# Utilisation d'une grille polaire adaptative pour la construction d'un modèle articulatoire de la langue

*Julie Busset*[1]

[1]LORIA/CNRS UMR7503
615, rue du Jardin Botanique. 54600 Villers-lès-Nancy, France
Julie.Busset@loria.fr

## ABSTRACT

The construction of articulatory models from medical images of the vocal tract, especially X-ray images, relies on the application of an articulatory grid before deriving deformation modes via some factor analysis method. One difficulty faced with the classical semi-polar grid is that some tongue contours do not intersect the grid what gives rise to incomplete input vectors, and consequently poor tongue modeling in the front part of the mouth cavity which plays an important role in the articulation of many consonants. First, this paper describes preparation of data, i.e. drawing or tracking articulator contours, compensation of head movements and the construction of the adaptive polar grid. Then, the results of the principal component analysis are presented and compared with those obtained with the semi-polar grid.

## 1. INTRODUCTION

La possibilité de générer avec le modèle articulatoire et la simulation acoustique les mêmes sons que ceux prononcés par un locuteur (ou au moins les fonctions de transfert du conduit vocal pas trop éloignées de celles observées) constitue l'hypothèse sous-jacente de la méthode d'analyse par synthèse de l'inversion acoustique articulatoire. Le modèle articulatoire, et par conséquent sa construction est la clé de voute de l'inversion. L'une des approches les plus fructueuses est l'application d'une « analyse factorielle » à un corpus d'images radiographiques du conduit vocal. Bien que cette technique d'imagerie médicale soit abandonnée à cause des risques pour la santé liés à l'exposition aux rayons X, il existe un grand nombre de corpus disponibles. De plus, les images radiographiques fournissent une « vue en deux dimensions » du conduit vocal à une vitesse d'acquisition qui permet de capturer tous les gestes articulatoires ce qui n'est pas le cas de l'imagerie par résonance magnétique.

Le but de l'analyse factorielle est de trouver les principaux modes de déformation des articulateurs de la parole. La forme du conduit vocal est obtenue à partir des images radiographiques de façon automatique ou non. Les contours ainsi obtenus sont transformés soit en vecteurs unidimensionnels, soit en concaténant les abscisses et les ordonnées, soit encore en appliquant une grille. La grille semi-polaire proposée par Maeda [Mae90] pour construire son modèle est probablement la plus répandue. Cependant, elle présente deux faiblesses. Tout d'abord, il y a des contours de la langue qui n'intersectent pas toutes les lignes de la grille ce qui donne des vecteurs incomplets pour l'analyse factorielle. C'est le cas des voyelles arrières comme le [u] illustré dans la figure 1. Pour contourner cette difficulté, nous avons comme possibilité soit d'étendre artificiellement le contour de la langue, soit de retirer ces contours de l'analyse factorielle avec la conséquence d'affaiblir le mode de déformation correspondant. Beautemps et al [Bea01] ont proposé une solution intéressante qui consiste à utiliser une grille dynamique dans la région de l'apex. L'avancement de la grille est ajouté comme paramètre articulatoire supplémentaire. La deuxième faiblesse de la grille semi-polaire est la nature des points utilisés dans l'analyse factorielle. [Ste02] explique que les points analysés devraient être des points de repère anatomiques complétés avec des points de repère mathématiques ou des pseudo points de repère pour obtenir des modes de déformations pertinents. Ce n'est pas le cas des grilles articulatoires traditionnelles parce que les points utilisés sont les intersections d'un objet déformable (la langue) avec les lignes d'une grille statique. Par conséquent ils ne représentent pas les mêmes points de chair au cours du temps. Nous proposons donc l'utilisation d'une grille polaire adaptative dont les extrémités sont la racine de la langue et l'apex de la langue (voir Figure 3) et qui s'adapte automatiquement à tous les contours de langue.

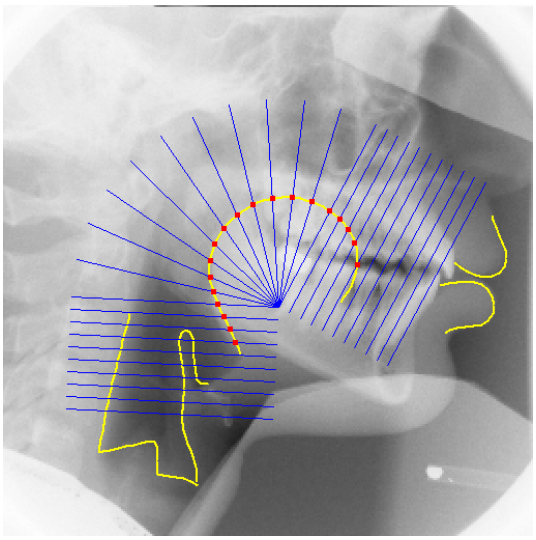

**Figure 1:** *Contour de la langue correspondant à la voyelle [u] intersecté avec la grille semi-polaire*

Nous commencerons par présenter les différentes méthodes utilisées pour obtenir les contours et les outils développés. A l'aide de ces contours nous construirons notre grille polaire qui dépend de la position de la mâchoire inférieure. Les différentes stratégies pour obtenir les défor-

mations linéaires seront étudiées, évaluées et comparées au cas de la grille semi-polaire.

## 2. PRÉPARATION DES DONNÉES

### *2.1. Corpus*

Le corpus utilisé pour évaluer les différentes présentations des données et les stratégies pour appliquer les différentes ACP (Analyses en Composantes Principales) est composé de quatre petits films soit 946 images radiographiques ($256 \times 256$ pixels) tous du même locuteur. Seules les images correspondant à de la parole, c'est-à-dire 672 images, ont été exploitées pour notre analyse.

### *2.2. Contours des articulateurs*

Le traçage des contours à la main est très fastidieux et plusieurs travaux ont été consacrés au suivi automatique des contours articulatoires [Lap96] [Thi99] [Fon06]. Nous avons développé un logiciel, appelé « Xarticulators », proposant plusieurs outils de suivi.

Les structures rigides, comme la mâchoire inférieure, peuvent être suivies par corrélation à partir d'une image de référence. Dans ce cas particulier, l'énergie photométrique de la mâchoire est suffisamment forte pour négliger celle de la langue, d'autant plus que la région suivie a été choisie pour minimiser l'influence de la langue. Ce simple suivi s'avère très efficace pour suivre la mâchoire inférieure, la partie supérieure du crâne (pour compenser le mouvement de la tête) et même l'os hyoïde (s'il ne va pas trop haut et n'intersecte pas d'autres organes ayant une énergie photométrique élevée). Le suivi nous fournit les paramètres de déplacement, c'est-à-dire la rotation et la translation.

Les structures se déformant au cours du temps, qui ont un contour visible et qui ne sont pas recouvertes par d'autres organes, c'est-à-dire les lèvres, le larynx et l'épiglotte, ont été suivies avec l'algorithme de Fontecave et Berthommier [Fon06]. Les contours sont tracés à la main pour une série d'images clé (environ 10% du nombre total d'images à traiter) et approchés par des courbes B-splines avec un nombre constant de points de contrôle. L'idée générale est de choisir une région rectangulaire dans l'image incluant l'objet à suivre. Les images sont découpées pour garder uniquement cette région afin d'enlever l'influence des autres organes. Elles sont indexées par rapport à la distance de leurs cœfficients DCT (Discrete Cosine Transform) à ceux des images clés. Pour toutes les images, les trois images clé les plus proches sont conservées avec leur distance aux images analysées. Finalement, le nouveau contour (donné par les points de contrôle) est la moyenne pondérée des contours des trois images clés les plus proches. L'évaluation visuelle du suivi montre que sur quelques images les contours obtenus sont très éloignés de la position souhaitée parce que ces images sont trop éloignés des images clés ; elles ont donc été ajoutées comme images clé.

Le contour de la langue est plus difficile à suivre automatiquement puisqu'il peut y avoir un ou deux contours visibles en fonction de la forme de la langue. Dans beaucoup d'images, la langue présente un sillon localisé dans le plan médio-sagittal qui donne des contours intéressants. Les deux bords supérieurs de la langue peuvent donner un autre contour (exceptionnellement deux si la langue n'est pas symétrique du tout). La figure 2 illustre l'existence de deux contours. Le contour que l'on considère est celui dans le plan médio-sagittal. Une autre difficulté est la présence des dents dans la cavité buccale qui cachent le contour de la langue. Les experts complètent souvent le contour de la langue en dessinant un contour convexe ; probablement parce que cette forme est plus naturelle. Cependant, les images aux ultrasons et les images IRM montrent que le contour de la langue est concave pas seulement pour les formes de langue rétroflexes mais pour beaucoup d'autres formes de langue moins extrêmes. Bien qu'il existe des algorithmes automatiques ou semi-automatiques, les difficultés présentées ci-dessus nous ont incités à tracer les contours de la langue à la main pour garantir leur pertinence. Cependant, l'interface graphique de Xarticulators offre des outils pour rendre ce travail plus facile.

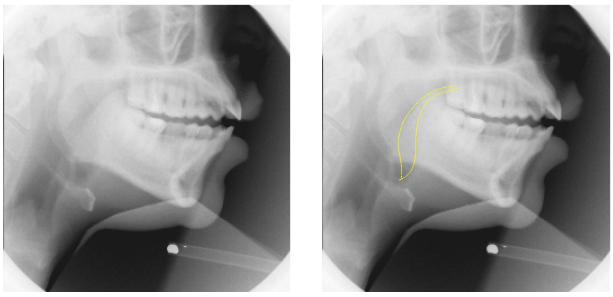

**Figure 2:** *A gauche : image radiographique avec deux contours visibles pour la langue, A droite : la même image avec les deux contours dessinés*

**Compensation des mouvements de la tête :** Pendant l'acquisition, le locuteur a bougé légèrement la tête malgré la contention qui lui était imposée. Il est donc nécessaire de compenser ce mouvement. Le mouvement de la tête est obtenu par le suivi de la partie supérieure du crâne par corrélation. La seule précaution est de choisir une région (utilisée comme un masque de corrélation), qui n'intersecte pas le filtre en aluminium pendant la séquence. Le suivi fournit les paramètres de déplacement de la tête, c'est-à-dire la rotation et la translation. Ce déplacement est soustrait de tous les contours des articulateurs avant toute analyse.

### *2.3. Construction de la grille adaptative*

Nous avons choisi de construire une grille polaire adaptative qui s'adapte automatiquement au contour de la langue. Les extrémités de la grille adaptative sont la racine de la langue et l'apex. Le centre de la grille correspond à une balise attachée à la mâchoire. La position du centre de la grille est calculée pour chaque image à partir des paramètres de déplacement de la mâchoire. Ces trois points sont utilisés pour définir une grille polaire. Les lignes de la grille sont espacées régulièrement (voir la figure 3).

## 3. STRATÉGIES POUR LA CONSTRUCTION DU MODÈLE

Nous avons appliqué l'analyse factorielle aux données obtenues à partir de la grille adaptative et de la grille semi-polaire.

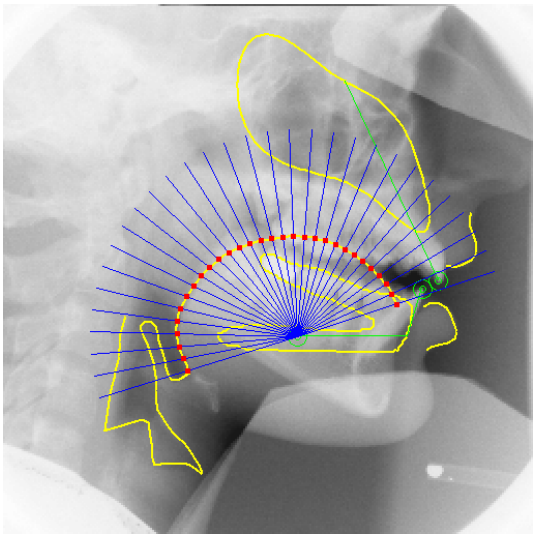

**Figure 3:** *La grille adaptative appliquée à un contour de langue, les régions utilisées pour compenser le mouvement de la tête et suivre la mâchoire inférieure, ainsi que des balises (centre de la grille et incisives) et les différents articulateurs.*

### *3.1. Grille polaire adaptative*

Comme dans le modèle de Maeda le principal paramètre est la mâchoire. Dans la plupart des modèles, le mouvement de la mâchoire est approché par une translation mesurée par la distance entre les incisives supérieures et inférieures. La première raison est que la dispersion de la position de l'incisive inférieure n'est pas trop éloignée d'une droite. Une deuxième raison pratique est que l'os de la mâchoire n'est pas toujours disponible ce qui empêche une approximation plus précise. Dans notre cas, les trois paramètres de déplacement de la mâchoire sont connus, c'est-à-dire l'angle de rotation et les deux coordonnées de la translation, puisque la mâchoire est suivie comme un tout.

**Soustraction de la contribution de la mâchoire :** Une fois que la composante de la mâchoire est connue, elle doit être enlevée des données de la langue avant d'appliquer l'analyse factorielle. Il existe deux possibilités. La première, qui est généralement adoptée, consiste à enlever la corrélation entre la mâchoire et la langue des données de la langue. La deuxième possibilité consiste à enlever mécaniquement (à l'aide des paramètres de translation et de rotation) le mouvement de la mâchoire. Il n'y a plus ainsi d'influence de la mâchoire sur le contour de la langue. D'autre part, il reste d'autres interactions plus complexes entre la langue et la mâchoire. La première stratégie est meilleure pour réduire la quantité de variance analysée. Cependant, ceci correspond à l'hypothèse implicite que le contenu articulatoire du corpus d'images radiographiques est phonétiquement équilibré ce qui est rarement vrai. Nous étudierons les deux stratégies.

**Présentation des données de la langue :** Les données fournies à l'analyse factorielle sont des vecteurs unidimensionnels dans le cas de la grille semi-polaire, c'est-à-dire les intersections du contour de la langue avec les lignes de la grille. Dans le cas de la grille polaire adaptative, il n'est pas possible d'utiliser des vecteurs unidimensionnels puisque les lignes de la grille ne sont pas constantes. Trois solutions sont envisagées :

**Approche 1 :** Utiliser les coordonnées $(x, y)$ des points d'intersection. Les coordonnées du centre de la grille sont soustraites des coordonnées de chaque point d'intersection.

**Approche 2 :** Utiliser les distances entre le centre de la grille et chaque point d'intersection. Il y a deux autres paramètres qui sont les angles correspondant à la racine de la langue et l'apex. Puisque les angles sont régulièrement espacés entre la racine de la langue et l'apex, on a la même information que dans le premier cas, mais avec moitié moins de données.

**Approche 3 :** Utiliser les distances entre le centre de la grille et chaque point d'intersection, et les angles correspondants, c'est-à-dire les coordonnées polaires des points par rapport au centre de la grille. La quantité de données est la même que dans le premier cas.

Nous avons testé ces trois approches parce qu'elles offrent des points de vue différents. La première solution est la plus naturelle, mais elle sépare les coordonnées $x$ et $y$ dans l'analyse. La deuxième solution semble plus adaptée à la nature de la langue, mais elle combine deux types de données : les distances et les angles. La dernière solution utilise des données redondantes parce que les angles peuvent être obtenus à partir des deux angles extrêmes mais conserve l'homogénéité entre les données (chaque point est défini par un angle et une distance). Les deux dernières solutions offrent davantage d'explications en utilisant la nature radiale de la langue qui est observée quand la langue se compresse pendant l'articulation de [u] par exemple. Elles devraient renforcer l'émergence de cette déformation dans l'analyse.

**Stratégies utilisant l'analyse en composantes principales :** L'analyse en composantes principales est appliquée à ces données. Pour la mâchoire, le même calcul est réalisé pour toutes les stratégies. Les différentes stratégies utilisées pour la langue sont :

**Stratégie 1 :** L'ACP est appliquée indépendamment sur les données de la mâchoire et de la langue. Cela signifie que la langue est analysée avec le centre exact de la grille.

**Stratégie 2 :** Tout d'abord, l'ACP est appliquée aux données de la mâchoire. Le mouvement de la mâchoire est calculé avec la première composante de l'ACP sur la mâchoire. Le mouvement obtenu est enlevé mécaniquement des données de la langue. Enfin, nous réalisons l'ACP sur les données de la langue.

**Stratégie 3 :** Cette stratégie est la même que la seconde exceptée que la corrélation entre la mâchoire et la langue est retirée des données de la langue avant d'appliquer l'analyse factorielle. Cette stratégie est très similaire à celle employée par Maeda pour retirer l'influence de la mâchoire des données de la langue.

### *3.2. La grille semi-polaire*

Une analyse factorielle guidée est appliquée aux données à partir d'une grille semi-polaire [Cai09] ; comme celle décrite par Maeda [Mae90]. Le mouvement de la mâchoire est approché par la distance entre les incisives inférieures et supérieures. L'analyse factorielle guidée consiste à choisir une mesure articulatoire, c'est-à-dire celle qui est la plus corrélée avec les mesures à expliquer, comme variable explicative et à soustraire sa contribution de la ma-

**Table 1:** *Les trois erreurs de reconstruction moyennes et les écarts-type des erreurs (en pixels) entre les contours de langue originaux et ceux reconstruits avec 6 facteurs.*

| | Stratégie 1 | | Stratégie 2 | | Stratégie 3 | |
|---|---|---|---|---|---|---|
| App. 1 | 0.62 | 0.28 | 0.71 | 0.32 | 0.72 | 0.32 |
| App. 2 | 0.72 | 0.39 | 0.57 | 0.29 | 0.61 | 0.33 |
| App. 3 | 0.59 | 0.28 | 0.66 | 0.32 | 0.65 | 0.29 |

**Table 2:** *Les trois plus mauvaises erreurs de reconstruction (en pixels) entre les contours de langue originaux et ceux reconstruits avec 6 facteurs.*

| | Stratégie 1 | | | Stratégie 2 | | | Stratégie 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| App. 1 | 2.0 | 1.7 | 1.6 | 2.1 | 1.9 | 1.8 | 2.1 | 1.9 | 1.9 |
| App. 2 | 2.8 | 2.3 | 2.1 | 2.2 | 1.9 | 1.8 | 2.3 | 2.3 | 2.0 |
| App. 3 | 1.8 | 1.7 | 1.6 | 2.1 | 1.9 | 1.8 | 1.8 | 1.7 | 1.7 |

trice de corrélation. Cette procédure est répétée pour atteindre une faible variance résiduelle. Etant donné que la pointe de la langue n'est pas toujours intersectée par les lignes de la grille utilisée pour générer les données, ses coordonnées ont été ajoutées. Cinq composantes ont été extraites ; une pour la mâchoire et quatre pour les différents modes de déformation de la langue.

## 4. EVALUATION DES MODÈLES ARTICULATOIRES DE LANGUE

Nous présentons ici l'évaluation des différentes stratégies utilisées pour construire le modèle articulatoire de la mâchoire et de la langue.

### *4.1. Cas de la mâchoire*

Nous avons appliqué l'ACP (Analyse en Composantes Principales) sur les données de la mâchoire (l'angle de rotation et les coordonnées de la translation). La variance expliquée est de 57% pour la première composante et de 32% pour la deuxième. Pour conserver un nombre de composantes linéaires le plus petit possible, nous avons choisi de retenir uniquement la première composante. Contrairement à d'autres approches la première composante linéaire contrôle la rotation et la translation. Nos expériences montrent que la mâchoire peut être approchée avec un seul facteur. Il n'y a pas d'amélioration significative globale avec plus de composantes.

### *4.2. Cas de la langue*

Concernant la langue, nos expériences (voir les tables 1 et 2) nous ont conduits à utiliser six composantes linéaires.

La meilleure solution obtenue est celle qui utilise comme données les deux angles extrêmes et les distances par rapport au centre de la grille (Approche 2) avec la deuxième stratégie (soustraction du mouvement de la mâchoire uniquement). L'erreur de reconstruction moyenne est de 0.57 pixel, l'écart-type est assez petit (0.29 pixel) et les trois pires erreurs de reconstruction sont en pixels de 2.2, 1.9, 1.8 (voir la figure 4). L'examen des pourcentages de variance expliquée montre que la variance cumulée des deux

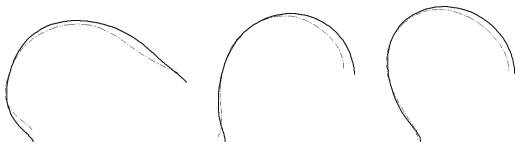

**Figure 4:** *Les trois plus mauvaises reconstructions (resp. 2.2 pixels, 1.9 pixels, 1.8 pixels) avec six composantes linéaires pour la stratégie 2 et l'approche 2 (en trait gras : le contour reconstruit et en pointillés : le contour original).*

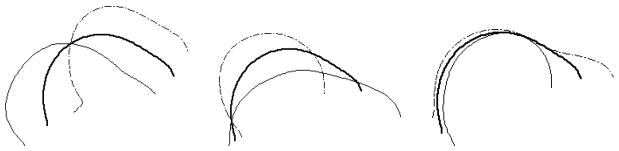

**Figure 5:** *Trois premières composantes principales. Pour chaque composante le contour moyen est le contour en gras et les deux autres correspondent à $\pm$ 3 écarts type.*

premiers facteurs est de 89% et la variance du troisième facteur est particulièrement petite (5.8%).

La figure 5 présente les trois premières composantes linéaires expliquant les déformations de la langue pour la seconde approche et la seconde stratégie, c'est-à-dire celle qui donne la plus petite erreur de reconstruction. Les deux premières composantes correspondent aux principaux modes de déformation de la langue (avant+haut/arrière+bas et arrondissement/aplatissement). La troisième composante correspond au mouvement vers l'avant de l'apex observé dans l'articulation de constrictions dentales. Les trois autres composantes contrôlent de petites déformations de la langue. Il faut noter que ces composantes dépendent du contenu phonétique de la base de données. Même si les composantes trouvées sont pertinentes, leur poids relatif ne rend probablement pas le comportement complet de la langue. Retirer la ou les deux dernières composantes linéaires dégrade la reconstruction de quelques images en dépit de la faible variance expliquée par ces composantes. Les dernières composantes prennent en partie en compte ces images qui apparaissent comme des « exceptions » à cause du contenu phonétique de la base de données.

Curieusement, soustraire la corrélation entre la mâchoire et la langue avant d'appliquer l'ACP (Stratégie 3) n'apporte pas de réelles améliorations. La première stratégie qui consiste à calculer les ACP indépendamment donne aussi de bons résultats pour les approches 1 et 3 (erreur de 0.62 et de 0.59 pixels avec six facteurs) avec un petit écart-type (0.28).

### *4.3. Comparaison des deux grilles*

Notre approche sépare la langue du pharynx inférieur contrairement au modèle de Maeda. Une comparaison de la précision sur toute la langue n'est pas pertinent. La moitié avant de la langue reconstruite avec l'utilisation de la grille adaptative est ainsi comparée avec la partie correspondante avec la langue reconstruite avec la grille semi-polaire. L'erreur moyenne de reconstruction avec la grille adaptative est de 0.69 pixel au lieu de 0.92 avec la grille semi-polaire. Nous pouvons remarquer que la région de l'apex est moins bien modélisée car l'erreur moyenne de

reconstruction sur la moitié avant de la langue est plus grande que l'erreur moyenne sur le contour entier. Les plus grandes erreurs de reconstruction dans le cas de la grille semi-polaire sont les images où la langue est très avancée. Cela s'explique par le fait que le dernier point intersectant la grille est très éloigné du point correspondant à l'apex ; le contour est approché linéairement entre ces deux points ce qui créer de grandes erreurs dans le cas où la langue est très avancée. Dans les autres cas les erreurs de reconstruction sont très faibles notamment quand la langue est très reculée. Pour la grille polaire, les contours les moins bien reconstruits sont les contours arrondis où la moitié avant concentre la majorité des erreurs.

## 5. CONCLUSION

Le mouvement de l'apex de la langue est meilleur avec la grille adaptative qu'avec la grille semi-polaire. En particulier, le mouvement de l'apex contactant la région alvéolaire est maintenant clairement visible.

L'évaluation que nous avons effectuée est purement géométrique. Nous préparons actuellement une évaluation plus complète associant la proximité acoustique du signal synthétisé par le modèle articulatoire avec le signal original. Cette évaluation est importante parce que l'impact acoustique des erreurs de modélisations dépend de l'endroit où elles se produisent par rapport aux constrictions du conduit vocal.

## 6. REMERCIEMENTS

Ce travail s'inscrit dans le cadre du projet Européen ASPI (IST-2005-021324) et des projets ANR DOCVACIM et ARTIS. Je tiens à remercier Yves Laprie, Shinji Maeda, Jun Cai, Fabrice Hirsch, Béatrice Vaxelaire, Michaël Aron et Marie-Odile Berger pour leurs discussions fructueuses.

## RÉFÉRENCES

[Bea01] Beautemps, D., Badin, P. and Bailly, G. (2001), "Linear degrees of freedom in speech production : Analysis of cineradio- and labio-film data and articulatory-acoustic modeling", in *Journal of the Acoustical Society of America*, vol. 109, pp. 2165–2180.

[Cai09] Cai, J., Laprie, Y., Busset, J. and Hirsch, F. (2009), "Articulatory Modeling Based on Semi-polar Coordinates and Guided PCA Technique", in *Interspeech, Brighton*.

[Fon06] Fontecave, J. and Berthommier, F. (2006), "Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data", in *Interspeech, Pittsburgh*.

[Lap96] Laprie, Y. and Berger, M. (1996), "Towards automatic extraction of tongue contours in x-ray images", in *Proceedings of International Conference on Spoken Language Processing 96*, Philadelphia (USA), vol. 1, pp. 268–271.

[Mae90] Maeda, S. (1990), "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, vol. 4, pp. 131–149.

[Ste02] Stegmann, M.B. and Gomez, D.D. (2002), "A brief introduction to statistical shape analysis", in *Technical University of Denmark*, Tech. Rep.

[Thi99] Thimm, G. and Luettin, J. (1999), "Extraction of articulators in x-ray image sequences", in *Eurospeech*, Budapest, pp. 157–160.