# Extraction d'un modèle articulatoire à partir d'une analyse tri-directionnelle de cinéradiographies d'un locuteur

Martine Cadot*, Yves Laprie**

*LORIA
martine.cadot@loria.fr,
http://www.loria.fr/~cadot
**LORIA
yves.laprie@loria.fr,
http://www.loria.fr/~laprie

**Résumé.** Nous expérimentons ici un processus d'identification des sons de la parole à partir d'images, et non d'enregistrements sonores comme habituellement réalisé. Il s'agit de l'analyse de séquences cinéradiographiques d'une personne prononçant plusieurs phrases. Des difficultés se présentent. La première, technique, est que ces données proviennent d'images annotées en plusieurs lieux, temps, et de manière semi-automatique ou manuelle. La deuxième, représentationnelle, est que les mouvements des articulateurs pendant la parole (langue, mâchoire, etc.) se situent dans un espace-temps complexe du fait des interdépendances mécaniques multiples et dynamiques. Le modèle articulatoire le plus connu est celui de Maeda (1990), obtenu à partir d'Analyses en Composantes Principales faites sur les tableaux de coordonnées des points des articulateurs d'un locuteur en train de parler. Nous proposons ici une analyse tri-directionnelle du même type de données, après leur transformation en une suite de tableaux de distances. Nous validons notre modèle par la prédiction des sons prononcés, qui s'avère presque aussi bonne que celle du modèle acoustique, et même meilleure quand on prend en compte la coarticulation.

## 1 Introduction

Pour parler, le locuteur met en mouvement un ensemble complexe d'articulateurs (voir figure 1) : la mâchoire qu'il ouvre plus ou moins, la langue à laquelle il fait prendre de nombreuses formes et positions, les lèvres qui lui permettent de laisser l'air s'échapper plus ou moins brutalement ou d'allonger le conduit vocal, etc. Un modèle articulatoire de la parole consiste en la formalisation de ces mouvements à l'origine de l'émission de parole. À terme, son pilotage devrait permettre de synthétiser la parole. En attendant, il permet déjà de réaliser des expériences de synthèse d'un signal acoustique à partir des paramètres du modèle articulatoire.

Citons trois exemples de son utilisation potentielle par les enseignants/chercheurs de l'équipe MultiSpeech du Loria :

— aider les étudiants en "Français Langue Étrangère" à prononcer certains sons inexistants dans leur langue maternelle en pointant les articulateurs critiques,
— aider les malentendants à suivre une émission vidéo en ajoutant dans un coin de l'image une "tête parlante" qui articule les phrases en même temps qu'elles sont prononcées,
— aider les enfants présentant des troubles du langage (Piquard-Kipffer et al., 2010).

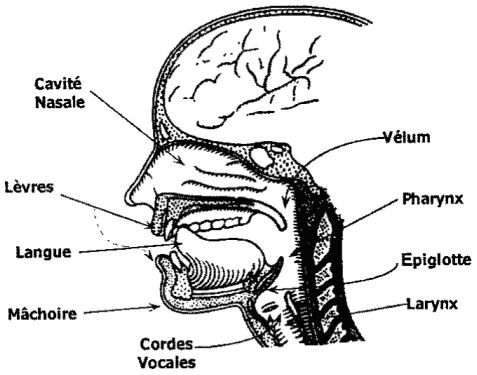


FIG. 1 – *Schéma de l'anatomie du conduit vocal (coupe sagittale d'après Flanagran 1972).*

Nous exposons dans cet article comment nous avons extrait un modèle articulatoire à partir de cinéradiographies d'un locuteur en train de parler. Chaque cinéradiographie est stockée sous la forme d'un fichier son et d'une suite de fichiers d'images sagittales de la tête (voir figure 2).

FIG. 2 – *Une radiographie avant annotation. Pour annoter le contour de la langue (courbe rouge de la figure 8 et courbe rose de la figure 5), on la repère comme la région ombrée superposée sur la machoîre et qui se déplace par rapport aux images juste avant ou juste après.*

Nous appuyant sur le travail fait pour extraire à partir d'un petit jeu de données (que nous appellerons **corpus 1**) un modèle articulatoire à 7 paramètres comme celui de Maeda (1990),

nous avons testé un modèle articulatoire de nature différente sur ces données, puis nous l'avons adapté à des données plus importantes (que nous appellerons **corpus 2**).

**Le modèle articulatoire à 7 paramètres.** Ce modèle a été conçu par Maeda (1990). Il a construit son modèle articulatoire (voir figure 3) au moyen d'ACP (Analyses en Composantes Principales) sur les coordonnées des points marqués sur chaque radiographie. Puis il l'a évalué acoustiquement en comparant les sons prononcés par le locuteur aux sons produits par un synthétiseur de sons piloté par son modèle.

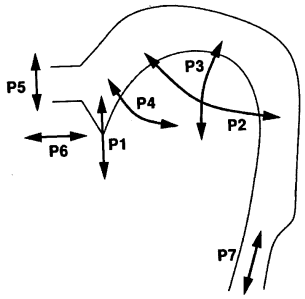


FIG. 3 – *Modèle articulatoire à 7 paramètres de Maeda.*

Dans la première partie de ses travaux de thèse, Busset (2013) a construit un modèle articulatoire à 7 paramètres également (voir figure 4) en reprenant les grandes lignes de la construction du modèle articulatoire de Maeda à partir d'une cinéradiographie d'un locuteur prononçant quelques séquences de mots. Dans ce cadre, elle a d'abord participé à l'annotation manuelle des radiographies avant de contribuer à développer des outils informatiques permettant d'automatiser en partie cette tâche délicate (voir figure 8). Elle a ensuite élaboré une stratégie d'enchaînement optimal d'ACP successives sur les contours des articulateurs. Puis les 7 paramètres ont été validés en utilisant un synthétiseur d'une manière proche de celle de Maeda.

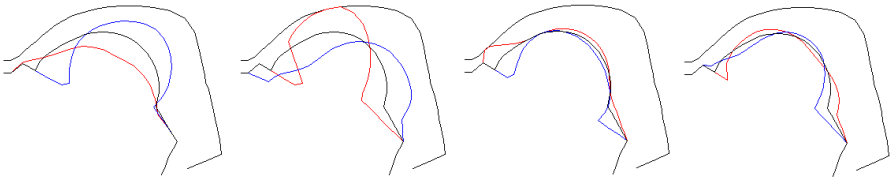

FIG. 4 – *Les 4 paramètres de la langue décrits dans Busset et Cadot (2013) ; la courbe noire au centre pour la forme neutre du conduit vocal, en rouge et en bleu les deux formes extrêmes.*

**Le modèle articulatoire issu d'analyse tri-directionnelle** La nouveauté de notre démarche consiste en l'utilisation d'une méthode d'analyse tri-directionnelle pour extraire le modèle de données et de méthodes d'apprentissage supervisé pour le valider. Les 2 premières directions

correspondent aux dimensions de l'image 2D, la troisième est le temps, représenté par la succession des images. Notre modèle exprime le lien entre une série de variables "explicatives" qui sont les facteurs obtenus à partir de l'analyse tri-directionnelle des positions des articulateurs repérées dans les images, et la variable son pour laquelle chaque son est représenté par un caractère alphanumérique de codage phonétique (voir dans le tableau 1 la liste des phrases prononcées et leur codage). Ses performances sont mesurées en comparant les sons réels aux sons prédits à l'aide des modèles d'apprentissage. Puis nous comparons les performances de notre modèle à celles du modèle acoustique formé des coefficients cepstraux obtenus à partir des fichiers audio (voir partie 4.1).

Cet article reprend une communication de Cadot et Laprie (2014) faite dans le cadre de l'atelier "Fouille de données complexes" de la conférence EGC'2014 à Rennes sur les données du corpus 2. La démarche relatée dans cet article complète et enrichit celle d'un travail précédent de Busset et Cadot (2013) qui reprenait les données du corpus 1 annotées lors de la première partie de sa thèse (Busset, 2013). Sur la figure 5, on peut voir les nombreux points marqués sur chaque image, chaque série d'une couleur différente correspondant à un articulateur différent. Les points indiqués par un losange sont ceux qui ont été mis dans le tableau de données.

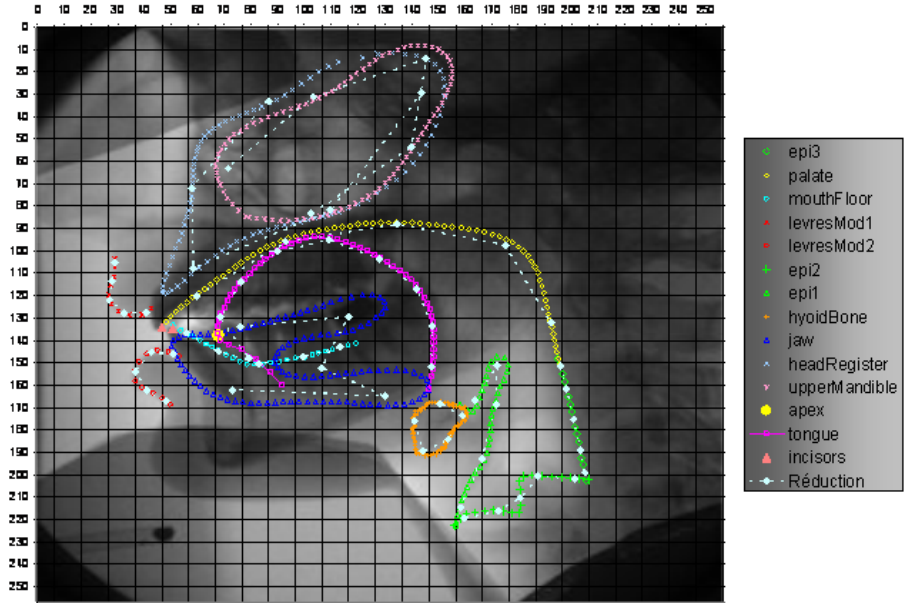


FIG. 5 – *Données d'une image du premier corpus. En carrés rose vif le contour de la langue avec un gros point jaune pour l'apex (bout de la langue), en petits ronds rouges la lèvre inférieure. En bleu pâle les losanges pleins et joints par des pointillés sont les points conservés pour l'analyse.*

Le corpus 1 était de petite taille, avec peu de sons différents, et une certaine répétition des phrases : il était divisé en 4 parties notées de H1 à H4, nous avons traité les parties H3 et H4

qui ne comportaient que 6 logatomes /aku/, /iku/, /uku/, /atu/, /itu/, /utu/ prononcés rapidement dans H4 et plus lentement dans H3. Le modèle obtenu était relativement simple, ce qui a permis à Julie Busset de le valider en interprétant un à un ses éléments (voir les éléments de validation du modèle pour la lèvre inférieure dans H4 en figure 6). Ce premier essai étant probant, nous sommes passés à l'échelle supérieure avec les données plus riches du corpus 2 et un modèle plus complexe, validé de façon automatique. Nous renvoyons le lecteur souhaitant plus de détails sur ces deux corpus et la motivation de leur création à Laprie et al. (2014b) et Laprie et al. (2014a), et pour une description plus détaillée du traitement des données du premier corpus à Busset et Cadot (2013) et Busset et Cadot (2012).

Notre exposé comporte quatre parties. Nous décrivons dans la première partie la construction du jeu de données numériques, dans la deuxième partie la méthodologie d'extraction du modèle articulatoire que nous avons choisie, dans la troisième l'évaluation par apprentissage automatique de ce modèle, et nous faisons le bilan dans la dernière.

# 2 Description et signification des données

Les données sont recueillies dans le but de construire un modèle, nous décrivons donc dans une première sous-section le type de modèle que nous visons et ses limites. Dans la deuxième sous-section, nous exposons le recueil des données, et dans la dernière sous-section comment elles ont été transformées en les données numériques que nous avons traitées.

## 2.1 Le modèle articulatoire à 7 paramètres et ses limites

Dans le modèle articulatoire de la parole construit par Maeda (voir figure 3), le conduit vocal, zone interne allant du pharynx aux lèvres, est schématisé en coupe sagittale, ainsi que les déformations que lui font subir les articulateurs. Elles sont résumées en 7 mouvements, qui sont les 7 paramètres du modèle de Maeda : P1, la mâchoire qui va de haut en bas, P2, P3, P4 la langue qui se déforme dans 3 directions, P5 et P6 les lèvres qui s'ouvrent et se ferment, s'avancent et reculent, et P7 pour le mouvement du larynx. En affectant différentes valeurs à ces 7 paramètres, on obtient différentes formes du conduit vocal, dont on déduit différents sons à l'aide du synthétiseur. Ce modèle est une représentation réaliste de la parole car ce sont les déformations du conduit vocal qui modulent les résonances du conduit vocal et produisent les différents sons de la parole.

Toutefois, avec ses 7 paramètres, ce modèle est une simplification forte de la réalité. Ses limites sont les suivantes :

1. les 7 paramètres sont insuffisants pour la production de certaines consonnes, notamment les "fricatives" (comme "s", "z", ...) ;
2. le modèle est dans un espace à 2 dimensions, celles de l'image, alors que le conduit vocal se meut dans un espace à 3 dimensions ;
3. il n'y a pas de correspondance univoque entre un son et une forme du conduit vocal, les sons peuvent s'articuler différemment selon le contexte phonétique et selon les personnes ;
4. les mouvements des différents articulateurs ne sont pas indépendants entre eux, il y a des contraintes mécaniques qui les lient.

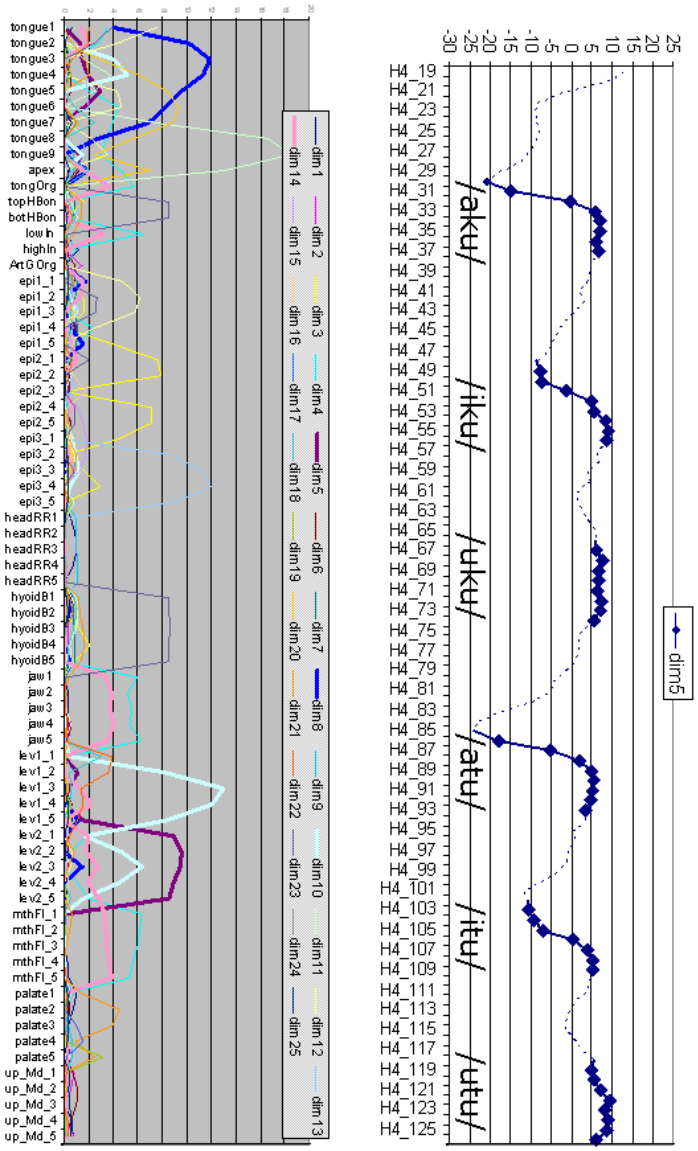


FIG. 6 – *Résultat du traitement 3-way MDS du premier corpus, partie H4 : à gauche les coordonnées des points des articulateurs dans les 25 dimensions conservées, avec en gras les 4 dimensions intervenant pour la lèvre inférieure (points* lev2_1 *à* lev2_5*) ; à droite les coordonnées des images successives dans la dimension 5, avec en bleu les points correspondant à du son, et le mot prononcé indiqué en dessous.*

Pour nous affranchir de la limite 4, nous avons construit notre modèle sur les distances entre points des articulateurs et non comme celui de Maeda, sur les coordonnées de ces points. Nous avons repoussé autant que possible la limite 1 en augmentant la taille du modèle, mais des contraintes techniques nous ont empêchés de repousser la limite 2 (les cinéradiographies sont de dimension 2) ainsi que la limite 3 (pas assez de données pour remplacer les sons par des séquences de 2 ou 3 sons, pas assez de locuteurs enregistrés et filmés).

Les contraintes techniques pourront certainement être dépassées ultérieurement avec les nouveaux corpus que nous constituons actuellement. Ce ne sont pas des cinéradiographies, qui ne se font plus à cause de la dangerosité des rayonnements ionisants, mais des données issues d'IRM (Imagerie par Résonance Magnétique), auxquelles nous pourrons appliquer les mêmes genres de traitement, et même les étendre à des images 3D.

## 2.2 La vidéo à l'origine des données

Une série de radiographies de la tête a été réalisée à 50 images par seconde pendant qu'un locuteur prononçait quelques phrases courtes, recopiées dans le tableau 1. Pour plus de détails sur ces données, on peut se reporter à Sock et al. (2011).

| Transcription orthographique | Transcription phonétique |
|---|---|
| Il a pas mal. | ilapamal@ |
| Les attablés. | lezatable |
| Très acariâtre. | tʁɛzakaʁjatʁ |
| Il zappe pas mal. | ilzappamal@ |
| Des abat-jour. | dezabaʒuʁ |
| Il l'a datée. | illadate |
| Crabe bagarreur. | kʁabbagaʁœʁ |
| Trois sacs carrés. | tʁwasakkaʁe |
| Pas de date précise. | paddatpʁesizœ |
| Blague garantie. | blaggaʁãti |
| Nous palissons. | nupalisõ |
| Il a pourri. | ilapuʁi |
| Couds ta chemise. | kutaʃmiz@ |
| Elle a tout faux. | ɛlatufo |
| Pour tout casser. | puʁtukase |

TAB. 1 – *Corpus 2 : les 15 phrases successives prononcées par le locuteur et leur transcription phonétique*

Dans le tableau 1 les sons sont codés par les 27 symboles phonétiques suivants :

@ a ã b d e ɛ œ f g i j ʒ k l m n o õ p ʁ s ʃ t u w z

Puis des contours ont été dessinés sur ces images afin de représenter au mieux les articulateurs (voir figure 5), et de repérer le plus finement possible leurs mouvements. On dispose aussi de la correspondance entre sons et images, un code de son étant associé à chaque image prise pendant que le locuteur parlait, et un code "silence" aux images prises en dehors de la

parole. Dans la mesure où le son est un fichier indépendant des fichiers images, cette mise en correspondance n'est pas donnée mais calculée, comme précisé dans Laprie et al. (2014a).

Seuls les contours des images ont été utilisés dans la phase de construction du modèle, les sons ayant servi dans la phase d'évaluation.

## 2.3 L'obtention des données numériques

La qualité du modèle extrait des données dépend de la qualité des données elles-mêmes, issues de l'annotation des images.

**La complexité du travail d'annotation** L'extraction des contours, en particulier celui de la langue qui est l'articulateur à la fois le plus déformable et le plus mobile, a suscité un certain nombre de travaux, par exemple Thimm et Luettin (1999), Laprie et Berger (1996), Jallon et Berthommier (2009). 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. Jallon 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.

Ainsi se dessine la stratégie de détourage des contours adoptée lors du développement du logiciel **Xarticulators** (Laprie et Busset, 2011). 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.

On a fait apparaître dans la figure 7 les contours annotés ainsi que les aides à l'annotation qui ont été développées. Ce sont ces données, de grande qualité, que nous avons utilisées ici.

| déformable | nb. min | nb. max | indéformable | nb. points |
|---|---|---|---|---|
| voile du palais | 69 | 107 | os hyoïde | 30 |
| épiglotte | 52 | 74 | plancher de la langue | 19 |
| larynx | 46 | 79 | palais | 39 |
| lèvre supérieure | 11 | 35 | mâchoire inférieure | 50 |
| lèvre inférieure | 13 | 45 | mâchoire supérieure | 23 |
| langue | 18 | 44 | | |

TAB. 2 – *Nombre de points des 11 contours dessinés dans les 1021 images : à gauche les articulateurs déformables, et à droite les indéformables*

**Un nombre de points différent par image** Les données se présentent sous la forme de 11 contours formés d'un nombre fixe de points pour les contours indéformables, et variable pour les contours déformables, comme indiqué dans le tableau 2. On dispose pour chaque image des coordonnées 2D des points de chaque contour.

# 3 Extraction du modèle articulatoire

Dans cette section, nous exposons d'abord comment Maeda a utilisé des ACP pour l'extraction de son modèle articulatoire, et les inconvénients de ce type d'analyse. Nous détaillons ensuite la méthode factorielle tri-directionnelle, que nous avons choisi d'utiliser, et enfin le traitement des données.

## 3.1 Utilisation d'Analyses en Composantes Principales

Pour extraire son modèle articulatoire, Maeda (1990) a utilisé des ACP à partir de données similaires. Par exemple, pour obtenir les 3 paramètres de la langue, P2, P3 et P4 (voir figure 3), le contour de la langue a été intersecté avec une grille dite semi-polaire s'adaptant à la forme du conduit vocal. Puis les coordonnées des points ont été disposées séquentiellement dans un tableau ayant autant de lignes que d'images, et les trois premières composantes d'une ACP ont donné les paramètres recherchés. De nombreuses variantes de ce modèle ont été proposées par la suite, portant essentiellement sur la création des tableaux de données soumis à des ACP. Par exemple pour le corpus 1, Laprie et Busset (2011) ont utilisé une grille polaire adaptative (en bleu dans la figure 8) ainsi que des coordonnées curvilignes pour mieux placer les points de la langue qui ont formé un premier tableau soumis à une ACP. Et avant de procéder à des ACP sur les tableaux de données des articulateurs suivants, ils ont supprimé les liaisons avec les articulateurs précédents par soustraction des corrélations.

Les ACP ont montré leur efficacité dans la construction du modèle articulatoire, mais aussi leurs limites. Il s'est avéré difficile d'extraire de nombreuses composantes d'un seul tableau de données : quand il contenait les points d'un seul articulateur, on atteignait la quasi-totalité de la variance expliquée avec 1, 2 et au maximum 3 composantes, et regrouper tous les points dans un même tableau ne permettait pas de dépasser les 7 composantes du modèle princeps. De plus à l'examen des contributions des points aux composantes, on a constaté qu'un nombre non négligeable de points se retrouvaient "éclatés", leur abscisse et leur ordonnée contribuant à des composantes différentes, ce qui rendait délicate leur interprétation. Ces problèmes sont inhérents à la méthode d'analyse choisie et nous ont conduits à en chercher une plus adaptée à la fois aux données et au type de modèle recherché.

## 3.2 La méthode de *3-way MDS*

Le MDS (MultiDimensional Scaling, et en français positionnement ou échelonnement multidimensionnel) fait partie des méthodes d'analyses factorielles des données, et est particulièrement adapté à l'analyse des données de type *dissimilarités* [1], non mesurables objectivement,

1. On appelle *dissimilarité* une *distance affaiblie*, notamment elle n'est pas astreinte à vérifier l'*inégalité triangulaire* qui impose pour tout triplet de points $(x, y, z)$ la relation $d(x, z) \leq d(x, y) + d(y, z)$.

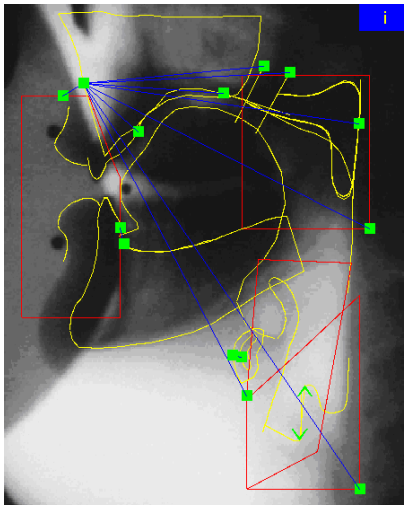


FIG. 7 – *Tracé semi-automatique des contours d'une image du corpus 2 à l'aide du logiciel "Xarticulators" : traits de construction en rouge et en bleu, balises en vert, contours obtenus en jaune.*

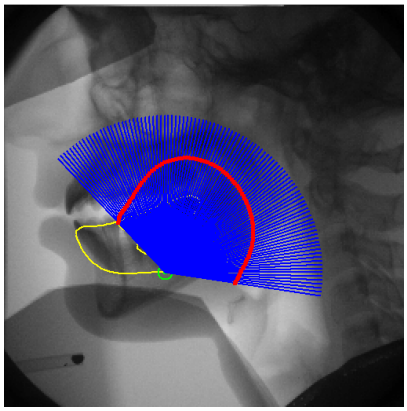

FIG. 8 – *Détermination de 100 points du contour de la langue par intersection de la courbe rouge (contour annoté) et du faisceau des 100 demi-droites bleues (dessinées automatiquement à partir de la balise verte et des 2 extrémités de la courbe rouge).*

correspondant à des impressions ressenties, recueillies en réponse à des questions comme celle-ci :

*Du point de vue de l'acidité, quelle distance ressentez-vous entre :*
— *la limonade A et la limonade B ?*
— *la limonade B et la limonade C ?*
— *la limonade A et la limonade C ?*

Le MDS a fait l'objet de nombreux articles et ouvrages. Pour plus de détails sur le MDS, nous renvoyons les lecteurs intéressés à Borg et Groenen (1997).

**Les principes du MDS.** La méthode d'analyse MDS est capable de positionner des objets dans un espace de dimension $p$ de telle sorte que leurs distances deux à deux soient les plus proches possible de leurs dissimilarités initiales. L'écart entre les deux tableaux de distances est appelé STRESS, et l'ajustement du modèle aux données est d'autant meilleur qu'il est proche de zéro.

La formulation mathématique à la base des MDS est la suivante : si pour deux objets numérotés par $i$ et $j$, on note $\delta_{ij}$ leur dissimilarité initiale, $d_{ij}$ leur distance dans l'espace euclidien de dimension p, et f une fonction monotone de ces dissimilarités, le *stress* brut est donné par la formule

$$Stress = \sum_{1 \leq i < j \leq n} (f(\delta_{ij}) - d_{ij})^2$$

C'est par le choix de la fonction f que le *stress* est minimisé.

**La mise en œuvre du MDS.** À l'initialisation, les points sont placés dans une position quelconque de l'espace de dimension $p$, et l'algorithme consiste à les déplacer un à un pour faire diminuer le *stress*, jusqu'au moment où sa valeur est jugée suffisamment petite. Comme pour l'ACP, le nombre de dimensions $p$ doit être fixé au départ par l'utilisateur. L'exemple traditionnel de cette méthode est son application à un tableau des distances en kilomètres entre des paires de villes. La méthode produit une carte 2D qui s'avère assez proche de la réalité, mis à part quelques distorsions liées au relief montagneux.

**Les différents types d'échelonnements tri-directionnels.** Le MDS a été étendu pour pouvoir prendre en considération une pile de tableaux de dissimilarités au lieu d'un seul. Ce qui est le cas par exemple si on obtient autant de tableaux de dissimilarités que de sujets d'un groupe de $q$ sujets. Aux deux dimensions du tableau s'ajoute une troisième dimension, qui est le sujet, d'où le nom *3-way MDS*, donné à la famille de méthodes MDS permettant de traiter les tableaux de dissimilarité en prenant également en compte les différences entre sujets. INDSCAL (pour INdividual Difference SCALing) est une des méthodes développées par Carroll et Chang (1970) pour traiter ce type de données. Sa méthode se situe entre deux variantes 3-way extrêmes du MDS (voir Carroll, 1972, page 106) :

— "identically same configuration" ; même espace $X$ de dimension $p$ pour chaque sujet $k$ avec une fonction $f_k$ différente pour chacun,
— "idiosyncratic configuration" ; un espace $X_k$ propre à chaque sujet $k$, avec la même fonction $f$ pour tous.

Pour le détail de ces méthodes de 3-way MDS, nous renvoyons les lecteurs intéressés au chapitre 21 de Borg et Groenen (1997).

## 3.3 Avantages du MDS par rapport à l'ACP pour notre problème

Nous avons fait état dans la partie 3.1 des problèmes rencontrés lors des ACP appliquées aux données du corpus 1. L'utilisation des MDS sur des tableaux de distances entre points fait que le problème d'éclatement des points par séparation de leur abscisse et de leur ordonnée, comme indiqué dans la partie 3.1, ne peut plus se produire. Ainsi le MDS respecte mieux la nature spatiale des données d'images que l'ACP. Le deuxième problème rencontré était le petit nombre de dimensions pertinentes obtenues par ACP. Pour éviter ce problème, au lieu de calculer pour chaque image les distances entre points des articulateurs nous avons calculé pour chacun de ces points les distances entre toutes ses positions dans les images, autrement dit dans sa trajectoire, et obtenu autant de tableaux de distances que de points d'articulateurs, et nous avons appliqué un MDS sur la pile de ces tableaux de distances.

## 3.4 Le traitement des données

Compte tenu de la taille des données à traiter nous avons choisi le logiciel R à toutes les étapes de ce travail : création des tableaux de données en entrée, création de tableaux de résultats en sortie, traitement des données "cepstrales" de référence, 3-way MDS, évaluation quantitative des résultats par arbres de décision et M-SVM. Pour l'examen des données d'entrée comme de sortie (parcours des tableaux de données pour contrôle des valeurs erronées, croisement de variables et graphiques), nous avons utilisé le tableur Excel pour ses possibilités de manipulations interactives.

**La création de la pile de tableaux.** Nous n'avons pris qu'un nombre réduit de points par articulateur, 3 pour les articulateurs indéformables sauf le palais qui en a 4, 3 pour chaque lèvre, 4 pour l'épiglotte, 5 pour le velum comme pour le larynx et 10 pour la langue. Pour réaliser cela, les contours de chaque articulateur ont été découpés en autant de parties que de points souhaités, en veillant à découper plus finement les parties les plus déformables, ou susceptibles de contacts. Par exemple pour la langue, les zones autour de la racine, du dos et de l'apex (pointe), ont des points plus rapprochés que le reste de la langue. Puis on a calculé dans chaque partie la moyenne des coordonnées des points, ce qui a donné les coordonnées du point cherché. Nous sommes arrivés ainsi à un nombre de 46 points pour 11 articulateurs des 1021 images ayant été annotées. On a ensuite calculé pour chaque point les distances euclidiennes entre ses positions sur les images prises deux à deux, ce qui a donné un tableau images$\times$images d'un demi-million environ de distances (en ne considérant que la moitié du tableau, car il est symétrique). Ce sont ces 46 tableaux de distances que nous avons analysés avec la méthode de 3-way MDS.

**La méthode 3-way MDS.** Nous avons choisi la fonction *smacofIndDiff* du package SMACOF (de Leeuw et Mair, 2009), avec le paramètre "idioscal" correspondant à la variante "idiosyncratic" décrite dans le troisième paragraphe de la partie 3.2, qui s'est avérée moins gourmande en mémoire vive que la méthode INDSCAL que nous avions utilisée avec succès pour les données 5 fois plus petites du corpus 1 analysé auparavant (Busset et Cadot, 2013). Et nous nous sommes limités à 200 itérations afin de réduire le temps d'exécution. Malgré cela, nous avons dû rapidement migrer d'un ordinateur 32 bits, double-processeur, 2 Go de RAM vers un ordinateur 64 bits, quadri-processeur, 8 Go de RAM. Nous avons pu ainsi obtenir les positions

des quelque 1000 images dans des espaces allant de 2 dimensions à 18. Pour obtenir les 18 dimensions, les 200 itérations demandées ont duré plus de 48 heures.
Nous appelons "modèles articulatoires" les matrices formées de 1021 lignes (une par image) et $q$ colonnes représentant les coordonnées de ces images dans l'espace à $q$ dimensions ($q$ allant de 2 à 18).

# 4 Évaluation de notre modèle

Nous attendons de notre modèle articulatoire qu'il soit une représentation la plus fidèle possible de la parole, l'idéal étant qu'il soit en sortie de synthétiseur aussi intelligible que l'est l'enregistrement audio de la parole pour des personnes n'ayant pas de problème d'audition.

Nous choisissons de mesurer sa qualité par son taux de reconnaissance des sons tels que définis dans la partie 2.2, et nous le comparons au taux de reconnaissance d'un "modèle acoustique", c'est-à-dire d'une matrice de coefficients acoustiques calculés aux instants successifs de l'enregistrement audio. Chaque évaluation sera donc faite successivement avec les deux modèles. Elle se fera d'abord de façon asynchrone, puis en prenant en compte des décalages temporels.

## 4.1 Extraction du modèle acoustique

Les coefficients *cesptraux*, obtenus par transformée de Fourier inverse du spectre de parole, fournissent les paramètres d'un modèle acoustique centré sur le conduit vocal, largement utilisé en traitement automatique de la parole (Busset, 2013). Nous avons extraits les 20 premiers coefficients cepstraux avec le package tuneR (Ligges, 2011) à partir de l'enregistrement audio réalisé pendant que les radiographies étaient acquises. C'est la matrice obtenue de 2119 lignes et 20 colonnes qui représente le "modèle acoustique" que nous souhaitons comparer à notre "modèle articulatoire" formé des 1021 images et $q$ colonnes où $q$ peut prendre une valeur allant de 2 à 18.

## 4.2 Mesure asynchrone de la qualité de reconnaissance des sons

**Transformation des données pour la mesure.** Pour le modèle articulatoire, nous avons adjoint aux matrices MDS à $q$ dimensions une colonne avec les sons correspondant à chaque image. Puis nous avons retiré toutes les lignes correspondant à des silences ainsi que celles correspondant à des sons très rares dans les données ("w", "j" et "n") c'est-à-dire présents 6 fois ou moins. Les matrices n'ont plus que 732 lignes et la colonne sons n'en a plus que 24 différents.

Pour le modèle acoustique, nous avons également adjoint la colonne des sons correspondants. Nous avons ensuite retiré les lignes de silence et celles des 3 sons retirés dans les matrices de facteurs MDS, afin de pouvoir mieux comparer les capacités de discrimination de sons des deux modèles. La matrice de coefficients cepstraux ainsi obtenue a 1450 lignes.

Pour comparer la qualité de représentation des sons de ces deux modèles, nous allons utiliser une même méthode sur les 2 matrices. Comme la deuxième matrice a environ deux fois plus de lignes que la première pour chaque son prononcé, si elle aboutit à une meilleure qualité de représentation des sons, on ne saura pas si c'est dû à la nature du modèle ou à

la taille de la matrice. Afin de rendre la comparaison plus pertinente, nous avons créé une troisième matrice de même taille que la première matrice à partir de la deuxième matrice, en lui retirant environ une ligne sur deux [2]. Ainsi la matrice "articulatoire" et la nouvelle matrice "acoustique" ont approximativement le même nombre de lignes par son prononcé.

**Les mesures d'évaluation utilisées.** Nous cherchons ici à comparer le taux de reconnaissance des sons du modèle articulatoire à celui du modèle acoustique. Pour cela nous utilisons des méthodes permettant de prédire la variable son à partir d'une vingtaine de variables numériques, qui sont les facteurs MDS pour le modèle articulatoire et les coefficients cepstraux pour le modèle acoustique. Notre but n'étant pas de trouver la meilleure méthode de discrimination pour chacun de ces 2 modèles, mais de comparer les résultats produits sur les deux modèles par des méthodes de discrimination, nous ne cherchons pas à optimiser le choix des méthodes de discrimination et de leurs paramètres pour obtenir le meilleur taux de reconnaissance possible, mais seulement pour les adapter au mieux aux données. Notamment le fait d'avoir une variable catégorielle à 24 modalités réduit drastiquement le nombre de méthodes possibles.

Nous avons utilisé le package Rpart (Therneau et Atkinson, 2015) pour obtenir des arbres de décision comme définis par Quinlan (1986), qui ont l'avantage de fournir des règles explicites de prédiction.

Nous avons complété notre étude en utilisant des méthodes de discrimination par SVM (Support Vector Machine) présentes dans le package KernLab (Karatzoglou et al., 2004). Les SVM sont une méthode de discrimination entre 2 classes. Pour discriminer les 24 sons, nous avons utilisé les M-SVM (*M-* pour multiples) qui en sont une extension à plus de 2 classes. Il en existe plusieurs variantes, décrites dans Lee et al. (2004). Dans ce package, nous avons pu utiliser les 4 types de M-SVM suivants :

— C-svc C classification
— C-bsvc bound-constraint svm classification
— spoc-svc Crammer, Singer, native multi-class [3]
— kbb-svc Weston, Watkins, native multi-class

Parmi les options, nous avons choisi la plus courante, qui est le noyau gaussien (option *rbfdot*) et la possibilité de ne saisir qu'un seul paramètre, C, que nous avons fait varier entre 1 et 100. Pour l'apprentissage, nous avons procédé par validation croisée : nous avons découpé au hasard les lignes de données en quatre parties de tailles équivalentes, l'*ensemble d'entrainement* correspondant à trois de ces parties, et *l'ensemble test* à la quatrième, et nous avons mis dans une colonne les sons prédits pour chaque ligne de la partie test. La colonne de sons prédits a été remplie au bout des 4 itérations pendant lesquelles chaque partie est devenue à son tour l'ensemble test.

Dans le graphique de la figure 9 sont représentées les 5 méthodes choisies pour les matrices de facteurs MDS, avec un nombre de facteurs allant de 4 à 18. On voit que la méthode par arbre de décision n'est pas très stable. L'examen des valeurs prédites montre que tous les sons ne sont pas prédits, le nombre de sons prédits augmentant avec le nombre de facteurs, sans jamais atteindre 24, qui est le nombre total de sons. Parmi les méthodes M-SVM, ce sont

2. Nous avons créé non pas une mais deux matrices de 732 lignes à partir de celle de 1450 lignes, obtenues en retirant différemment une ligne sur deux environ : l'une en retirant plutôt la première ligne après le changement de sons, et l'autre en retirant plutôt la deuxième ligne après le changement de sons. Leurs résultats s'étant avérés très proches, nous n'avons transcrit ici que les résultats de la première version.

3. Pour les non "native multi-class", chaque classe est comparée à l'ensemble de toutes les autres.

les "faux" M-SVM, c'est-à-dire utilisant des stratégies de type “une classe contre toutes les autres” qui donnent les meilleures prédictions, non seulement pour C=30, mais aussi pour les autres valeurs de C, comme l'établit la figure 10. L'examen des logs d'apprentissage sur les 4 sous-ensembles montre que les chutes de performances portent parfois sur un seul ensemble test.

On peut conclure que le taux de prédiction croît quand le nombre de facteurs MDS passe de 4 à 14, puis il stagne autour de 0,81 pour plus de 14 facteurs.

Les mêmes méthodes appliquées aux matrices de données cepstrales donnent des résultats de qualité similaire : par rapport aux matrices MDS, les taux de prédiction sont dans l'ensemble moins bons (maximum 0,76, pour un nombre de coefficients compris entre 16 et 20) avec les matrices cepstrales de taille 732, mais meilleurs (maximum 0,86 pour 20 coefficients) avec celles de taille 1450 (voir lignes en traits pleins, figure 13).

## 4.3 Prise en compte des enchaînements temporels

Jusqu'ici, l'aspect de séquentialité temporelle n'a pas été pris en compte, pas plus que la coarticulation. La séquentialité nous avait permis de valider par interprétation experte le traitement du corpus 1 (voir partie droite de la figure 6), nous avons privilégié des méthodes d'apprentissage automatique pour la validation du traitement du corpus 2. Néanmoins les distances ayant été calculées suivant l'ordre des images, il serait intéressant de chercher dans les résultats du traitement ce qu'est devenu cet aspect séquentiel.

La coarticulation est le phénomène qui lie le mouvement des articulateurs à l'enchaînement des sons : la position des articulateurs pour un son donné dépend non seulement du son à prononcer, bien sûr, mais également de la position des articulateurs juste avant, donc des sons précédents, et juste après, donc des sons suivants. L'association entre un son et une position des articulateurs, donc une image, est “floue”, ce que nous avons choisi de représenter par un décalage possible entre l'image du son et le son lui-même.

Pour la capter, nous avons utilisé des décalages dans le son prédit de la façon suivante : si le son prédit pour une image correspond au son de l'image précédente, le décalage est noté -1, s'il correspond à celui de 2 images plus loin, il est noté +2. Et pour un ensemble de décalage donné, par exemple (-2, -1, 0, 1), on juge que le son prédit est juste s'il est le même qu'attendu pour la même image, pour l'image précédente, ou celle encore avant, ou pour l'image suivante.

**Les résultats.** On voit dans la figure 11 que le taux de prédictions exactes augmente ainsi jusqu'à plus de 0,94, soit une amélioration de 0,13 quand on autorise des décalages allant jusqu'à 3 images avant ou après, ce qui montre que notre modèle rend compte de la coarticulation présente dans les données.

On a retrouvé un phénomène de décalage également avec les matrices cepstrales, mais d'une ampleur moindre, comme on peut le voir dans la figure 12 pour les matrices de 1450 lignes. Dans la figure 13, les scores sans décalage ont été représentés par un trait plein, et un trait en pointillés représente les décalages de -3 à 3. On voit que les décalages font gagner moins de 0,07 en moyenne, que ce soit pour les matrices de 732 lignes (en bleu) ou celles de 1450 lignes (en rouge).

Quelle que soit l'explication de ce phénomène pour le modèle acoustique, à savoir la persévérance et l'anticipation du son (Bonnet, 1986) ou tout simplement l'imprécision de la seg-

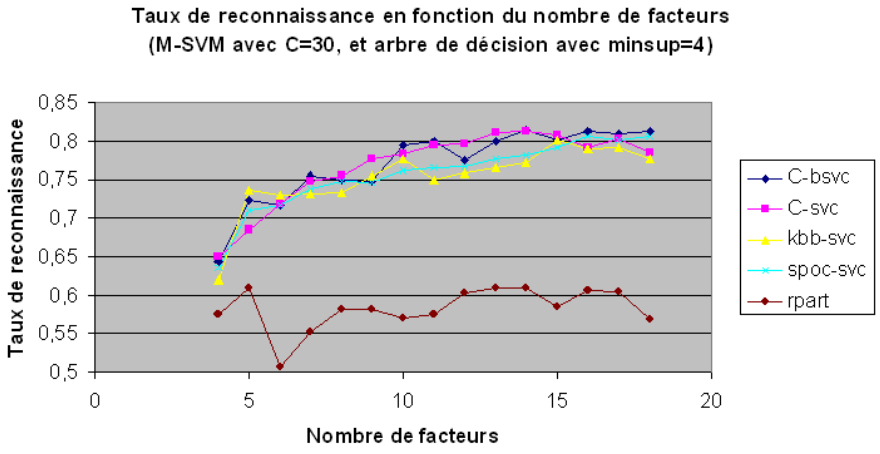


FIG. 9 – *Taux de reconnaissance du modèle articulatoire par différentes méthodes de M-SVM et par arbre de décision*

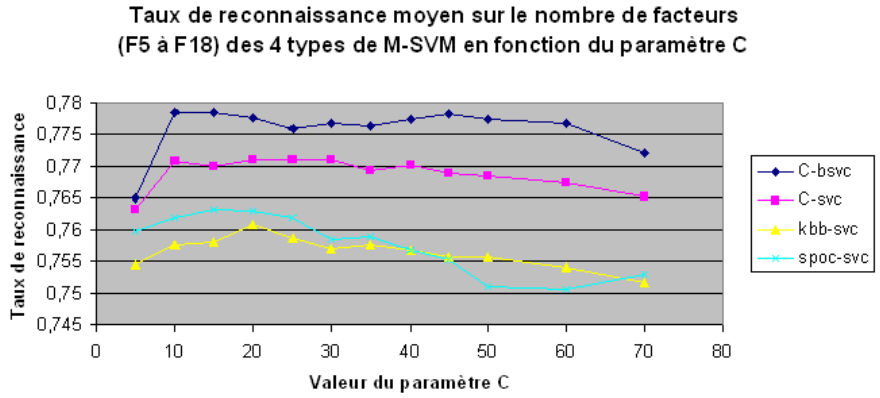


FIG. 10 – *Modèle articulatoire, taux de reconnaissance en fonction du paramètre C selon 4 types de M-SVM.*

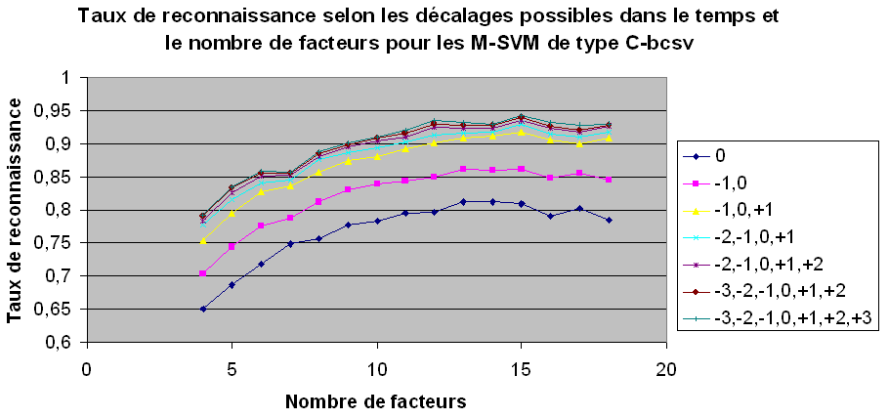


FIG. 11 – *Modèle articulatoire sur 732 images, taux de reconnaissance avec divers décalages pour M-SVM type C-bcsv.*

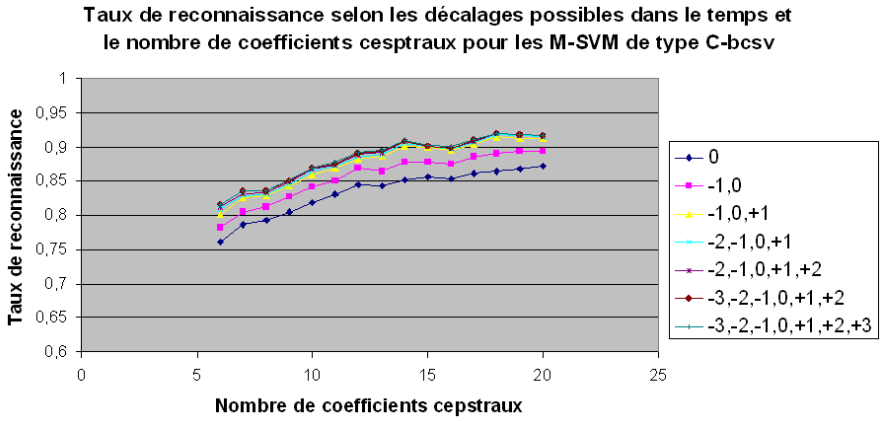


FIG. 12 – *Modèle acoustique sur 1450 lignes, taux de reconnaissance avec divers décalages pour M-SVM type C-bcsv.*

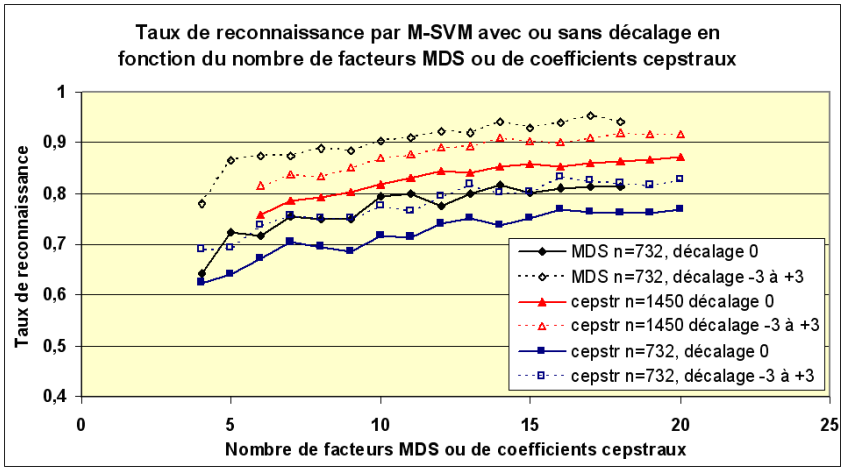


FIG. 13 – *Modèles articulatoire et acoustique, taux de reconnaissance avec décalages par M-SVM type C-bcsv.*

mentation, pour le modèle articulatoire la coarticulation y joue certainement un rôle non négligeable.

Au final, en prenant en compte les décalages, le modèle articulatoire prédit mieux les sons, toutes choses égales par ailleurs, que le modèle acoustique.

# 5 Discussion, conclusion, perspectives

Nous venons d'exposer comment nous avons utilisé une méthode 3-way MDS dans le but d'extraire un modèle articulatoire des images successives d'une personne en train de parler.

Nous avons rencontré un certain nombre de difficultés lors de son application :

1. le choix du nombre de points à prendre par image pour que les programmes tiennent en mémoire, sans trop de perte d'informations sur les mouvements des articulateurs,
2. la difficulté d'interprétation des dimensions MDS, due au remplacement de INDSCAL par IDIOSCAL pour les mêmes raisons de place insuffisante en mémoire,
3. la distribution déséquilibrée des sons qui gêne le fonctionnement de certains discriminateurs,
4. le choix un peu artificiel de sons séparés pour la reconnaissance d'un modèle articulatoire : prononcer "la" est-il équivalent à prononcer "l" puis "a" ?
5. où placer la prise en compte de l'aspect de séquentialité temporelle dans l'analyse : dans la construction du modèle ou dans son évaluation ?

Malgré ces difficultés nous arrivons à un modèle articulatoire basé sur des images animées qui contient au moins autant d'informations sur les sons - sinon plus en se plaçant toutes choses égales par ailleurs - que le modèle acoustique basé sur un enregistrement audio. Ces

bons résultats nous invitent à continuer dans cette voie, en tentant d'améliorer dans différentes directions :

— revoir la programmation des fonctions R utilisées pour solutionner les points 1 et 2, en prenant plus de points par image, et en réutilisant INDSCAL au lieu d'IDIOSCAL,
— essayer de changer de méthode de discrimination pour répondre aux points 3 et 4,
— essayer d'incorporer les dépendances temporelles dans le modèle 3-way MDS pour le point 5, ou inversement une méthode MDS dans un modèle de séries temporelles.

# Références

Bonnet, C. (1986). *Manuel de psychophysique*. Collection U. Paris : Armand Collin.

Borg, I. et P. Groenen (1997). *Modern Muldtidimensional Scaling*. Springer series in Statistics. New York: Springer-Verlag.

Busset, J. (2013). *Inversion acoustique articulatoire a partir de coeffcients cepstraux*. Thèse de doctorat, Université de Lorraine.

Busset, J. et M. Cadot (2012). Démêler les actions des articulateurs en jeu lors de la production de parole avec le logiciel C.H.I.C. : Analyse de séquences de radiographies de la tête. In *6th International Conference Implicative Statistic Analysis - A.S.I. 6 - 2012*, Caen, France, pp. 291–305.

Busset, J. et M. Cadot (2013). Fouille d'images animées : cinéradiographies d'un locuteur. In *Atelier FOuille de données Spatio-Temporelles et Applications - FOSTA*, Toulouse, France, pp. 1–12.

Cadot, M. et Y. Laprie (2014). Méthodologie 3-way d'extraction d'un modèle articulatoire de la parole à partir des données d'un locuteur. In *Atelier Fouille de Données Complexes des 14èmes Journées Francophones "Extraction et Gestion des Connaissances"*, Rennes, France, pp. 1–12.

Carroll, D. (1972). Individual differences and multidimensional scaling. In R. Shepard, A. Romney, et S. Nerlove (Eds.), *Multidimensional Scaling*, Volume 1: Theory, pp. 105–155. Seminar Press.

Carroll, D. et J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrica 35*, 283–319.

de Leeuw, J. et P. Mair (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software 31*(3), 1–30.

Jallon, J. F. et F. Berthommier (2009). A semi-automatic method for extracting vocal-tract movements from x-ray films. *Speech Communication 51*(2), 97–115.

Karatzoglou, A., A. Smola, K. Hornik, et A. Zeileis (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

Laprie, Y. et M.-O. Berger (1996). Towards automatic extraction of tongue contours in x-ray images. In *International Conference on Spoken Language Processing 96*, Philadelphia, USA, pp. 268–271.

Laprie, Y. et J. Busset (2011). A curvilinear tongue articulatory model. In *International Seminar on Speech Production 2011 - ISSP'11*, Montréal, Canada.

Laprie, Y., R. Sock, B. Vaxelaire, et B. Elie (2014a). Comment faire parler les images aux rayons X du conduit vocal ? *SHS Web of Conferences 8*, 14.

Laprie, Y., B. Vaxelaire, et M. Cadot (2014b). Geometric articulatory model adapted to the production of consonants. In *10th International Seminar on Speech Production (ISSP)*, Köln, Germany.

Lee, Y., Y. Lin, et G. Wahba. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association 99*(465), 67–81.

Ligges, U. (2011). tuneR: Analysis of music. Technical report, Department of Statistics, University of Dortmund, Germany.

Maeda, S. (1990). Compensatory articulation during speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. Volume 4, pp. 131–149. Kluwer Academic Publisher, Amsterdam.

Piquard-Kipffer, A., D. Lelarge, L. Pierron, et F. Monnay (2010). Création de livres numériques pour enfants présentant des troubles du langage. Démonstration de la création de livres numériques à la conférence IHM 2010 au Luxembourg du 20 au 23 septembre 2010.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*, 81–106.

Sock, R., F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Ferbach-Hecker, L. Ma, J. Busset, et J. Sturm (2011). An X-ray database, tools and procedures for the study of speech production. In *Proceedings of the 9th International Seminar on Speech Production (ISSP2011)*, Montréal, Canada, pp. 41–48.

Therneau, T. M. et E. J. Atkinson (29 juin 2015). An introduction to recursive partitioning using the rpart routines. Documentation du package r, Mayo Foundation.

Thimm, G. et J. Luettin (1999). Extraction of articulators in xray image sequences. In *EUROSPEECH*, Budapest, pp. 157–160.

# Summary

For several reasons it is difficult to analyze the sequences of radiographs of a person talking. The first is technical: these data are images annotated in several places, times, in a semi-automatic or manual way. The second is representational: the movements of the articulators during speech (tongue, jaw, etc.) are complex to describe because of multiple mechanical and dynamic interdependencies. When speaking, a speaker sets in motion a complex set of articulators: the jaw which opens more or less, the tongue which takes many shapes and positions, the lips that allow him to leave the air escaping more or less abruptly, etc.. The best-known articulary model is the one of Maeda (1990), derived from Principal Component Analysis made on arrays of coordinates of points of the articulators of a speaker talking. We propose a 3-way analysis of the same data type, after converting tables into distances. We validate our model by predicting spoken sounds, which prediction proved almost as good as the acoustic model, and even better when coarticulation is taken into account.