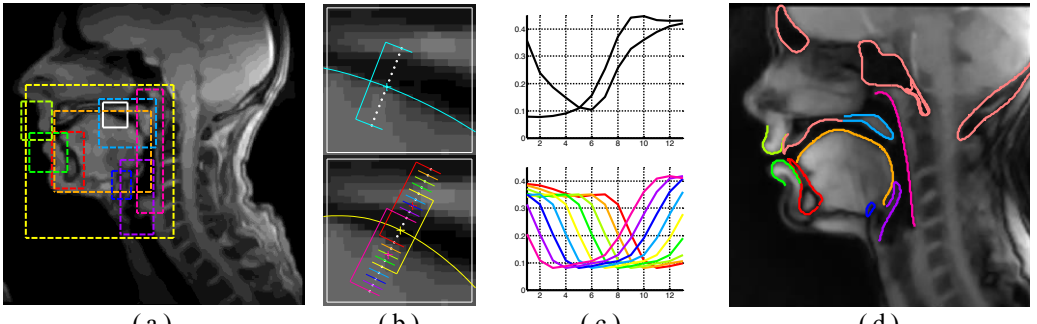
5 Evaluation

5.1 Données IRM pour l'évaluation

Nous avons testé les différentes méthodes présentées ci-dessus sur trois ensembles de séquences d'image IRM dynamiques obtenues par différentes techniques. Chaque ensemble était représenté par $n_{cl} = 60$ images sélectionnées par classification hiérarchique (*cf.* 4.1).

Le corpus [STRO] est composé de 18 combinaisons /pVCV/ avec V = [a i u], C = [b d m n ʁ l] prononcés par un locuteur français PB (Voyelle-Consonne-Voyelle). Il a été obtenu par la méthode d'échantillonnage synchronisé décrite par Masaki *et al.* (1999), dans laquelle le locuteur répète 128 fois chaque séquence en synchronie avec un bip délivré par l'imageur (voir Fig. 1a pour un exemple d'image). Le locuteur a été enregistré en 2002 dans les laboratoires ATR Human Information Processing Research Laboratories (Kyoto, Japan) (imageur : Marconi Eclipse 1,5 T, 468 images médio-sagittales, 28,9 images/sec reconstruites, résolution de 1 mm/px, champ de vue 256×256 mm², épaisseur de coupe 5 mm, profondeur d'image 8 bits/pixel, technique d'écho de gradient, temps de répétition (TR) 900 ms, temps d'écho (TE) 3 ms, angle de bascule (Flip) 30°).

Le corpus [VCV30] comprend 5 répétitions de CV avec C = [p t k] et V = [a i u] et de [fa sa ʁa ʒu zi ma la bao] prononcées par une locutrice allemande NB au Biomedizinische NMR Forschungs GmbH Göttingen. Les images fournies par une séquence d'écho de gradient radial hautement sous-échantillonné (3T Siemens Prisma Fit MRI System, 3225 images, 30 im/sec, 1,41 mm/px, 192×192 mm², épaisseur de coupe 8 mm, 12 bits/px, TR 1,96 ms, TE 1,28 ms, Flip 5°, nombre de projections radiales par image reconstruite 17 (NP)) sont reconstruites conjointement avec les profils de sensibilité des antennes en résolvant un problème inverse non linéaire suivant la méthode décrite dans Uecker *et al.* (2010).



(a) (b) (c) (d)

FIGURE 1 : (a) image [STRO] avec un exemple de cadres (cadre global jaune) ; (b, haut) zoom sur le cadre blanc de (a) avec contour de langue tracé et segment de normale au contour sous-tendant le profil d'intensités ; (b, bas) zoom avec contour prédit et illustration des segments de normale pour l'exploration des profils; (c, haut) exemple de profils moyens ; (c, bas) profils correspondant à (b, bas) ; (d) image [VCV55] avec exemple de contours prédits.

Le corpus [VCV55] comprend les mêmes répétitions de CV produites par la même locutrice. Le taux d'acquisition est de 55 images/sec (3200 images, TR 2,0 ms, NP 9). Notons qu'aucune différence visible n'apparaît entre les images de VCV30 et VCV55 (à l'exception de la mâchoire dans les transitions rapides). Les images ont été ensuite sur-échantillonnées à 0,71 mm/px (voir exemple à la Fig. 1d).

Notons que toutes les images ont été débruitées avant tout autre traitement à l'aide d'ondelettes de Daubechies de type 1. Elles sont ensuite recalées par rapport au palais grâce à la comparaison à un motif standard précédemment explicitée. Elles sont finalement recadrées comme indiqué en 4.2.

5.2 Résultats d'évaluation

Nous avons testé plusieurs combinaisons des méthodes décrites ci-dessus. Tous les résultats ont été obtenus par une méthode de validation croisée (*leave-one-out cross validation, cf.* Arlot & Celisse (2010)) qui calcule l'erreur d'estimation pour chaque élément test en utilisant les $n_{cl} - 1$ autres éléments pour établir les modèles. Cette procédure a été appliquée sur les ensembles d'apprentissage déterminés par classification hiérarchique pour chaque corpus. Nous avons utilisé deux métriques pour mesurer les différences entre contours : la RMS des distances points à points des contours, et la RMS des distances des points de contours prédits aux contours tracés.

Deux types de modèles linéaires de contours déformables ont été testés : (1) un modèle *global* M_{gbl} qui représente l'ensemble des coordonnées des contours avec un seul jeu de $n_{int\_gbl} = 12$ composantes donnant une erreur RMS de distance point à point moyennée sur les organes entre 0,86 et 1,06 mm (et de 0,60 – 0,74 mm pour les distances points à contour) suivant les corpus, et (2) un ensemble de modèles d'organes *locaux* M_{orgs}, donnant des erreurs de 0,23 – 0,26 mm (0,20 – 0,22) avec n_{cnt_orgs} totalisant entre 44 et 53 composantes (partiellement corrélées entre organes) qui expliquent chacune au moins 0.1% de la variance par organe (et au total au moins 99,5% de la variance par organe).

De manière analogue, nous avons testé deux types de modèles MLR : un modèle *global* MLR_{gbl} qui prédit les coordonnées des contours par l'intermédiaire de M_{gbl} à partir des $n_{int\_gbl} = 25$ composantes représentant environ 95% de la variance des intensités de l'image cadrée sur l'ensemble du conduit vocal avec une erreur de 1,48 – 1,56 mm (0,85 – 0,91) suivant les corpus, et des modèles *locaux* MLR_{orgs} qui prédisent les contours de chaque organe par l'intermédiaire des M_{orgs} à partir des n_{int_orgs} (entre 3 et 40) composantes des images cadrées sur chaque organe donnant une erreur de 1,23 – 1,41 mm (0,71 – 0,82).

Nous avons également testé l'amélioration apportée à la prédiction des modèles MLR par notre méthode ASMM. L'application de l'ASMM à chacun des contours prédits par le modèle *global* MLR_{gbl} permet de réduire les erreurs à 1,34 – 1,37 mm (0,64 – 0,67), tandis que les erreurs des modèles M_{orgs} sont ramenées à 1,14 – 1,31 mm (0,59 – 0,65). Pour les organes séparés, les erreurs varient de 0,59 à 2,04 mm (0.33 – 1.03), avec une gamme de 1,73 à 2,04 mm (0.91 – 1.03) pour la langue qui a l'erreur la plus grande. Un exemple de contours obtenus est donné à la Fig. 1d. Notons que les estimations des erreurs RMS sont globales et cachent des disparités entre phonèmes, positions sur les organes, et aussi corpus et sujets ; les erreurs les plus importantes se retrouvent sur les extrémités de la langue par exemple (jusqu'à 2,91 mm point à point et 2,43 mm point à contour).

Notons que la méthode des démons a donné des erreurs de 1,90 – 2,09 (1,02 – 1,12), plus élevées que les méthodes MLR avec ASMM, et qui ne sont pas suffisamment réduites par les ASMM.

Nous avons également testé plusieurs méthodes pour les organes rigides (mâchoire et hyoïde). La méthode MLR suivie d'ASMM donne une erreur de 0,72 – 1,15 mm (0,42 – 0,59) pour la mâchoire, et de 0,93 – 1,53 mm (0,59 – 1,07) pour l'hyoïde. Ces résultats sont assez similaires à ceux obtenus pour les organes déformables. Les résultats les meilleurs sont obtenus par des méthodes différentes en fonction des corpus, souvent le démon suivi d'un ASM, mais ne sont pas nettement meilleurs que pour la méthode MLR + ASMM : pour la mâchoire les erreurs sont de 0,62 – 0,86 mm (0,40 – 0,45), et pour l'hyoïde de 0,93 – 1,37 (0,59 – 0,83). Les méthodes basées sur la correspondance à un motif standard sont moins performantes.

6 Conclusion et perspectives

Nous avons développé une méthode de prédiction des contours individuels médiosagittaux des principaux organes orofaciaux impliqués dans la parole et la déglutition (le palais dur, la langue, les lèvres supérieure et inférieure, le velum, la paroi arrière pharyngée, l'épiglotte, ainsi que la mâchoire et l'os hyoïde) à partir d'images IRM dynamiques. Cette méthode est basée sur l'apprentissage de modèles d'apparence et de modèles de forme, ainsi que d'un modèle multilinéaire dont les résultats sont ensuite corrigés par un ASM modifié utilisant les informations locales de profils d'intensité de pixels sur les normales aux contours. Cette nouvelle méthode donne des erreurs moyennes points à contour entre 0,57 et 0,70 mm tous organes confondus. Ces performances, nettement meilleures que celles décrites dans les articles qui donnent des évaluations chiffrées, sont atteintes au prix d'un tracé manuel de tous les contours pour un corpus d'apprentissage d'une soixantaine d'images. Cet inconvénient est cependant minime si l'on souhaite traiter des corpus de centaines de milliers d'images pour le même locuteur. Nous notons aussi que les résultats sont très sensiblement semblables pour deux méthodes d'IRM aussi différentes que la méthode d'échantillonnage synchronisé qui oblige le sujet à répéter 128 fois le segment de phrase et la méthode d'écho de gradient radial hautement sous-échantillonné suivi de reconstruction par inversion non linéaire qui permet des acquisitions en temps-réel jusqu'à 55 images/sec. ou plus.

Cette nouvelle méthode ouvre des perspectives très intéressantes. Nous allons tout d'abord la tester pour des tâches de déglutition qui risquent de s'avérer plus difficiles à traiter, parce que les organes sont souvent en contact entre eux ou avec les aliments, et donc les contrastes plus faibles. Les contours obtenus pour la parole permettront d'établir des modèles articulatoires plus élaborés et d'analyser plus finement la variabilité et la coarticulation en parole. Les quantités de données possibles permettront également d'établir par apprentissage automatique des cartes d'association entre diverses modalités pour un même locuteur (p. ex. articulation, son) ou entre locuteurs.

Remerciements

Ce travail a bénéficié du support de l'ANR par les projets ANR-13-TECS-0011-06 «e-SwallHome» et ANR-11-INBS-0006 «Infrastructure d'avenir en Biologie Santé».

Références

ARLOT, S. & CELISSE, A. (2010). A survey of cross-validation procedures for model selection. 40-79.

BRESCH, E. & NARAYANAN, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging, 28*(3), 323-338.

COOTES, T.F., TAYLOR, C.J., COOPER, D.H. & GRAHAM, J. (1995). Active shape models - Their training and application. *Computer Vision and Image Understanding, 61*(1), 38-59.

ERYILDIRIM, A. & BERGER, M.-O. (2011). A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In *19th European Signal Processing Conference (EUSIPCO 2011)* pp. 61-65. Barcelona, Spain.

KIM, J., KUMAR, N., LEE, S. & NARAYANAN, S.S. (2014). Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In *10th International Seminar on Speech Production, ISSP10* (S. Fuchs, M. Grice, A. Hermes, L. Lancia & D. Mücke, Eds.), pp. 222-225. Cologne, Germany.

KROON, D.-J. & SLUMP, C.H. (2009). MRI Modality transformation in demon registration, *IEEE International Symposium on Biomedical Imaging, ISBI '09* (pp. 963-966). Boston, MA: IEEE Signal Processing Society.

MASAKI, S., TIEDE, M.K., HONDA, K., SHIMADA, Y., FUJIMOTO, I., NAKAMURA, Y. & NINOMIYA, N. (1999). MRI-based speech production study using a synchronized sampling method. *Journal of the Acoustical Society of Japan (English), 20*(5), 375-379.

NIEBERGALL, A., ZHANG, S., KUNAY, E., KEYDANA, G., JOB, M., UECKER, M. & FRAHM, J. (2013). Real-Time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine, 69*, 477-485.

OLTHOFF, A., ZHANG, S., SCHWEIZER, R. & FRAHM, J. (2014). On the physiology of normal swallowing as revealed by magnetic resonance imaging in real time. *Gastroenterology Research and Practice, 2014*, 10.

PROCTOR, M.I., BONE, D., KATSAMANIS, A. & NARAYANAN, S.S. (2010). Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Interspeech 2010 (11th Annual Conference of the International Speech Communication Association)* (T. Kobayashi, K. Hirose & S. Nakamura, Eds.), pp. 1576-1579. Makuhari, Japan.

SERRURIER, A. & BADIN, P. (2008). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *Journal of the Acoustical Society of America, 123*(4), 2335-2355.

SILVA, S. & TEIXEIRA, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech & Language, 33*(1), 25-46.

UECKER, M., ZHANG, S., VOIT, D., KARAUS, A., MERBOLDT, K.-D. & FRAHM, J. (2010). Real-time magnetic resonance imaging at a resolution of 20 ms. *NMR in Biomedicine 23*, 986–994.