Extraction semi-automatique des mouvements du conduit vocal à partir de données cinéradiographiques

Julie Fontecave et Frédéric Berthommier

ICP – Institut de la Communication Parlée
INPG, 46 avenue Félix Viallet, 38000 Grenoble, France
fonte,bertho@icp.inpg.fr

ABSTRACT

Since high speed X-ray films still provide the best dynamic view of the entire vocal tract, large existing databases have been preserved and are available for the speech research community. We propose a new technique for automatic extraction of vocal tract movements from these data. At first, the method was developed for the extraction of tongue movements in Wioland recorded in Strasbourg in 1977. Then, the same technique was adapted to other articulators and other X-rays films, taking into account their specificities. Finally, a quantitative evaluation of the estimate error and a comparison with Thimm and Luettin (1999) are achieved.

1. INTRODUCTION

La radiographie a été pendant longtemps l'une des principales techniques d'acquisition de données articulatoires en offrant la possibilité d'obtenir une vue sagittale complète des articulateurs du conduit vocal, de la glotte jusqu'aux lèvres. Devenue dynamique à la fin des années 1950, sous le terme de cinéradiographie, elle permet l'observation des mouvements des articulateurs de la parole avec une résolution temporelle importante, de l'ordre de 60 ips (images par seconde). Depuis quelques années, pour des questions de déontologie, on n'enregistre plus de nouveaux films radiologiques du conduit vocal. La cinéradiographie ayant fait la preuve de son utilité pour la recherche scientifique, il est nécessaire de pouvoir continuer à utiliser les données existantes en préservant les films. C'est dans ce contexte que Munhall et coll. [1] ont réalisé la base ATR « X-ray film database for Speech Research », à partir de films enregistrés par Rochette (Université Laval), et Stevens et Perkell (M.I.T.). Soutenu par le programme « Ingénierie des Langues » du CNRS, l'Institut de Phonétique de Strasbourg et l'Institut de la Communication Parlée de Grenoble ont aussi élaboré une base de données cinéradiographiques du français incluant les séquences Wioland et Flament [2].

L'extraction de données géométriques à partir de films radiologiques est généralement réalisée manuellement, mais on doit faire face à de grandes quantités de données pour traiter la moindre séquence. L'extraction automatique des contours de la langue fût envisagée par Laprie et Berger [3] pour exploiter au mieux ces grandes bases. Mais jusqu'à présent, seuls les travaux de Thimm et Luettin [4] ont aboutis au traitement complet d'un film issu de la base ATR (Laval43).

En vue d'améliorer cette situation, nous avons mis en place une méthode semi-automatique applicable film par film et qui combine le marquage manuel et la reconstruction automatique du mouvement. Cette technique [5] est basée sur une adaptation de l'algorithme de rétro-marquage [6], dont le principe est d'associer des paramètres implicites et extraits du signal vidéo à des paramètres géométriques contrôlés et définis a posteriori, plutôt que d'extraire directement des données géométriques. Pour estimer les mouvements de langue, la méthode se décompose en 3 étapes : (1) le traitement manuel d'un nombre restreint d'images clefs qui permet de définir des paramètres géométriques (ici le contour de la langue), (2) une étape automatique d'indexation de la base à partir de ces mêmes images clefs réduites et cadrées, qui a pour but d'associer à chacune des images de la base le marquage géométrique et (3) des traitements postérieurs de régularisation. A noter que le rétro-marquage peut être rendu entièrement automatique lorsque les informations géométriques sont extractibles dans les images clefs (voir un exemple, dans http://www.icp.inpg.fr/~bertho/m2p/jep06/main.wmv, sur les mouvements de la main). Mais dans le cas de la langue, cette tâche très difficile même pour l'expert humain est dévolue au marquage manuel dans des conditions de facilitation que nous décrirons par la suite.

A l'heure actuelle, cette méthode a aisément été appliquée avec succès sur plusieurs films radiographiques et adaptée pour tirer profit des particularités de ces bases. D'abord sur Wioland [7] pour la mise au point, puis sur le film Flament [8], composé de près de 5000 images, avec lequel nous nous intéressons plus particulièrement à la pointe de la langue et au voile du palais. Enfin, l'application de la méthode sur l'une des séquences de la base de données d'ATR, Laval43, a permis la comparaison directe de nos résultats avec ceux de Thimm et Luettin [4] cités précédemment.

2. MÉTHODE QUASI-AUTOMATIQUE D'EXTRACTION DE MOUVEMENTS

2.1. Extraction des mouvements de la langue à partir de la base Wioland

Cette séquence enregistrée en 1977 [7] et numérisée récemment comprend 5673 images du conduit vocal

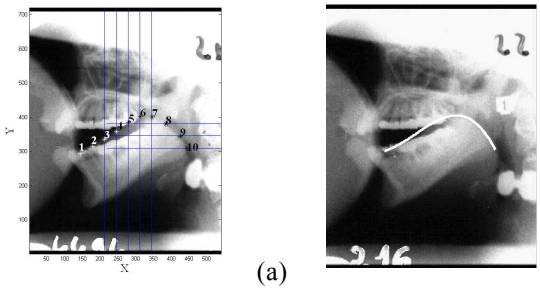
provenant de 64 séquences vidéos (64 phrases prononcées par une locutrice française), enregistrées à 66 ips. Durant la phase de mise au point, notre méthode a d'abord permis de récupérer les mouvements de la langue, puis elle a aisément pu être étendue à d'autres parties du conduit vocal (lèvres, vélum…).

L'étape manuelle consiste à décrire, pour 100 images clefs choisies aléatoirement, la position du contour de la langue avec 10 points (Fig. 1a), dont 8 par intersection avec des lignes verticales et horizontales (base et dos) et 2 points libres pour la pointe, soit 12 degrés de liberté (ddl). Cette étape est réalisée avec une interface qui permet, grâce à un curseur actionné manuellement, de voir la langue en mouvement, et dans de nombreux cas, d'associer un contour quasiment indiscernable sur l'image clef statique. Le choix des lignes de marquage et des points libres est fait de telle sorte qu'il n'y ait pas de données manquantes. A ce stade, pour chaque image clef, on dispose d'une configuration géométrique brute pour la langue en reliant les 10 points.

Puis, pour chaque image de la séquence, l'index de l'image clef la plus proche est assigné. La mesure de similarité est la distance Euclidienne entre les coefficients DCT (Discrete Cosinus Transform) basses fréquences des deux images. Au préalable, les images sont restreintes à un cadre minimal d'observation de l'articulateur cible pour tout le film, ceci de façon à minimiser l'interférence avec d'autres articulateurs ou des parasites (e.g. les inscriptions manuelles visibles figure 1). Après indexation, on aboutit ainsi à un premier marquage de la base entière en associant l'information géométrique disponible uniquement pour les images clefs. Des traitements postérieurs, filtrage temporel et moyennage de configurations voisines obtenues par multi-indexation permettent d'améliorer sensiblement cette première estimation entachée d'une erreur de quantification et des erreurs dues à l'indexation. Un lissage spline est également appliqué sur les points estimés. Nous aboutissons ainsi à une reconstruction complète du mouvement observable par superposition sur le film d'origine (Fig. 1b).

En terme de temps de traitement, le marquage manuel de 10 points sur 100 images clefs est estimé à 2 heures minimum. Ensuite, le temps d'exécution de la méthode pour la base complète (5673 images) dure quelques minutes.

Une évaluation objective a été mise en place à l'aide d'un deuxième jeu de 100 images tests marquées. L'erreur RMS par ddl est calculée entre les marques manuelles et les marques estimées par la méthode quasi-automatique. L'erreur $Etot_1$ considérée au final est la moyenne de cette erreur sur les 12 degrés de liberté. Elle est égale à 11 pixels (à comparer avec 350 pixels de longueur totale) équivalents à 3 mm après une calibration approximative car cette information n'est pas disponible directement sur le film.



(a) (b)

Figure 1 : (a) Excepté pour la pointe, les points sont marqués à l'intersection entre le contour de la langue et les lignes verticales ou horizontales.
(b) Le résultat est observé sur une vidéo (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/langue-wioland.wmv) par superposition des configurations géométriques estimées.

2.2. Estimation des mouvements du conduit vocal

Nous avons appliqué la même méthode aux autres parties du conduit vocal, essentiellement lèvres et vélum. Pour chaque articulateur, les images d'origine sont découpées de façon à inclure l'élément à marquer pour tout le film, et à exclure les interférences. Les paramètres de la méthode (nombre d'images clefs, degrés de liberté, nombre de coefficients DCT nécessaires pour l'indexation) sont définis de façon indépendante pour chaque articulateur. Les parties fixes du conduit vocal (palais, pharynx) sont également marquées de façon à reconstruire les mouvements du conduit vocal complet (figure 2).

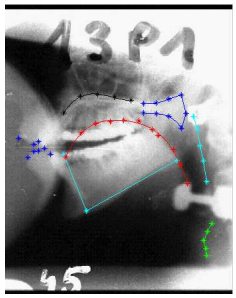


Figure 2 : Marquage complet du conduit vocal dans Wioland (http://www.icp.inpg.fr/~bertho/m2p/jep06/conduit-wioland.wmv)

Les mouvements du conduit vocal pourront être associés avec l'audio afin d'étudier les aspects dynamiques de la relation entre configuration géométrique du conduit vocal et acoustique.

3. BASE CINÉRADIOGRAPHIQUE FLAMENT : POINTE ET VOILE DU PALAIS

Pour traiter la base cinéradiographique Flament, enregistrée dans des conditions proches de Wioland, 13 ddl ont été définis pour marquer le contour de la langue : 9 points à 1 ddl pour la base et le dos et 2 points à 2 ddl pour la pointe. La pointe de la langue est nettement plus

visible dans ce film et une adaptation a été réalisée afin de mieux capturer ses mouvements rapides et parfois relativement indépendants. Elle consiste en un double marquage associant une estimation globale des 13 ddl comme précédemment, et une seconde spécifique de la pointe incluant 5 ddl seulement. Cette dernière est calculée à partir d'un cadre focalisé sur la pointe (Fig. 3). Le nombre d'images clefs a aussi du être étendu à 200. La fusion de ces deux estimations est réalisée par substitution des 5 ddl de la pointe dans l'estimation globale (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/langue-flament.wmv). L'erreur de reconstruction $Etot_1$ est estimée à 10 pixels pour une longueur moyenne de langue de 375 pixels.

Le suivi des mouvements de la pointe permet en particulier de détecter les instants de contact de la langue, que l'on peut rapprocher de l'audio et des événements consonantiques.

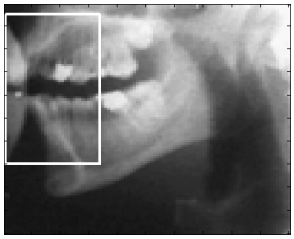


Figure 3 : La position de la pointe est estimée localement à partir d'un cadre spécifique.

D'autre part, le corpus est dédié à la question des nasales du Français [8] et le vélum est bien visible sur le film. Un traitement spécifique à cet articulateur a aussi été réalisé avec succès (voir http://www.icp.inpg.fr/~bertho/m2p/jep06/velum-flament.wmv).

4. BASE CINÉRADIOGRAPHIQUE D'ATR : UN COMPARATIF SUR LA LANGUE

La base de données d'ATR est la plus grande disponible pour les recherches en parole, avec 25 films différents totalisant une durée de 55 minutes et près de 100000 images. Nous avons extrait ces images à partir du DVD mais il n'est pas pour l'instant possible d'en réaliser le traitement complet à cause de l'étape de marquage manuel propre à chaque film. Notre étude concerne la séquence Laval43 dans un but comparatif.

4.1. Méthode d'extraction développée à l'IDIAP

Le film Laval43 (environ 4000 images) a été marqué en totalité par Thimm et Luettin [4] à l'IDIAP. En résumé, les résultats, disponibles en détail sur le site http://www.idiap.ch/machine_learning.php?project=64 ont été obtenus à partir d'une technique de normalisation d'histogrammes et d'une méthode d'extraction de contours. La méthode, notée TL, fait aussi appel à des images clefs choisies aléatoirement, sur lesquelles l'application d'un détecteur de Canny permet de récupérer le contour de la langue. La procédure de suivi de contours utilise l'appariement avec ces images clefs et l'information temporelle. Les résultats concernent plusieurs articulateurs du conduit vocal, mais ne permettent pas de reconstruire sa forme complète, en particulier car la pointe de la langue est souvent manquante. Nous nous intéresserons ici aux résultats concernant la langue, dans le but de comparer directement cette méthode à la nôtre, notée FB. Ces résultats ont été récupérés et conditionnés, de même que le film Laval43, pour permettre une comparaison objective. Nous disposons ainsi d'un jeu de splines définissant le contour de la langue pour chaque image. Nous le noterons S_{TLi} dans la suite de l'article.

4.2. Comparaison de méthodes

La méthode de rétro-marquage a été appliquée avec 200 images clefs, 9 points à 1 ddl et 2 points à 2 ddl pour la pointe, soit 13 ddl. Nous disposons ainsi d'un second jeu de splines, noté S_{FBi}.

Pour comparer les 2 estimations à partir de ces 2 jeux de splines, 2 types de mesures sont considérés :

- une mesure relative D, qui calcule la distance entre les 2 splines, proposée par Thimm [9]. Il s'agit de l'aire comprise entre les 2 courbes splines, normalisée par la somme de leurs longueurs.

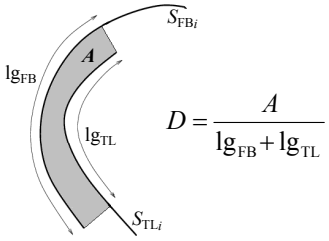


Figure 4 : Mesure de distance entre 2 splines

- une mesure de $Etot_1$ basée sur des images tests mesurant l'écart entre le marquage manuel et les 2 estimations (Fig. 5). Pour cette mesure, compte-tenu des données manquantes pour la pointe (dues à la difficulté d'estimation par une approche contour), nous n'avons pris en compte que les 8 points définissant le contour du dos et de la base (Fig. 5).

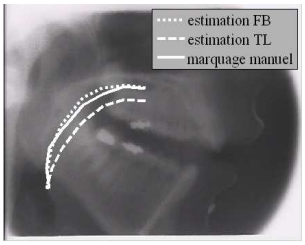


Figure 5 : Comparaison sur une image test d'un marquage manuel de la langue avec 2 marquages estimés (la pointe est exclue de cette comparaison)

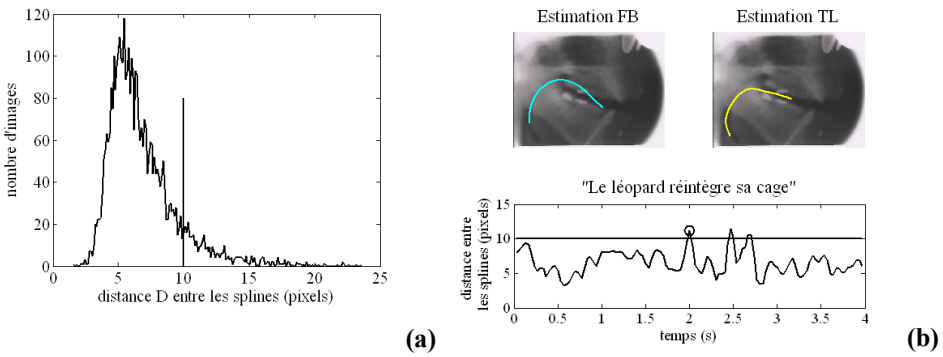


Figure 6 : (a) Répartition de la différence *D* entre les 2 estimations, et définition d'un seuil de décrochage *D*>10
(b) Décrochage observé au milieu de la séquence considérée, avec les deux contours estimés à cet instant

La différence moyenne *D* entre les 2 estimations est de 6.8 pixels. La distribution de *D* (Fig. 6a) montre que pour 10% de la base, il existe un décrochage entre les 2 méthodes que nous caractérisons par un seuil $D > 10$ pixels. Au dessus du seuil, l'écart moyen est de 12.7 pixels. Ce décrochage est observable visuellement sur la figure 6b. A noter que le contour associé par rétro-marquage est correct dans ce cas, et qu'il inclut la pointe.

Avec la méthode de rétro-marquage, l'erreur $Etot_1$ calculée sur 8 ddl varie en fonction du nombre d'images clefs et des traitements postérieurs (Fig. 7). Nous obtenons ainsi une erreur inférieure à 8 pixels (condition 175 clefs, 4v af) et autour de 20 pixels pour Thimm et Luettin (1999), sachant que cette section de la langue a une longueur estimée de 250 pixels.

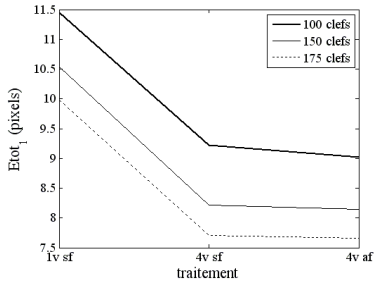


Figure 7 : Evolution de l'erreur $Etot_1$ pour notre méthode de marquage avec différents traitements postérieurs (indexation simple ou multiple, sans ou avec filtrage)

5. CONCLUSION

Nous montrons que le rétro-marquage basé sur les paramètres vidéo DCT basses fréquences offre la possibilité de suivre le mouvement du tractus vocal, même lorsque les contours ne sont pas entièrement visibles, notamment pour la pointe. C'est un progrès, mais la question reste ouverte de savoir si nos mesures sont suffisamment précises pour être mises en correspondance avec les caractéristiques temporelles et spectrales de la parole afin d'en obtenir de nouvelles informations. Par contre, en ce qui concerne le mouvement du velum, la cinéradiographie apportera sans doute des données précieuses qui ne sont pas accessibles autrement.

REMERCIEMENTS

Nous remercions Pascal Perrier pour les films cinéradiographiques Wioland et Flament, numérisés dans le cadre du programme « Ingénierie des Langues » du CNRS. Nous remercions Kevin Munhall pour nous avoir adressé une version de la base ATR sur DVD.

BIBLIOGRAPHIE

[1] K.G. Munhall, E. Vatikiotis-Bateson & Y. Tohkura. X-ray Film database for speech research. *Journal of the Acoustical Society of America,* 98 : 1222-1224, 1995.

[2] A. Arnal, P. Badin, G. Brock, P.-Y. Connan, E. Florig, N. Perez, P. Perrier, P. Simon, R. Sock, L. Varin, B. Vaxelaire & J.-P. Zerling. Une base de données cinéradiographiques du français. *XXIIIèmes Journées d'Etude sur la Parole*, pages 425-428, 2000.

[3] Y. Laprie & M.-O. Berger. Extraction of Tongue Contours in X-Ray Images with Minimal User Interaction. In *Proc. Int. Conf. on Spoken Language Processing,* volume 1, pages 268-271 , 1996.

[4] G. Thimm & J. Luettin. Extraction of articulators in X-ray image sequences. In *Proc. Eur. Conf. on Speech Communication and Technology*, pages 157-160, 1999.

[5] J. Fontecave & F. Berthommier. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database. In *Proc. Eur. Conf. on Speech Communication and Technology*, pages 1081-1084, 2005.

[6] F. Berthommier. Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 3, pages 789-792, 2004.

[7] F. Wioland. Faits de jointure en français. Implications aux niveaux articulatoire et acoustique. Incidences sur le plan des fonctions linguistiques. *Doctorat d'Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg , 1985.

[8] B. Flament. Recherche sur la mise en relief en français. Approche théorique et essai de caractérisation phonétique à partir de données de la mingographie et de la radiocinématographie. *Doctorat d'Etat*, Institut de Phonétique – Université des Sciences Humaines de Strasbourg, 1984.

[9] G. Thimm. Segmentation of X-ray image sequences showing the vocal tract. *IDIAP Research Report*, IDIAP, Suisse, 1999.