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

Martine Cadot, Yves Laprie

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,
http://www.loria.fr/ cadot
**LORIA
yves.laprie@loria.fr,
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). Des applications pratiques d'un tel modèle ont déjà été mises en œuvre par les enseignants/chercheurs de l'équipe Parole du Loria, par exemple, pointer pour les étudiants en "Français Langue É trangère" les les articulateurs en jeu lors de la prononciation des sons, doubler les enregistrements vidéo pour les malentendants par une "tête parlante" plus réaliste.

Nous exposons dans cet article 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 2) 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 (pour un exemple de sons transcrits en phonétique, voir le tableau 1). L'acoustique intervient de surcroît

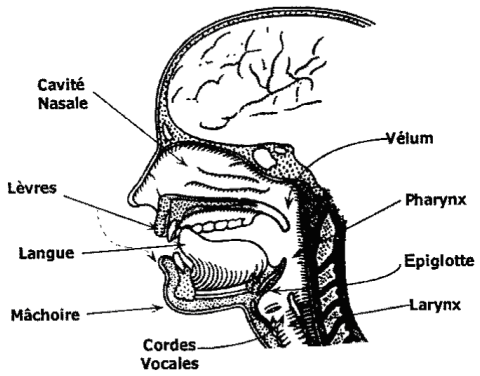

FIG. 1 – *Schéma de l'anatomie du conduit vocal (d'après Flanagran 1972).*

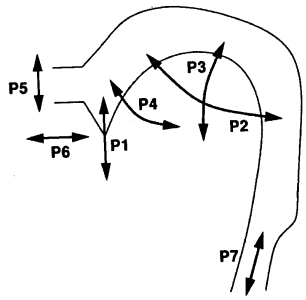

FIG. 2 – *Modèle articulatoire à 7 paramètres de Maeda.*

dans notre évaluation car nous mettons en parallèle les performances de notre modèle et celles du modèle acoustique formé des coefficients cepstraux.
La démarche relatée dans cet article complète et enrichit celle d'un travail précédent de Busset et Cadot (2013). Nous avions alors un corpus de taille inférieure, avec peu de sons différents, et une certaine répétition des phrases. Le modèle étant de petite taille, nous l'avons validé de façon experte, en interprétant un à un ses éléments. Ce premier essai étant probant, nous sommes passés à l'échelle supérieure avec des données plus riches et un modèle plus important, validé de façon automatique.

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. 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 Motivation du recueil des données

Dans le modèle articulatoire de la parole construit par Maeda (voir figure 2), le conduit vocal, zone interne allant de l'arrière de la gorge 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 l'air venu des poumons et produisent les différents sons de la parole. Toutefois, avec ses 7 paramètres, c'est une simplification forte de la réalité, et certains sons sont plus difficiles à simuler ainsi que d'autres.

2.2 La vidéo à l'origine des données

Une série de radiographies de la tête a été réalisée pendant qu'un locuteur prononçait quelques phrases courtes, recopiées dans le tableau 1 (pour plus de détail sur ces données, se reporter à Sock et al. (2011)). Puis des contours ont été dessinés sur ces images afin de représenter au mieux les articulateurs (voir figure 1), et de repérer le plus finement possible leurs mouvements.

On dispose aussi de la correspondance entre sons et images. Les 27 symboles suivants ont été choisis pour représenter les différents sons présents dans les données, selon le tableau 1 :

@ 9 A ã b d e E f g i j k l m n o õ p R s S t u w z Z

Ces sons ne sont pas utilisés dans la phase de construction du modèle, mais 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. Le repérage des contours des articulateurs s'est appuyé sur toute une série d'outils mis en oeuvre au sein de l'équipe Parole du LORIA. Cette méthodologie d'annotation semi-automatique (voir Laprie et Busset (2011)), a débouché sur plusieurs méthodes qui ont été appliquées à une première bande vidéo, puis évaluées et comparées de diverses façons avant d'être testées à l'aide d'un synthétiseur proche de celui de Maeda (voir la thèse de Busset (2013) pour plus de détails). Ces méthodes ont été appliquées sur une deuxième bande vidéo, plus riche, et les données produites ont été retravaillées lors du stage de Clément (2013). Ce sont ces données, de grande qualité, que nous avons utilisées ici.

Transcription orthographique	Transcription acoustique
Il a pas mal.	ilApAmAl@
Les attablés.	lezAtAble
Très acariâtre.	tREzAkARijAt
Il zappe pas mal.	ilzApAmAl@
Des abat-jour.	dezAbAZuR
Il l'a datée.	ilAdAte
Crabe bagarreur.	kRAbAgAR9R
Trois sacs carrés.	tRwsAkARe
Pas de date précise.	pAdAtpResiz9
Blague garantie.	blAgARãti
Nous palissons.	nupAlisõ
Il a pourri.	ilApuRi
Couds ta chemise.	kutASmiz@
Elle a tout faux.	ElAtufo
Pour tout casser.	puRtukAse

TAB. 1 – *Les 15 phrases successives prononcées par le locuteur et leur transcription acoustique*

Elles se présentent sous la forme de 11 contours formés d'un nombre fixe de points pour les uns, et variable pour les autres. On dispose pour chaque image des coordonnées 2D des points de chaque contour, comme indiqué dans le tableau 2.

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 3-way, 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 2), le contour de la langue a été repéré sur chaque radiographie par une centaine de points. 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, Laprie et Busset (2011) ont utilisé une grille polaire adaptative 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 les ont nettoyés de leur liens avec les articulateurs précédents par soustraction des corrélations.

déformable	nb min	nb max	indéformable	nb de points
voile du palais	69	107	os hyoide	30
epiglotte	52	74	plancher de la langue	19
larynx	46	79	palais	39
lèvre supérieure	11	35	machoîre inférieure	50
lèvre inférieure	13	45	machoîre 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*

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 axes, on a constaté que les abscisses et les ordonnées d'un nombre non négligeable de points se retrouvaient sur des axes différents, 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 Méthode de 3-way 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* [1], non mesurables objectivement, correspondant à des impressions ressenties, recueillies par 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 possibles 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 δ_{ij} leur dissimilarité initiale, d_{ij} leur distance dans l'espace

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

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 oeuvre du MDS. Au départ de l'algorithme, 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.

Dans notre cas, le fait d'utiliser les distances entre points au lieu de leurs coordonnées permet d'éviter que les abscisses et ordonnées d'un même point ne se retrouvent sur deux axes factoriels différents.

Les differents types de 3-way. Le MDS a été étendu pour pouvoir prendre en considération plusieurs 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 méthode MDS permettant de le traiter en prenant 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" ; même espace X de dimension p pour chaque sujet k avec une fonction f_k différente pour chacun,
- "idiosyncratic" ; 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 interessés au chapitre 21 de Borg et Groenen (1997).

3.3 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" et par 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 du tableau. 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 l'apex (pointe), de la racine et du dos 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 la distance euclidienne entre ses positions sur deux images, ce qui a donné un tableau 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 section 3.2, qui s'est avérée moins gourmande en mémoire vive que la méthode INDSCAL que nous avions utilisée avec succès pour des données 5 fois plus petites dans 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, 2Gio de RAM vers un ordinateur 64 bits, quadri-processeur, 8 Gio 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 exposons maintenant les mesures de qualité de ces représentations dans une seule direction, celle de la discrimination du son. Notre indice de qualité est la proportion de sons correctement identifiés, que nous appelons taux de reconnaissance. Pour pouvoir apprécier le taux de reconnaissance du modèle articulatoire, nous le comparons à celui du modèle acoustique. 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 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 radios étaient prises. C'est la matrice obtenue de 2119 lignes et 20 colonnes qui représente le "modèle acoustique" que nous souhaitons comparer aux "modèles articulatoires" formés des 1021 images et q colonnes pour q allant de 2 à 18.

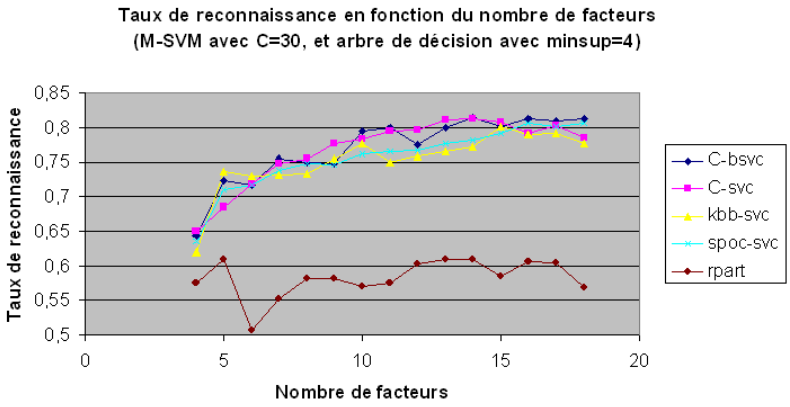

FIG. 3 – *Taux de reconnaissance en fonction du nombre de facteurs.*

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 correspondants 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 de 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 finale de coefficients cepstraux contenant 1450 lignes, nous en avons fait une version réduite en retirant environ une ligne sur deux [2] afin d'avoir la même distribution de sons que la matrice de facteurs MDS. Nous avons ainsi obtenu une deuxième matrice avec seulement 732 lignes.

Les mesures d'évaluation utilisées. Nous avons utilisé le package Rpart (**?**) 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 **?**. Dans ce package, nous avons pu utiliser les 4 types de M-SVM suivants :

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.

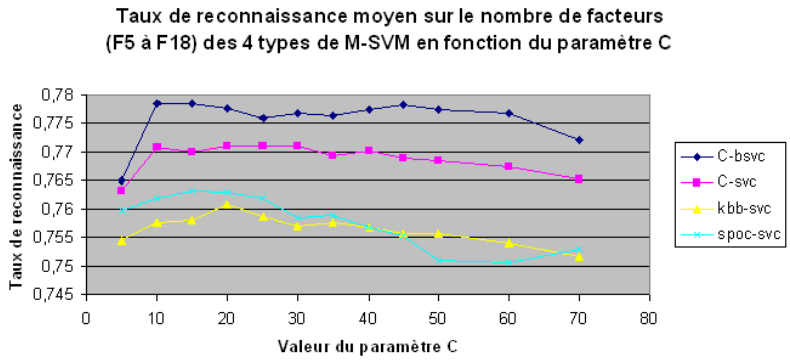

FIG. 4 – *Taux de reconnaissance en fonction de C selon 4 types de M-SVM.*

– 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 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 3, on a représenté 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 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 4. 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 reconnaissance 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 résultats. Les mêmes méthodes appliquées aux matrices de données cepstrales donnent des résultats de même nature que ceux que nous venons de décrire. Par rapport aux matrices MDS, les taux de reconnaissance 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

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

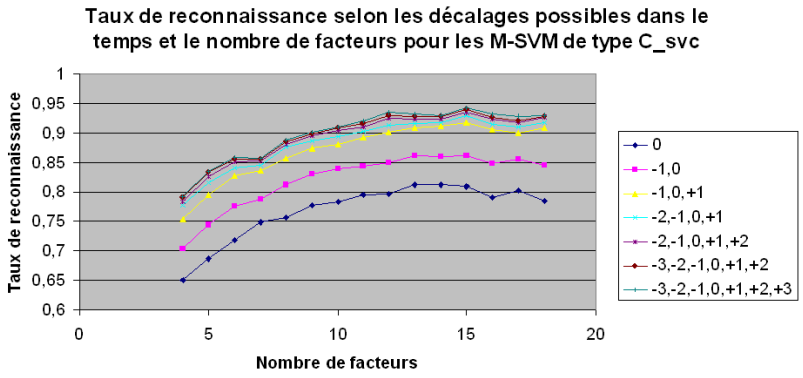

FIG. 5 – *Taux de reconnaissance MDS avec décalages pour M-SVM type C-bcsv.*

meilleurs (maximum 0,86 pour 20 coefficients) avec celles de taille 1450 (voir lignes en traits pleins, figure 6).

4.3 Prise en compte des proximités temporelles

Jusqu'ici, l'aspect temporel n'a pas été pris en compte, pas plus que la co-articulation. Nous le faisons ici en opérant des décalages dans la reconnaissance : 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. On voit dans la figure 5 que le taux 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 phénomène se retrouve pour les matrices cepstrales, mais avec une ampleur moindre. Dans la figure 6, 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). Ce qui peut s'expliquer par le fait que le phénomère de résilience des sons doit être moins prégnant que celui de coarticulation.

Au final, en prenant en compte les décalages, le modèle articulatoire prédit légèrement mieux les sons 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 performant des données dont nous disposons. Nous avons rencontré un certain nombre de difficultés lors de l'application de cette méthode, nous les listons ici :

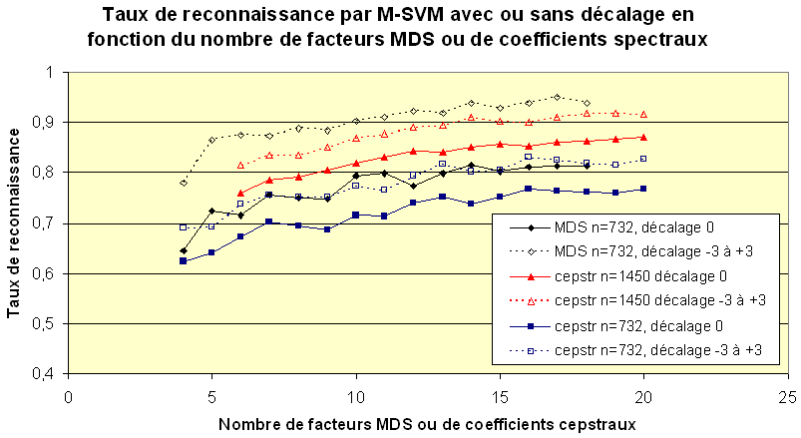

FIG. 6 – *Taux de reconnaissance avec décalages de M-SVM type C-bcsv.*

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 temporel dans l'analyse : dans la construction du modèle ou dans son évaluation ?

Malgré ces difficultés nous arrivons à un modèle articulatoire qui contient autant d'informations sur les sons que le modèle acoustique. 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 discrimination pour répondre aux points 3 et 4,
– essayer d'incorporer les dépendances temporelles dans le modèle 3-way MDS pour le point 5.

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.

Clément, R. (2013). Construction de modèles articulatoires du conduit vocal pour la production de la parole. Rapport de stage, Université de Lorraine, Master Ingénierie de la Mesure et de l'Image, Spécialité Mesure, Performance et Certifcation.

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.

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.

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.