Méthodologie 3-way d'extraction d'un modèle articulatoire de la parole à partir des données d'un locuteur

Martine Cadot, Yves laprie

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

Résumé. Pour parler, le locuteur met en mouvement un ensemble complexe d'articulateurs : 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, etc. 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 3-way du même type de données, après leur transformation en 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 co-articulation.

1 Introduction

Construire un modèle articulatoire de la parole, c'est être capable d'indiquer les mouvements des articulateurs (mâchoires, lèvres, etc.) à l'origine de celle-ci (voir figure 1, à gauche). Les applications pratiques d'un tel modèle sont nombreuses [1].

Nous exposons ici comment nous avons extrait un modèle articulatoire à partir de données recueillies auprès d'un locuteur. Ce travail se situe dans la lignée des travaux initiés par Maeda (1990). Il a construit son modèle articulatoire (voir figure 1, à droite) au moyen d'analyses en composantes principales sur des données de même type. Puis il l'a évalué de façon acoustique en comparant les sons réels aux sons produits par un synthétiseur de sons piloté par son modèle. La nouveauté de notre démarche consiste en l'utilisation d'une méthode d'analyse 3-way pour extraire le modèle, et de méthodes d'apprentissage supervisé pour le valider. Notre évaluation se fait en comparant de façon phonétique les sons prédits aux sons réels. L'acoustique intervient de surcroît dans notre évaluation car nous mettons en parallèle les performances de notre modèle et celles du modèle acoustique formé des coefficients *cepstraux* [2].

1. Pour plus de détails, voir http ://parole.loria.fr/.
2. Ces coefficients représentant le signal de parole ont été extraits à l'aide de la bibliothèque R de Ligges (2011)

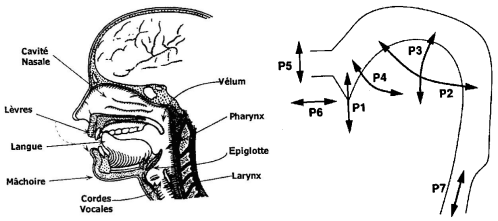

FIG. 1 – *Schéma de l'anatomie du conduit vocal, à gauche d'après Flanagran 1972, à droite modèle de Maeda 1970*

2 La référence : un modèle articulatoire extrait par ACP

Le modèle de référence (Maeda, 1990) est une représentation au moyen de 7 paramètres (Fig. 1, à droite) des mouvements des articulateurs déformant le conduit vocal lors de la production de parole. Il a été obtenu par ACP à partir de la cinéradiographie d'une personne en train de parler. Pour obtenir les 3 paramètres de la langue, par exemple, un tableau de nombres T a été construit de la façon suivante : sur chaque radiographie R_i le contour de la langue a été marqué par une série de n_i points P_{ij}, et leurs coordonnées x_{ij} et y_{ij} ont fourni la ligne T_i de longueur $2 \times n_i$. Et les trois premières composantes d'une ACP opérée sur une version corrigée du tableau ont donné les paramètres P2, P3 et P4 de la langue. De nombreuses variantes de ce modèle ont été proposées par la suite, portant essentiellement sur la normalisation des tableaux de données soumis à des ACP, (Laprie et Busset, 2011; Busset, 2013).
Les ACP ont montré leur efficacité dans la construction du modèle articulatoire, mais aussi leurs limites : difficile d'aller au-delà de 7 facteurs pertinents, tous les articulateurs ayant des mouvements liés, et difficile d'interpréter finement les contributions des points aux axes quand abscisses et ordonnées d'un nombre non négligeable de points se retrouvent sur des axes différents. C'est pour pouvoir dépasser ces limites que nous nous sommes interessés à une méthode 3-way MDS.

3 Extraction de notre modèle par 3-way MDS

Les principes du MDS. Le MDS (MultiDimensional Scaling, et en français "positionnement multidimensionnel", fait partie des méthodes d'analyses factorielles des données, et est particulièrement adapté à l'analyse des données de type *dissimilarités* [3], non mesurables objectivement, correspondant à des impressions ressenties, à des distances subjectives, etc. (Borg et Groenen, 1997). La méthode d'analyse MDS est capable de positionner des objets dans un espace euclidien de dimension p de telle sorte que leurs distances deux à deux soient les plus proches possibles de leurs dissimilarités initiales. La formulation mathématique à la base des MDS est la suivante : si pour deux objets numérotés par i et j, on note δ_{ij} leur dissimilarité initiale, d_{ij} leur distance dans l'espace euclidien de dimension p, et f une fonction monotone de

3. 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 $d(x, z) \leq d(x, y) + d(y, z)$.

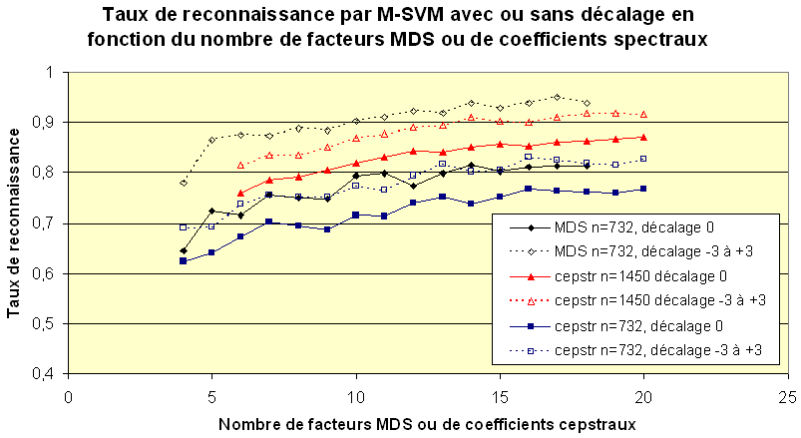

FIG. 2 – *Taux de reconnaissance avec décalages de M-SVM type C-bcsv.*

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.

Le 3-way MDS pour nos données Nos données, coordonnées des points des articulateurs sur les radiographies ont été transformées en autant de tableaux de distances entre images que de points. Les méthodes de 3-way MDS permettent de traiter simultanément plusieurs tableaux par MDS, d'après Borg et Groenen (1997); Carroll (1972); Carroll et Chang (1970).

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 Carroll (1972), qui s'est avérée moins gourmande en mémoire vive que la méthode INDSCAL de Carroll et Chang (1970) que nous avions utilisée pour des données plus simples dans Busset et Cadot (2013). Avec 200 itérations, nous avons pu obtenir à partir des 50 points les positions des 732 images dans des espaces allant de 2 dimensions à 18 (64bits, 4 coeurs, 8 Gio de RAM, pour dim=18, temps=48h).

4 Evaluation, conclusion, perspectives

Pour l'évaluation, nous avons comparé (Fig. 2) la capacité d'apprentissage des sons de notre modèle MDS issu de 732 images par des M-SVM (bibliothèque R de Karatzoglou et al. (2004)), à deux variantes du modèle acoustique (i.e. à base de matrices cepstrales). Un décalage de -3 à 3 (trait pointillé) signifie que le son reconnu par M-SVM est situé à moins de 4 images avant ou après le son à reconnaitre, le décalage étant nul (trait plein) quand on a bien reconnu le son correspondant à l'image. On voit qu'avec la prise en compte des décalages notre modèle articulatoire (MDS 732 images) avec seulement 12 facteurs surpasse le modèle acoustique de même niveau (732 lignes) et même de niveau supérieur (1450 lignes).

Ces bons résultats nous invitent à continuer dans cette voie, en tentant d'améliorer dans différentes directions : 1) revoir la programmation des fonctions R utilisées, et prendre plus de points par image, 2) incorporer les dépendances temporelles directement dans le modèle 3-way MDS, plutot que les prendre en compte par des décalages dans l'évaluation.

Références

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 (2013). Fouille d'images animées : cinéradiographies d'un locuteur. In *Atelier FOuille de données Spatio-Temporelles et Applications - FOSTA*, Toulouse, 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.

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

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.

Summary

For speaking, a speaker sets in motion a complex set of articulators: the jaw that 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 co-articulation is taken into account.