SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

Comment faire parler les images aux rayons X du conduit vocal

Yves LAPRIE[1], Rudolph SOCK[2&3], Béatrice VAXELAIRE[2], Benjamin ELIE[1]
[1]Institut de Phonétique de Strasbourg, Université de Strasbourg France
[2]CNRS, INRIA, UL, LORIA UMR 7503, Nancy, France
[3]Université Pavla Jozefa Safarika, Faculté des Lettres Košice, Slovaquie
Yves.Laprie@loria.fr

1. Introduction

La production de la parole est un phénomène dynamique qui repose sur la réalisation de gestes articulatoires par le locuteur. Son étude nécessite donc des moyens d'acquisition, souvent issus de l'imagerie médicale (Marchal & Cavé 2009), afin de visualiser et de mesurer ces gestes. La cinéradiographie a représenté un progrès décisif durant la seconde moitié du vingtième siècle, car elle offrait une fréquence d'échantillonnage suffisante pour observer la plupart des gestes articulatoires, en particulier ceux de la langue. Compte tenu de la dangerosité de cette technologie, due à l'exposition du sujet aux rayons X, les enregistrements ont été arrêtés. Cependant il existe un fonds de films aux rayons X important, varié du point de vue des langues concernées, souvent de bonne qualité, qui représente donc une matière première d'un intérêt potentiel considérable.

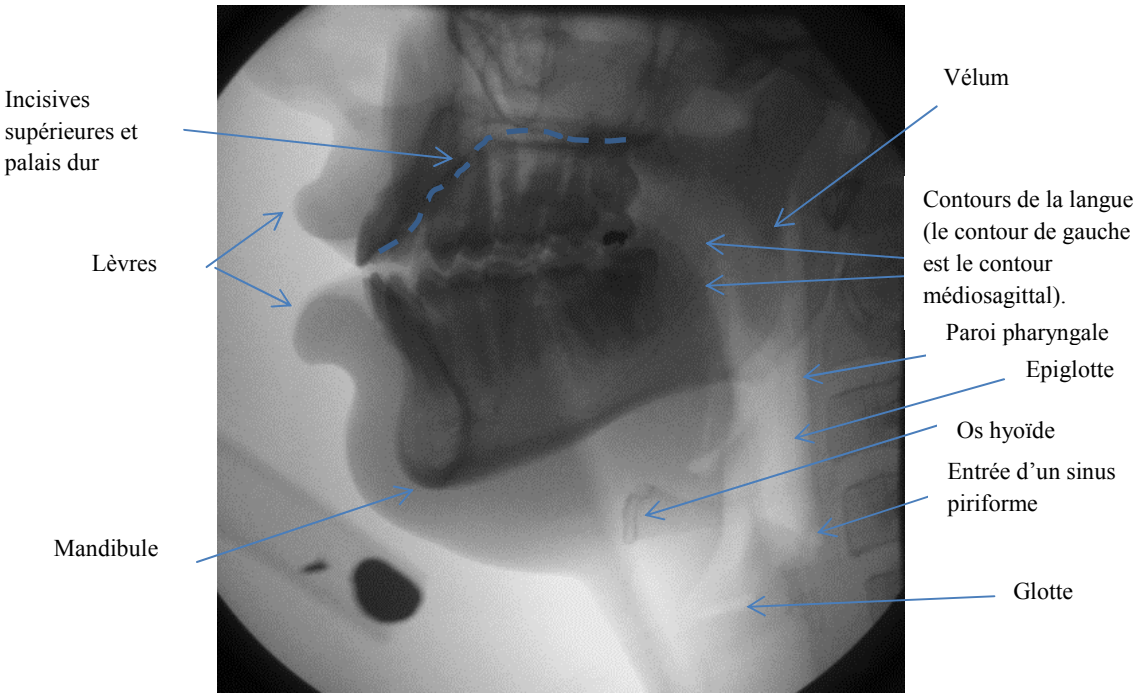

Figure 1 : Image aux rayons X du conduit vocal

Les données cinéradiographiques disponibles dans notre laboratoire revêtent un caractère unique au monde, tant par leur nombre (près de 50 films dont *une vingtaine est de très bonne qualité*) que par le

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801344

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

spectre linguistique qu'elles couvrent : plus d'une dizaine de langues de familles linguistiques très variées, d'Afrique, d'Amérique Latine, d'Asie et d'Europe. Ces données portent sur la production de la parole, comprenant des images cinéradiographiques du conduit vocal et l'enregistrement acoustique synchronisé. Chaque film examine une problématique linguistique spécifique, au moins. Tous ces films ont été partiellement traités à la main par des experts phonéticiens qui, pour un certain nombre d'images radiographiques, ont dessiné un tracé précis des contours du conduit vocal dans le plan médian (tracé sagittal) et, dans certains cas, un tracé du pavillon labial (La Figure 1, extraite de l'un des films de notre base de données, montre les différents articulateurs de la parole). Toutefois, à raison d'une moyenne de 3000 images par film, on comprend aisément que toutes les données n'aient pas pu être traitées. Le développement de *techniques d'extraction semi-automatique ou automatique* des contours du conduit vocal a donc été un aspect incontournable et central à notre projet. Nous y reviendrons dans le paragraphe 3.2.

En outre, les tracés sagittaux existants étaient jusqu'ici stockés sur des supports papiers ; de ce fait, ils étaient difficilement exploitables d'un point de vue statistique et informatique. Pour permettre une *large distribution de ces données*, et pour faciliter leur exploitation, les films ont été numérisés puis intégrés dans une base de données DOCVACIM (Sock et al., 2011). La recherche dans la base de données se fait par phonèmes, par corpus et par langue, en entrant le nom d'un corpus et/ou en tapant la séquence sonore qui présente un intérêt pour l'utilisateur de la base. Les informations ainsi obtenues s'affichent de manière succincte et signalétique.

Une telle base de données représente une source d'informations articulatoires de premier ordre, et nos efforts ont tout particulièrement porté sur le développement de modèles articulatoires géométriques du conduit vocal (paragraphe 4). Ces modèles peuvent être utilisés afin d'étudier l'articulation des sons de la parole d'un point de vue dynamique et de la variabilité interlocuteur. Ils peuvent aussi être intégrés à un synthétiseur articulatoire, comme nous le présenterons à la fin de cet article.

2. Généralités sur l'élaboration du corpus

Dans toutes nos expériences, qu'elles soient dans le domaine de la cinéradiographie ou non, nous essayons, dans la mesure du possible, de ne retenir que de véritables mots (lexèmes ou morphèmes) de la langue étudiée, en l'occurrence le français dans cette étude. L'utilisation éventuelle, mais restreinte, de logatomes reste cependant justifiée par leurs ressemblances structurelles (segmentale et suprasegmentale) avec les mots de la langue cible, dans la mesure où les sons de ces logatomes rentrent dans le cadre de l'habitus des locuteurs. Mis à part ces contraintes phono-tactiques de la langue, se pose le problème de limitation dans le nombre, souvent exponentiel, des items à analyser. Nous nous appuyons donc souvent sur des hypothèses de départ en regardant, en priorité, les bornes des réalisations (Abry & Boe, 1981) maximales et minimales d'opposition et de contraste. À titre d'exemple, nous nous intéressons tout d'abord, lorsqu'il s'agit d'étudier le phénomène des espaces articulatoire et acoustique vocaliques du français, aux voyelles extrêmes /i a u/ de cette langue, cela pour obtenir les limites du domaine spatio-temporel vocalique en priorité. Ces domaines de réalisations vocaliques une fois définis, il devient possible d'y situer spatialement et temporellement, et de manière plus cohérente, les autres éléments vocaliques du français. Il en sera de même pour l'étude d'oppositions consonantiques simples/doubles, par exemple ; l'étude en priorité des consonnes occlusives /p t k/ nous offre non seulement la possibilité d'une délimitation plus aisée, et donc plus fiable, sur le signal acoustique, elle nous permet également d'obtenir une bonne mise en relation entre les événements acoustiques et les événements articulatoires (contacts et relâchements des articulateurs). Cette fiabilité dans les mesures articulatoires et acoustiques est particulièrement appréciable lorsque l'on raisonne en termes de millimètres et de millisecondes, respectivement. L'inclusion d'autres consonnes dans nos investigations se fait, en général, dans une phase d'observation ultérieure. Signalons toutefois que le corpus idéal est rarement établi et que les décisions définitives, en ce qui concerne le choix des items, sont largement déterminées par les mots disponibles dans la langue cible. Les items, en général, sont des paires minimales ou des séquences de sons extraites

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

de mots qui nous offrent la possibilité, toute chose étant égale par ailleurs, de tirer des conclusions plus sûres sur une analyse comparative des valeurs mesurées.

Pour des raisons de variantes intra-individuelles et pour satisfaire aux exigences des traitements statistiques – palliant ainsi partiellement le nombre relativement restreint de données que l'on puisse obtenir en cinéradiographie, nous faisons répéter notre corpus une dizaine de fois, lors de nos expériences acoustiques et/ou cinématiques connexes (voir, par ex., Hardcastle *et al*., 1996 ; Sock *et al*., 2005). De tels résultats, fondés sur des traitements statistiques plus élaborés, nous permettent ainsi de conforter les enseignements que nous tirons des analyses cinéradiographiques. Insistons toutefois sur la haute résolution spatiale des données cinéradiographiques, ce qui semble suffisant pour pouvoir en tirer des enseignements robustes.

Les phrases sont présentées en ordre aléatoire, de manière à ce que les termes d'un même item ne se suivent jamais, cela pour éviter tout phénomène d'ancrage, dû à la répétition de l'item.

La phrase porteuse permet la réalisation *plus ou moins naturelle* des productions des locuteurs ; ainsi, on évitera une analyse de mots isolés qui présentent souvent des caractéristiques temporelles et/ou fréquentielles inexploitables par les modèles de durée, de production ou encore par des systèmes de synthèse. Nos phrases sont donc choisies d'après un certain nombre de critères : elles sont simples, courtes et elles appartiennent à la langue courante pour pouvoir être prononcées par n'importe quel locuteur.

2.1. Spécificité de l'élaboration du corpus cinéradiographique

Les données de cette étude portent sur deux corpus. Le premier est formé de quatre petits films acquis pour un locuteur masculin à une cadence de 25 images par seconde, représentant au total un peu moins de 700 images, en ne considérant que les images correspondant aux instants où de la parole est effectivement produite. Deux des films étaient formés de phrases destinées à l'étude de la coarticulation labiale et les deux autres de logatomes VCV pour les 3 voyelles /i,a,u/ et les consonnes /t,k/.

Le second est un corpus de 58 phrases courtes, de 4 à 6 syllabes. Chaque phrase est chargée de sens. Le corpus propose un certain nombre de séquences du type VCV consonne simple, ou du type VCCV consonne double ou groupe de deux consonnes différentes.

Les deux structures syllabiques VCV et VCCV se trouvent dans un entourage vocalique identique et unique (/aCa/ ou /aCCa/). C'est la condition *ceteris paribus*, évoquée *supra*.

La structure VCV, comportant la consonne simple, est placée en contexte :

(a) vocalique labialisé, l'élément labialisé placé avant ou après la consonne ;
(b) consonantique nasal, l'élément nasal étant placé avant ou après la consonne ;
(c) d'une semi-consonne, donnant des groupes de consonnes courants en français.

Ce corpus a été élaboré pour une étude originelle plus générale sur l'observation comparée entre les vitesses d'élocution normale et rapide (voir *infra* pour les consignes données aux locuteurs pour la variation de leur vitesse d'élocution lors des enregistrements).

Nous avons par ailleurs utilisé les contours qui ont servi à construire le modèle articulatoire de (Maeda 1979). Ces contours ont été tracés à la main en 1978, puis numérisés ultérieurement.

2.2. Généralités sur le choix des locuteurs

Les sujets retenus sont toujours des locuteurs natifs, comprenant selon leur disponibilité, des voix d'hommes et des voix de femmes (ces dernières ne présentant aucun inconvénient d'analyse instrumentale pour nos études articulatoires, voire acoustiques, sur la dimension temporelle, comme ce serait le cas dans le domaine spectral). Ils ont, en général, une bonne diction, ne présentant aucun

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

antécédent pathologique du conduit vocal et possédant une audition normale. Ils sont tous volontaires, et ne sont jamais au courant des objectifs de nos recherches avant les enregistrements.

Nous gardons une fiche signalétique minutieuse sur la biographie de nos locuteurs, et au besoin (et suivant les expériences) nous l'intégrons dans nos études comme élément pertinent de certains résultats obtenus.

Pour ce qui concerne les investigations cinéradiographiques correspondant aux corpus utilisés dans ce travail, nous avons retenu deux sujets français.

Le premier est un locuteur masculin F.H. ayant vécu en Moselle puis en Alsace où il a poursuivi ses études. Il a 27 ans lors de l'enregistrement du premier corpus.

Le second est une locutrice M.M. ayant vécu en Franche-Comté, et fait ses études supérieures puis suivi une formation théâtrale professionnelle en Alsace. Elle a 24 ans à la réalisation du film.

2.3. Acquisition des données

2.3.1. Remarques sur la condition de laboratoire

Etant donné la nature précise des mesures que nous obtenons de nos échantillons de parole, et étant donné les enseignements que nous désirons tirer de nos résultats expérimentaux – en termes de contraintes spatio-temporelles finement définies, nos acquisitions sont systématiquement effectuées en *conditions de laboratoire*. Ces conditions ont le mérite, lorsqu'il s'agit d'obtenir des *données acoustiques seules*, de garantir un *meilleur rapport signal/bruit*, en comparaison avec toute acquisition qui pourrait se faire en condition de *parole purement spontanée*. L'obtention de données cinéradiographiques spontanées en parole est, bien entendu, impossible. Du coup, cela soulève nécessairement la question de la naturalité de nos séquences produites, sous forme de phrases lues ou non et, par-là, de leur pertinence pour toute interprétation en termes phonologiques ou linguistiques.

Quelques solutions peuvent être proposées pour remédier à ce problème :

1) Il est possible de comparer certains de nos résultats obtenus en cinéradiographie avec ceux constatés, pour de la parole plus ou moins spontanée, en conditions peu contraignantes ; c'est ce que nous faisons (cf. *supra*).

2) Nous nous rapprochons de la parole spontanée grâce au recours à des phrases, certes courtes, mais dotées de sens.

Dans tous les cas de figure, l'on doit se dire que plus l'on se rapproche de la parole spontanée, moins l'on réussit à contrôler les conditions d'acquisition et les contextes segmentaux et suprasegmentaux à partir desquels les mesures seront prélevées. Le prix à payer semble donc de gagner en précision et en fiabilité dans nos mesures, quitte à perdre en naturalité des séquences de parole produites (mais voir *infra* pour les autres précautions que nous prenons en laboratoire, par rapport à cette problématique de la parole naturelle et spontanée).

2.3.2. Acquisition de données cinéradiographiques

Les données cinéradiographiques occupent une place importante dans le domaine de l'étude de la production de la parole. En effet, cette méthode permet de fixer les *positions articulatoires* durant la production de la parole, c'est-à-dire de visualiser et d'analyser l'*action simultanée* de tous les organes, grâce aux vues de profil, depuis le larynx jusqu'aux lèvres, avec le *signal acoustique* synchrone. Cette technique permet de filmer des énoncés entiers et d'identifier sur les clichés les zones articulatoires « cibles » des différents sons émergeant dans le conduit vocal. Nous réussissons ainsi à obtenir des données qui mettent en relation l'articulatoire avec l'acoustique pour nos investigations.

Pour l'acquisition des données cinéradiographiques de cette investigation, le sujet est assis sur une chaise adaptée, munie d'un serre-tête. Nous utilisons la parole lue, et les « consignes » données aux locuteurs

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

sont de ne pas marquer de pauses silencieuses à l'intérieur des phrases. La vitesse d'élocution est régulière et on n'observe pas de variations de vitesse d'élocution à l'intérieure de chaque condition de vitesse d'élocution.

Les phrases sont enregistrées d'abord en vitesse d'élocution normale ou conversationnelle, celle qui paraît la plus naturelle au locuteur. Nous exigeons ensuite, après entraînement, une vitesse d'élocution rapide (soit environ deux fois plus que sa production normale), tout en restant dans le cadre de la « normalité ». Les ruptures de vitesse ou de rythme sont ainsi évitées. Ces modalités d'enregistrement ont contribué à la régularité de la gestion de la vitesse d'élocution. Le résultat obtenu est, dans le meilleur des cas, un facteur d'accélération d'environ 1.5 en moyenne, mesuré après l'enregistrement par ajustement auditif d'un métronome sur le rythme syllabique. Nous obtenons, après les enregistrements, les vitesses d'articulation suivantes, en nombre de syllabes/seconde :

1) En vitesse d'élocution normale, 4, 6 syll./s pour le locuteur A.E. et 4, 2 syll./s pour le locuteur M.M. ;
2) En vitesse d'élocution rapide, 6, 3 syll./s pour le locuteur A.E. et 5, 8 syll./s pour le locuteur M.M.

Ces résultats indiquent que les deux locuteurs ont bien réussi à effectuer la tâche d'accélération de la vitesse d'élocution qui leur était demandée.

3. Dépouillement des films

3.1. Spécificités des images aux rayons X du conduit vocal

Afin de mieux comprendre la nature et les difficultés du travail de dépouillement des images aux rayons X, il convient d'abord d'expliquer l'origine d'une image aux rayons X et ses caractéristiques.

Une image aux rayons X est obtenue à l'aide d'une source de rayons X dirigée vers la tête du locuteur qui est placée devant un récepteur, en l'occurrence la pellicule du film. La direction de la source est perpendiculaire au plan sagittal et les rayons sont plus ou moins stoppés selon la densité des organes qu'ils doivent traverser, ce qui produit des variations d'intensité photométrique sur l'image. Chaque image est donc la projection de la tête, qui est un objet tridimensionnel, sur le plan de l'image qui est bien sûr un objet bidimensionnel. L'information de profondeur qui correspond à la troisième dimension est perdue lors de cette projection. La formation de ces images est à l'origine des difficultés que représente l'extraction des contours des articulateurs, puisque tous les organes situés sur un même rayon se projettent en un seul point. Les organes les plus opaques donnent par conséquent les contours les plus marqués, et peuvent même masquer partiellement ou totalement certains contours. Ainsi l'os de la mandibule et les dents masquent partiellement la langue, et les plombages éventuels en matériaux opaques aux rayons X cachent complètement la langue si elle passe devant. Par ailleurs, la netteté d'un contour dépend aussi de la forme de l'objet projeté sur l'image, et le contour est d'autant plus marqué que le bord de l'objet coïncide avec la direction des rayons. Ainsi, quand le sillon médiosagittal de la langue est pas ou peu marqué, la langue donne lieu à un seul contour. En revanche, quand le sillon médiosagittal est marqué (voir Figure 2), ce qui est le cas à l'arrière de la langue pour /i/ ou /u/, le contour le plus visible est souvent celui du bord de la langue, et le sillon médiosagittal correspond à un contour moins net. De la même façon, l'interprétation des contours visibles dans le bas du larynx est rendue délicate par la présence des deux sinus piriformes. De bonnes connaissances anatomiques sont donc nécessaires pour identifier correctement les contours des articulateurs sur l'image.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

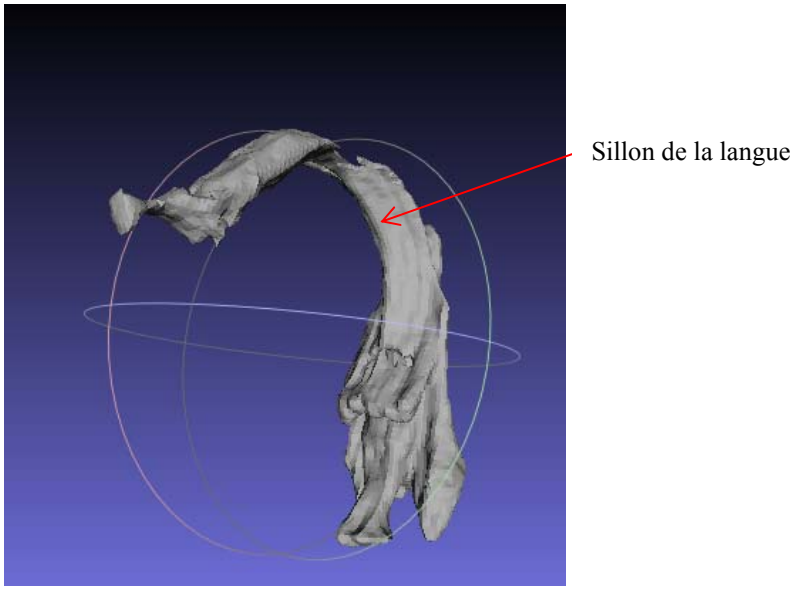

Figure 2 : Visualisation du volume d'air du conduit vocal depuis le larynx jusqu'aux lèvres.
Le sillon de la langue est indiqué. Il apparaît sous la forme d'une convexité sur la colonne d'air ce qui correspond bien à une concavité pour la langue.

Pour chaque image il est nécessaire d'extraire les contours : de la mandibule, de l'os hyoïde, de l'épiglotte, de la langue, des lèvres inférieure et supérieure, du palais dur, du vélum, de la paroi pharyngale, du larynx et de la glotte. Le contour du palais peut servir de contour de référence pour annuler les déplacements de la tête du locuteur pendant l'enregistrement. Il est cependant préférable de retenir une structure indéformable qui n'entre pas en contact avec la langue, cela pour limiter les imprécisions dans la détection de la région de référence qui sert pour tous les contours. Pour cette raison nous avons retenu une région située entre le haut des fosses nasales et le crâne.

3.2. Outils de suivi de contour et de dépouillement de films

Il faut donc extraire douze contours pour chaque image, ce qui donne pour un film d'environ mille images (soit vingt secondes de parole à une fréquence d'échantillonnage de 50 Hz) 12000 contours. L'ampleur du travail a souvent découragé les tentatives de traçage systématique des contours et ce problème a donc suscité un certain nombre de travaux (par exemple Thimm & Luettin (1999), Laprie & Berger (1996), Fontecave & Berthommier (2009)) pour automatiser l'extraction des contours, en particulier celui de la langue qui est l'articulateur à la fois le plus déformable et le plus mobile. La plupart d'entre eux s'est appuyée sur les approches utilisées en traitement d'images pour le suivi d'objets déformables. Mais les spécificités des images aux rayons X présentées plus haut expliquent que ces tentatives n'aient jamais été très concluantes. La plupart des techniques issues de la vision par ordinateur exploitent en effet souvent l'intensité du contraste correspondant au contour. Dès que les dents, la mandibule ou les plombages masquent partiellement ou totalement la langue, le suivi a tendance à capturer un contour suffisamment marqué qui n'est plus celui de la langue. Cette erreur est difficilement récupérable puisque les images suivantes présentent souvent des contours de la même nature. Fontecave et Berthommier (2009) ont proposé une technique intéressante utilisant une base d'images clés dans lesquelles les contours de la langue ont été tracés manuellement. L'extraction du contour de la langue d'une image consiste à trouver les trois images clés les plus proches et à interpoler le contour de la langue à partir de ceux des images clés. Il s'agit donc d'un suivi semi-automatique puisque l'utilisateur doit détourer manuellement le contour de la langue pour les images clés. Le point fort de cette technique est de garantir des contours dont la forme est compatible avec celle de la langue. Le point faible est qu'il est nécessaire que les images

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

clés couvrent simultanément la variabilité des formes de la langue et celle des organes qui sont susceptibles de la recouvrir. Ces deux conditions sont difficiles à satisfaire en pratique, sauf à multiplier le nombre d'images clés sans lesquelles le contour de la langue a été tracé à la main, ce qui réduit fortement l'intérêt de la méthode.

En revanche cette technique prend tout son intérêt lorsque l'organe à suivre est peu souvent masqué et déformable, ce qui est le cas des lèvres, du vélum, de l'épiglotte et du larynx. Par ailleurs, elle se prête évidemment bien à un schéma de corrections itératives consistant à ajouter dans la base des images clés celles qui ont fait échouer le suivi.

Dans le cas des structures osseuses et rigides, dans notre cas la mandibule, l'os hyoïde, le palais dur, une technique de corrélation est tout à fait adaptée. Elle consiste à corréler une image inconnue avec une image de référence de la structure rigide à suivre, en lui appliquant un ensemble de déplacements couvrant les positions possibles de la structure à suivre. Il s'agit d'un suivi automatique, puisque l'utilisateur n'intervient jamais dans le suivi. Dans les deux cas le suivi est d'autant plus efficace que la région de l'image sur laquelle est appliqué le suivi est circonscrite à la région dans laquelle le contour peut apparaître. Une étape préliminaire simple consiste donc à définir cette région pour toute la séquence d'images à traiter.

On voit ainsi se dessiner la stratégie de détourage des contours que nous avons adoptée lors du développement du logiciel *Xarticulators*. Elle consiste à exploiter des outils de suivi automatique pour les structures osseuses, des outils de suivi semi-automatique pour les organes faiblement masqués, et enfin des outils de détourage manuel pour la langue. Le choix de recourir à un détourage manuel pour la langue est dicté à la fois par les difficultés évoquées plus haut et par l'impact acoustique de la forme de la langue. De petites erreurs sur le contour de la langue peuvent en effet avoir une influence forte sur l'acoustique, en particulier à proximité de la constriction principale du conduit vocal ; il est donc essentiel de disposer de contours de langue aussi fiables que possible.

3.3. Logiciel de dépouillement

Le logiciel *Xarticulators* a été conçu de manière à exploiter les outils de suivi automatique ou semi-automatique. Bien sûr, les films aux rayons X ont, à quelques exceptions près, été acquis à l'aide de machines analogiques sur de vrais films. Le film peut donc sauter de temps en temps ou avoir été rayé avant d'être numérisé. Par ailleurs, bien que la tête du locuteur ait été en grande partie immobilisée, le locuteur a pu bouger légèrement et donc déplacer sa tête. De la même façon, les filtres en aluminium destinés à renforcer les contours des lèvres notamment ne sont pas solidaires de la tête et peuvent donc bouger indépendamment de la tête du locuteur. Pour toutes ces raisons il est nécessaire de pouvoir facilement corriger les résultats des suivis, et nous avons donc ajouté de nombreux outils pour rendre le détourage manuel et les corrections le plus simple possible. Dans le cas de la langue, il est souvent difficile de localiser le contour de l'apex quand une image est sortie de son contexte. La perception du geste qu'effectue la langue aide beaucoup à en localiser avec précision le contour au niveau de l'apex. *Xarticulators* offre donc des outils pour rejouer le film juste avant ou juste après l'image à traiter.

Comme l'objectif est d'extraire les contours des articulateurs de la parole dans une séquence d'images, *Xarticulators* comprend de nombreux outils destinés à manipuler des séquences d'images, sauver les contours ou encore les exporter dans d'autres formats.

3.4. Protocole de dépouillement d'un film

Il est très important d'assurer l'homogénéité et la qualité des contours qui sont extraits à partir des images pour que l'exploitation des contours soit possible. Nous avons donc établi un protocole de dépouillement précis et les outils logiciels permettant de le mettre en œuvre. La première étape est de définir une région de recalage sur une image de référence, choisie parmi les premières images du film et suffisamment contrastée. Cette région sera suivie sur toutes les images du film, afin de pouvoir annuler les mouvements de la tête du locuteur ; elle couvre une partie des fosses nasales et de la base du crâne en faisant attention

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

de ne pas couvrir ni les yeux qui peuvent cligner durant la prise de vue, ni les filtres éventuels dont la position est indépendante de la tête. Le contour du palais dur et des incisives supérieur est aussi dessiné à la main dans cette image de référence. Ce contour est attaché par une contrainte de position, à la région de recalage. Cela signifie que sa position sera déduite de celle de la région de recalage dans toutes les autres images avec un double avantage : (i) une fois la région de recalage suivie par corrélation, le contour du palais est connu, (ii) si la région de recalage est mal localisée par le suivi automatique, il est possible de la déplacer à la main et de vérifier sur la position du palais que la correction est pertinente.

Le suivi par corrélation est aussi utilisé pour la mandibule, en suivant une région de la mandibule présentant un contraste suffisant. Les résultats du suivi de ces deux régions suivies par corrélation sont vérifiés avec beaucoup d'attention, et corrigés le cas échéant, puisque la position de la première est utilisée pour contraindre les régions où est appliqué le suivi semi-automatique, et que le déplacement géométrique de la seconde est soustrait de la position de la langue, lors de la détermination des modes de déformation de la langue.

Nous utilisons les contraintes de position relative par rapport à la région de recalage pour définir les régions utilisées pour guider le suivi semi-automatique. La première étape consiste à détourer à la main l'objet à suivre sur les images clés choisies pour couvrir les variations de forme et de position. Comme le suivi construit le contour par interpolation, à partir des points de contrôle des splines, il faut que les contours des images clés soient tous homologues entre eux. Dans le cas du vélum, par exemple, cela signifie que le raccordement avec le palais dur doit avoir lieu au même endroit, et cela en garantissant une tangente commune entre le vélum et le palais. De la même façon, le point d'arrêt du vélum dans les fosses nasales doit être le même pour toutes les images clés. Pour cette raison, des traits de construction, dont la position est définie relativement à celle de la région de recalage, ont été ajoutés. La Figure 3 montre la région de recalage et celles utilisés pour circonscrire le suivi semi-automatique.

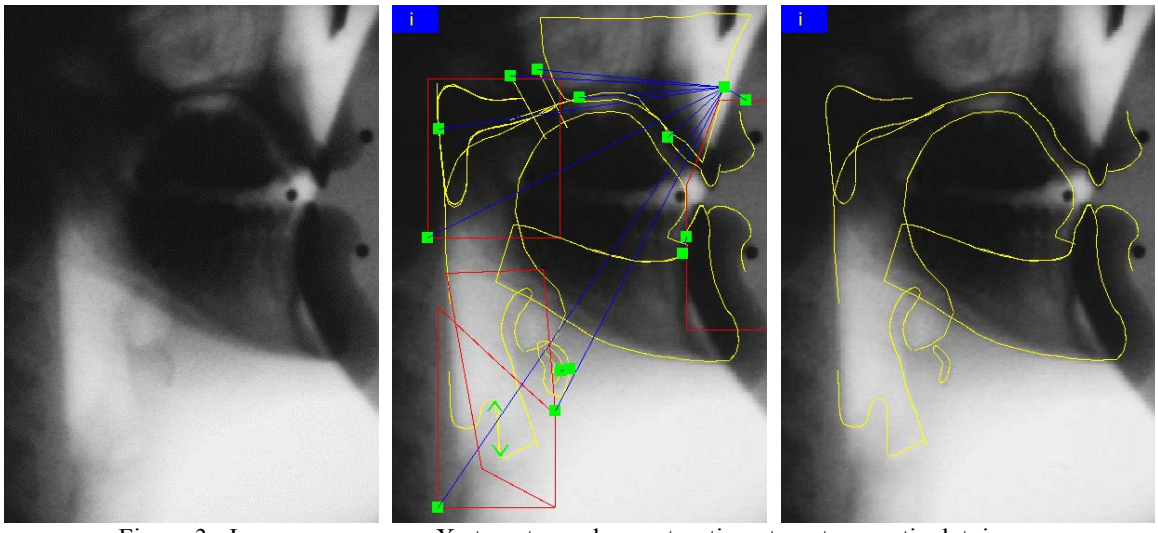

Figure 3 : Images aux rayons X et contours de construction et contours articulatoires
L'image de gauche est l'image originale. Celle de droite est le résultat du dépouillement. Il faut noter que le contour de la mandibule ne couvre que la partie avant ce qui est suffisant pour connaître complètement l'ouverture de la mâchoire. L'image du centre montre les contours et régions utilisés lors du dépouillement. Chaque région rouge correspond à l'utilisation du suivi semi-automatique pour un contour précis (vélum, lèvres, larynx, épiglotte). La position de tous les contours et régions utilisés pour le dépouillement est déduite de celle de la région de recalage (la forme grossièrement triangulaire située en haut à droite de l'image). Les carrés verts et lignes bleues sont les ancres et liens de dépendance avec la région de recalage. Pour le dépouillement du vélum trois contours (3 segments de droite) sont utilisés pour garantir que tous les contours sont homologues entre eux.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

3.5. Films dépouillés

Le logiciel *Xarticulators* a été utilisé pour le dépouillement de quatre films destinés à l'étude de la coarticulation en français, un film plus long d'une quinzaine de phrases destiné à l'étude de l'anticipation en français, et enfin huit films portant sur l'étude de la gémination en berbère, soit environ trois mille cinq cents images. Les travaux de dépouillement se poursuivent actuellement en concentrant les efforts sur les films dont la qualité est la meilleure et l'intérêt scientifique le plus manifeste.

Afin de visualiser les résultats du dépouillement, le logiciel *Xarticulators* permet de reconstruire des films sur lesquels les contours sont superposés et les étiquettes phonétiques affichées. Soulignons à ce sujet qu'il faut synchroniser le film avec le signal acoustique. Dans certains cas les films ont été enregistrés en intégrant une tirette visible sur l'image, et dont le mouvement brusque produit un bruit facile à repérer sur le signal acoustique ou son spectrogramme. Quand ce n'est pas le cas, nous avons utilisé les consonnes /l,p,b,m/ qui sont facilement repérables sur les images (par l'intermédiaire du contact entre la pointe de la langue et le palais pour /l/, et la fermeture complète des lèvres pour les autres), et également facilement repérables sur le spectrogramme.

4. Construction de modèles articulatoires

Les contours peuvent être exploités directement pour mesurer la constriction principale du conduit vocal, l'ouverture ou la protrusion des lèvres, l'ouverture de la mâchoire… Ils peuvent aussi servir à la construction de modèles articulatoires afin de modéliser la forme du conduit vocal (Beautemps et al. 2001). Le premier objectif est bien sûr la synthèse articulatoire pour produire un signal sonore à partir d'une suite de phonèmes, mais aussi d'étudier la formation des constrictions dans le conduit vocal, du point de vue de la précision géométrique nécessaire à la réalisation des indices acoustiques caractéristiques des sons de la parole.

Nous avons étudié deux aspects de ce problème :

1. Est-il possible d'approcher la forme du conduit vocal d'un locuteur à l'aide du modèle construit pour un autre locuteur ? Il s'agit donc de savoir s'il est possible de copier la forme géométrique du conduit vocal d'un locuteur à l'aide de celui qui a été construit pour un autre locuteur. Cette question a été abordée pour les voyelles parce qu'il n'y a pas de contact entre la langue et le palais qui est propre à chaque locuteur.
2. Est-il possible d'approcher la forme du conduit vocal à l'aide d'un modèle géométrique contrôlé par un petit nombre de paramètres pour les consonnes ? Cette fois il s'agit d'approcher, avec une représentation paramétrique aussi concise que possible, la forme du conduit vocal en assurant qu'elle est proche de l'originale au voisinage de la constriction qui donne les propriétés acoustiques des consonnes.

Avant de détailler les résultats obtenus, nous allons présenter la stratégie suivie pour construire un modèle articulatoire (Laprie & Busset 2011) qui contrôle chacun des articulateurs explicitement à l'aide de modes linéaires contrôlant soit le mouvement d'articulateurs rigides, ici la mandibule, soit les modes de déformation de la langue, des lèvres, du velum et de la région larynx/épiglotte. Ce modèle a été construit à partir des quatre petits films acquis pour le locuteur F.H. représentant au total un peu moins de 700 images, en ne considérant que les images correspondant aux instants où de la parole est effectivement produite (cf. paragraphes 2.1 et 2.2).

La mandibule est modélisée en premier parce que la langue et la lèvre inférieure lui sont attachées du point de vue anatomique. Puisqu'il s'agit d'un objet rigide (bidimensionnel dans le cas des images aux rayons X) il suffit d'une translation et d'une rotation pour parfaitement définir sa position, c'est-à-dire 3 paramètres. Une analyse en composantes principales a été appliquée sur ces 3 paramètres afin de réduire leur nombre. La variance expliquée par la première composante est de 75%, et de 19% pour la seconde.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

Il est donc possible de ne retenir que deux paramètres pour approcher la mâchoire, tout en conservant une très bonne fidélité. Le mouvement de la mandibule est soustrait des contours de la langue et de la lèvre inférieure avant d'analyser les déformations de ces articulateurs. L'analyse en composantes principales est ensuite appliquée aux contours de la langue, des lèvres, du vélum, et du larynx. La langue, qui est l'articulateur le plus important, est analysée depuis la racine jusqu'au plancher buccal. Avec 6 composantes linéaires, la précision de la reconstruction atteint 0.52 mm sur les images analysées.
Ce modèle ne peut être appliqué tel quel aux images d'un autre locuteur pour des raisons de taille et d'orientation relative de la cavité buccale par rapport au pharynx. Une procédure d'adaptation a donc été développée pour prendre en compte ces aspects (Laprie & Busset 2011) avant d'évaluer la précision de la reconstruction pour un autre sujet.
Nous avons mesuré la précision de la reconstruction sur les données cinéradiographiques, représentant au total 520 images que Maeda (1979) avaient utilisées pour construire son modèle articulatoire. Ces images présentent l'intérêt de correspondre à une locutrice, ce qui signifie que l'adaptation est moins directe que s'il s'était agi d'un locuteur, puisque la longueur globale du conduit vocal, comme le rapport entre les tailles du pharynx et de la bouche, changent entre locuteur et locutrice. Par ailleurs, ces images correspondent essentiellement à des voyelles, ce qui implique que la comparaison ne prend en compte que la forme de la langue sans interférence avec la forme du palais. L'erreur de reconstruction est de 0,55 mm en moyenne, ce qui montre que le modèle construit sur un locuteur peut être utilisé pour copier avec succès la forme de la langue d'un autre locuteur, pour ce qui concerne les voyelles.

Le second aspect étudié concerne l'approximation de la forme du conduit vocal au voisinage de la constriction, parce qu'elle joue un rôle acoustique décisif pour la production des consonnes. L'étude que nous avons réalisée ne concerne les données d'un seul locuteur (celles du locuteur M.M. décrites au paragraphe 2.2) car nous ne voulions pas réaliser une double adaptation, portant à la fois sur des paramètres de taille des cavités pharyngale et buccale d'une part, et sur la forme du palais et des incisives supérieurs d'autre part.

Il est beaucoup plus difficile d'approcher la forme de la langue dans le cas des consonnes pour plusieurs raisons. D'abord, les formes que prend la langue pour réaliser les fortes constrictions que requiert une consonne sont nettement plus complexes que celles des voyelles parce que la précision géométrique nécessaire est plus grande. Par ailleurs, la langue ne se déforme plus sous le seul effet des muscles qui la contrôlent, mais aussi sous l'effet du contact avec le palais dur, ou les incisives supérieures pour les occlusives. Les modes de déformation calculés par l'ACP vont donc indument intégrer l'effet de « collision » entre la langue et le palais, avec pour conséquence une forme de langue souvent trop lisse et l'impossibilité de réaliser la fermeture complète du conduit vocal, au lieu de constriction des occlusives non labiales. Nous avons donc exploré plusieurs pistes pour atteindre une meilleure précision de reconstruction.

La première consiste à donner plus de poids statistique aux consonnes, lors de l'application de l'analyse en composantes principales. Nous avons ainsi repris une idée utilisée communément en phonétique articulatoire qui consiste à définir des cibles articulatoires virtuelles au-delà du palais (voir par exemple (Birkholz et al. 2011)). Dans notre cas, nous avons modifié les contours de la langue de la base de données en cas de contact avec le palais, en prolongeant le contour au-delà du palais. Le contour jusqu'au palais est conservé sans modification, et l'extension au-delà du palais préserve la forme de la langue et ne dépasse jamais 10mm. Cette modification n'est pas suffisante à elle seule car le nombre de contours de langue correspondant à des consonnes est faible, face à celui des voyelles même si le corpus est phonétiquement équilibré. Nous avons donc donné plus de poids aux images des consonnes dans le calcul de la variance. Il est apparu que la contribution des contours de la consonne /l/ est essentielle, car ces contours présentent une forme de la pointe de la langue très marquée. Leur contribution a donc été particulièrement renforcée de manière à accroître le poids des modes de déformation représentant la pointe de la langue.

L'analyse en composantes principales construit une base de vecteurs utilisés comme modes de déformation. Les poids de chaque vecteur pour représenter le contour de la langue sont obtenus en

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

projetant le contour représenté par un vecteur de points sur les vecteurs de la base. Le contour de la langue, reconstruit à partir des composantes linéaires en utilisant ces poids, donne la meilleure approximation globale possible de la langue. Dans le cas de très fortes constrictions, voire de contacts avec le palais, il est important que le contour de la langue soit précis dans cette région. Il faut donc un critère d'évaluation qui accorde plus d'importance aux points de la langue au lieu de constriction. Pour cette raison, nous utilisons une distance pondérée et la procédure d'optimisation de Powell (Flannery et al. 1993) pour trouver les poids de chaque mode de déformation. Une seconde raison nécessite cette optimisation en lieu et place de la projection. Les contours de la langue reconstruits peuvent dépasser le palais, ce qu'il faut bien sûr interdire. Un algorithme de détection et résolution de collisions (entre la langue et le palais) est donc utilisé pour « couper » la partie de la langue qui dépasse éventuellement le palais. Le contour de la langue n'est donc plus une simple combinaison linéaire des vecteurs de base et nécessite en conséquence cette étape d'optimisation.

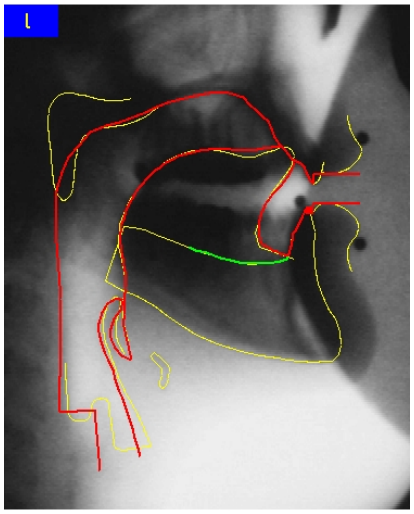

Figure 4 : Exemple d'approximation du contour de la langue d'un /l/.
Les contours jaunes sont ceux qui ont été extraits du film. L'approximation du conduit vocal est représentée par la ligne rouge.

Nous avons testé le modèle articulatoire sur les 1015 images de la base de données présentée plus haut. La Figure 4 illustre l'approximation de la forme de la langue dans le cas d'un /l/ dont le lieu de constriction est approché correctement par le modèle articulatoire. Les résultats sont présentés dans la Table 1. Comme attendu, le même nombre de composantes que celui utilisé pour la première base de données donne une erreur de reconstruction plus importante (0,830 mm au lieu de 0,52mm) mais l'erreur au niveau de la constriction reste du même ordre de grandeur (0,567 mm), ce qui est très encourageant dans la perspective de la synthèse articulatoire.

Nombre de composantes	Erreur de reconstruction moyenne globale en mm	Erreur de reconstruction au lieu de constriction en mm
12	0,307 (0,226)	0,205 (0,146)
8	0,366 (0,347)	0,236 (0,239)
6	0,830 (0,599)	0,567 (0,575)

Table 1 : Erreur de reconstruction de la langue en fonction du nombre de composantes linéaires utilisées. L'écart type de l'erreur est indiqué entre parenthèses.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

5. Conclusion et perspectives : « faire parler les données cinéradiographiques »

La synthèse articulatoire et plus généralement les techniques de simulation de la production de parole font le lien entre les espaces articulatoire et acoustique. Elles permettent en particulier d'étudier les mécanismes d'anticipation et de coordination gestuelle du point de vue de leur impact acoustique. Cette solution proposée dès les années soixante-dix par Cooper et al. (1977) a rapidement été abandonnée parce que les techniques de simulation et les données articulatoires étaient insuffisantes pour assurer la pertinence des résultats. Dans notre cas nous disposons de données géométriques bidimensionnelles, du signal acoustique d'origine, et après dépouillement des films du contour de chacun des articulateurs, et enfin de modèles d'approximation de la forme du conduit vocal.

Nous avons étudié la possibilité de synthétiser de la parole à partir des images aux rayons X. Dans un premier temps nous sommes partis des contours extraits des images, et non pas de leur approximation géométrique, afin de valider les outils de synthèse articulatoire.

Un film aux rayons X fournit une image toutes les 20 ou 40 ms selon que la cadence d'enregistrement est 50 ou 25 Hz. Cela est bien sûr insuffisant pour la plupart des articulations consonantiques. La durée du burst d'une occlusive, par exemple, varie de quelques millisecondes à trente millisecondes environ. Il faut donc élaborer une stratégie de suréchantillonnage des images qui produit des transitions géométriques rapides et cohérentes du point de vue la production de la parole. La seconde difficulté est de coordonner l'évolution de la forme du conduit vocal avec les différentes sources d'excitation situées soit à la glotte quand les plis vocaux vibrent, ou en aval d'une constriction forte provoquant un bruit de turbulence dans le cas des fricatives ou occlusives (Scully 1987).

La segmentation phonétique du signal de parole enregistré lors de l'acquisition du film aux rayons X et la fréquence fondamentale calculée sur le signal d'origine fournissent les informations temporelles nécessaires pour assurer ces deux niveaux de coordination.

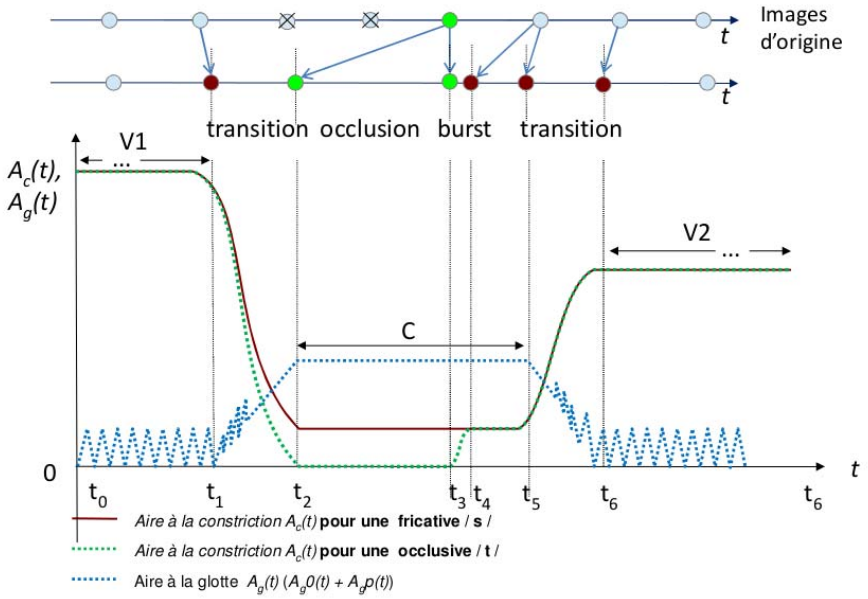

Figure 5 : Schéma de coordination temporel pour la une occlusive ou une fricative.
Les deux lignes du haut représentent les images d'origine, avant et après réorganisation temporelle. La partie basse de la figure montre l'évolution en fonction du temps de l'aire à la constriction (trait brun pour la fricative, traits pointillés verts pour l'occlusive) et l'aire à la glotte (traits pointillés bleus). L'aire à la glotte correspond aux mouvements des plis vocaux qui vibrent avant et après la consonne ce qui explique sa variation rapide et périodique. Les images d'origine peuvent être dupliquées comme celle marquée par un disque vert pendant la constriction, ou encore celles marquées par un disque brun

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

qui servent à définir les transitions rapides. Les images marquées par une croix ne sont pas utilisées.

Nous nous sommes appuyés sur le travail de Maeda (1982,1996) pour élaborer les schémas temporels. La Figure 5 illustre le schéma construit pour une occlusive ou une fricative sourde. Nous avons réorganisé les instants d'apparition des images, en les copiant, ou éventuellement en les supprimant, de manière à assurer une évolution temporelle de l'aire à la constriction pertinente. Une fois le schéma temporel élaboré la fonction d'aire est calculée à partir des contours et la simulation acoustique proprement dite peut être lancée. Soulignons que la détermination de la fonction d'aire à partir des contours est une étape importante puisqu'elle conditionne directement les résultats de la simulation acoustique. Nous avons d'ores et déjà synthétisé avec succès des logatomes de la forme VCV (voir Figure 6) issus de notre base de données et nous travaillons actuellement sur la synthèse de consonnes néessitant une plus grande précision géométrique, notamment les consonnes rhotiques pour lesquelles la proximité entre le vélum et l'arrière ou le dos de la langue joue un rôle essentiel.

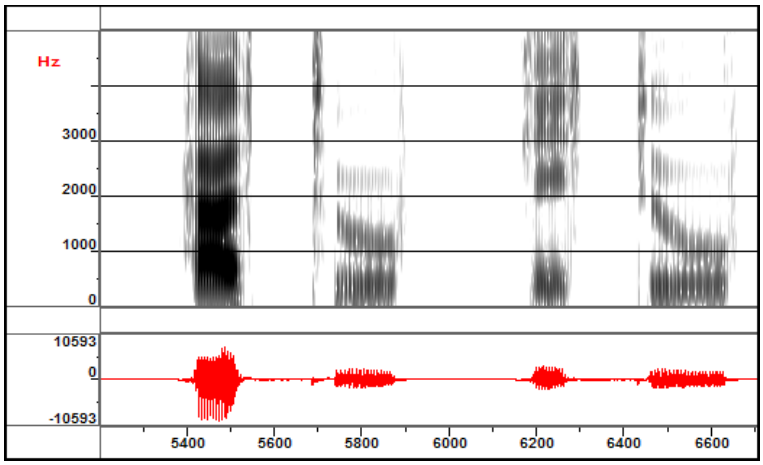

Figure 6 : Spectrogramme des logatomes /atu/ et /itu/ resynthétisés à partir desdonnées cinéradiographiques.

Dans certains cas l'image bidimensionnelle seule ne suffit pas, ou du moins ne donne pas une information géométrique pertinente. C'est particulièrement vrai pour le vélum dont la partie basse se termine en pointe au centre du haut la cavité pharyngale et ne ferme pas le conduit vocal comme on pourrait le croire sur l'image aux rayons X.
Les images IRM (Imagerie par Résonance Magnétique) sont donc très utiles pour déterminer la géométrie précise de cette région du conduit vocal. Nous avons donc acquis plusieurs séries d'images tridimensionnelles du conduit vocal afin de construire un modèle géométrique plus précis.
Toute la méthodologie présentée dans cet article (constitution de bases de données, dépouillement, construction de modèles articulatoires, synthèse à partir des images) est beaucoup plus générale que le seul traitement de données cinéradiographiques. Nous enregistrons actuellement des films IRM à une cadence d'environ 25 Hz qui donneront qui donneront lieu au même type de travaux que ceux décrit dans cette communication. Pour l'instant il s'agit de films bidimensionnels acquis pour une coupe médiosagittale mais nous étudions la possibilité d'acquérir ces films pour plusieurs coupes sagittales de manière à disposer d'une information géométrique plus complète.

6. Bibliographie

(Abry & Boë 1981) Abry C. and Boë L.J., Sur les notions d'opposition et de contraste dans l'élaboration d'un corpus, Bulletin de l'Institut de Phonétique de Grenoble, vol 10, pp. 1-12

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801344

(Beautemp et al. 2001) Beautemps D., Badin P., and Bailly G., Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," Journal of the Acoustical Society of America, vol. 109, no. 5, pp. 2165–2180

(Birkholz 2011) Birkholz P, Kröger BJ, Neuschaefer-Rube C. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. IEEE Transactions on Audio, Speech and Language Processing, 19(5), pp. 1422-1433

(Flannery et al. 1993) Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical Recipes, 2nd Edition Cambridge University Press

(Hardcastle et al. 1996) Hardcastle W.J., Vaxelaire B., Gibbon F. Hoole P. and Nguyen N.,Tongue kinematics in /NO/ clusters and singleton /N/: A combined EMA/EPG study, Proceedings of the Australian Speech Science and Technology Conference, Adelaide

(Jallon & Berthommier 2009) Jallon J. F. and Berthommier F., A semi-automatic method for extracting vocal-tract movements from x-ray films, Speech Communication, vol. 51, no. 2, pp. 97–115, 2009.

(Laprie & Busset 2011) Laprie Y., Busset J. Construction and evaluation of an articulatory model of the vocal tract. – In: 19th European Signal Processing Conference - EUSIPCO-2011

(Laprie & Berger 1996) Laprie Y. and Berger M.-O., Towards automatic extraction of tongue contours in x-ray images," in Proceedings of International Conference on Spoken Language Processing 96, vol. 1, Philadelphia (USA), pp. 268–271.

(Laprie & Busset 2011) Laprie Y., Busset J. – "Construction and evaluation of an articulatory model of the vocal tract". – In: 19th European Signal Processing Conference - EUSIPCO-2011.

(Laprie et al. 2013) Laprie Y, Loosvelt M. , Maeda S., Sock R., Hirsch F., "Articulatory copy synthesis from cine X-ray films", Proc. Of InterSpeech 2013

(Maeda 1979) Maeda S., Un modèle articulatoire de la langue avec des composantes linéaires, Actes 10èmes Journées d'Etude sur la Parole, Grenoble, pp. 152-162

(Maeda 1982) Maeda S., A digital simulation of the vocal tract system, Speech Communication, Vol. 1, 199-229.

(Maeda 1990) Maeda S., "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. Hardcastle and A. Marchal, Eds. Amsterdam: Kluwer Academic Publisher, pp. 131–149.

(Maeda 1996) Maeda S., "Phoneme as concatenable units: VCV synthesis using a vocal tract synthesizer", in "Sound Patterns of Connected Speech: Description, Models and Explanation", Arbeitsberichte des Institut für Phonetik und digitale Spachverarbeitung der Universität Kiel (31), 145-164.

(Marchal & Cavé 2009) L'imagerie médicale pour l'étude de la parole, Alain Marchal et Christian Cavé (réd.). Hermes Science Publications

(Scully 1987) Scully C., Linguistic units and units of speech production. Speech Communication 6(2), pp. 77-142

(Sock R. 2001) La Théorie de la Viabilité en production-perception de la parole. Psychologie et Sciences Humaines (Mardaga., p. 285-316). Liège: Keller D. Durafour JP. Bonnot JF & Sock.

(Sock et al. 2011) Sock R. , Hirsch H., Laprie Y., Perrier p., Vaxelaire B., Brock G., Bouarourou F., Fauth C., Ferbach-Hecker V., Ma L., Busset J., Sturm J. – An X-ray database, tools and procedures for the study of speech production. – In : 9th International Seminar on Speech Production (ISSP 2011) , L. Ménard, S.R. Baum, V.L. Gracco, D.J. Ostry (réd.), pp. 41–48.

(Thimm &Luettin 1999) Thimm G. and Luettin J., Extraction of articulators in xray image sequences, in Proc. EUROSPEECH, Budapest, pp. 157–160.