Normalisation articulatoire du locuteur par méthodes de décomposition tri-linéaire basées sur des données IRM

Julián Andrés Valdés Vargas[1], Pierre Badin[1], G. Ananthakrishnan[2], Laurent Lamalle[3]
(1) GIPSA-lab (Département Parole & Cognition), UMR 5216 CNRS – Grenoble University
(2) Centre for Speech Technology, KTH (Royal Institute of Technology), Stockholm, Sweden
(3) SFR1 RMN Biomédicale et Neurosciences (Unité IRM Recherche 3 Tesla), INSERM, CHU de Grenoble
{julian-andres.valdes-vargas,Pierre.Badin}@gipsa-lab.grenoble-inp.fr, agopal@kth.se,Llamalle@ujf-grenoble.fr

RESUME

Le but de cette étude était de caractériser, modéliser et comparer les différentes stratégies articulatoires linguales pour un groupe de locuteurs. Des modèles individuels par analyse en composantes principales (ACP) et des méthodes de décomposition multilinéaires ont été appliqués aux contours de langue extraits d'un corpus d'imagerie par résonance magnétique (IRM) de sept locuteurs prononçant 63 voyelles et consonnes du français. En moyenne sur les sept locuteurs, en utilisant quatre composantes, l'erreur quadratique moyenne de prédiction (RMSE) était de 0,13 cm pour les modèles individuels ACP et de 0.29 cm pour le modèle 'parallel factor' (PARAFAC), avec des pourcentages de variance expliquée de 91% et 62%, respectivement. Un modèle de régression multilinéaire permet également de prédire avec 10 composantes les contours de langue d'un sujet cible à partir de ceux d'un sujet source avec approximativement 65% de la variance expliquée et une RMSE de 0.38 cm. Tous les modèles ont été évalués par une procédure de validation croisée.

ABSTRACT

Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods

The aim of this study was to characterise, to model and to compare the different lingual articulatory strategies of a group of speakers. Individual principal component analysis (PCA) models and multi-linear decomposition methods have been applied to the tongue contours extracted from a magnetic resonance imaging (MRI) corpus of seven speakers articulating 63 French vowels and consonants. On the average over the seven speakers, using 4 components, the Root Mean Square prediction Error (RMSE) was 0.13 cm for the individual PCA models while the RMSE for the parallel factor model (PARAFAC) was 0.29 cm, accounting for a percentage of variance explanation of 91% and 62%, respectively. A multi-linear regression (MRL) model could predict, with 10 components, the tongue contour of a target subject from a given source subject, with about 65% of the variance explained and an RMSE of 0.38 cm. All the models have been assessed by a leave-one-out cross-validation procedure.

MOTS-CLES: Modélisation articulatoire, normalisation du locuteur, analyse factorielle, IRM.

KEYWORDS: Articulatory modelling, speaker normalisation, factor analysis, MRI.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 529–536,
Grenoble, 4 au 8 juin 2012.

1. Introduction

The Speech & Cognition Department at GIPSA-lab has developed acoustic-to-articulatory inversion methods to provide speakers with a visual articulatory feedback (Ben Youssef *et al.*, 2011), based on a fairly complete orofacial clone. This clone is made of a set of models of articulators (jaw, tongue, velum, lips, etc.) based on articulatory data acquired on a single speaker (Badin & Serrurier, 2006). Therefore, the clone represents faithfully the characteristics of a specific speaker, but not necessarily those of other speakers that may have different morphologies and different articulatory control strategies. Thus, one important issue is the normalisation problem: how can the speaker-specific models of the orofacial clone be adapted to other speakers? This problem is particularly challenging as it implies discovering how different speakers with different morphologies can produce articulated sounds that are considered equivalent for speech communication purposes.

Several studies based on measurements using Electromagnetic Articulography (EMA) and Magnetic Resonance Imaging (MRI) have been led in this field. Harshman *et al.* (1977) made a Parallel Factor analysis (PARAFAC) study on X-ray data of five American English speakers. The tongue postures were decomposed in two factors which explained 92.7% of the variance. In another study, Hoole (1998) provided a two factor PARAFAC solution for the German vowel system in three different consonant contexts /p t k/. Two-factor independent models were successfully extracted by Principal Component Analysis (PCA) for each consonant context. The explained variance amounted to about 92.3% and the Root Mean Square reconstruction Error (RMSE) to 1.24 mm for each model. On the other hand, the extracted two-factor PARAFAC solution for the complete dataset presented an increase of RMSE compared to the individual models, the explained variance now amounting to 80% and the RMSE to 1.9 mm. In another study, Hoole (1999) showed how the PARAFAC model error could be further analysed to extract an additional component. His approach consisted in examining the error of the two-factor PARAFAC model by subtracting the articulatory data predicted from the original data. Then, a PCA was employed to extract an extra-component. The final model explained over 90% of the variance. PARAFAC was performed by Hoole *et al.* (2000) on a set of MRI data of nine German speakers uttering seven German vowels in five different contexts. Two factors accounted for about 87 % of the variance with a RMSE of about 2.2 mm. Geng & Mooshammer (2000) provided a two factor PARAFAC solution. The speech material consisted of six German speakers uttering fifteen German vowels in /t/-context recorded by EMA. Two factors led to a variance explanation of about 96% and an RMSE of about 2 mm. A two-factor model resulted in a stable solution that explained about 70% of the variance in a study made by Zheng *et al.* (2003). The data consisted of MRI images of five American English speakers pronouncing nine English vowels. Hu (2006) presented a study on the Chinese dialect called Ningbo. Seven speakers pronouncing ten vowels were recorded by means of EMA. Two factors explained about 90% of the variance. More recently, Ananthakrishnan (2010) proposed a two factor PARAFAC model that accounted for 71% of the variance explanation for three French speakers articulating 13 vowels.

The present study attempts to extend this type of modelling from vowels to consonants. We first describe the set of data acquired to perform the different experiments; then we describe the performance of individual speaker models and compare them in terms of variance explained, RMSE and individual articulatory strategies. Next, we present an

attempt to build a single model to drive the tongue contours of all the speakers based on multi-linear decomposition methods. We perform a PARAFAC solution up to 10 components and a more practical solution using Multiple Linear Regression (MLR) with a large number of components.

2. Data

In this study, midsagittal Magnetic Resonance Images (MRI) of seven French speakers (two males: *PB*, *YL*, and five females: *HL, AA, MG, AK, MGO*) have been collected. The subjects were asked to pronounce and sustain 63 different articulations for 16 seconds each. The corpus consisted of the 10 French oral vowels /i e ɛ a y ø œ u o ɔ/, the 3 nasal vowels /ã ɛ̃ ɔ̃/ and the 10 consonants /p t k f s ʃ m n ʁ l/ articulated in symmetric VCV context of five vowels /a e ɛ i u/. The contour of the tongue was manually traced. The present study is limited to the contour from the tongue tip to the base of the epiglottis, which is resampled with N = 150 equidistant points to model what we call *Tongue upper contour*.

3. Individual articulatory models (PCA)

PCA is a two-way factor analysis approach often used for dimensionality reduction and analysis of data sets to summarize their main characteristics. Consider articulatory measurements for the speaker *s:* $1 \leq s \leq S$, which consists of $X_s = [x_1, x_2, \ldots, x_A]$, being x_a $(1 \leq n \leq N)$ a row vector of measurements for the articulation *a:* $1 \leq a \leq A$. Such that X_s is decomposed into a set of control parameters $\pi_s^{[A \times Cmp]}$ (set of *Cmp* components that explain the variations in articulations) and the articulatory model $C_s^{[N \times Cmp]}$ (coefficients that explain the contribution of each articulator point to the components) by the following equation: $X_s = \pi_s * C_s^T + \gamma_s$, where γ_s is the residual error.

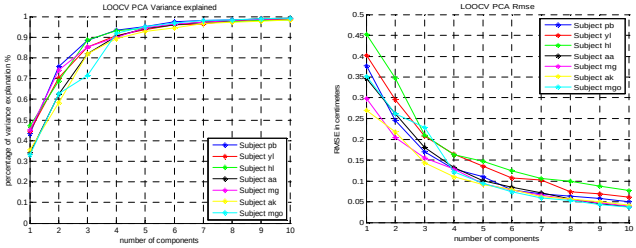

FIGURE 1 - Performance of the LOOCV PCA individual models as a function of number of components for the tongue upper contours of the seven speakers *PB, YL, HL, AA, MG, AK* and *MGO*. Left: variance explained (%). Right: RMSE (cm).

The models were made and assessed by means of a leave-one-out cross validation (LOOCV) procedure. One observation of the data was left out; the model was built from the remaining data and used to predict the left-out articulation, this process was repeated for each articulation on the set. LOOCV was useful to decide how many predictors to use. For instance, the cross-validated mean-square error will tend to

decrease if valuable predictors are added, but increase if worthless predictors are added. Indeed, increasing the number of predictors might lead to an over-fitted or degenerated model (Riu & Bro, 2003). Figure 1 displays the variance explained and RMSE relative to the reconstruction of the tongue for the whole corpus of vowels and consonants. We have found that, on average over our seven speakers, the PCA model with the first four components explains an amount of 91% of the data variance, with an RMSE of 0.13 cm.

3.1. Differences between speaker control strategies

Using a procedure based on a guided PCA analysis of tongue contours, Badin and Serrurier (2006) have shown that the first four components account for the largest amount of tongue movement variance. In this section we describe the results of the Guided PCA analysis of our seven speakers. The *jaw height* parameter *JH* was defined as the normalized value of the measured lower incisor height; it was used as the first control parameter of the tongue model (the associated model coefficients were obtained by the MLR of all the vertex coordinates against *JH*). The next two parameters, *tongue body TB* and *tongue dorsum TD* were extracted by PCA from the coordinates of the midsagittal tongue contour, excluding the tongue tip region, from which the *JH* contribution had been removed (the associated model coefficients were obtained by MLR, as for *JH*). The next parameter called *tongue tip TT* was extracted by PCA from the midsagittal tongue tip contour coordinates, from which the *TB* and *TD* contributions had been removed (the associated coefficients were also obtained by MLR).

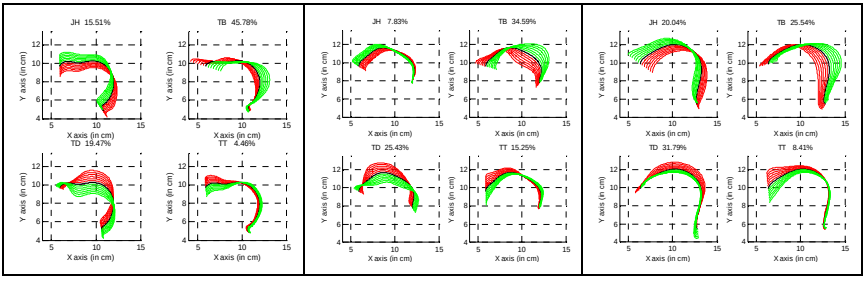

FIGURE 2 - Illustration of the first four components and their variance explained extracted by Guided PCA for the tongue contour of speakers *PB, AA* and *YL (*from left to right respectively). Each predictor is varied from -3 to +3 with a 0.5 step. X and Y axis are cms.

Hence, in order to understand the articulatory characteristics of each subject, we compared their four guided PCA components explained above. Figure 2 illustrates the associated nomograms for the subject PB, AA and YL. The main effect of *JH* is a rotation of the tongue around a point located in its back. In our case, the *JH* parameter of subjects *MGO*, *MG*, *AA* and *AK* is associated with a movement of the front of the tongue without movement in the back. Oppositely, subjects *HL*, *PB* and *YL* move the back of the tongue when *JH* moves. The tongue body parameter *TB* controls front-back

displacements while the *TD* parameter is related to flattening-arching movements. It appears that the *TB* component of subjects *HL*, *AK* and *YL* is a horizontal movement of the tongue body while it is a diagonal movement for subjects *PB*, *MG*, *AA* and *MGO*. Besides, *TB* explains more variability than *TD* for most subjects, but that behaviour is swapped for subject *YL*. In other words, subject *YL* uses more his tongue dorsum component than his tongue body component compared to the other subjects. On the other hand, the *TT* parameter controls precisely the tongue tip motions. We have observed that subjects *AA*, *AK*, *MG*, *MGO* and *PB* are able to move their tongue tips more independently from the tongue back than the subjects *HL* and *YL* do.

4. Multi-linear decomposition methods

4.1. PARAFAC model

PARAFAC is a factor analysis approach often used to decompose multi-way data. In our specific case, the dimensions of the three-way data are related to the articulations, articulator points and subjects, respectively. The data of a given subject X_s is decomposed as:

$$X_s = \pi * \Phi_s * C^T + \gamma_s$$

where γ_s is the residual error. $\pi^{[A\ x\ Cmp]}$ is the set of universal components that models the variations in articulations over the S subjects, the articulatory model $C^{[N\ x\ Cmp]}$ is a matrix of coefficients that models the contribution of each component, over the S subjects, to the articulator points. The extra matrix Φ_s provides speaker-specific weights to the contribution of the components.

4.2. PARAFAC model with vowels

In order to make a fair comparison of our results with those given by the literature, we restricted our modelling to the 10 French oral vowels. Using a two factor PARAFAC model, the average reconstruction error, over our seven speakers, was 0.25 cm for the 150 articulator points while the RMSE for tongue contours under-sampled to 3 points was 0.21 cm, accounting for a variance of 75.1% and 85.8%, respectively.

Type	Study	No. Subjects	Corpus	No. Points	Variance Exp
EMA	Hoole(1998)[5]	7	15 vowels	4 sensors	80.0%
	Geng(2000) [6]	6	15 vowels	4 sensors	96.0%
	Hu(2006) [7]	7	10 vowels	3 sensors	90.0%
X ray	Harshman(1977)[8]	5	10 vowels	13 points	92.7%
MRI	Hoole(2000) [9]	9	7 vowels	13 points	87.0%
	Zheng(2003) [10]	5	9 vowels	13 points	76.2%
	Ananth(2010) [3]	3	13 vowels	150 points	71.0%
Our Results					
MRI	Valdes(2012)	7	10 vowels	3 points	85.8%
		7	10 vowels	150 points	75.1%

TABLE 1 – Comparison of our results with the literature using 2 PARAFAC components.

Table 1 shows that, on the overall, our results are comparable with those reported in the literature. The challenge is to extend this analysis to a corpus consonants (63 articulations), as explained in the following sections.

4.3. PARAFAC model extended to consonants

In section 3, it was shown that the individual speaker models (PCA) need four components to explain about 91% of the variance. Figure 3 displays the variance explanation and RMSE related to the reconstruction of the tongue upper contour of our seven subjects by a PARAFAC model assessed by means of LOOCV. It appears that 25 components are not enough to explain the variance that the individual PCA models reach with 4 components. We see that to drive all articulatory models from the same set of PARAFAC control parameters, we need at least the same number of components as the total number of components for each subject when using individual PCA models (7x4). We conclude that PARAFAC is not able to take into account the dimensionality reduction that could be expected from the fact that the speakers have produced the same set of phonemes, even though they used different articulatory strategies. This problem is very likely related to the fact that inter-speaker variability cannot be efficiently represented in linear terms.

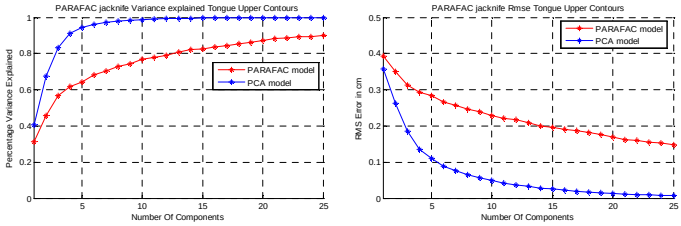

FIGURE 3 – LOOCV PARAFAC model as a function of number of components for the tongue upper contour of the seven speakers *PB, YL, HL, AA, MG, AK* and *MGO*. Left: variance explained. Right: RMSE.

4.3.1. Multiple linear regression between control parameters of couple of subjects

In the previous sections we attempted to model the tongue contour by using a reduced set of control parameters common to all speakers. This section presents an alternative approach, aiming at solving the problem of driving the contours of one target speaker from those of a source speaker, using a large number of PCA components. This solution does not allow interpreting the semantics of the components, but provides a practical solution to the normalisation problem. In this experiment, we attempted to predict the PCA control parameters of a target subject π_{TS} from the PCA control parameters of a source subject π_{SS}. Formally, a MLR model, given *Cmp* components, is expressed by: $\pi_{TS\ i} = \beta_1\pi_{SS1} + \beta_2\pi_{SS2} + \ldots \beta_i\pi_{SSi} + Y_i$, for $i = 1, 2 \ldots Cmp$, β being the coefficients of the linear regression.

We have built MLR models between each possible combination of couple of subjects.

Figure 4 shows the evaluation for subject *PB*. It appears that, the model gave strong signs of being over-fitted from the tenth component on. So, we discarded the meaningless components. Nevertheless, with the 10 first components, the MLR model is able to predict the tongue contour of subject TS from subject SS accounting for about 65% of the variance and with a RMSE of 0.38 cm.

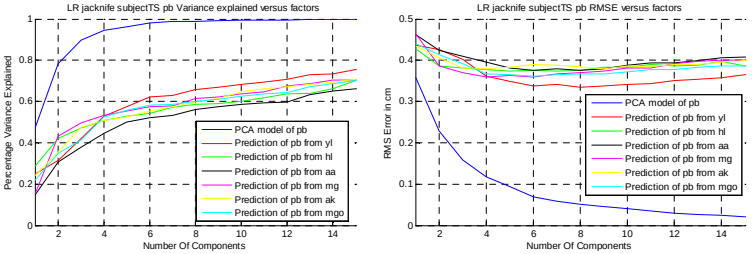

FIGURE 4 - RMSE of LOOCV MLR models between control parameters of *PB* and the other subjects as a function of number of components.

5. Conclusions and perspectives

We applied individual PCA models and multi-linear decomposition methods to model the tongue upper contours of 63 French phonemes extracted from an MRI database of 7 French speakers. As far as we know, this is one of the few studies that includes both vowels and consonants. The primary focus of this study was to establish a model that represents different speaker articulatory strategies. The experiments carried out showed that such a kind of model is possible, using 4 components, with an RMSE of 0.13 cm for the individual PCA models and 0.29 cm for the PARAFAC model, accounting for a variance of 91% and 62%, respectively. We also performed a more practical solution in which a large number of components were used to make a given target subject more likely predictable from a source subject. Using 10 components, the RMS error was 0.38 cm accounting for about 65% of the variance explanation.

The present study shows that linear methods may not offer a good solution to model tongue variations among different speakers, especially in the presence of consonants. There is indeed an inter-speaker variability due to speaker independent control strategies that might not be possible to model with linear methods. Thus, future work is to be directed at using non linear methods.

Acknowledgements

We sincerely thank all our kind and patient subjects. We thank also S. Masaki, S. Takano, I. Fujimoto, and Y. Shimada (ATR, Kyoto, Japan) for the MRI data on the first subject. This work has been partially supported by the French ANR-08-EMER-001-02 grant *ARTIS* (Articulatory inversion from audio-visual speech for augmented speech presentation).

Bibliography

ANANTHAKRISHNAN, G., BADIN, P., VALDÉS, J. A., & ENGWALL, O. (2010). Predicting unseen articulations from multi-speaker articulatory models., (pp. 1588-1591). Makuhari, Japan.

BADIN, P., & SERRURIER, A. (2006). Three-dimensional modeling of speech organs: Articulatory data and models. *In IEICE Technical Report , 106* (177), 29-34.

BEN YOUSSEF, A., HUEBER, T., BADIN, P., & BAILLY, G. (2011). Toward a multi-speaker visual articulatory feedback system. *In Interspeech 2011* , 589-592.

GENG, C., & MOOSHAMMER, C. (2000). Modeling the German stress distinction., (pp. 161-164). Kloster Seeon, Germany.

HARSHMAN, R., LADEFOGED, P., & GOLDSTEIN, L. (1977). Factor analysis of tongue shape. *Journal of the Acoustical Society of America , 62* (3), 693-707.

HOOLE, P. (1998). Modelling tongue configuration in German vowel production. Dans R. Mannell, & J. Robert-Ribes (Éd.), *Australian Speech Science and Technology Association Inc.*, (p. paper 1096). Sydney, Australia.

HOOLE, P. (1999). On the lingual organization of the German vowel system. *Journal of the Acoustical Society of America , 106* (2), 1020-1032.

HOOLE, P., WISMUELLER, A., LEINSINGER, G., KROOS, C., GEUMANN, A., & INOUE, M. (2000). Analysis of the tongue configuration in multi-speaker, multi-volume MRI data., (pp. 157-160). Kloster Seeon, Germany.

HU, F. (2006). On the lingual articulation in vowel production: case study from Ningbo Chinese. Dans H. C. Yehia, D. Demolin, & R. Laboissière (Éd.). Ubatuba, SP, Brazil: UFMG, Belo Horizonte, Brazil.

RIU, J., & BRO, R. (2003). Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems , 65* (1), 35-49.

ZHENG, Y., HASEGAWA-JOHNSON, M., & PIZZA, S. (2003). Analysis of the three-dimensional tongue shape using a three-index factor analysis model. *The Journal of the Acoustical Society of America , 113* (1), 478-486.