Adaptation of orofacial clones to the morphology and control strategies of target speakers for speech articulation

Julian Andres Valdes Vargas

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité: **Signal, Image, Parole, Telecom (SIPT)**

Arrêté ministériel: 7 août 2006

Présentée par

Julián Andrés VALDÉS VARGAS

Thèse dirigée par **Pierre BADIN**

préparée au sein du **Département Parole & Cognition (DPC) de GIPSA-Lab**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique & Traitement du Signal (EEATS)**

Adaptation de clones orofaciaux à la morphologie et aux stratégies de contrôle de locuteurs cibles pour l'articulation de la parole

Adaptation of orofacial clones to the morphology and control strategies of target speakers for speech articulation

Thèse soutenue publiquement le **28 juin 2013**
devant le jury composé de :

M Michel DESVIGNES
Professeur INP, GIPSA-Lab, Grenoble, (Président)

M Yves LAPRIE
DR CNRS, LORIA, Nancy, (Rapporteur)

M Rudolph SOCK
Professeur, IPS, Université de Strasbourg, (Rapporteur)

M Thierry LEGOU
IR1 CNRS, Laboratoire Parole et Langage, Marseille, (Examinateur)

M Pierre BADIN
Directeur de recherche, GIPSA-Lab, Grenoble, (Directeur de thèse)

Université Joseph Fourier / Université Pierre Mendès France / Université Stendhal / Université de Savoie / Grenoble INP

Abstract

The capacity of producing speech is learned and maintained by means of a perception-action loop that allows speakers to correct their own production as a function of the perceptive feedback received. This auto feedback is auditory and proprioceptive, but not visual. Thus, speech sounds may be complemented by augmented speech systems, i.e. speech accompanied by the virtual display of speech articulators shapes on a computer screen, including those that are typically hidden such as tongue or velum. This kind of system has applications in domains such as speech therapy, phonetic correction or language acquisition in the framework of Computer Aided Pronunciation Training (CAPT). This work has been conducted in the frame of development of a visual articulatory feedback system, based on the morphology and articulatory strategies of a reference speaker, which automatically animates a 3D talking head from the speech sound. The motivation of this research was to make this system suitable for several speakers. Thus, the twofold objective of this thesis work was to acquire knowledge about inter-speaker variability, and to propose vocal tract models to adapt a reference clone, composed of models of speech articulator's contours (lips, tongue, velum, etc), to other speakers that may have different morphologies and different articulatory strategies.

In order to build articulatory models of various vocal tract contours, we have first acquired data that cover the whole articulatory space in the French language. Midsagittal Magnetic Resonance Images (MRI) of eleven French speakers, pronouncing 63 articulations, have been collected. One of the main contributions of this study is a more detailed and larger database compared to the studies in the literature, containing information of several vocal tract contours, speakers and consonants, whereas previous studies in the literature are mostly based on vowels. The vocal tract contours visible in the MRI were outlined by hand following the same protocol for all speakers.

In order to acquire knowledge about inter-speaker variability, we have characterised our speakers in terms of the articulatory strategies of various vocal tract contours like: tongue, lips and velum. We observed that each speaker has his/her own strategy to achieve sounds that are considered equivalent, among different speakers, for speech communication purposes. By means of principal component analysis (PCA), the variability of the tongue, lips and velum contours was decomposed in a set of principal movements. We noticed that these movements are performed in different proportions depending on the speaker. For instance, for a given displacement of the jaw, the tongue may globally move in a proportion that depends on the speaker. We

also noticed that lip protrusion, lip opening, the influence of the jaw movement on the lips, and the velum's articulatory strategy can also vary according to the speaker. For example, some speakers roll up their uvulas against the tongue to produce the consonant /ʁ/ in vocalic contexts. These findings also constitute an important contribution to the knowledge of inter-speaker variability in speech production.
In order to extract a set of common articulatory patterns that different speakers employ when producing speech sounds (normalisation), we have based our approach on linear models built from articulatory data. Multilinear decomposition methods have been applied to the contours of the tongue, lips and velum. The evaluation of our models was based in two criteria: the variance explanation and the Root Mean Square Error (RMSE) between the original and recovered articulatory coordinates. Models were also assessed using a leave-one-out cross validation procedure. The purpose of using such a method was to verify the capabilities of models to generalize by evaluating their performance on data that were not used for training. In order to model the tongue, lips and velum contours with a common set of components for all speakers, several multilinear decomposition methods were performed and compared. Join PCA gave the best results among other techniques. In conclusion, we have found that there is a considerable reduction in terms of number of components when using joint PCA, compared to the total number of components needed by the individual PCA models of all speakers. These modelling results constitute an important extension, of the studies available in the literature, to more speakers, more articulations (consonants) and more articulors (lips, velum).

Keywords: MRI, vocal tract contours, inter-speaker variability, speech articulatory modelling, speaker normalisation, factor analysis, linear decomposition

Résumé

La capacité de production de la parole est apprise et maintenue au moyen d'une boucle de perception-action qui permet aux locuteurs de corriger leur propre production en fonction du retour perceptif reçu. Ce retour est auditif et proprioceptif, mais pas visuel. Ainsi, les sons de parole peuvent être complétés par l'affichage des articulateurs sur l'écran de l'ordinateur, y compris ceux qui sont habituellement cachés tels que la langue ou le voile du palais, ce qui constitue de la parole augmentée. Ce type de système a des applications dans des domaines tels que l'orthophonie, la correction phonétique et l'acquisition du langage. Ce travail a été mené dans le cadre du développement d'un système de retour articulatoire visuel, basé sur la morphologie et les stratégies articulatoires d'un locuteur de référence, qui anime automatiquement une tête parlante 3D à partir du son de la parole. La motivation de cette recherche était d'adapter ce système à plusieurs locuteurs. Ainsi, le double objectif de cette thèse était d'acquérir des connaissances sur la variabilité inter-locuteur, et de proposer des modèles pour adapter un clone de référence, composé de modèles des articulateurs de la parole (lèvres, langue, voile du palais, etc.), à d'autres locuteurs qui peuvent avoir des morphologies et des stratégies articulatoires différentes.

Afin de construire des modèles articulatoires pour différents contours du conduit vocal, nous avons d'abord acquis des données qui couvrent l'espace articulatoire dans la langue française. Des Images médio-sagittales obtenues par Résonance Magnétique (IRM) pour onze locuteurs francophones prononçant 63 articulations ont été recueillis. L'un des principaux apports de cette étude est une base de données plus détaillée et plus grande que celles disponibles dans la littérature. Cette base contient, pour plusieurs locuteurs, les tracés de tous les articulateurs du conduit vocal, pour les voyelles et les consonnes, alors que les études précédentes dans la littérature sont principalement basées sur les voyelles. Les contours du conduit vocal visibles dans l'IRM ont été tracés à la main en suivant le même protocole pour tous les locuteurs.

Afin d'acquérir de la connaissance sur la variabilité inter-locuteur, nous avons caractérisé nos locuteurs en termes des stratégies articulatoires des différents articulateurs tels que la langue, les lèvres et le voile du palais. Nous avons constaté que chaque locuteur a sa propre stratégie pour produire des sons qui sont considérées comme équivalents du point de vue de la communication parlée. La variabilité de la langue, des lèvres et du voile du palais a été décomposé en une série de mouvements principaux par moyen d'une analyse en composantes principales (ACP). Nous avons remarqué que ces mouvements sont effectués dans des proportions différentes en fonction du locuteur. Par exemple, pour un déplacement donné de la mâchoire, la

langue peut globalement se déplacer dans une proportion qui dépend du locuteur. Nous avons également remarqué que la protrusion, l'ouverture des lèvres, l'influence du mouvement de la mâchoire sur les lèvres, et la stratégie articulatoire du voile du palais peuvent également varier en fonction du locuteur. Par exemple, certains locuteurs replient le voile du palais contre la langue pour produire la consonne /ʁ/. Ces résultats constituent également une contribution importante à la connaissance de la variabilité inter-locuteur dans la production de la parole.

Afin d'extraire un ensemble de patrons articulatoires communs à différents locuteurs dans la production de la parole (normalisation), nous avons basé notre approche sur des modèles linéaires construits à partir de données articulatoires. Des méthodes de décomposition linéaire multiple ont été appliquées aux contours de la langue, des lèvres et du voile du palais. L'évaluation de nos modèles repose sur deux critères: l'explication de la variance et l'erreur quadratique moyenne. Les modèles ont également été évalués en utilisant une procédure de validation croisée. Le but de l'utilisation de telle procédure était de vérifier la capacité de généralisation des modèles en évaluant leurs performances sur des données qui n'ont pas été utilisées pour leur construction. Afin de modéliser la langue, les lèvres et le voile du palais avec un ensemble commun de composantes pour tous les locuteurs, plusieurs méthodes de décomposition linéaires multiple ont été utilisées et comparées. L'ACP conjointe a donné les meilleurs résultats. En conclusion, nous avons constaté une réduction considérable en termes de nombre de composantes nécessaires lors de l'utilisation d'ACP conjointe, par rapport au nombre total de composantes nécessaires par les modèles ACP individuels de tous les locuteurs. Ces résultats de modélisation constituent une extension importante des études disponibles dans la littérature, à des locuteurs plus nombreux, incluant de plus nombreuses articulations (en particulier les consonnes) et de plus nombreux articulateurs (lèvres, voile du palais).

Mots-clés: IRM, contours du conduit vocal, variabilité inter-locuteur, Modélisation articulatoire, normalisation du locuteur, analyse factorielle, décomposition en composantes linéaires

Acknowledgement

First of all, I would like to express my sincere gratitude to my advisor Dr. Pierre Badin for his guidance during this thesis work.

Special thanks go to Prof. Michel Desvignes (GIPSA-Lab, Grenoble, France) for accepting to be the president of the jury, to Dr. Yves Laprie (LORIA, Nancy, France) and Prof. Rudolph Sock (IPS, Strasbourg, France) for accepting to evaluate my thesis as reviewers, and Thierry Legou for accepting to participate as examiner.

I would also like to thank LORIA (Nancy, France) to fund an extension of four supplementary months to work on the redaction of this manuscript.

I am grateful to Laurent Lamalle (CHU, Grenoble) for his help in MRI recording. Besides, I sincerely thank all our kind and patient speakers.

Thanks to all GIPSA-Lab, especially the members of the Speech and Cognition Department. I am also grateful for all the people that I have had the opportunity to meet, especially for: Emre Çolakoğlu, Ronald Meijers, Zuheng Ming, Will Barbour, Kelem Gomes and Noel Hanna. The time that we spent together as colleagues and friends was fun and enriching for me.

Sylvain et Amélie, vous avez été là pour moi au début de mon séjour en France. Je vous remercie de votre accueil, vos conseils et votre guide dans la vie pratique française. Vous étiez ma première motivation pour venir en France et apprendre le Français.
Felipe y María, mil gracias por ser mis consejeros y amigos.
Mauricio Toro, tener gente como vos a mi lado me ha hecho sacar lo mejor de mí mismo. Tu amistad me es invaluable.
Carlos Galvis Salzburg, gracias por darme la patadita que necesitaba para comenzar mi doctorado. Creo que no lo hubiera hecho sin la conversación de aquel día.
Mi banda latina en Grenoble, ustedes son lo más importante que tengo aquí. No se imaginan el bien que me hace tenerlos cerca. Cada momento memorable es un sello en mi vida. Pañuelitos, gracias por darle música a mi vida.
Padres (Hermes Valdés y Elizabeth Vargas), ustedes son el motor de mi vida. Espero que estén orgullosos de mí. Los amo.
Dios, gracias papá. Gracias por ese milagro que cambio mi vida para siempre.

Content

List of figures

List of tables

Acronyms and terms

CAPT: Computer Aided Pronunciation Training

EM: Expectation-Maximization

EMA: Electro-Magnetic Articulography

F1: Premier formant

F2: Second formant

F3: Third formant

FULLTNG: full tongue contour

GMM: Gaussian Mixture Model

HMM: Hidden Markov Model

LOOCV: leave-one-out cross validation procedure

MLR: Multilinear Regression

MRI: Magnetic Resonance Imaging

PARAFAC: Parallel factor

PCA: Principal Analysis Component

RMSE/RMS error: Root Mean Square Error

UPPERTNG: upper tongue contour

VCV: Consonant-Vowel- Consonant

Introduction

Motivation of research

Speech production requires a precise mastery of the various articulators in the vocal tract (i.e. lips, jaw, tongue, velum, epiglottis, etc.). This skill is learned and maintained by means of a perception-action loop that allows the speaker to correct his/her production as a function of the perceptive feedback received (Matthies et al., 1996; Bailly, 1997). The feedback that a speaker receives from his/her own production is auditory and proprioceptive, but not really visual. On the other hand, Erber (1975) has demonstrated the contribution of lip vision to the perception of speech, while Badin et al. (2010b) have recently shown the contribution of tongue vision to the recognition of consonants in adverse audio signal to noise ratios. Besides, a number of studies have explored the importance of perceptive feedback in domains such as speech therapy, phonetic correction or language acquisition (Badin et al., 2010b). Thus, visual articulatory feedback systems, which aim at supplying the speaker with a visual return of the articulation just pronounced, have proved to be suitable to improve speech intelligibility (Badin et al., 2010a))

The Speech & Cognition Department at GIPSA-lab has therefore developed an acoustic-to-articulatory inversion system. Such a system is able to create a visual articulatory feedback from the acoustic signal (Ben Youssef et al., 2011a). This system is based on a fairly complete orofacial clone made of articulator models like the jaw, lips, tongue, velum, etc., that can play augmented speech, i.e. speech accompanied by the virtual display of articulators, including those that are typically hidden such as the tongue or the velum. This system is potentially useful for applications in the domain of Computer Aided Pronunciation Training (CAPT) and speech rehabilitation.

However, the clone of our visual articulatory feedback system developed at GIPSA-lab is based on articulatory data acquired on a single speaker (Badin & Serrurier, 2006). Therefore, the clone represents faithfully the characteristics of one specific speaker, but not necessarily those of other speakers that may have different morphologies and different articulatory control strategies. Thus, the twofold objective of this thesis work was to acquire knowledge about inter-speaker variability, and to propose vocal tract models to adapt a reference clone to a variety of speakers.

The main difficulty to model vocal tract contours is the variability in terms of morphology and articulatory strategies of different speakers. One important issue is what we call the normalisation problem: how can speaker-specific models of the orofacial clone be adapted to other speakers? This task 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.

Organization of the manuscript

The thesis manuscript is organised as follows.

Chapter 1 (State of the art on articulatory normalisation) describes previous studies about articulatory normalisation. This chapter discusses the results of several studies, principally made for vowels, based on different recording methods, languages, number of speakers, measured articulator points and corpuses.

Chapter 2 (Articulatory speech data) focuses on the acquisition and edition of MRI data recorded for eleven French native speakers. In addition, one of the speakers was recorded three times for the same corpus. The French corpus recorded for the eleven speakers consisted of ten oral vowels, 3 nasal vowels and 10 consonants articulated in 5 symmetric vocalic contexts. Firstly, this chapter describes the articulatory phonetic characteristics of the French language. Secondly, the properties of our corpus are presented. Furthermore, a comparison between corpuses described in the literature and the corpus of this work is presented. This comparison is made in terms of number of speakers, articulator measurements, and size of the corpuses. One of the main contributions of this study is a more detailed and larger database of the vocal tract contours than those reported in the literature. This chapter also details how the vocal tract contours are hand-traced and how the unknown parts are predicted. For instance, the sublingual cavity and the tongue tip are not obviously identified for all the articulations.

Chapter 3 (Articulatory characterisation and individual models of speakers) aims to characterize our speakers as regards articulatory control strategies. In other words, this chapter presents a qualitative description and comparison of each speaker's data. During the reading of this chapter, the reader will go through several experiments that compare speakers in terms of: individual speakers' models of different contours (i.e. lips, tongue and velum), vocal tract measurements and vocalic formants. Finally, an important issue called intra-speaker variability is studied.

Chapter 4 (Individual and multilinear models of the tongue, lips and velum contour) presents results of different linear and multilinear models built with a corpus of vowels and consonants in vocalic context. Firstly, we explain how the mean is subtracted from the data. Then, the results of individual speaker models are presented and compared to each other in terms of relative variance explained – i.e. ratio of the variance of reconstructed data over the variance of original measured data – and the Root Mean Square Error (RMSE). Third, the results of various multilinear decomposition models of the tongue, lips and velum are also presented. In order to have a reference starting point for the tongue models, our models are first limited to a repertoire of only French vowels and compared to the studies in the domain. Then, multilinear models are built for a more extended corpus of vowels and consonants in vocalic context.

Finally, **Chapter 5** (Conclusions and perspectives) presents the final conclusions and summarizes the contributions of this study. It also proposes tentative studies for future works.

Note: related project ARTIS

The work presented in this thesis work contributed to the French ANR-08-EMER-001-02 ARTIS project which involves collaboration between GIPSA-Lab, LORIA, IRIT, and TSI-Télécom ParisTech. The main objective of this research project is to develop a system of visual articulatory feedback that can deliver augmented speech by means of a virtual talking head.

Chapter 1. State of the art on articulatory normalisation

1.1. Introduction

The main difficulty to model the vocal tract contours of several speakers is the variability in terms of morphology and articulatory strategies. One important issue is what we call the normalisation problem in two different aspects: morphology of the vocal tract and articulatory strategies. According to the literature, the problem of morphological normalisation is approached in two different manners: the first strategy is purely geometric (Hashi et al., 1998; Engwall, 2004; Geng & Mooshammer, 2009; Apostol et al., 2004). The idea of this approach is to reduce the cross-speaker variability by applying scaling transformations. The second method is also based in applying scaling transformations but observing the acoustic consequences of it (Mathieu & Laprie, 1997; Boë et al., 2000). On the other hand, the goal of articulatory normalisation is to build models that extract common articulatory patterns used by different speakers (Harshman et al., 1977; Hoole, 1998; Hoole, 1999; Hoole et al., 2000; Geng & Mooshammer, 2000; Zheng & Johnson, 2003; Hu, 2006; Ananthakrishnan et al., 2010). In other words, the purpose of articulatory normalisation is to control the models of different speakers using the same number of components.

Proposing solutions to the normalisation problem, concerning the two aspects explained above, 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.

In this chapter we first describe the linear decomposition methods used in previous studies. Second, studies about articulatory normalisation, in the literature, are presented. Finally, studies about geometric and acoustic normalisation, presented in the literature, are described.

1.2. Linear decomposition methods

This section describes all the linear decomposition methods used in different studies in the literature.

1.2.1. PCA

PCA is a two-way factor analysis approach often used for dimensionality reduction and analysis of data sets to extract regularities (Pearson, 1901). Consider articulatory

measurements $X_s = [x_1, x_2, \ldots, x_A]$ for the speaker s which consists of vectors of measurements ($1 \leq n \leq N$) for the articulations from 1 to A. Such a 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 measurement to the components) by the following equation:

$X_s = \pi_s * C_s^T + \xi_s$, where ξ is the residual error.

Figure 1-1 illustrates PCA with a schematic representation.

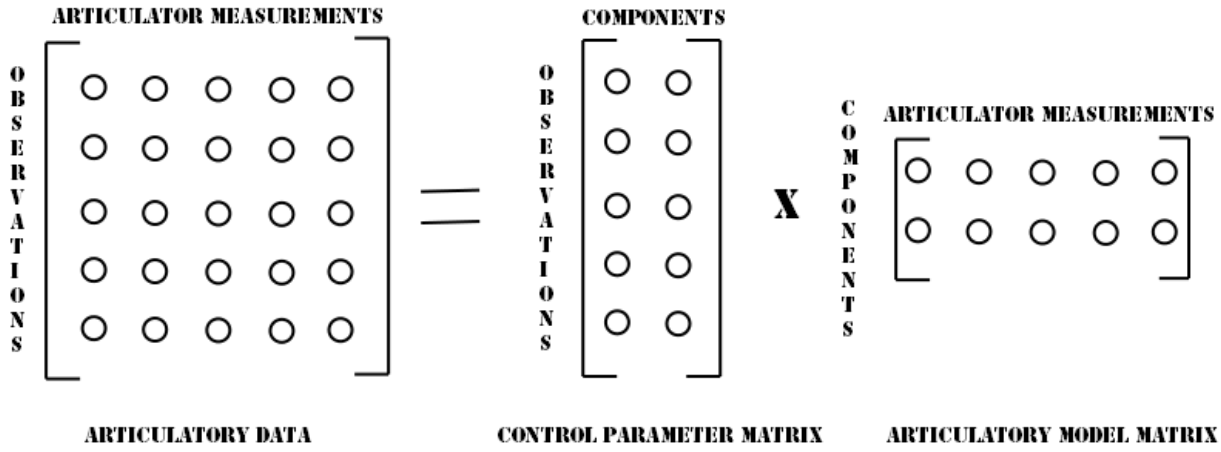


Figure 1-1 – Schematic representation of PCA

1.2.2. Parallel factor (PARAFAC)

PARAFAC is a three-way factor analysis approach which is often used to decompose 3-dimensional data (Harshman, 1970; Harshman & Lundy, 1994). In our specific case, the three dimensions are related to articulations, articulator measurements and speakers, respectively. The difference between PARAFAC and PCA is that PARAFAC extracts patterns for several speakers, while PCA only decomposes the data of an individual speaker. The data of a given speaker X_s, decomposed by PARAFAC, can be seen as:

$X_s = \pi * \Phi_s * C^T + \xi_s$ where ξ is the residual error.

π is the matrix of universal control parameters which represents the common patterns extracted for all speakers. The matrix Φ is a diagonal matrix which provides speaker-specific weights to the contribution of each component. The matrix of coefficients C, also called the universal articulatory model, represents the contribution of each articulator measurement to the universal components extracted for all speakers. Figure 1-2 illustrates PARAFAC with a schematic representation.

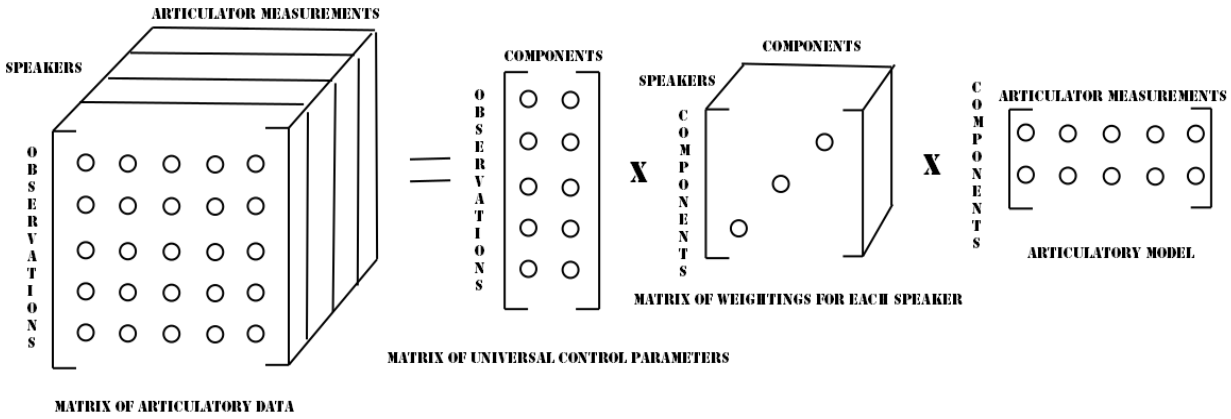


Figure 1-2 – Schematic representation of PARAFAC

1.2.3. Tucker

The Tucker decomposition, also called three-mode PCA is an extension of PARAFAC (Tucker, 1966). The matrices of universal control parameters (Π), speaker-specific weights (Φ) and coefficients (C) represent the same as PARAFAC. Oppositely, these matrices can be decomposed with different number of components each. In other words, TUCKER allows the extraction of different number of patterns for each dimension. In PARAFAC all the dimensions are decomposed with the same number of components. The data of a given speaker X_s, decomposed with TUCKER, can be represented as:

$$\sum_{i=1}^{w1}\sum_{m=1}^{w2}\sum_{n=1}^{w3} \pi * \Phi s * C * G + \xi_s \text{ where } \xi \text{ is the residual error.}$$

The extra matrix G is called the core matrix which contains the factor loadings for all three modes of variation. This matrix explains the interaction between the components extracted for each mode of variation. The Figure 1-3 illustrates TUCKER with a schematic representation using the Kronecker multiplication (⊗).
The Kronecker multiplication is computed as follows:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

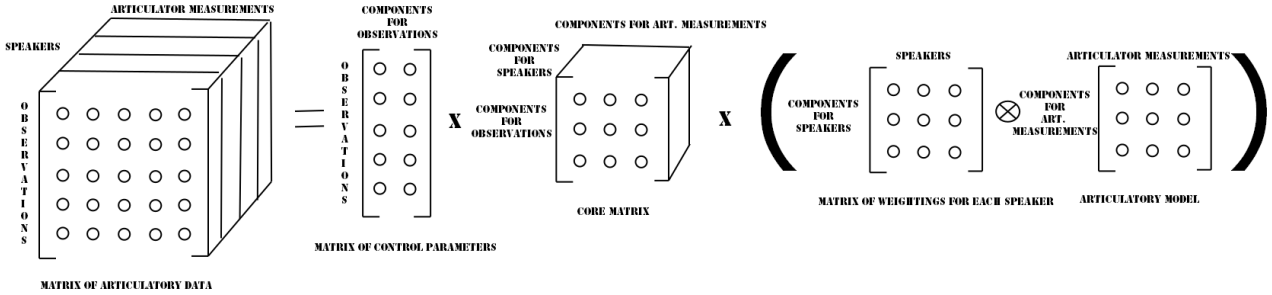


Figure 1-3 - Schematic representation of TUCKER

1.2.4. Joint PCA

This method has been proposed by Ananthakrishnan et al. (2010), but named as two-level PCA. In this study it will be called joint PCA. Joint PCA is an extension of PCA to decompose the data of several speakers instead of an individual set of data. In this technique, data are decomposed using the regular PCA but forced to extract a universal set of control parameters for all speakers. The data of several speakers are put together as $X = [X_{s1}; X_{s2}; \ldots; X_{sy}]$ in which each speaker is a set of articulatory measurements $X_s = [x_1, x_2, ..., x_A]$. A graphical representation of this method is given by the Figure 1-4.

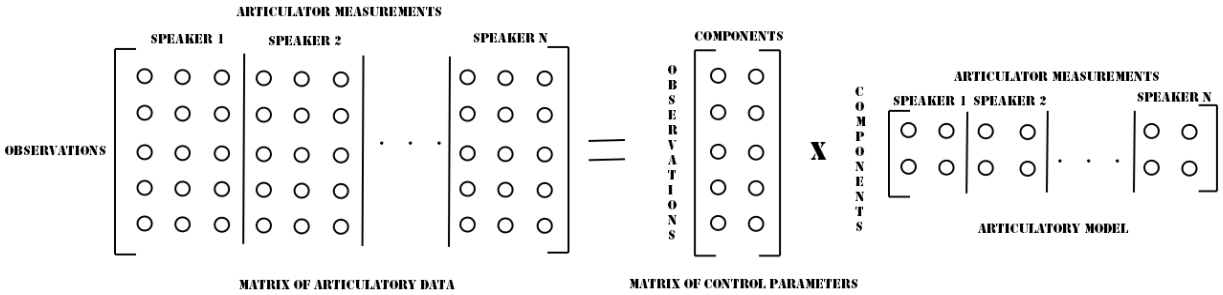


Figure 1-4 – Schematic representation of joint PCA

1.3. Previous studies on articulatory normalisation based on linear models

Several studies based on measurements using Electromagnetic Articulography (EMA) and Magnetic Resonance Imaging (MRI), principally made for vowels, have explored the problem of speaker articulatory normalization.

Harshman et al. (1977) , performed a Parallel Factor analysis (PARAFAC) on the vowel production of five American English speakers. The tongue postures were decomposed in two factors which explained 92.7 % of the variance.

Hoole (1998) provided a two factor PARAFAC solution on the German vowel system in three different consonant context (/p, t/, k/). He elaborated models for individual consonant contexts and for multiple consonant Contexts. For the /p/ and /k/ context a

two-factor independent models were successfully extracted. The explained variance amounted to about 92.3 % and the RMS error to 1.24 mm for both models. However, the two-factor model for the /t/- context ran into problems and the resulting solution gave strong signs of being degenerated. The extracted two-factor solution for the complete dataset presents an increase in model error compare to the individual models, the explained variance now amounted to 80 % and the RMS error to 1.9 mm.

In Hoole (1999) is shown how the PARAFAC model error can be further analyzed to extract an additional component. The approach consists of examining the error of the two-factor PARAFAC model by subtracting the articulatory data predicted from the original data (model error of each speaker). The error datasets measured the displacement of the tongue required to move each articulator point from its position predicted by the PARAFAC model to its actual position. Then, PCA was employed to extract an extra-component. The final model explained over 90 % of the variance. However, the fact of having two final models (two factor PARAFAC model and single factor PCA model) for each speaker is not desirable in terms of normalization.

Hoole et al. (2000) performed PARAFAC on a set of MRI data of nine German speakers, seven German vowels in five different contexts. Two factors accounted for about 87 % of the variance explanation with an RMS error of about 2.2 mm.

Geng & Mooshammer (2000) provided a two factor PARAFAC solution. The speech material consisted of six German speakers and fifteen German vowels in /t/-context recorded by EMA. Two factors amounted for 96 % of the variance explanation and about 2 mm of reconstruction error.

Zheng & Johnson (2003) showed that a two-factor model results in a stable solution that explains about 70 % of the variance. The 3D coordinate data consisted of MRI images of five American English speakers pronouncing nine English vowels.

Hu (2006) presented a study on Chinese dialect. Seven Ningbo speakers were recorded pronouncing ten vowels by means of EMA. Three transducers were mounted on the tongue.Two factors explained about 90 % of the variance.

More recently, Ananthakrishnan et al. (2010) modelled 13 French vowels using PARAFAC, TUCKER and joint PCA with two components. The average reconstruction error with PARAFAC, over three speakers, was 3.9 mm, accounting for a variance of 71 %. The reconstruction error for TUCKER and joint PCA was 4 mm and 3.6 mm, respectively. The modelling was also extended to a corpus of 73 articulations including consonants /p t k f s ʃ m n ʁ l/ in different vocalic contexts. In this case, the RMS error for the models with PARAFAC, TUCKER and joint PCA was about 3.5 mm for all the methods. The number of components needed increased to 5. However, the results in terms of variance explanation were not given.

Apart from the studies of Hoole (1998) and Geng & Mooshammer (2000) which include experiments with some consonants, only Ananthakrishnan's study involved a larger set of consonants.

1.4. Previous studies on geometric and acoustic normalisation

Several studies have explored the problem of morphological normalisation. The first group of studies is based on applying scaling transformations to the vocal tract shape to reduce the cross-speaker variability:

Hashi et al. (1998) applied a normalisation procedure to the articulatory data of 20 English and 8 Japanese speakers with different vocal tract shapes and sizes. In this normalization scheme, scaling was applied to point-parametrized vowel postures (i.e. tongue shapes represented by resampled points of the tongue contour). The purpose was to minimize the cross-speaker variability. The final results offered a reduction of cross-speaker variance in the y dimension of the normalized space. They concluded that the variability due to different palatal heights in the y dimension could be ''successfully removed'' from the normalized data by means of a scaling transformation. The median reduction in the cross-speaker standard deviation was on the order of 1 mm. They mentioned that this decrease in variance may seem small but should be interpreted with reference to the relatively small articulatory range for the y dimension in the oral cavity. However, the normalisation procedure did not give comparable results to reduce the cross-speaker variability in the x dimension. They attributed this result to the difference between articulatory speaker strategies. Some speakers usually positioned their tongues in an anterior position, across all vowels, while other speakers habitually positioned their tongues in a posterior position. Thus, the method was successful to reduce variability caused by morphological divergences but not the one regarding dissimilarities in articulatory strategies. The results were consistent within both languages: English and Japanese.

Engwall (2004) made a study based on MRI of nine speakers. The idea was to define scaling factors, in an automatic way, which can adapt a 3D tongue model to new speakers. This technique was able to estimate a tongue shape that was not included in the training set with a precision of 1.5 mm for the midsagittal plane and 1.7 mm for the whole 3D tongue.

Geng & Mooshammer (2009) showed a method to normalise vowel systems by minimizing the variability between different speakers with respect to an average tongue shape. The data of this study consisted of 15 German vowels, in the consonantal context /t/, recorded for seven speakers. Overall, the cross-speaker

variability was reduced for most vowels. However, for certain cases there was even more cross-speaker variability after normalization.

Apostol et al. (2004) proposed a transformation to reduce the inter-speaker variability of eight French vowels. The transformation is based on the frequency ratios of the formants across the different speakers. On the whole, the global variability clearly decreases after speaker transformation. Nevertheless, for the vowel /y/ no reduction of variability was observed, and for the vowel /o/ the variability increases. These limitations can be related with different vocal tract strategies, among the speakers, that the transformation is not able to take into account.

The second group of studies, apart from applying scaling transformations, observe the acoustic consequences of having different vocal tract sizes:

Mathieu & Laprie (1997) adapted Meada's two-dimensional articulatory model (1988) to a new speaker. The study was based on MRI data of eleven French vowels. The adaptation consists in modifying two scale factors which control the sizes of pharynx and mouth cavities. The results were evaluated by measuring the error between formants of the original data and data predicted from the model. Before adaptation the mean error, over the whole set of vowels, was 46 Hz for F1, 209 Hz for F2 and 184 Hz for F3. After the adaptation procedure the mean error, over the whole set of vowels, was 49 Hz for F1, 125 Hz for F2 and 170 Hz for F3. The vocalic space of the given speaker was covered by the adapted model. In other words, the adapted model was validated by checking its capabilities to preserve the phonemic identity of the source speaker.

Boë et al. (2000) used the Maeda model and applied scale factors to simulate different vocal tract sizes (children, females and males). Results show that the adult's vocal tract is not a uniform scaled version of a child vocal tract. They also showed that in order to simulate the vocal tract growth, articulatory adjustments were made on constriction size as well as constriction location.

1.5. Conclusion

In this chapter previous studies about articulatory normalisation were presented. We observed that most of the studies in the domain propose PARAFAC solutions using 2 components to represent the tongue movement of several speakers. The studies are mostly made for vowels and represent a variance explained that ranges between 76.2% and 96%. Besides, previous studies about geometric normalisation were described. Overall, the various geometric methods presented in the literature are successful to reduce variability caused by morphological divergences but not the one regarding dissimilarities in articulatory strategies. The next chapter describes the data recorded and used to build linear models of the vocal tract contours.

Chapter 2. Articulatory speech data

2.1. Introduction

An important element of any statistical articulatory model is the data. In order to decide what kind of information must be included in the corpus, one first has to clarify the aim of the models that data are used for. In our particular case, we want to build linear models to extract articulatory patterns common to several French speakers. There are three main issues to be considered in the construction of the speech corpus for our articulatory models: (1) the inter-speaker variability, which refers to how much the speakers differ from each other, should be large enough to extract articulatory patterns that are as general as possible; (2) the articulatory phonetic coverage, which is related to the set of speech utterances produced by the speakers, should cover as much as possible the range of articulatory movements present in the French language; and (3) the database size, which is related to the articulatory data available, should be large enough for the articulatory models to estimate reliable statistical parameters. However in practice, a recording session is limited to about two hours. Thus, in order to fulfil the time condition, the size of the corpus has to be limited to certain vowels and consonants in vocalic context. The properties of this corpus will be detailed in the next sections.

In this chapter, the French language is first described in terms of articulatory phonetics. Secondly, the methods for articulatory data acquisition that have been mostly used in previous studies are presented. The criteria to choose the recording method for this study are also discussed. Third, the characteristics of our corpus are described. Fourth, a grid system used to obtain different representations of the tongue contour is explained. Finally, this chapter explains how the various vocal tract contours are edited and sometimes estimated.

2.2. French articulatory phonetics

This study was restricted to French language, as it is already a tough challenge to study variability and normalisation in one language. The speakers recorded for this study were thus chosen as native French speakers. Before moving forward to the description of corpus, one needs to describe the French language in terms of articulatory phonetics.

2.2.1. The principal organs of articulation

The first step is to identify the organs that are taken into account in speech production (see Figure 2-1), as described by McFarland (2009).

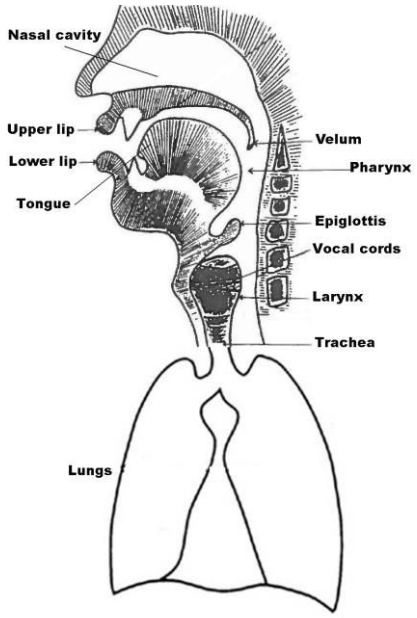


Figure 2-1- Organs implied in speech production (modified from (Léon, 2012))

Starting from the bottom, one can find the lungs. They are responsible for expelling the air required to generate sounds. The air goes then through the trachea which ends in a box called the larynx cartilage. Suspended in the larynx there are two bands of elastic tissue, called the vocal folds (often called vocal cords). If the vocal folds are abducted, voiceless sounds are produced, as in /p/. If the vocal folds are adducted and vibrate, they produce voiced sounds as in /a/.

The vocal tract above the vocal folds is made of mainly three parts, the pharyngeal region, the oral cavity and the nasal cavity. The connexion between the oral and nasal cavity is controlled by the velum. The pharyngeal region is the zone between the vocal folds and the velar region. The oral cavity goes from the velum up to the lips. The oral cavity contains the tongue, the upper and lower teeth and the palate. Finally, the jaw is a bony structure that carries both tongue and lower lips.

2.2.2. Vowels and consonants

The first fundamental difference between vowels and consonants is that vowels are produced with an open vocal tract and without blocking the airstream, while consonants are characterised by a constriction or an occlusion (Derivery, 1997). Besides, in the French language, vowels are monosyllabic whereas a consonant alone does not represent a syllable. That is not the case in English in which we can find

syllabic consonants. For example, the word people is pronounced [pi:pl] with two syllables.

According to Léon (2012) and Derivery (1997), on the one hand, vowels in French are characterised by four properties: height, backness, nasalisation and roundness. On the other hand, consonants can be distinguished by two properties: manner of articulation and place of articulation. The manner of articulation describes general mechanisms involved in the production of a given speech sound (i.e. friction or occlusion release). The place of articulation refers to the position of maximal narrowing in the vocal tract. Following sections describe vowels and consonants in the French language regarding the articulatory properties mentioned above.

2.2.2.1. Vowels

French vowels can be classified according to their height. For instance, vowels articulated close to the palate are considered as close vowels. In contrast, vowels produced with a low position of the tongue are counted as open vowels. There are intermediary classification stages between open and close vowels that go from close-mid vowels to open-mid vowels. Besides, vowels can be classified, according to their backness, as front or back vowels. For instance, vowels for which the tongue is positioned forward are considered front vowels. Oppositely, when the tongue is positioned towards the back, they are classified as back vowels. Moreover, vowels are classified according to their nasalisation as oral or nasal. Oral vowels are produced with the velum raised to prevent airstream going through the nasal cavity. Oppositely, nasal vowels are pronounced with a low velum to permit the air going through both cavities: oral and nasal. Furthermore, vowels can be also characterized by roundness of the lips as rounded or not rounded. The Figure 2-2 shows a graphical representation of the French oral vowels concerning the properties explained above.

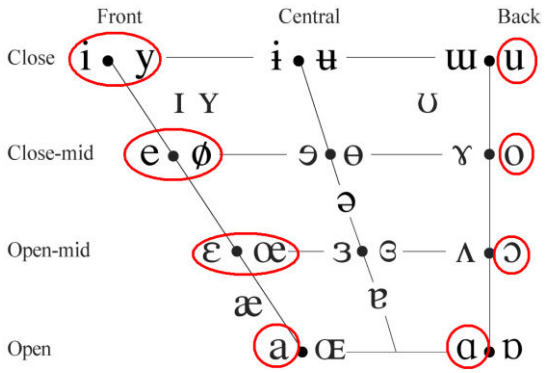


Figure 2-2 - Articulatory classification of oral vowels in the French language (Based on the International Phonetic Alphabet (IPA revised to 2005)). The red circles represent vowels included in the French language. Where symbols appear in pairs, the one to the right represents a rounded vowel

2.2.2.2. Consonants

According to Léon (2012) and Derivery (1997), consonants can be classified in six manners of articulation: voiced, voiceless, oral, nasal, occlusive and fricatives. They are either considered voiced or voiceless as regards the vibration or lack of vibration in the laryngeal region, respectively. According to the velum operation, consonants can be either identified as oral or nasal. The velum is closed for oral consonants and opened for nasal consonants. Besides, consonants are classified in occlusive and fricatives. Occlusive consonants are produced by blocking the air flow in the vocal tract. Similarly, fricative consonants are produced by constraining the air through a narrow zone between two articulators.

Moreover, consonants are classified as regards the place of articulation as it is shown in the Figure 2-3. Names of the places of articulation represent two regions in contact to produce the constriction. For instance, an apico-dental consonant is produced by constricting the apical region of the tongue against the upper teeth. The Table 2-1 summarised the articulatory classification of consonants regarding both aspects: place of articulation and manner of articulation (Léon, 2012; Derivery, 1997):

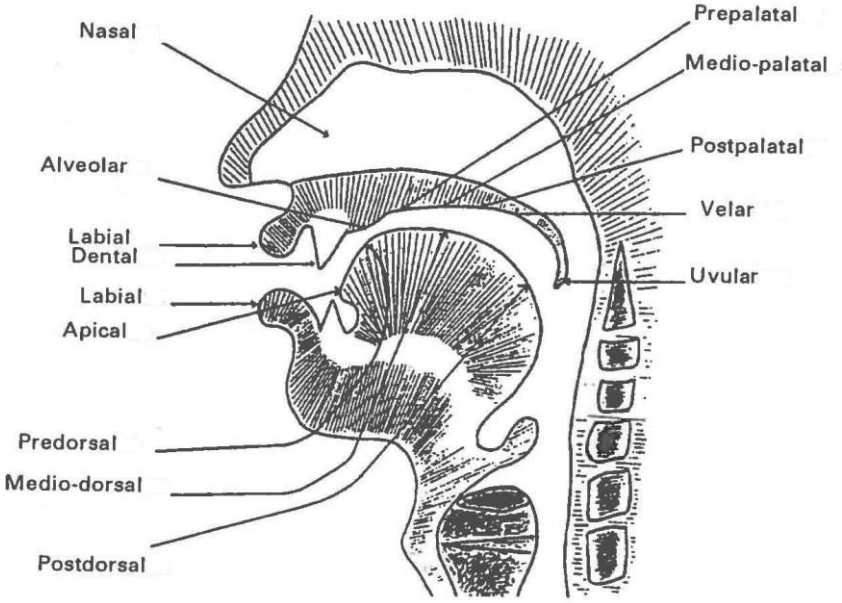


Figure 2-3 - Place of articulation of consonants (modified from (Léon, 2012))

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Table 2-1 - Articulatory classification of consonants in the French language regarding manner and place of articulation (Based on the International Phonetic Alphabet (IPA revised to 2005)). The red circles represent consonants included in the French language. Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible

According to Derivery (1997), the consonants /k/ and /g/ can belong simultaneously to two different categories of place of articulation. They can either be palatal when followed by a front vowel or velar when followed by a back vowel.

2.2.2.3. Variations of /R/

In French there are at least five variations of the consonant /ʁ/ (Léon, 2012). In the French spoken in France, the consonant /ʁ/ is uvular. However, in certain places like Quebec (Canada) it is articulated in the apico-alveolar zone, as the /r/ in many other romance languages like Spanish and Italian. It can be either pronounced with a single flap or multiple flaps. Other variations of the standard /ʁ/ exist and are characterized by either a strong or weak articulation. A weak articulation of /ʁ/ generally connotes a popular speaking way (called faubourienne in Paris). In Chapter 3, particular strategies of our speakers to produce the consonant /ʁ/ are presented.

2.3. Methods for articulatory data acquisition

Previous articulatory normalisation studies have been mainly based on three recording methods: X-ray radiography, Magnetic Resonance Imaging (MRI) and Electro-Magnetic Articulography (EMA). X-Ray was used for the first times by Meyer (1907) and Mosher (1927). The data collected by means of this method were useful to obtain pictorial representations of the tongue contour. However, the information in pictorial representation was relatively complete but not sufficient to accurately identify the vocal tract contours in the images. MRI has been used in several speech articulation studies since 1986 (Rokkaku et al., 1986). This method provides detailed information of the vocal tract. Nevertheless, the recorded speaker has to maintain the articulation for several seconds because of the relatively slow acquisition speed that characterises MRI systems. Aalto et al. (2011) hypothesized that sustained phonation may induce

greater variability into the data. However, they were not able to prove it. Another disadvantage of MRI is that speakers are recorded lying in supine position. The gravitational effects of this posture might have some influence on articulation (Tiede et al., 2000; Stone et al., 2007). Engwall (2003; 2006) stated that the supine position, with the speaker facing upward, affects the position and shape of the tongue, often decreasing the passage in the pharynx. However, this gravitational effect is moderate. EMA offers a good solution to track articulatory movements. The main drawback of EMA is that small electromagnetic receiver coils have to be glued on the articulators of interest. Thus, it is difficult to keep the sensors fixed during long recording sessions. Besides, sensors and wires in the mouth may perturb somehow the natural articulation. Furthermore, some regions of the vocal tract are not easily reached to glue the receiver coils (i.e. the velum, back of the tongue, etc.). Besides, Legou et al. (2008) used a method called static palatography. This method is based on the observation of the tongue print using a black paste spread on the tongue surface. The images obtained are basically related to the palate contour. The Table 2-2 shows a compact comparison of X-ray, MRI and EMA:

	EMA	**MRI**	**X-ray**
Whole Vocal Tract	No	Yes	Yes
Tongue imaging	Pellets	Full-length	Full-length
Time resolution	500 Hz	0-24 Hz	30-60 Hz
Health hazard	No	No	Yes
Quality of signal	Good	Good	Good
Head movement	Restricted	Restricted	Free/Restricted
Portable	No	No	No
Expensive	Yes	Yes	Yes

Table 2-2 - Comparison of 3 recording methods (modified from (Ridouane, 2006))

Despite its disadvantages, MRI offers more complete information of the vocal tract compared to EMA and more legible data compared to X-Ray. This has motivated the choice of MRI as the method for articulatory data acquisition in this study.

2.4. Experimental setup and protocol

During a recording session the speaker is asked to go through three different stages. First at all, the speaker is installed under the most comfortable conditions as possible on a bed that is shifted into the recording machine. Since the MRI machine produces

sounds that could be damaging for human ears, the speaker is protected with earplugs. However, during a recording session the speaker can call the MRI operator at any time by using an interphone. Second, all the recording properties of the machine have to be set up. Meanwhile, the speaker is requested not to move since the alignment recording properties of the machine are being defined. Only after this point, the speaker is instructed to pronounce and maintain the vocal tract shape of certain articulations between 8 and 43 seconds each. The articulations included in the corpus are described in the section 2.5.

2.4.1. MRI protocol

Stacks of sagittal MR images (slices) were recorded, for each articulation, using three different imaging systems. These systems were set up using the following parameters[1]: Echo Time (TE) which represents the time in milliseconds between the application of the 90° pulse and the peak of the echo signal. The Flip Angle (FA) refers to the angle of excitation relative to the main magnetic field direction. The Repetition Time (TR) is related to the amount of time that exists between successive pulse sequences applied to the same slice. The slice thickness represents the thickness of an imaging slice. The Field of View (FOV) is defined as the size of the spatial encoding area of the image. The spatial resolution determines how clearly defined the image looks. It is measured in millimetres per pixel (mm/pixel). Finally, the acquisition time refers to the period of time required to collect the image data. Table 2-3 shows the MRI recording protocol used for all speakers (some information is missing for speakers PB_1998 and PB_2002 because they were recorded long time ago and some values were not registered).

[1] http://fonar.com/glossary.htm

Speakers	Imaging System	TE	FA	TR	Slice Thickness
PB_2002					5 mm
PB_1998	1-Tesla MRI scanner Philips GyroScan T10-NT				3.6 mm
YL	Philips Gyroscan 1.5 Tesla scanner	3.5 ms	12°	5.6 ms	1.25 mm
HL	Philips Gyroscan 1.5 Tesla scanner	3.5 ms	12°	7.7 ms - 8.6 ms	1.25 mm
PB_2011	Philips Gyroscan 1.5 Tesla scanner	10.74 ms	80°	4.26 ms	4 mm
LH	Philips Ashieva 3T TX scanner	10.74 ms	80°	4.26 ms	4 mm
RL	Philips Ashieva 3T TX scanner	10.74 ms	80°	4.26 ms	4 mm
LD	Philips Ashieva 3T TX scanner	10.74 ms	80°	4.26 ms	4 mm
BR	Philips Ashieva 3T TX scanner	10.74 ms	80°	4.26 ms	4 mm
AA	Philips Gyroscan 1.5 Tesla scanner	10.74 ms	80°	4.26 ms	4 mm
MG	Philips Gyroscan 1.5 Tesla scanner	10.74 ms	80°	4.26 ms	4 mm
AK	Philips Gyroscan 1.5 Tesla scanner	10.74 ms	80°	4.26 ms	4 mm
MGO	Philips Gyroscan 1.5 Tesla scanner	10.74 ms	80°	4.26 ms	4 mm

Speakers	FOV	Display matrix	Sagittal images	Spatial resolution	Acquisition time
PB_2002	256 × 256 mm	256 × 256	5	1 mm / pixel	
PB_1998	256 × 256 mm	256 × 256	25	1 mm / pixel	35 - 43 sec
YL	230 × 16.0 × 178.4 mm	240 × 240	130 axial	0.958 mm / pixel	15.7 sec
HL	230 × 16.0 × 178.4 mm	240 × 240	130 axial	0.958 mm / pixel	21.8 - 24.1 sec
PB_2011	256 × 256 mm	256 × 256	2	1 mm / pixel	16.2 sec
LH	256 × 256 mm	256 × 256	1	1 mm / pixel	16.2 sec
RL	256 × 256 mm	256 × 256	1	1 mm / pixel	16.2 sec
LD	256 × 256 mm	256 × 256	1	1 mm / pixel	16.2 sec
BR	256 × 256 mm	256 × 256	1	1 mm / pixel	16.2 sec
AA	256 × 256 mm	256 × 256	1	1 mm / pixel	8.1 sec
MG	256 × 256 mm	256 × 256	1	1 mm / pixel	8.1 sec
AK	256 × 256 mm	256 × 256	1	1 mm / pixel	8.1 sec
MGO	256 × 256 mm	256 × 256	1	1 mm / pixel	8.1 sec

Table 2-3 – MRI recording protocol of all speakers. Male speakers PB, YL, LH, RL, LD, BR (in black) and female speakers HL, AA, MG, AK, MGO (in red)

2.4.2. MRI markers

One of our speakers was recorded using a technique based on MRI markers (Badin et al. ,2012). This method consisted in attaching markers on the tongue and lip contours that were visible in the MRI midsagittal images. The purpose of this technique was to track the evolution of flesh points and to obtain information about the biomechanical properties of the tongue and lip contours. However, since only one of our speakers was recorded using this method, there are no models using MRI markers in this manuscript.

2.5. Articulatory corpus

In this study, MRI data have been collected for eleven French speakers (six males: PB, YL, LH, RL, LD, BR and five females: HL, AA, MG, AK, and MGO). Three data sets, containing the same information, were recorded for speaker PB (PB_1998, PB_2002 and PB_2011). Ideally, one would have liked to record more speakers to make this study more general. However, in practice, speakers' hunting is not always easy and the preparation of a recording session takes certain time. In the same way, one would have liked to record all possible combinations of consonants in vocalic contexts existing in the French language (i.e. the combination of all the consonants in Table 2-1 in all possible vocalic contexts). But, during a recording session, the speaker being recorded must not feel too much fatigue. Therefore, a recording session is limited to about two hours for the most resistant speakers. These facts have forced the reduction of the size of the corpus. For example, it was decided to keep only 13 vowels among the 16 oral and nasal vowels in French. The reason of that reduction is that most French speakers were not able to distinguish the pronunciation between some vowels. For instance: the nasal vowels /ɛ̃/ and /œ̃/, the oral vowels /ə/ and /œ/ and the oral vowels /a/ and /ɑ/. Thus, it was decided to keep only /ɛ̃/, /œ/ and /a/. Engwall et al. (2003) found out that a limited VCV corpus, that covers the whole articulatory space, can capture the same articulatory features as a more complete corpus. Besides, Beautemps et al. (2001) shown that a model based on a reduced corpus could reconstruct the data with an accuracy close to that obtained with the model based on the whole corpus. Thus, also in this study the corpus was reduced to only one representative consonant for each place of articulation. The medio-palatal /ɲ/ and the postdorsal-velar /ŋ/ were totally excluded. The consonant /ɲ/ was excluded because it was replaced by the consonant /n/; indeed, the consonant /ɲ/ can be considered as the consonant /n/ produced in a more palatal manner. Besides, the consonant /ŋ/ was not taken into account because it does not belong to the French language itself. The consonant /ŋ/ is related to the termination *'ing'* borrowed from the English language. The number of articulations recorded for each speaker varied between 63 and 74. Due

to technical problems in the MRI machine, speaker AA was not able to record the vocalic context /o/. Therefore this speaker has 10 less articulations in common with other speakers. The corpus in common for all speakers consisted of 63 articulations: 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 vowel-consonant-vowel (VCV) context of the five vowels /a e ɛ i u/.

2.5.1. Comparison of our corpus with the corpuses in the literature

Table 2-4 shows the comparison between our corpus and those reported in the literature. This comparison is made in terms of number of speakers, size of the corpuses and number of articulator measurements. As one can see, this study includes more speakers and the corpus is larger than those in the literature because it is composed by vowels and consonants. One of the main contributions of this study is the acquisition of data for the complete vocal tract contours, which allows us to model and study the synergy between different organs involved in speech. Another important contribution is the inclusion of consonants in vocalic contexts, which also implies a challenge for the modelling.

Recording method	Study	No. Speakers	Size of corpus	No. Points
EMA	Hoole (1998)	7	15 vowels	4 sensors
	Geng & Mooshammer (2000)	6	15 vowels	4 sensors
	Hu (2006)	7	10 vowels	3 sensors
X ray	Harshman et al. (1977)	5	10 vowels	13 points
MRI	Hoole et al. (2000)	9	7 vowels	13 points
	Zheng & Johnson (2003)	5	9 vowels	13 points
	Ananthakrishnan et al. (2010)	3	13 vowels	150 points
Our Results				
MRI	Valdes (2013)	11	13 vowels, 10 consonants in 5 vocalic contexts	150 points

Table 2-4 - Comparison between our corpus and the corpuses in the literature

2.6. Edition and estimation of midsagittal articulatory contours and landmarks

Once the articulatory data have been acquired, we could then proceed to their processing. The following sections explain how the vocal tract contours were manually traced and how parts that are difficult or impossible to see were estimated.

2.6.1. Edition

In our study, we will consider that an articulation is represented by the complete set of vocal tract contours. These contours have thus been edited by hand for the midsagittal image of each articulation[2]. Rigid structures are considered as non deformable shapes (see Figure 2-4). Thus, they are only drawn once for the whole corpus of a given speaker. These rigid structures are: the skull bones (nasal bone, sphenoid, foramen and occipital), the palate, the jaw and the hyoid bone. In order to be able to draw the palate and the jaw contours, which also include the incisors, some reference MRI recordings were used. At the acquisition stage, the speakers have been instructed to record six different reference postures (see Figure 2-5): incisors in contact, incisors in contact with tongue tip pushed against the posterior part of the incisors, closed jaw, opened jaw, advanced jaw with the upper lip wrapping the upper teeth and retracted jaw with lower lip wrapping the lower teeth. These reference MRI positions were important because they allowed us to identify the shape of the incisors to draw the palate and the jaw. When editing the position of the palate for a given articulation, the skull bones move together with the palate. They are used as a reference guide to position the palate contour, and thus to take into account the movements of the speakers, at least those along the sagittal direction. Besides, the anatomical landmarks of the upper lip (N2), the lower lip (LL1), the tongue tip (TT) and the attachment between the jaw and the mouth floor (jawAttach) were also drawn. These anatomical landmarks were useful to identify the corresponding regions. For example, N2 and LL1 were used to mark the beginning of the upper and lower lip, respectively. In the same way, TT and jawAttach were used by the expert to indicate the position of the tongue tip and the attachment between the jaw and the mouth floor, respectively.

[2] The data of the different speakers have been edited by several experts. Speakers YL, LH, RL, LD, BR, HL, AA and MG were entirely traced by the author of this thesis (Julián Andrés Valdés Vargas). The speakers PB-2011, AK and MGO were traced by Julián Valdés in collaboration with an intern student at GIPSA-lab (Arielle Koncki). Speakers PB-1998 and PB-2002 were traced by Pierre Badin and Gopal Ananthakrishnan (PhD student at KTH (Sweden)). However, the final tracings of all speakers were verified by Julián Valdés to follow the edition policies described in this section.

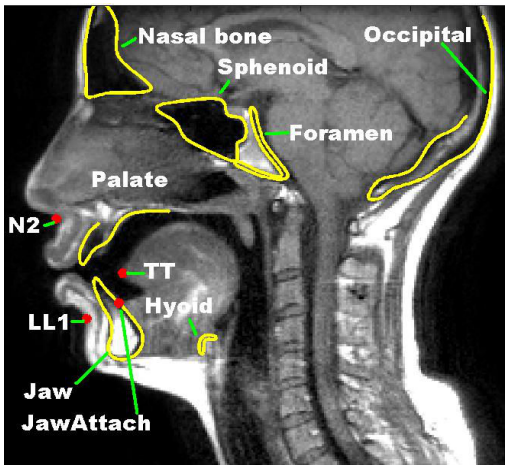


Figure 2-4 - Rigid structures and anatomical landmarks of speaker AA pronouncing the articulation /u/

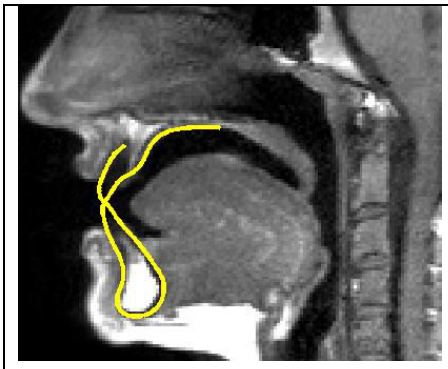
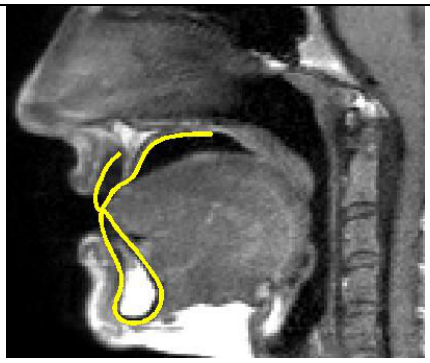
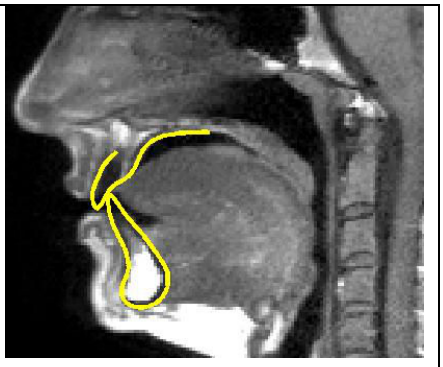
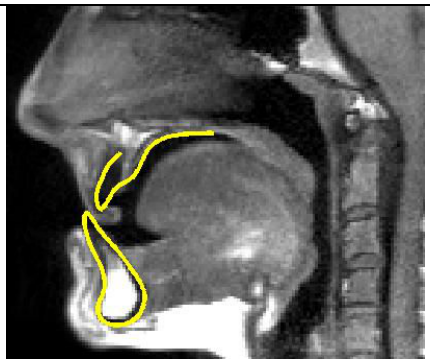
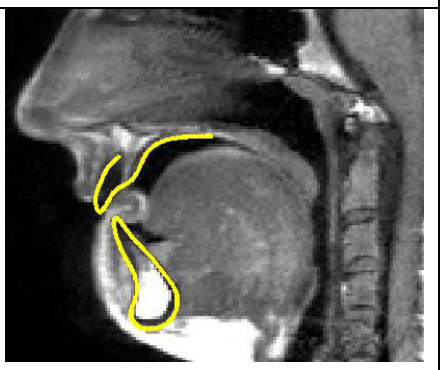
		
incisors in contact	**incisors in contact with tongue tip pushed against the posterior part of the incisors**	**Jaw clenched**
		
opened jaw	**advanced jaw with the upper lip wrapping the upper teeth**	**retracted jaw with lower lip wrapping the lower teeth**

Figure 2-5 - MRI images for reference postures of the speaker AA

After the rigid structures and the anatomical landmarks were drawn for a given speaker, we could then proceed to the edition of all the midsagittal images corresponding to each articulation in the corpus. In other words, the rigid structures and the anatomical landmarks were positioned at the right place and the deformable contours like the lips, the tongue, the epiglottis, the pharynx and the velum were outlined by hand. The Figure 2-6 illustrates the process of hand tracing of each contour. The upper lip was outlined from the forehead, including the nose, up to the attachment with the upper incisors. On the other hand, the lower lip was traced from the neck, including the chin, up to the low part of the jaw. Nevertheless, the entire coordinates of the upper and lower lip were not used. The horizontal line which crosses the anatomical landmark N2 was used to cut the final upper lip. Similarly, the final lower lip contour was cut using the anatomical landmark LL1. The palate was positioned, using the skull bones as guide references, from the attachment with the upper lip up to the velum contour. The velum was outlined from the last point of the palate up to a point in the nasal cavity which corresponds to about the same x coordinate that the starting point. Besides, the Pharynx was traced from the last point of the velum up to the beginning of the larynx. By means of rotations and translations, the rigid contours of the jaw and hyoid bone were positioned at their corresponding places. The hyoid bone was positioned at the low part of the tongue and close to the epiglottis. The tongue was outlined from a point below and close to the hyoid bone up to the epiglottis. However, the concave cavity of the tongue between jawAttach and TT, called sublingual cavity, was not always visible. Thus, this part of the tongue was not always traced as a concave shape but outlining the contour as visible. The epiglottis was traced from the last point of the tongue up to the glottis. The contours of the low part of the vocal tract (glottis, back-larynx and trachea) were usually traced based on the position of the last three segments of the spinal cord. However, these organs were blurry in most of the MRI and thus difficult to trace. Besides, the spinal cord was outlined from the foramen bone up to the lowest part of the vocal tract. The anatomical landmarks of the tongue tip (TT) and jaw attachment (jawAttach) were positioned at last to indicate the corresponding regions. Since TT and jawAttach were not always obviously identified, they had to be estimated for some articulations of the corpus. The next section explains how it was done. The coordinates of the traced articulations were stored in centimetres (cm) and translated to fix the tip of the upper incisors to the (x, y) coordinate (5, 10). The decision of the anchor point (5, 10) was arbitrary but useful to align the different articulations between them. The Figure 2-7 shows an example of one articulation completely hand-traced.

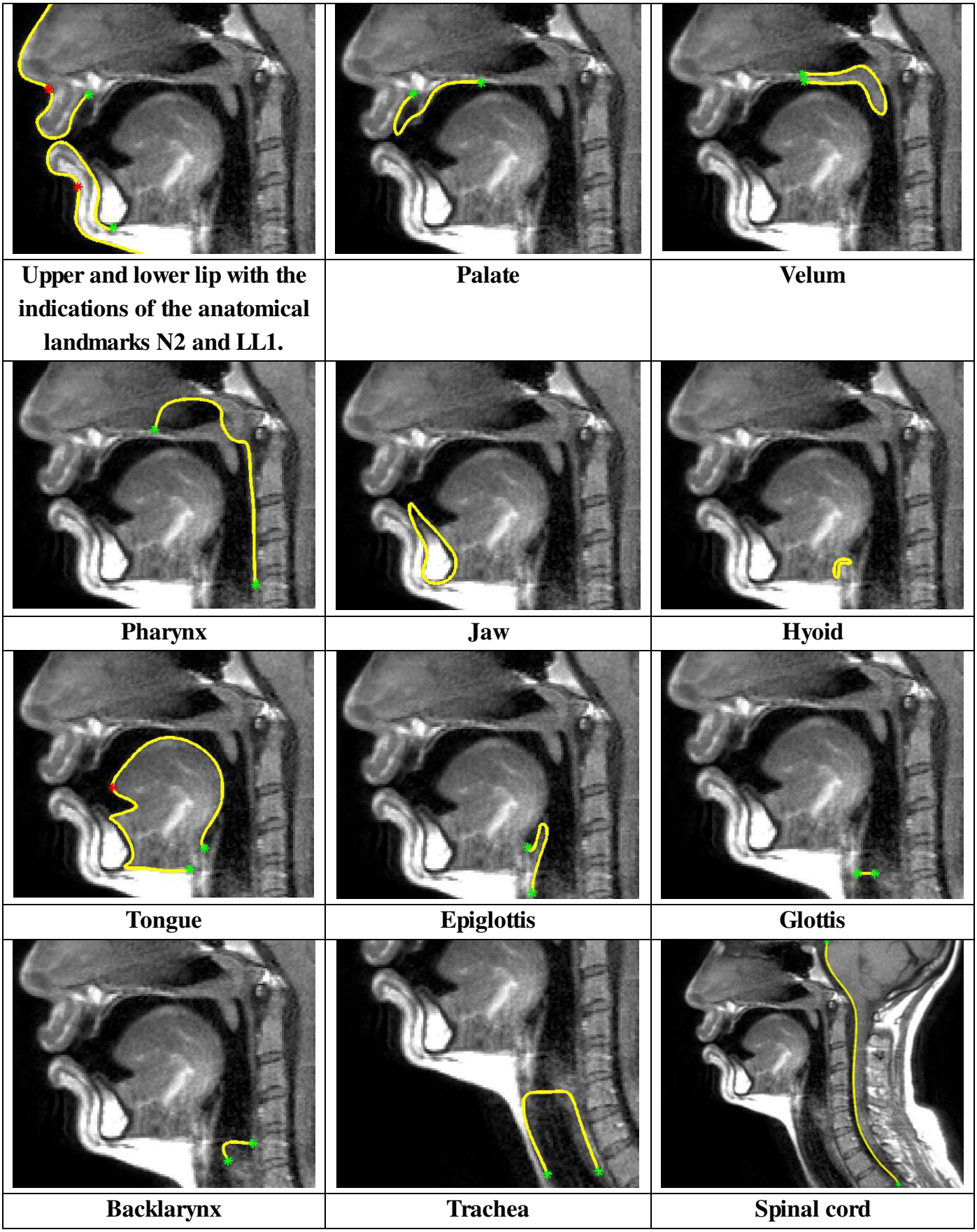


Figure 2-6 – Illustration of hand tracing of each contour. The green points indicate the starting and ending point of each contour. The red points represent the anatomical landmarks

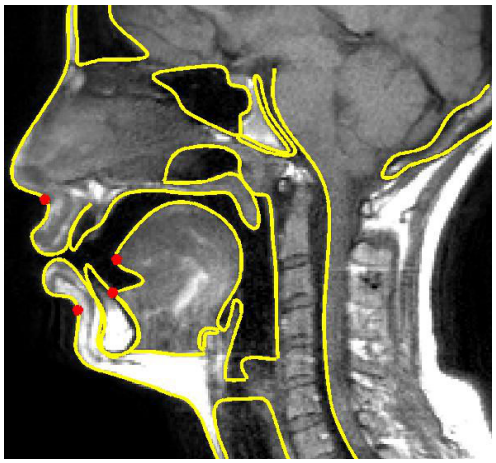
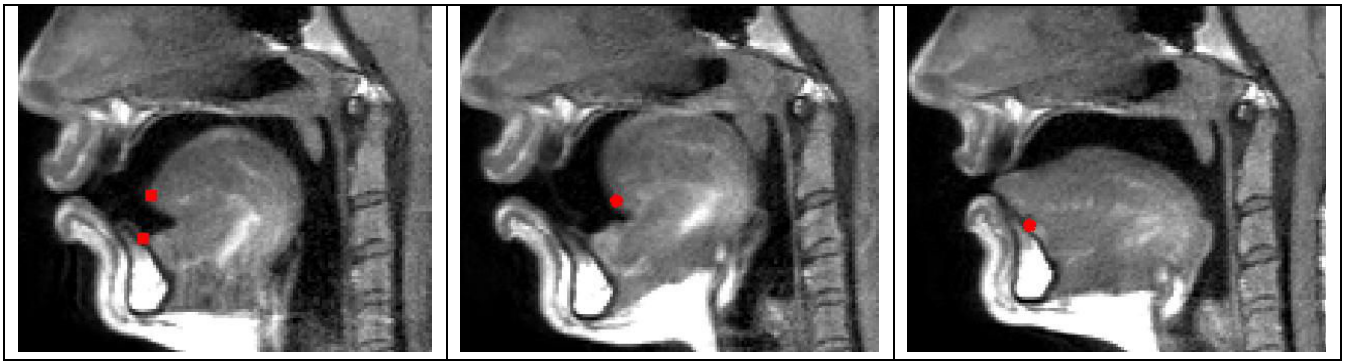


Figure 2-7 - Complete manually edited contours of articulation /u/ of the speaker AA

2.6.2. Estimation of non visible landmarks

The estimation of the TT and jawAttach points was made by first identifying the set of articulations for which these regions were visible. TT and jawAttach were first located on the visible articulations and then estimated for the not visible articulations. The estimation was computed as the average position in the visible articulations. When tracing a given articulation, the estimated points were a guide to decide the final TT and jawAttach positions. However, the final decision was also taken by consideration of the expert eye. The Figure 2-8 shows an example of one articulation in which TT and jawAttach were visible and two articulations in which TT and jawAttach were estimated. The position of TT was usually not clear in rounded articulations and the jawAttach was mostly unknown in articulations with low tongue.



Visible TT and jawAttach (Articulation /u/)	**Estimation of not clear TT (articulation /ku/)**	**Estimation of unknown jawAttach (Articulation /se/)**

Figure 2-8 - Articulations with visible and not visible TT and jawAttach landmarks of the speaker AA and their estimations

2.7. The grid system

In Chapter 4, articulatory models of the tongue contour are presented. We want then to be able to use different representations of data that would possibly constitute an improvement of the tongue modelling performance. Thus, in order to represent the points of the tongue contour in different manners and compute different articulatory measurements, the grid system proposed by Beautemps et al. (2001) was used (see Figure 2-9).

The grid system was set up once for each speaker. First, the grid center was defined as the mean of the center of gravity of all articulations. Second, one of the central lines of the grid was oriented to be parallel to the average position of the pharyngeal wall, while the other central line was oriented to be parallel to the palate. The

Figure 2-10 shows an example of the center and orientation of the grid for speaker AA. Moreover, the 28th grid line corresponds to the tongue tip and the 6th grid line is attached to the beginning of the epiglottis.

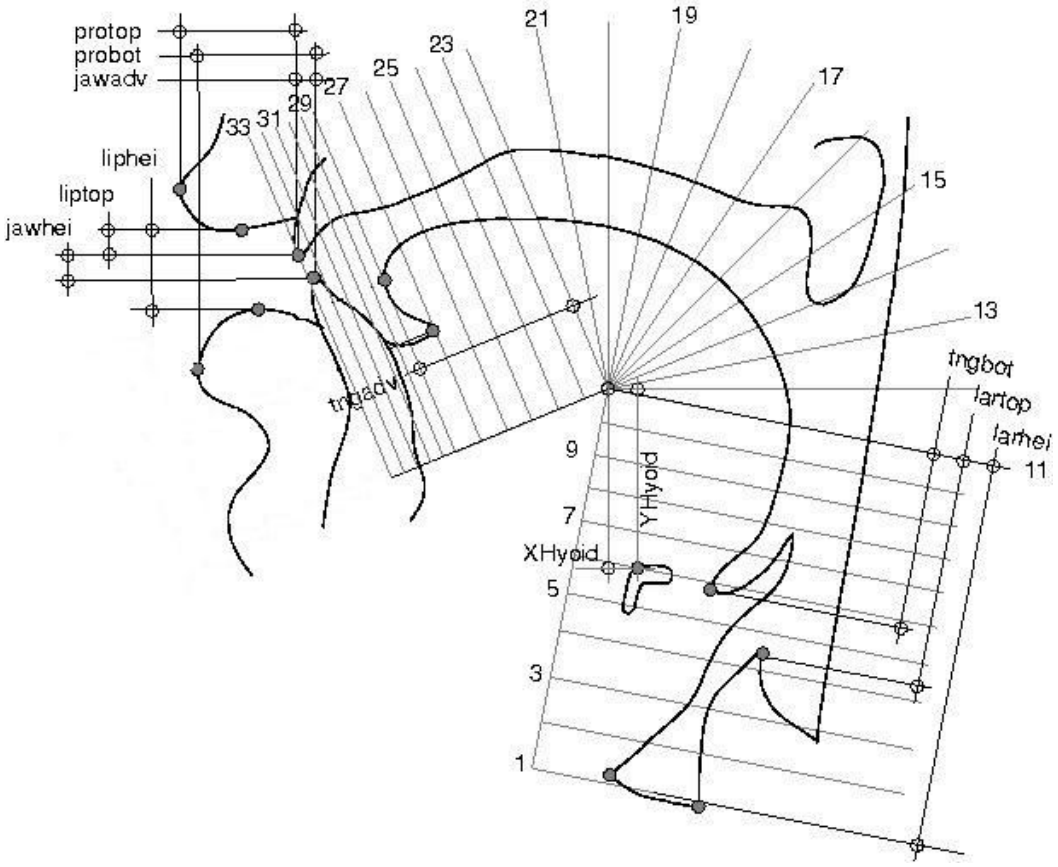


Figure 2-9- illustration of the grid system to represent the tongue contour (from Beautemps et al., 2001)

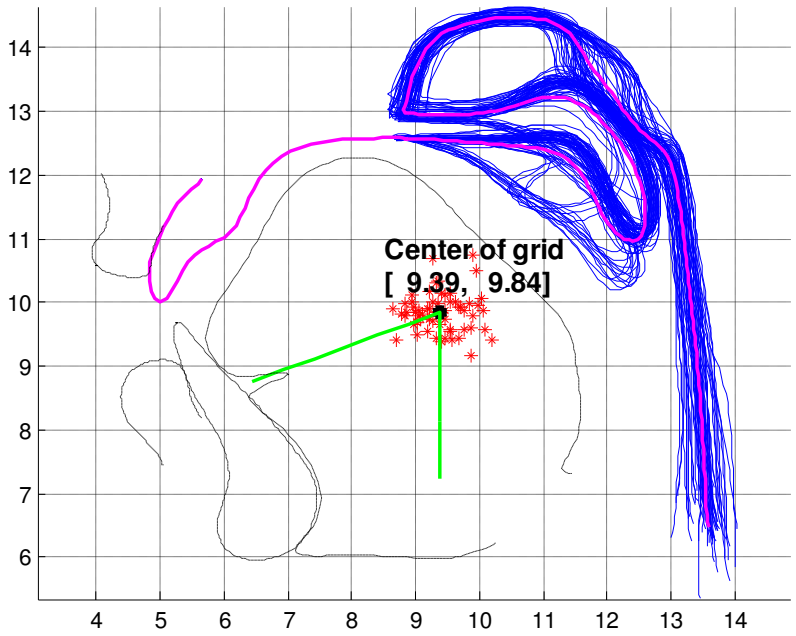


Figure 2-10 – Center and orientation of the central lines of the grid for speaker AA. Center of gravity for each one of the articulations of speaker AA (red stars), center of grid (black point), central grid lines (green), superposition of pharyngeal wall of each articulation (blue), mean of the pharyngeal wall (magenta)

2.7.1. Articulatory measurements

In this study we used several vocal tract parameters that were directly measured from the traced articulations using the grid system, as illustrated in Figure 2-9. The protrusion of the upper lip (proTop) was measured as the distance between the upper incisor and the most advanced point of the upper lip. Similarly, the protrusion of the bottom lip (proBot) was measured as the distance between the lower incisor and the most advanced point of the lower lip. The jaw advance displacement (JawAdv) was measured as the horizontal distance between the upper and lower incisor edges. Besides, the distance between the upper and lower lip (lipHei) was measured as the distance between the lower point of the upper lip and the highest point of the lower lip. The movement of the upper lip (lipTop) was measured as the distance between the upper incisor edge and the lowest point of the upper lip. The vertical displacement of the jaw (jawHei) was measured as the distance between the upper incisor and the lower incisor edges. The advancement of the tongue (tngadv) was measured as the distance between the 22nd grid line to the 28th grid line which is attached to the tongue tip. The tongue bottom (tngbot) was calculated as the distance between the 6th grid line, which is attached to the beginning of the epiglottis, and the 11th grid line. Finally, larTop was measured as the distance between the grid center and the upper part of the larynx. Similarly, larHei was measured as the distance between the grid center and the lower part of the larynx.

2.7.2. Tongue contour representation and sampling

The tongue contour was represented in four manners. The first one is a representation of 200 equidistant (x, y) points of the full tongue contour (FullTng). FullTng is the contour from the jawAttach to the base of the epiglottis. The first 50 equidistant (x, y) points represent the sublingual cavity located between the jawAttach up to the tongue tip. In some cases, the data related to the sublingual cavity may be missing, as explained in section 2.6.1.The other 150 points are related to the contour from the tongue tip up to the base of the epiglottis. Second, the upper tongue contour (UpperTng) was represented by 150 equidistant (x, y) points. UpperTng is the contour from the tongue tip up to the base of the epiglottis. The other two representations are based on a resampling of UpperTng using the grid system. The INTRXY representation refers to the 23 (x, y) intersection points between the grid lines and the tongue contour, from the 6^{th} grid line to the 28^{th} grid line. The coordinates INT are related to the 23 distances between the central lines of the grid to the tongue contour, from the 6^{th} grid line to the 28^{th} grid line. This last representation of the tongue includes the tngbot and tngadv parameters in order to be able to model the up-down and front-back movements of the tongue.

2.8. Conclusion

In this chapter, we have seen how the French language is structured in terms of articulatory phonetics. The vowels can be classified as regards four properties: height, backness, roundness and nasalisation. On the other hand, consonants are characterised by their manner of articulation and their place of articulation.

In order to collect data, MRI has been chosen as the recording method for this study. We have compared MRI with other recording methods like X-ray and EMA. The final decision was MRI because it offers more complete information of the vocal tract compared to EMA and more legible data compared to X-Ray. This chapter described how the data were recorded in MRI and which protocols were used.

Since the linear decomposition methods presented in the following chapters are intended to extract common articulatory patterns that can be reliable for analysis of articulatory features in the French language, the corpus has been selected to cover as much as possible the whole articulatory space. Besides, male and female speakers with different vocal tract sizes have been chosen to extract articulatory patterns that are as general as possible. Finally, this chapter explained how the various vocal tract contours were edited and estimated in case of necessity.

The final data set contained the information of 11 French speakers (6 males and 5 females), and 63 articulations including vowels and consonants in vocalic context. The data of (x, y) coordinates for 12 vocal tract contours (lips, palate, velum, pharynx,

jaw, hyoid bone, tongue, epiglottis, glottis, backlarynx, trachea, and spinal cord) were included. This data set was used to build and compare vocal tract models of ours speakers, as presented in following chapters. Moreover, some examples of MRI images can be found in the annex A at the end of this manuscript. A set of 12 articulatory measurements were also taken from the vocal tract of all speakers. These measurements were used to compute statistics and make comparisons.

The next chapter will concentrate on characterizing our speakers in terms of articulatory strategies. The articulatory strategies of the tongue, lips and velum of each speaker are compared. The speakers are also statistically compared in terms of various articulatory measurements.

Chapter 3. Articulatory characterisation and individual models of speakers

3.1. Introduction

Before modelling any vocal tract contour, one has to be aware that articulatory variability in speech can be ascribed to two sources: differences in anatomical conformation and differences between articulatory strategies employed by several speakers. This chapter aims to characterize our speakers as regards these two aspects. In other words, this chapter presents a qualitative description and comparison of each speaker's data. The reader will be guided through several analyses that compare speakers in terms of morphology, articulatory strategy of different vocal tract contours and vocal tract measurements. Finally, the results of three individual models made for the same speaker, but with data recorded at three different moments, are presented to expose the problem of intra-speaker variability.

3.2. Individual linear decomposition models and evaluation

The individual models explained in this chapter are built by means of the PCA method described in Chapter 1. They are validated in terms of the relative variance explained. The variance explained is given by the ratio of variance of predicted data (X_p) over the variance of original measured data (X), as shows in the following equations:

$$\mathrm{VARIANCE}(X) = \frac{\sum_1^n \sum_1^m (X_i - \overline{X_i})^2}{n.m}$$

$$\mathrm{VARIANCE\_EXPLAINED}(X, X_p) = \frac{\mathrm{VARIANCE}(X_p)}{\mathrm{VARIANCE}(X)}$$

Being n the number of observations and m the number of articulator measurements.

3.3. Speakers' tongue control strategies

This section presents comparisons between the different tongue control strategies used by our speakers. First, it is explained how PCA is guided to impose specific control parameters. Then, our speakers are compared as regards extracted linear components. The analyses presented on this section are limited to the upper tongue contour; the contour from the tongue tip to the base of the epiglottis.

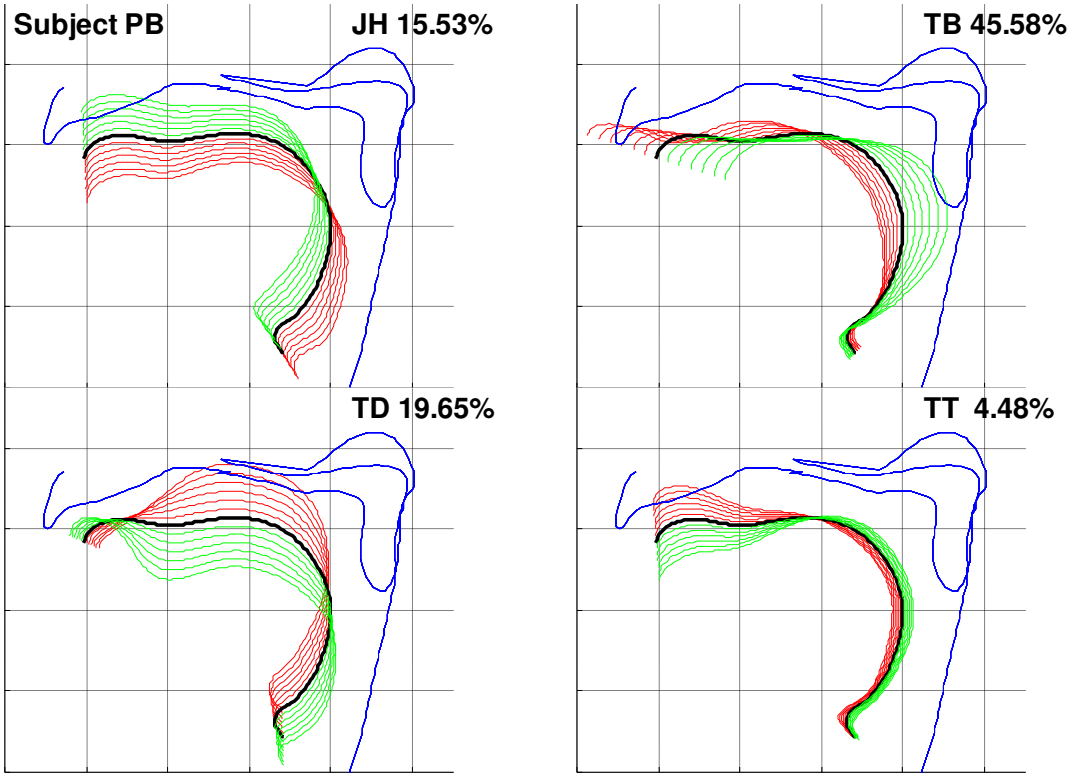
3.3.1. Guided PCA of the upper tongue contour

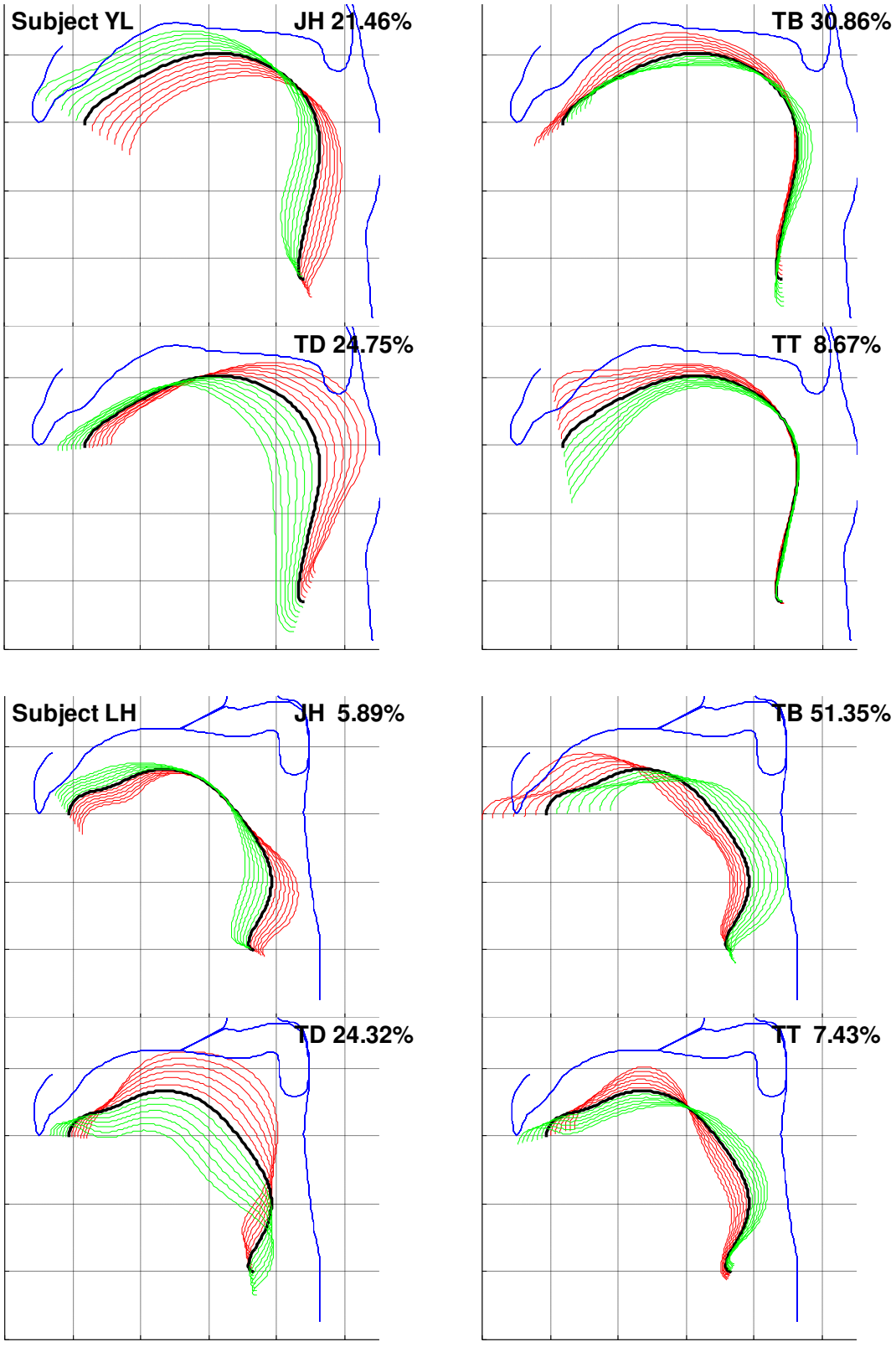
Using a procedure based on a guided PCA analysis of tongue contours, Badin & Serrurier (2006) have shown that the first four components account for the largest amount of tongue movement variance. In this section we describe the results of a Guided PCA analysis for our eleven speakers. Two alternatives were explored to decide how to measure the jaw height parameter (JH). The first option was to use the three degrees of freedom of the jaw (x, y translation and rotation) as proposed by Edwards & Harris (1990). However, the correlation computed between the y-coordinate of the lower incisor and the angle of rotation of the jaw has shown a strong relation up to 0.92. Thus, the rotation of the jaw was not taken into account. The JH parameter was defined as the normalized value of the y-coordinate for the lower incisor (Badin & Serrurier, 2006); it was used as the first control parameter of the tongue model (the associated model coefficients were obtained by the Linear Regression (henceforth LR) 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 LR, 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 LR).

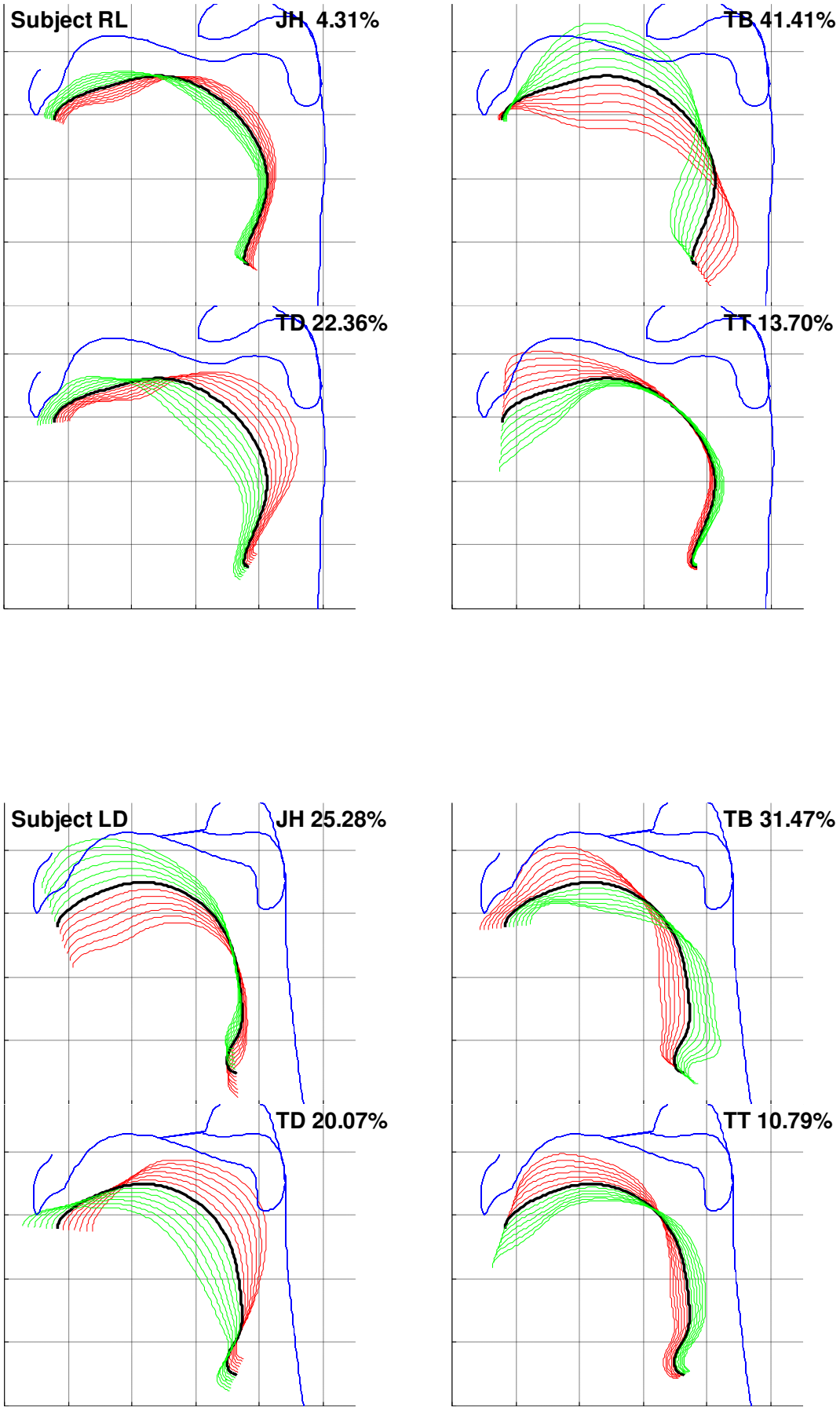
3.3.2. Comparison of Guided PCA components between speakers

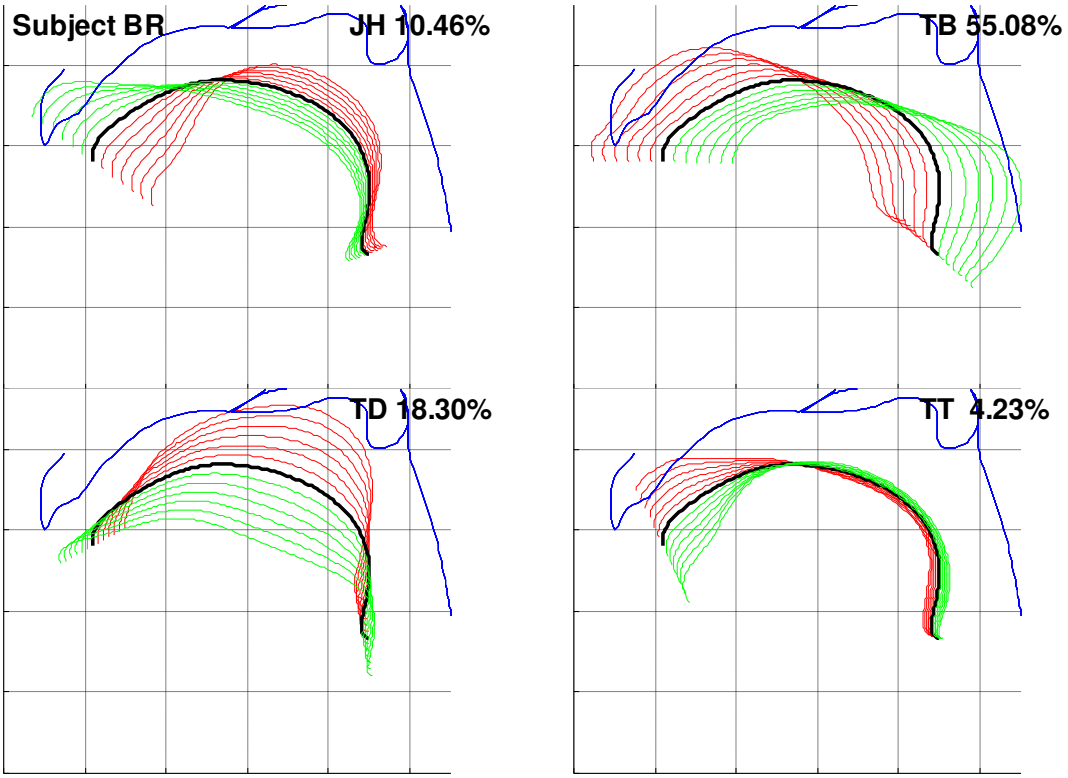
In order to understand the articulatory characteristics of each speaker, we compared their four guided PCA components determined as explained above. A graphical representation in which each predictor (JH, TB, TD and TT) is varied within a range constitutes a nomogram. Figure 3-1 illustrates the associated nomograms for all speakers. The main effect of JH is a rotation of the tongue around a point located in its back. JH of speakers MGO, MG, AA, AK and LD is associated with a movement of the front of the tongue without movement in the back. Oppositely, speakers HL, PB, LH, RL, BR and YL move the back of the tongue when JH moves. TB controls front-back displacements while TD is related to flattening-arching movements. It appears that TB of speakers LH, BR, LD, HL and AK is related to a horizontal movement of the tongue body while it is a more diagonal movement for speakers PB, YL, MG, RL, AA and MGO. Besides, TT controls precisely the tongue tip motion. We have observed that speakers BR, RL, AA, AK, MG, MGO and PB are able to move their tongue tips more independently from the tongue back than speakers HL, LD, LH and YL do.

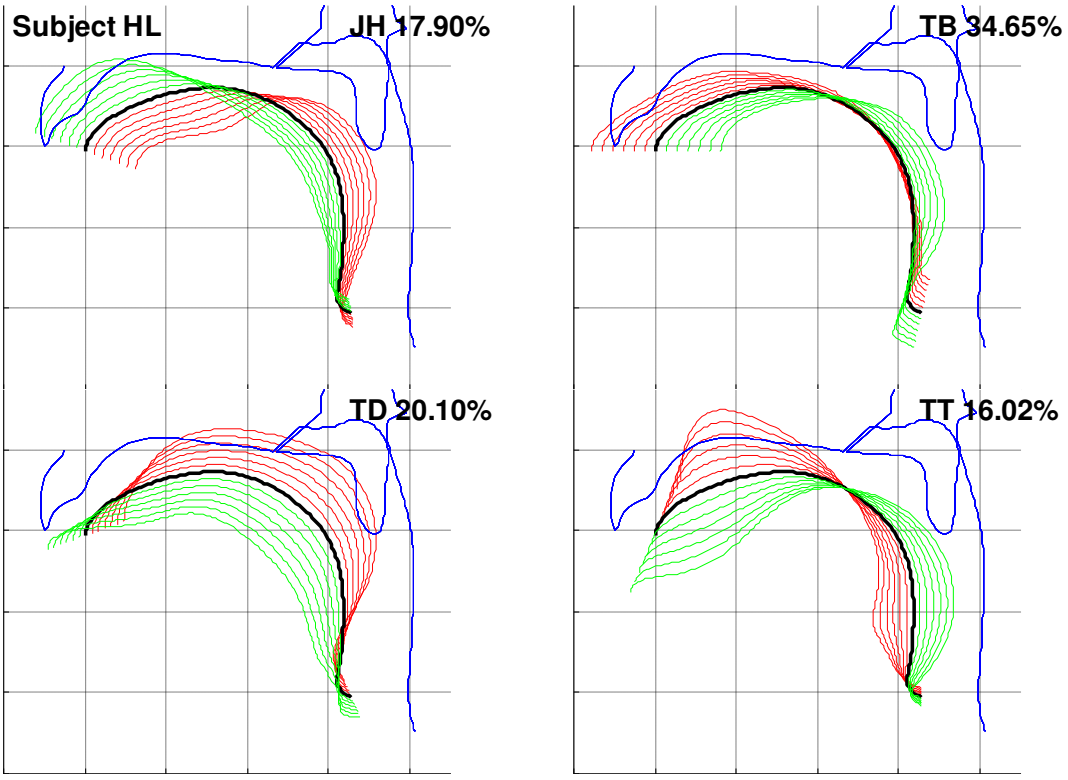
Figure 3-2 shows the percentage of variance explained by each component for all speakers. We see that, among our speakers, JH explains the maximum variance for speaker LD and the minimum for speaker RL. TB explains the maximum variance for speakers BR and the minimum for speakers YL and LD. TD explains the maximum variance for speaker MG and the minimum for speaker BR. Finally, TT explains the maximum variance for speaker AK and the minimum for speaker BR.

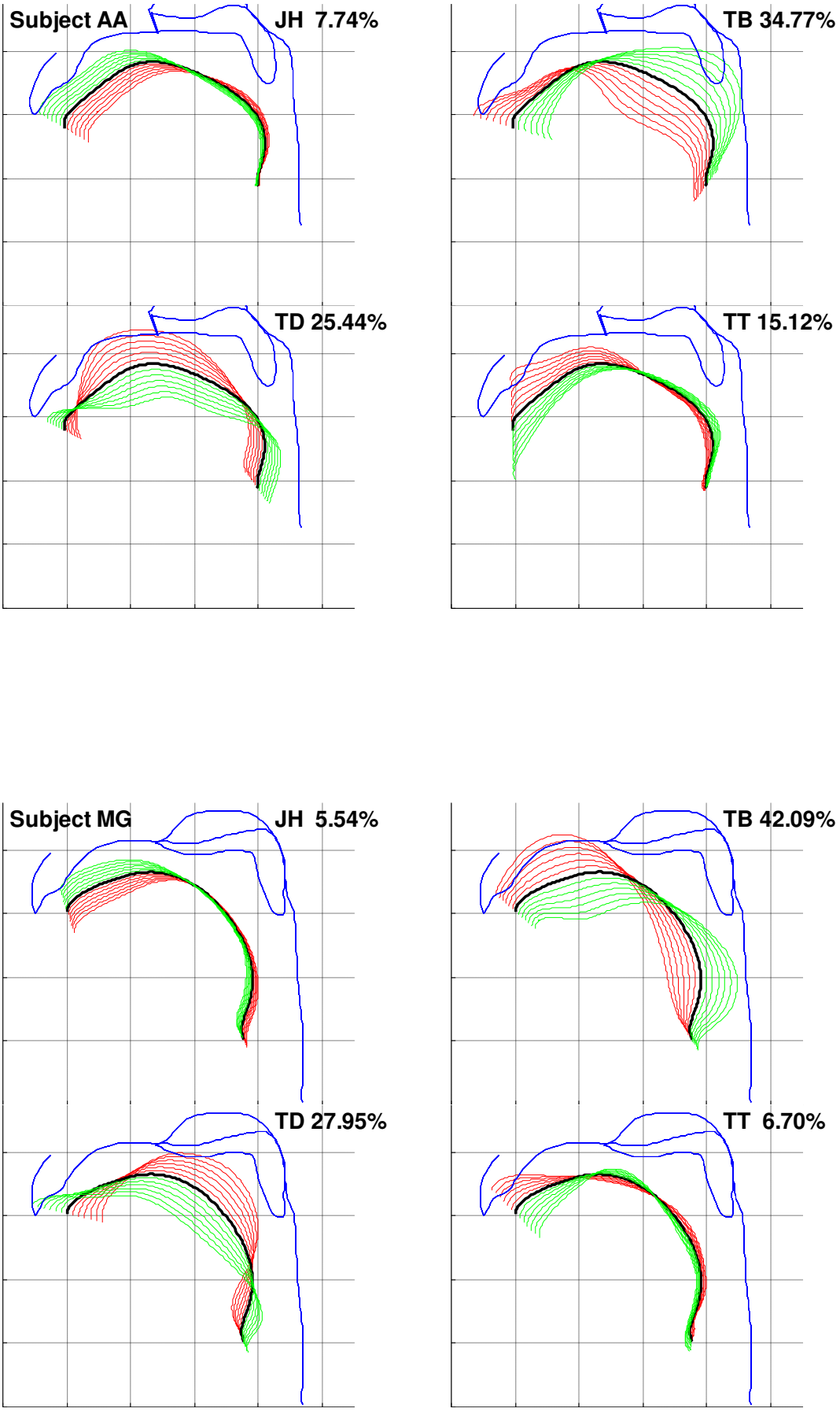



Subject YL
JH 21.46%
TB 30.86%
TD 24.75%
TT 8.67%
Subject LH
JH 5.89%
TB 51.35%
TD 24.32%
TT 7.43%


Subject RL
JH 4.31%
TB 41.41%
TD 22.36%
TT 13.70%
Subject LD
JH 25.28%
TB 31.47%
TD 20.07%
TT 10.79%


Subject BR
JH 10.46%
TB 55.08%
TD 18.30%
TT 4.23%


Subject HL
JH 17.90%
TB 34.65%
TD 20.10%
TT 16.02%


Subject AA
JH 7.74%
TB 34.77%
TD 25.44%
TT 15.12%
Subject MG
JH 5.54%
TB 42.09%
TD 27.95%
TT 6.70%

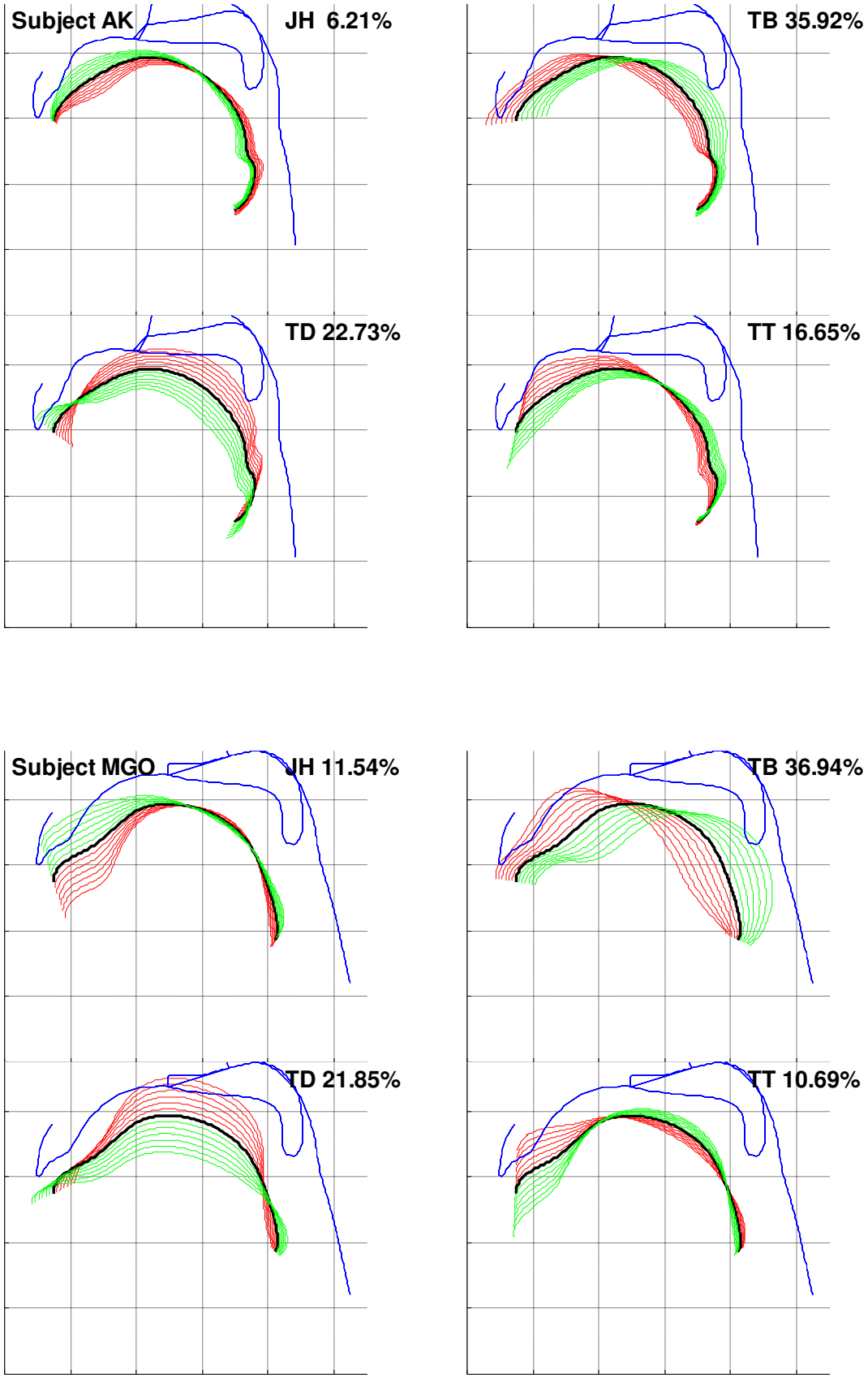


Figure 3-1 - Nomograms of the four upper tongue contour components determined by Guided PCA for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. Each predictor (JH, TB, TD and TT) is varied from -3 to +3 with a 0.5 step. The reference wall (palate, velum and pharynx) is shown in blue. The relative data variance explained by each component is displayed

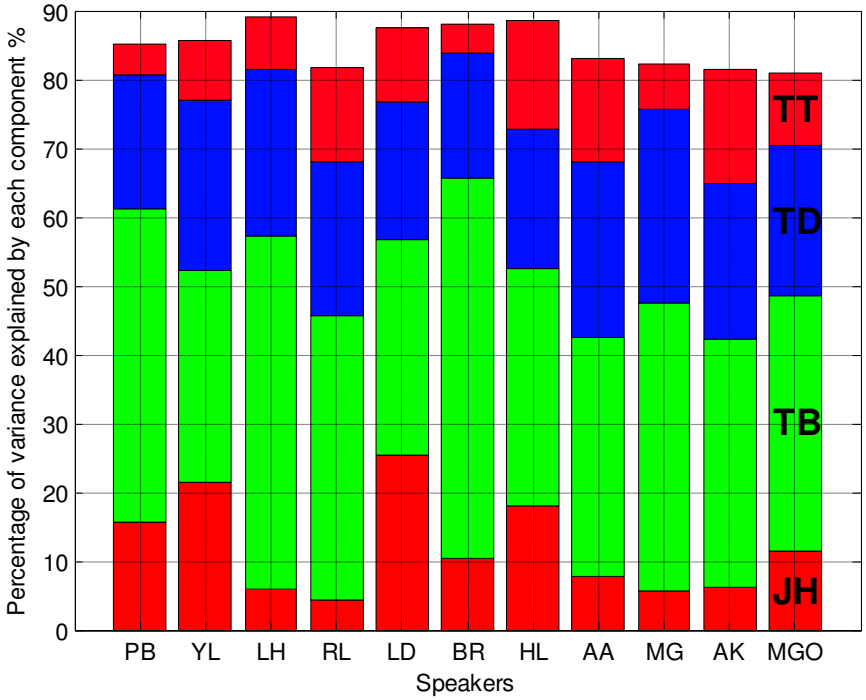


Figure 3-2 – Percentage of variance explained by each guided PCA component of the tongue models, for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. JH contribution (red section at the bottom), TB contribution (green section), TD contribution (blue section), TT (red section at the top)

3.4. Synergy between jaw and tongue

As the jaw is one of the major tongue carriers (Badin & Serrurier, 2006), it is important to describe the synergy between jaw and tongue. Therefore, this section presents and compare the characteristics of the jaw movement and its relation with the tongue movement, between different speakers.

First, we will describe the independent movement of the jaw. Second, we will explain the relation that the jaw movement has with the tongue.

The Figure 3-3 shows statistics of the amplitude of vertical and horizontal jaw movements. On each box, the central mark is the median; the borders of the box are the 25th and 75th percentiles. The lines extend to the most extreme data points that are not considered outliers, and outliers are represented individually with a red cross. As explained in Chapter 2, the vertical displacement of the jaw (JawHei) was measured as the vertical distance between the upper incisor and the lower incisor. On the other hand, the jaw advance (JawAdv) was measured as the horizontal distance between the upper and lower incisor. In Figure 3-3, speakers are ordered from the one with minimum movement amplitude to the one with maximum amplitude. We see that speaker MG makes the smallest use of vertical jaw movement among our speakers. Oppositely, speaker LD makes the maximal use of vertical jaw movement. These observations may be somehow validated by observing JH on the nomograms of

Figure 3-1. In the case of JawAdv, speaker LH has the minimum horizontal displacement and speaker YL has the maximum.

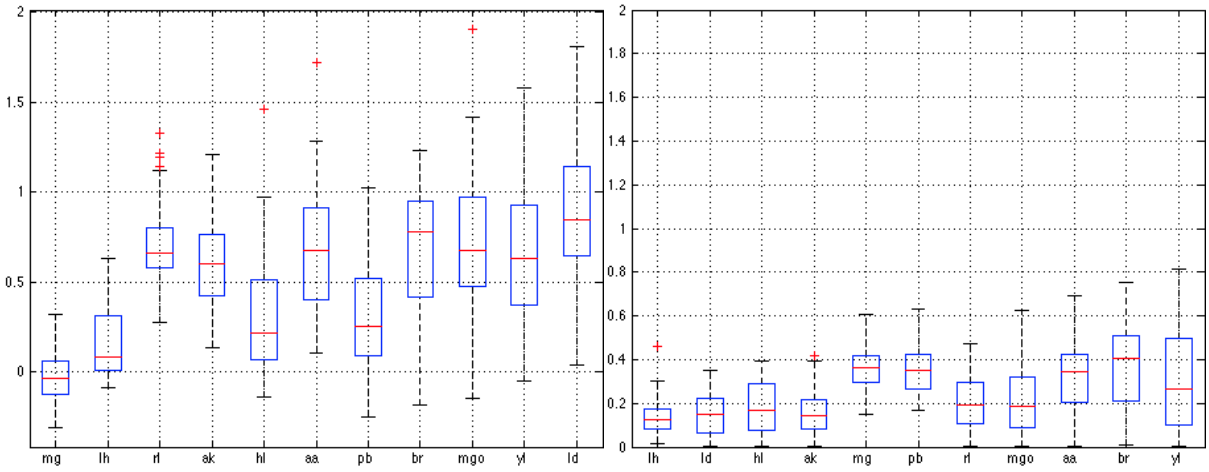


Figure 3-3 - Statistics of jaw movement amplitude. Left: vertical displacement of the jaw (cm). Right: jaw advance (cm)

It is known that for a given displacement of the jaw, the tongue may globally move in a proportion that depends on the speaker (Bailly et al., 1998). The Figure 3-4 shows the slopes of the LR between the jaw height (JH) parameter, explained in section 3.3.1, and the coordinate-y of 150 points corresponding to the upper tongue contour, for each speaker. Note that this study is only based on the y coordinate of the tongue which is related to vertical movements. Thus, following results should not be directly compared with the analysis of section 3.3 which takes into account the x and y coordinate of the tongue contour. Overall, one can observe that the jaw motion has a special influence on the tongue tip and the pre-dorsal region of the tongue (point indices between 0 and 50). Over the tongue tip (point indices between 0 and 25), the slope reaches: speaker PB = ~1.2, speaker YL = ~0.8, speaker LH = ~1.07, speaker RL = ~0.1, speaker LD = ~1.0, speaker BR = ~0.9, speaker HL = ~0.5, speaker AA = ~0.4, speaker MG = ~1.5, speaker AK = ~0.1 and speaker MGO = ~0.9. Thus, the jaw movement has almost no influence on the vertical tongue tip motion for speaker RL and AK. For speakers RL and AK, the tongue tip moves about 0.1 times the jaw movement. On the other hand, speaker MG moves vertically her tongue tip 1.5 times the movement of her jaw. Apparently, speaker MG uses some kind of compensation strategy: she has the most reduced JawHei range (as seen on Figure 3-3), but compensates this small range by moving her tongue tip about 1.5 times her jaw movement. Another important remark is that JH for speakers AK, HL and LD seems to have more influence on the pre-dorsal region of the tongue (point indices between 25 and 50) than the tongue tip.

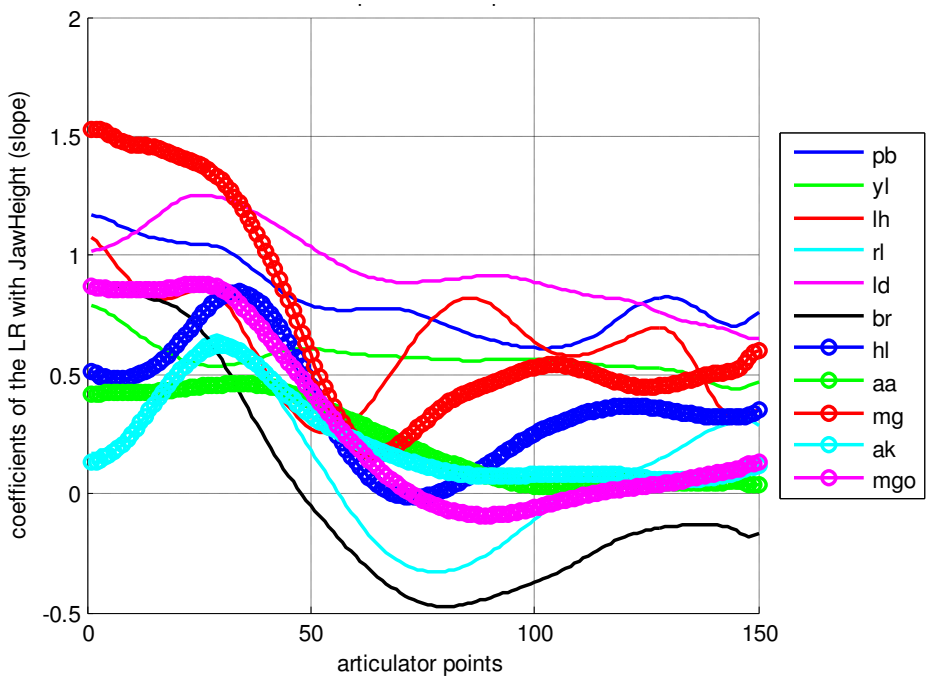


Figure 3-4 - Slopes of the linear regression between the jaw height parameter (JH) and the Y coordinate of 150 points corresponding to the upper tongue contour for male speakers PB, YL, LH, RL, LD, BR (with lines) and female speakers HL, AA, MG, AK, MGO (with circles). The first articulator measurement represents the tongue tip and the 150th point represents the back of the tongue

3.5. Speakers' lip control strategies

Apart from the vocal tract organs studied above, the lips also constitute an important part of speech articulation. This section compares different strategies employed by our speakers to control their lips.

3.5.1. Guided PCA of the upper and lower lip contour

Using a procedure based on a guided PCA analysis, Badin et al. (2002), Bailly et al. (2008) and Beautemps et al. (2001) used five components to linearly decompose the contours of the upper and lower lip. However, in these studies the last two linear components accounted for a little variance explained. Badin et al. (2012) used three components to represent the lips variance. In this study we use the same approach of Badin et al. (2012) based on a linear decomposition of three components for the upper and lip contours.

As for the upper tongue contour, JH was imposed as the first parameter to control simultaneously the upper and lower lip contours (the associated model coefficients were obtained by the LR of all the lip contour points coordinates against JH). The next two parameters for each lip, lip protrusions (ULP and LLP) and lip heights (ULH and LLH) were extracted by PCA from the coordinates of the lip contours, from which the JH contribution was first removed (the associated model coefficients were obtained by

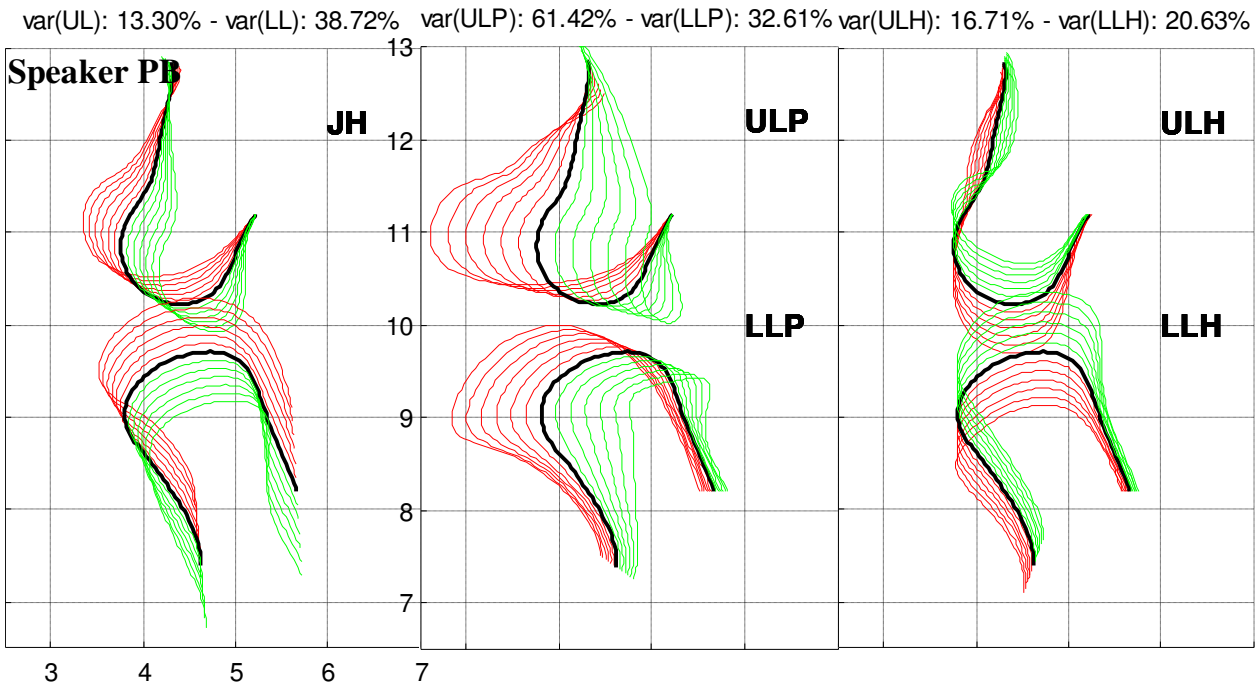
LR, as for JH). ULP and LLP were defined as the normalized x-coordinates of the measured most advanced point of the upper and lower lip, respectively. ULH and LLH were defined as the normalized y-coordinates of the measured lowest and highest point for the upper and lower lip, respectively.

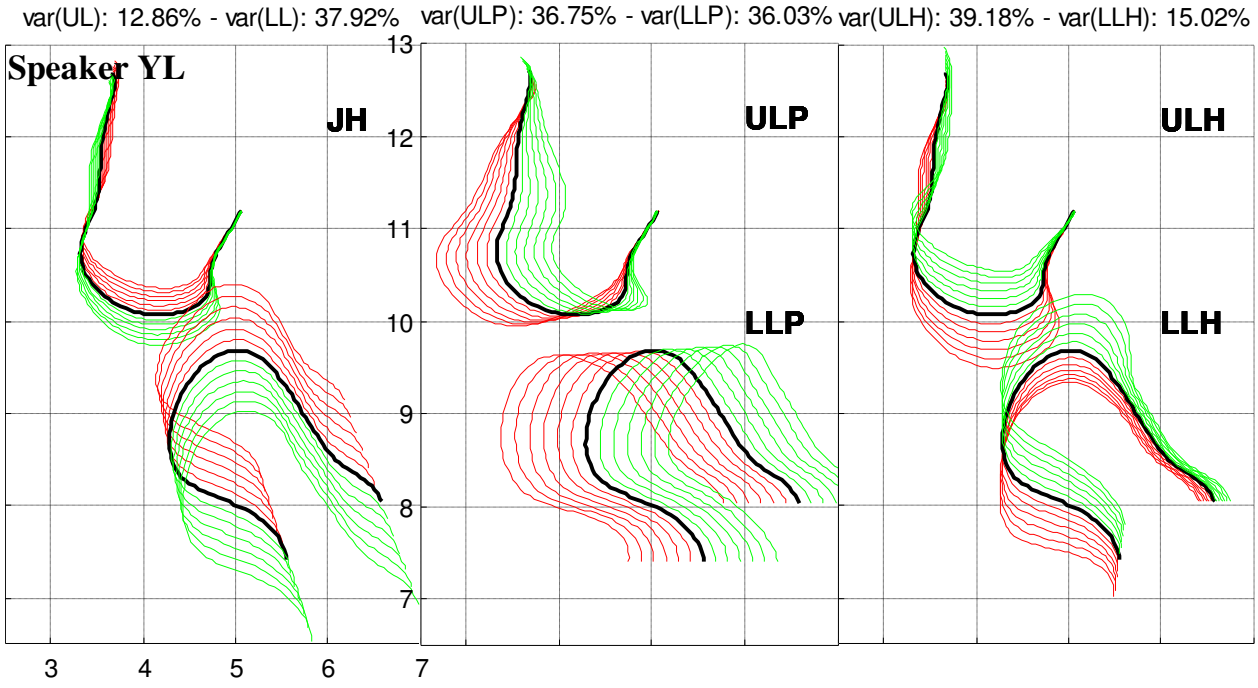
3.5.2. Comparison of Guided PCA components between speakers

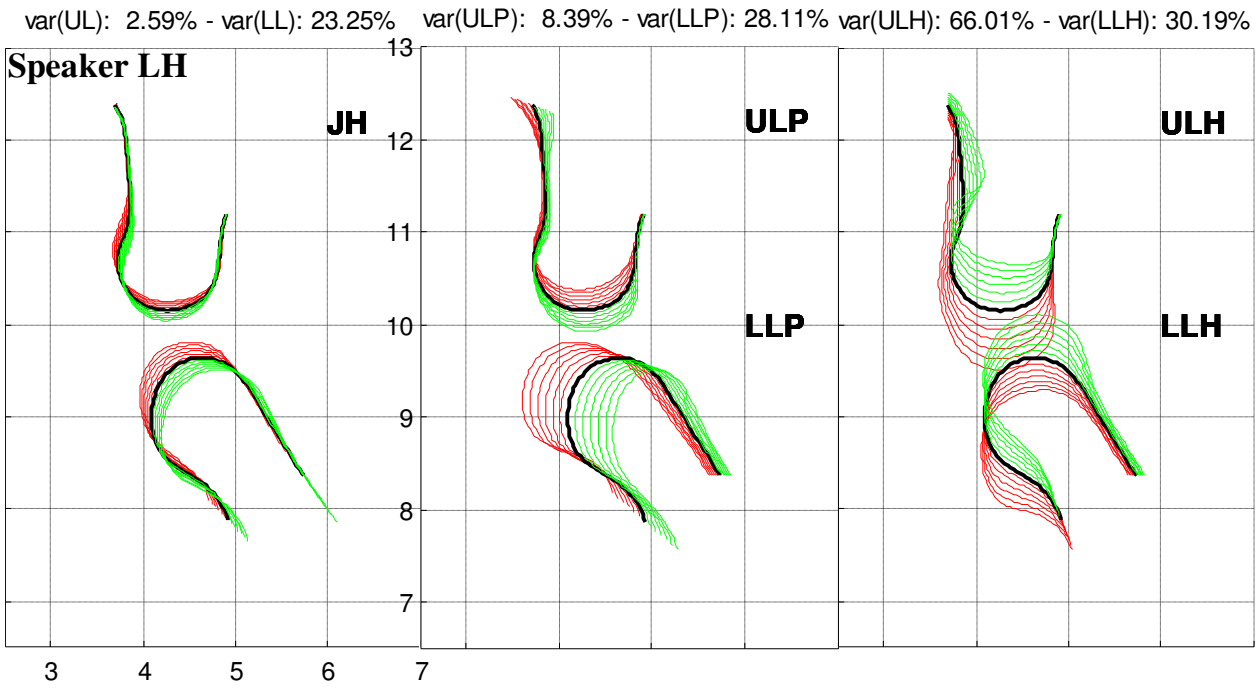
The Figure 3-5 illustrates the nomograms associated with the lips motion for all speakers. In general, JH has more influence on the lower lip than on the upper lip. The variance of the lower lip movement explained by JH ranges between 23.25% and 51.28% over our speakers, while that of the upper lip ranges between 1.68% and 25.19%. However, for some speakers like LH, RL, HL, AA and MG the jaw movement has very little influence on the upper lip action compared to other speakers. For speakers PB, YL, LD, BR, AA, MG and MGO the ULP parameter has more influence on the upper lip than LLP has on the lower lip. The opposite occurs for speakers LH, RL, HL, and AK. Moreover, speakers LH and AK have a very little upper lip protrusion compared to the other speakers. ULH has usually more influence on the upper lip than the LLH parameter has on the lower lip, except for the speakers PB and LD.

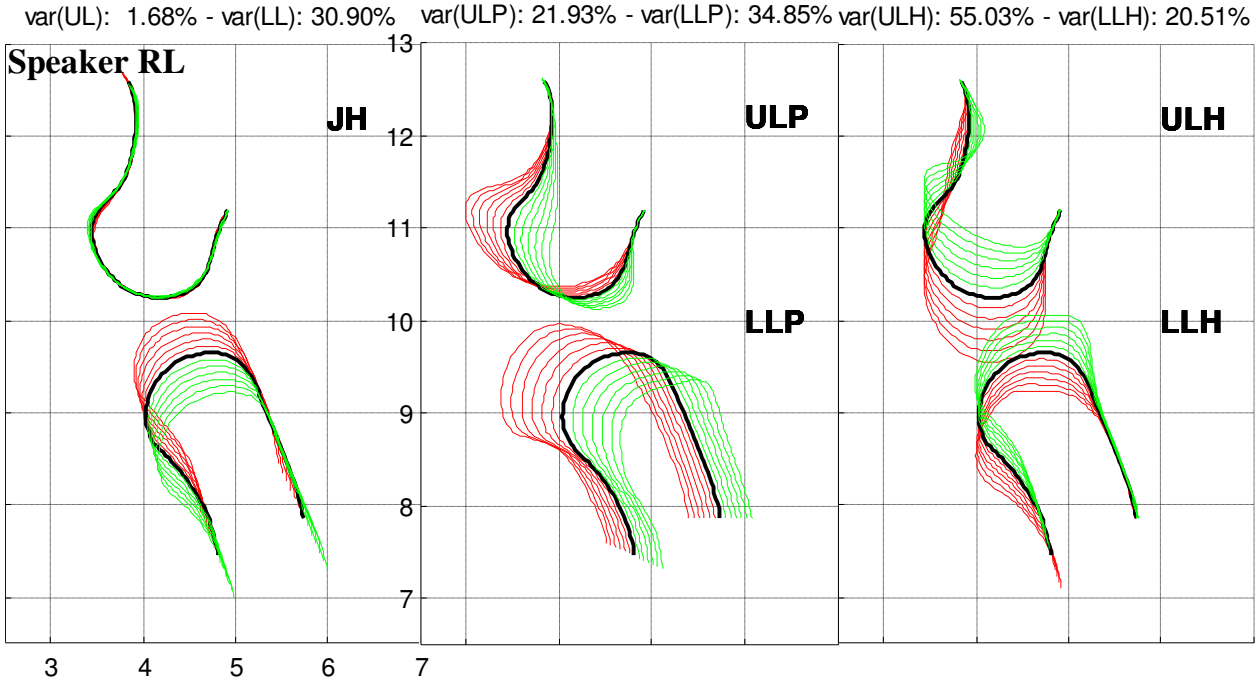
Figure 3-6 shows the percentage of variance explained by each component for all speakers. For the upper lip models we see that, among our speakers, JH explains the maximum variance for speaker LD and the minimum for speaker RL. ULP explains the maximum variance for speakers PB, AA and LD, and the minimum for speakers AK and LH. ULH explains the maximum variance for speaker LH and the minimum for speaker LD. Note that speaker LD presents more variance explained for JH, more for ULP and less for ULH compared to other speakers.

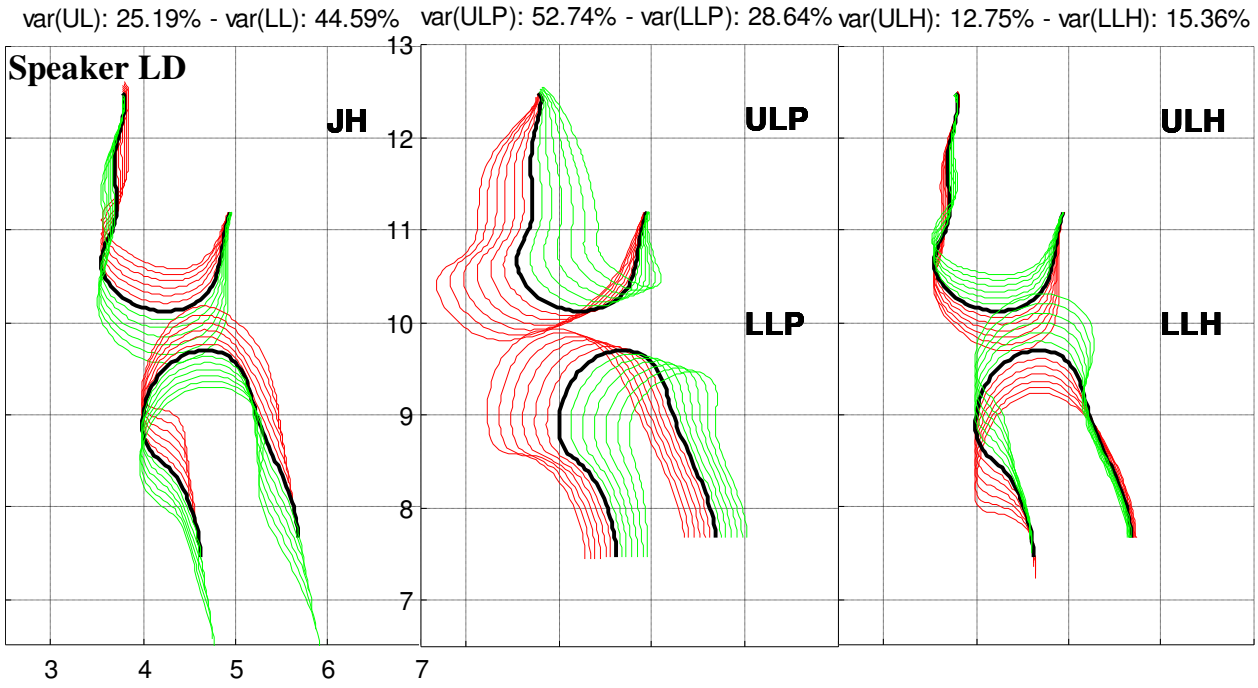
For the lower lip models we see that, among our speakers, JH explains the maximum variance for speaker MGO and the minimum for speaker LH. LLP explains the maximum variance for speaker AA and the minimum for speaker MGO. LLH explains the maximum variance for speakers LH and BR, and the minimum for speaker HL. Note that speaker MGO presents more variance explained for JH and less for LLP compared to other speakers.

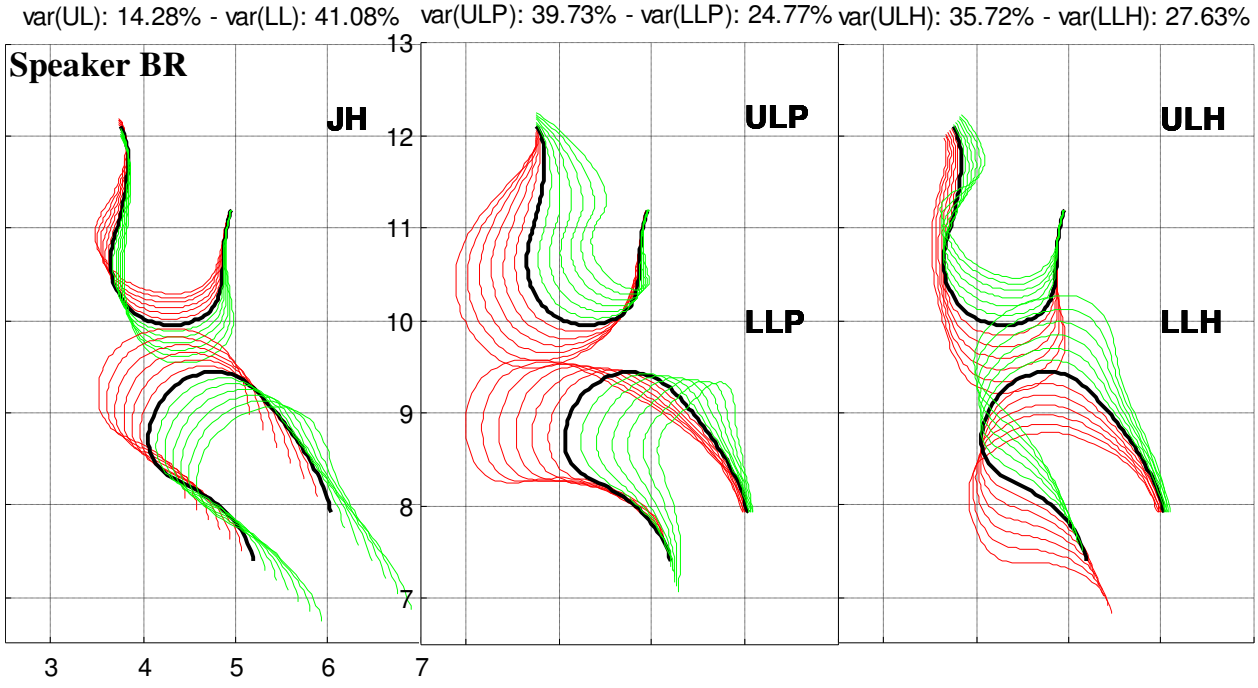

var(UL): 13.30% - var(LL): 38.72%
var(ULP): 61.42% - var(LLP): 32.61%
var(ULH): 16.71% - var(LLH): 20.63%
Speaker PB
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7

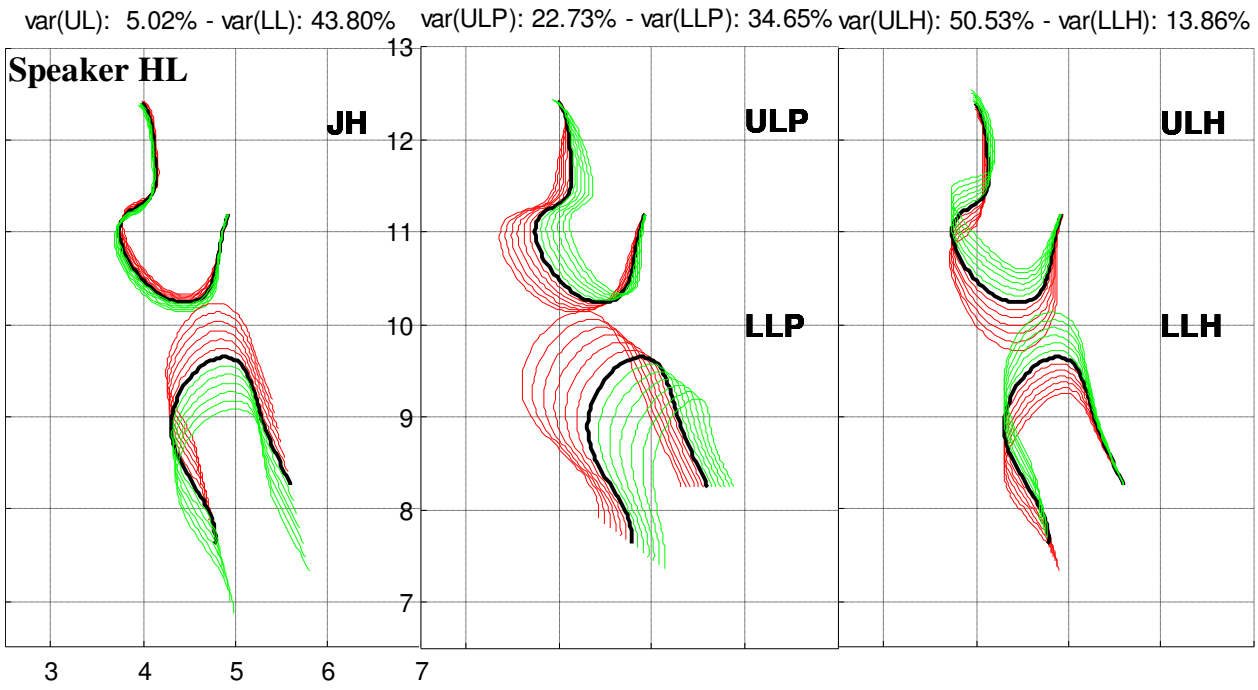

var(UL): 12.86% - var(LL): 37.92%
var(ULP): 36.75% - var(LLP): 36.03%
var(ULH): 39.18% - var(LLH): 15.02%
Speaker YL
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7

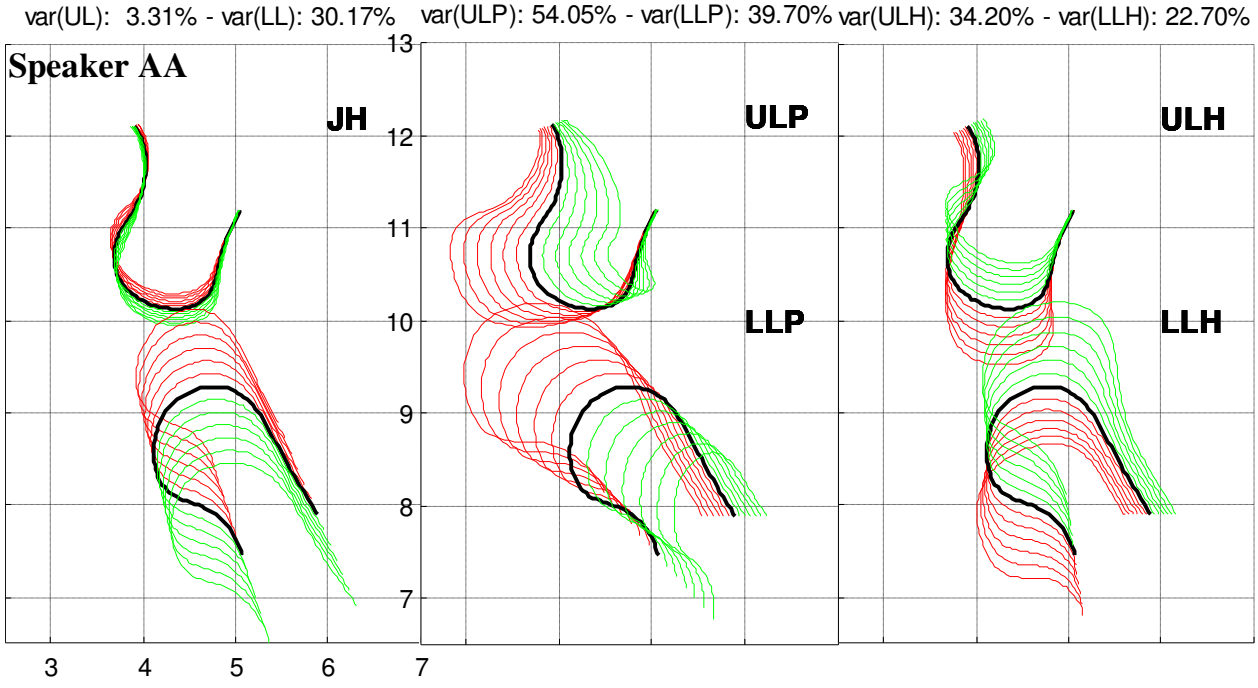

var(UL): 2.59% - var(LL): 23.25%
var(ULP): 8.39% - var(LLP): 28.11%
var(ULH): 66.01% - var(LLH): 30.19%
Speaker LH
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7

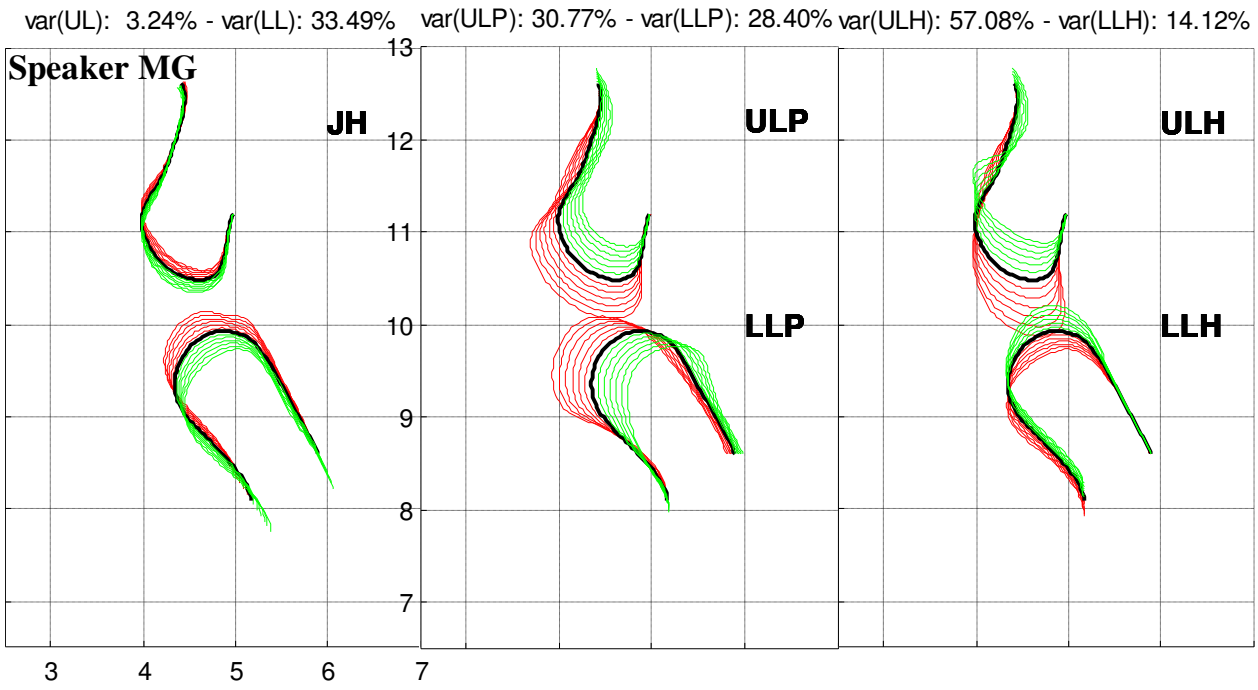

var(UL): 1.68% - var(LL): 30.90%
var(ULP): 21.93% - var(LLP): 34.85%
var(ULH): 55.03% - var(LLH): 20.51%
Speaker RL
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7

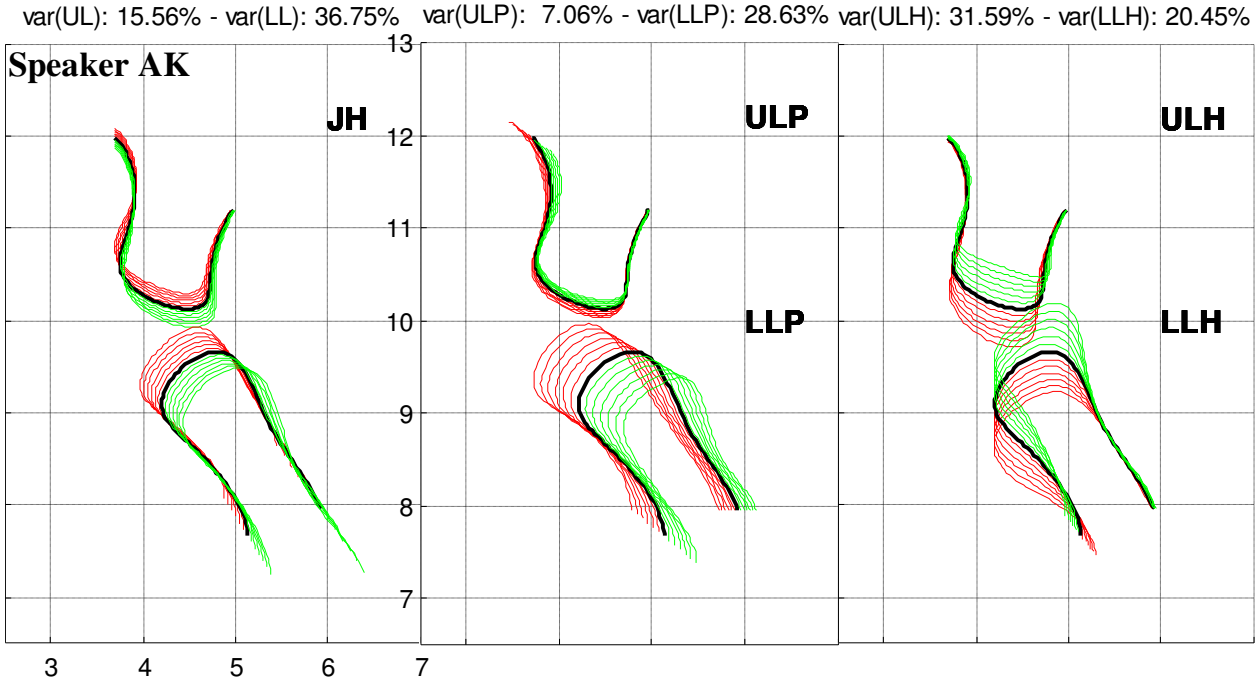

var(UL): 25.19% - var(LL): 44.59%
var(ULP): 52.74% - var(LLP): 28.64%
var(ULH): 12.75% - var(LLH): 15.36%
Speaker LD
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7


var(UL): 14.28% - var(LL): 41.08%
var(ULP): 39.73% - var(LLP): 24.77%
var(ULH): 35.72% - var(LLH): 27.63%
Speaker BR
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7


var(UL): 5.02% - var(LL): 43.80%
var(ULP): 22.73% - var(LLP): 34.65%
var(ULH): 50.53% - var(LLH): 13.86%
Speaker HL
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7


var(UL): 3.31% - var(LL): 30.17%
var(ULP): 54.05% - var(LLP): 39.70%
var(ULH): 34.20% - var(LLH): 22.70%
Speaker AA
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7


var(UL): 3.24% - var(LL): 33.49%
var(ULP): 30.77% - var(LLP): 28.40%
var(ULH): 57.08% - var(LLH): 14.12%
Speaker MG
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7


var(UL): 15.56% - var(LL): 36.75%
var(ULP): 7.06% - var(LLP): 28.63%
var(ULH): 31.59% - var(LLH): 20.45%
Speaker AK
JH
ULP
LLP
ULH
LLH
13
12
11
10
9
8
7
3
4
5
6
7

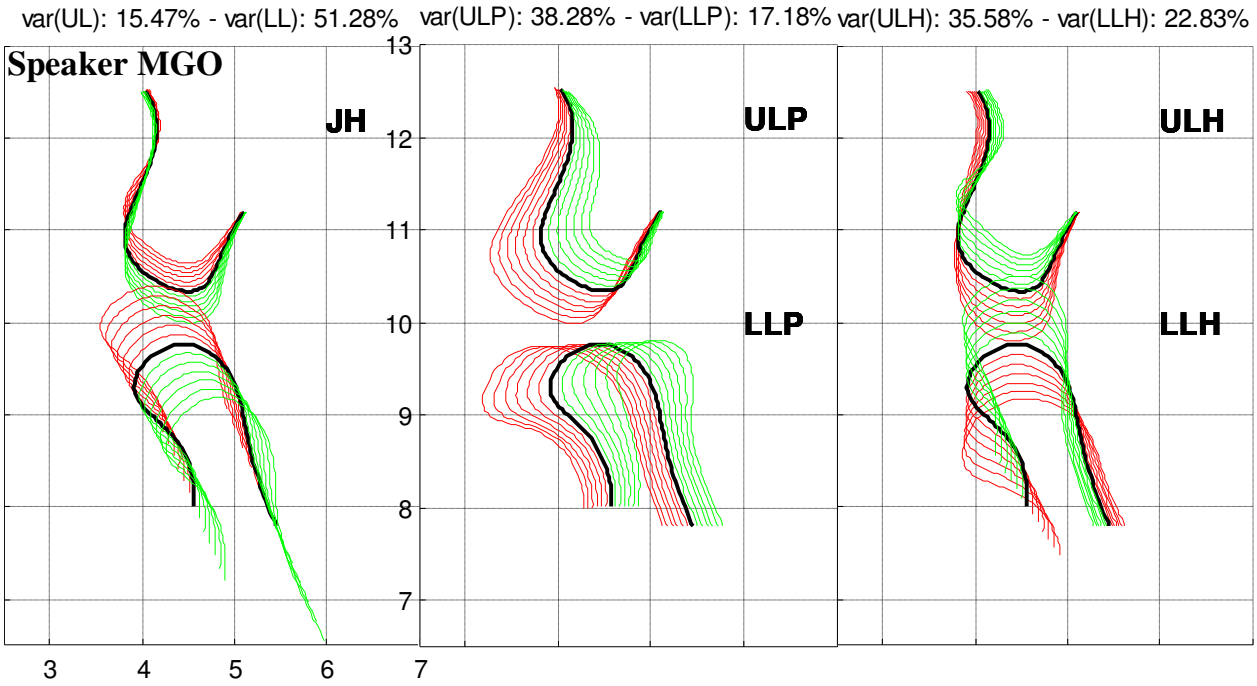
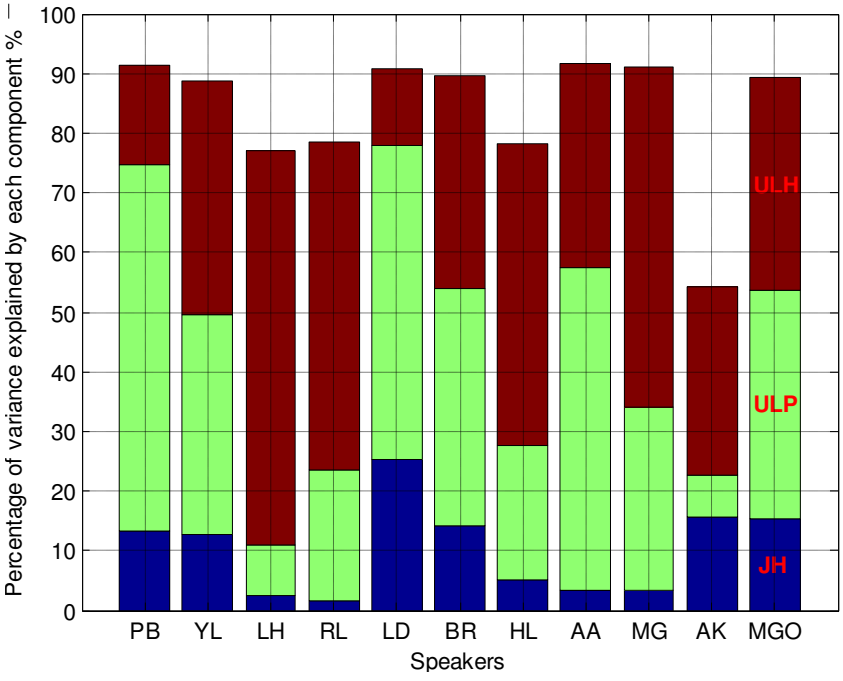


Figure 3-5 - Nomograms of lip components determined by Guided PCA for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. Each predictor (JH, ULP, LLP, ULH and LLH) is varied from -3 to +3 with a 0.5 step. The relative data variance explained by each component is displayed



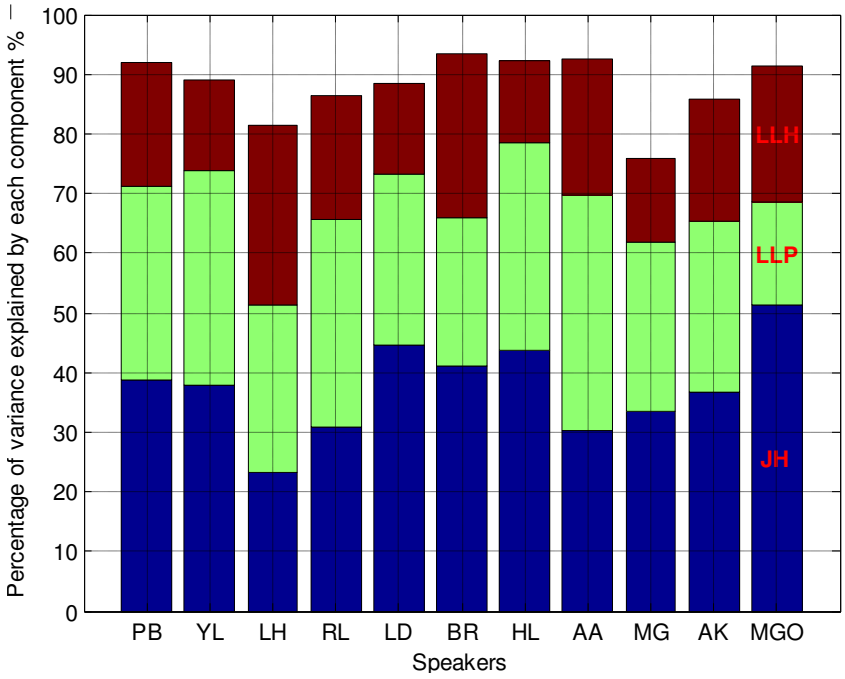


Figure 3-6 - Percentage of variance explained by each guided PCA component of the upper lip and lower lip models, for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. Top: Upper lip models in which JH contribution (blue section at the bottom), ULP contribution (green section), ULH contribution (brown section at the top). Bottom: Lower lip models in which JH contribution (blue section at the bottom), LLP contribution (green section), LLH contribution (brown section at the top)

3.5.3. Protrusion

Figure 3-7 shows statistics related to the protrusion amplitude of the upper lip (proTop) and lower lip (proBot). This statistical analysis is a complement for the nomograms of the predictors ULP and LLP presented on Figure 3-5. The display conventions of Figure 3-7 are similar to those of Figure 3-3. For each observation, proTop was measured as the horizontal distance between the upper incisor and the most advanced point of the upper lip. Similarly, proBot was measured as the horizontal distance between the lower incisor and the most advanced point of the lower lip, as explained in Chapter 2. Speakers are ordered from the one with minimum movement amplitude to the one with maximum amplitude. In the case of proTop, speaker LH performs the minimum upper lip protrusion movement among our speakers. Oppositely, speaker PB has the maximum range of movement. In the case of proBot, speaker MG has the minimum range of displacement, while speaker AA has the maximum.

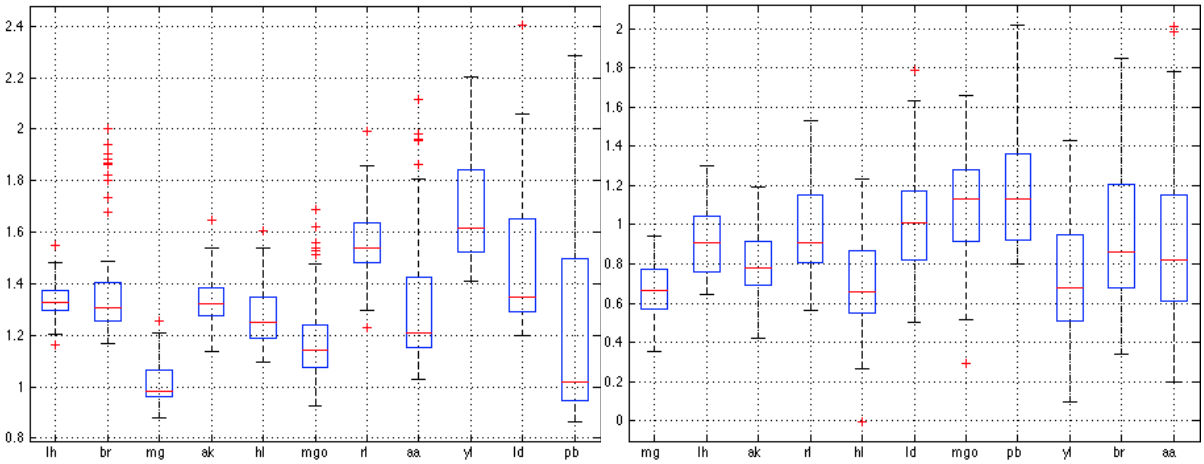


Figure 3-7 - Statistics of the lip protrusion amplitude. Left: upper lip protrusion (cm). Right: lower lip protrusion (cm)

3.5.4. Lip aperture

Figure 3-8 shows statistics of the lip aperture (lipHei). lipHei was measured as the vertical distance between the lower point of the upper lip and the highest point of the lower lip. Speakers are ordered from the one with minimum lip opening to the one with maximum opening. Overall, the lip opening ranges between 0 cm and about 1.3 cm for almost all speakers. Note that speaker AA performs a wider lip opening, up to about 2 cm, compared to other speakers.

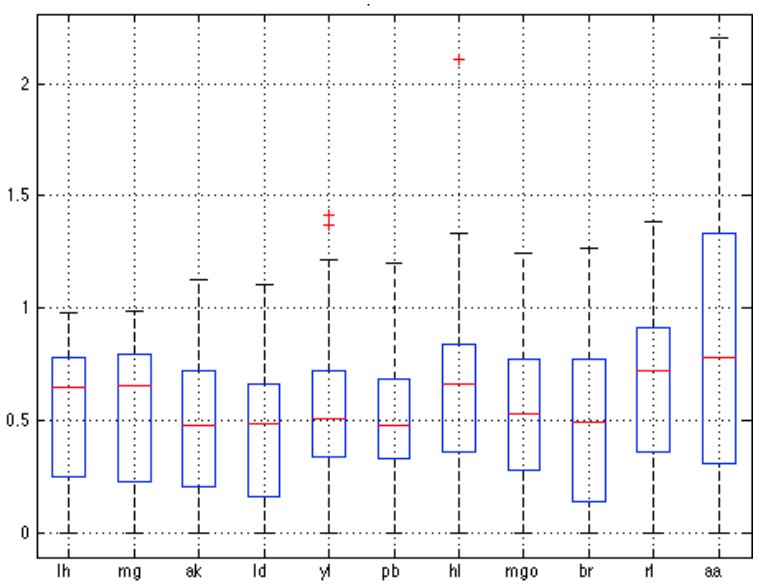


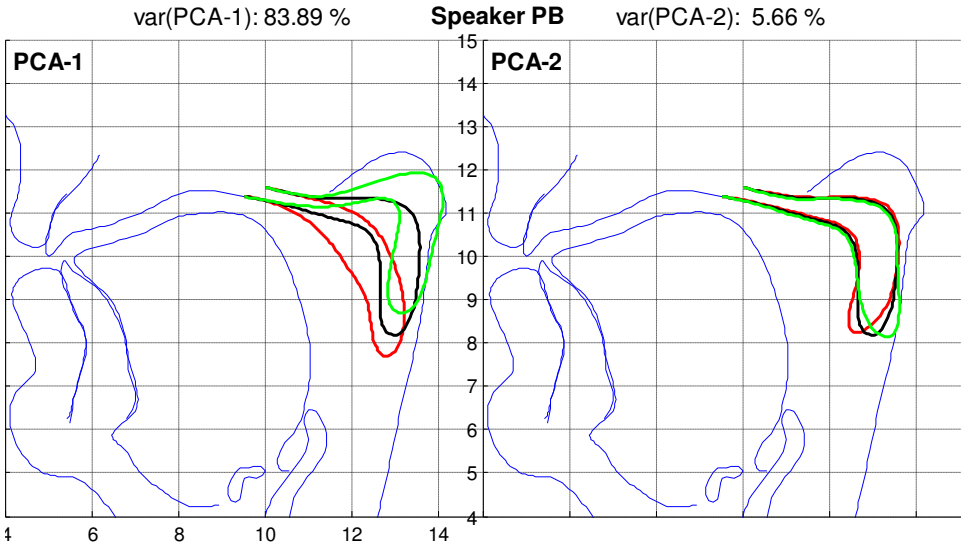
Figure 3-8 – Statistics of the lip aperture amplitude (cm)

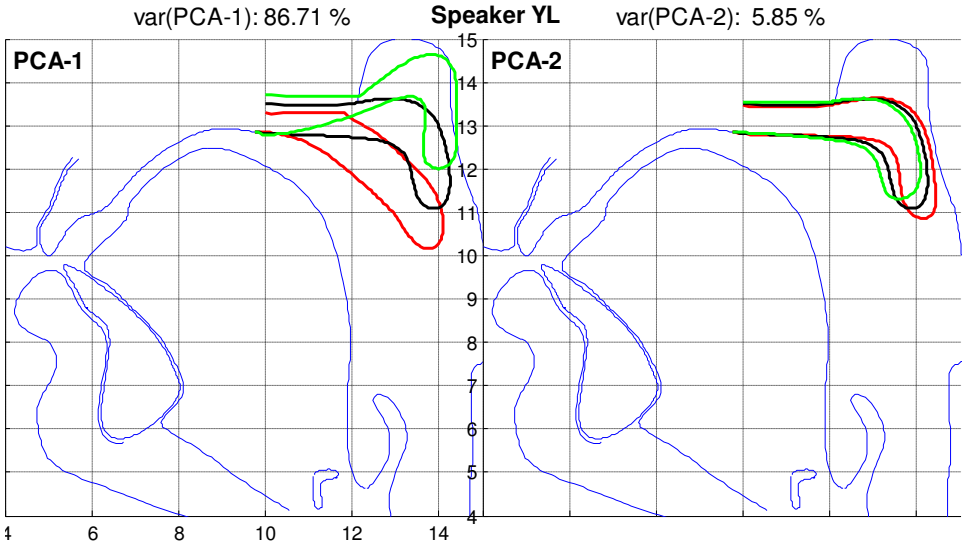
3.6. Velum control strategies

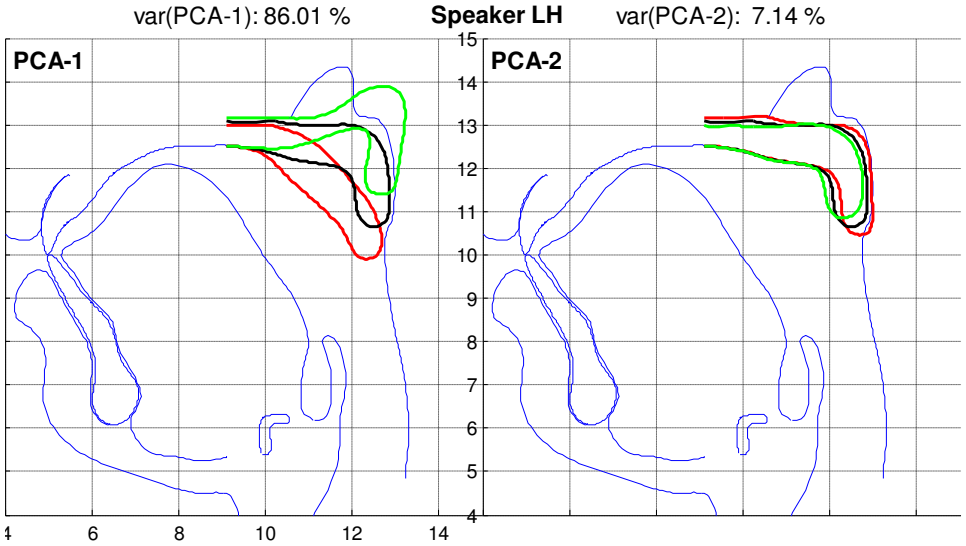
This section compares the articulatory strategies used by our speakers to move their velum. Figure 3-9 illustrates the first two principal components of the velum contour extracted by PCA. Overall, the first component represents an oblique movement

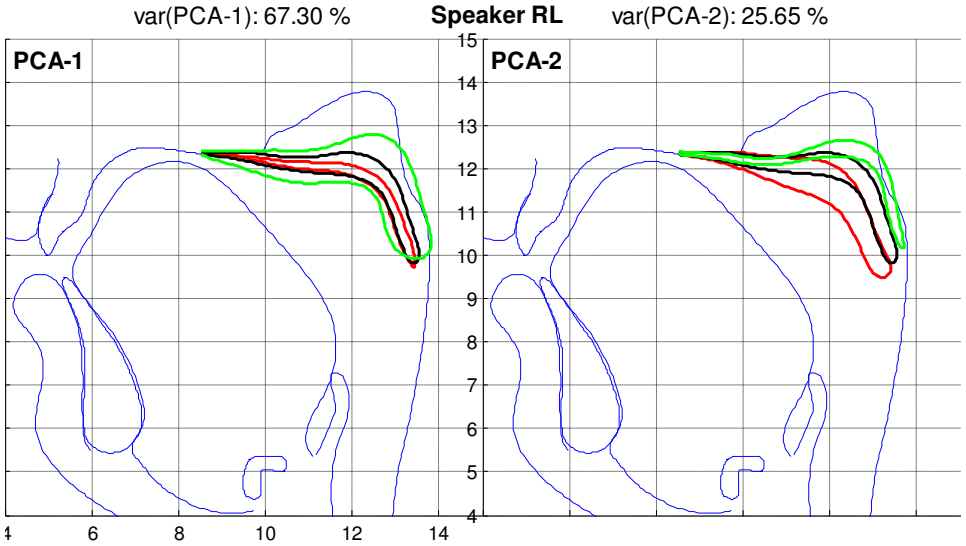
related to the levator veli palatini muscle, as stated by Serrurier & Badin in several studies (2005; 2008). The second component corresponds mostly to a closure of the nasopharyngeal port by a back to front movement (Serrurier & Badin, 2005). The second component may be either related to the sphincter action of the superior pharyngeal constrictor (Serrurier & Badin, 2008) or to the mechanical perturbation from the tongue (Serrurier & Badin, 2005). Note that for speaker RL the components explained above are swapped, which means that speaker RL makes more use of the horizontal constricting closure of the nasopharyngeal port than of the oblique movement of the velum. Besides, one can observe that the first component of speaker RL, as well as the second component of speakers LD and MG, corresponds to a contraction-expansion movement. Besides, the first component of speakers HL and AK represents a simple up-down movement, instead of an oblique vertical movement as for other speakers.

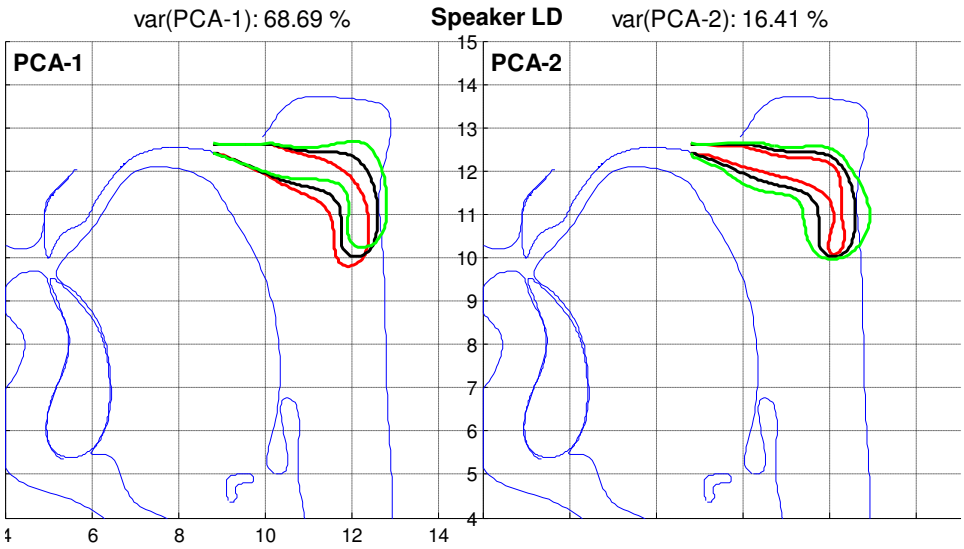
Figure 3-10 shows the percentage of variance explained by each component for all speakers. We see that, among our speakers, PCA-1 explains the maximum variance for speakers BR, YL and LH, and the minimum for speaker RL. PCA-2 explains the maximum variance for speaker RL and the minimum for speaker BR. Note that speaker RL presents less variance explained for PCA-1 and more for PCA-2 compared to other speakers.

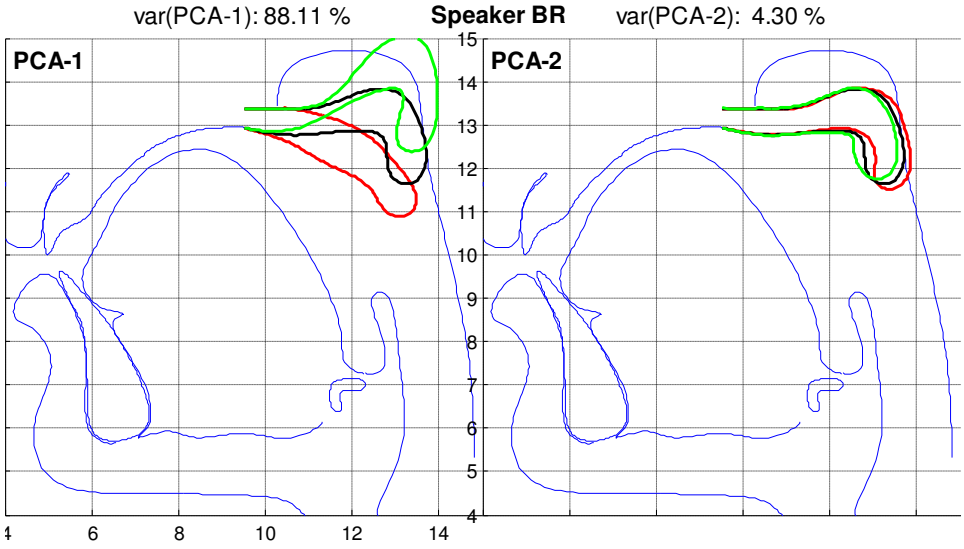


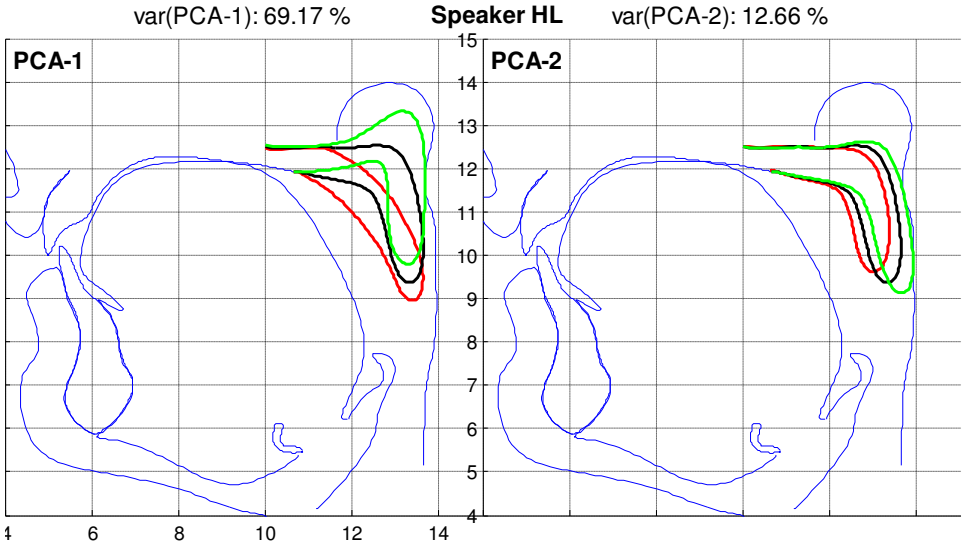

var(PCA-1): 86.71 %
Speaker YL
var(PCA-2): 5.85 %
PCA-1
PCA-2

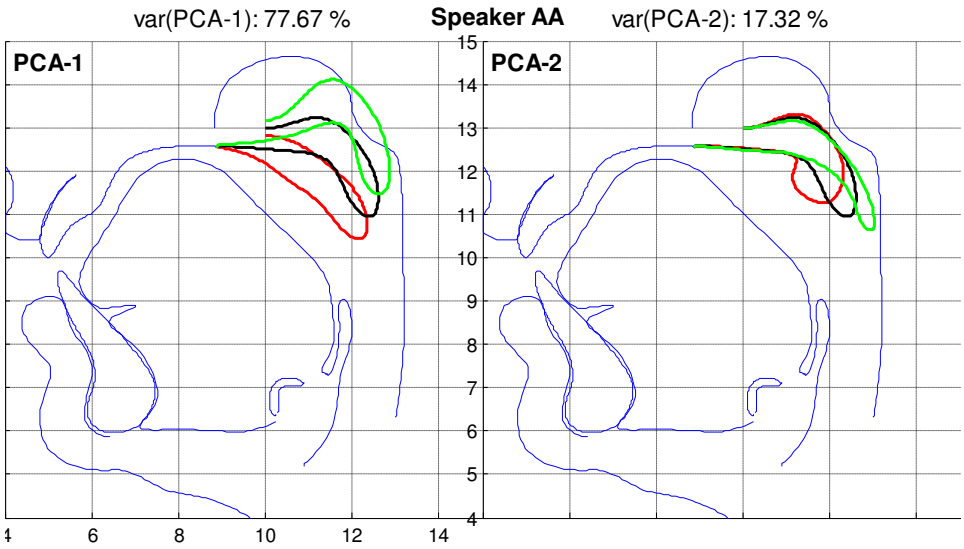

var(PCA-1): 86.01 %
Speaker LH
var(PCA-2): 7.14 %
PCA-1
PCA-2

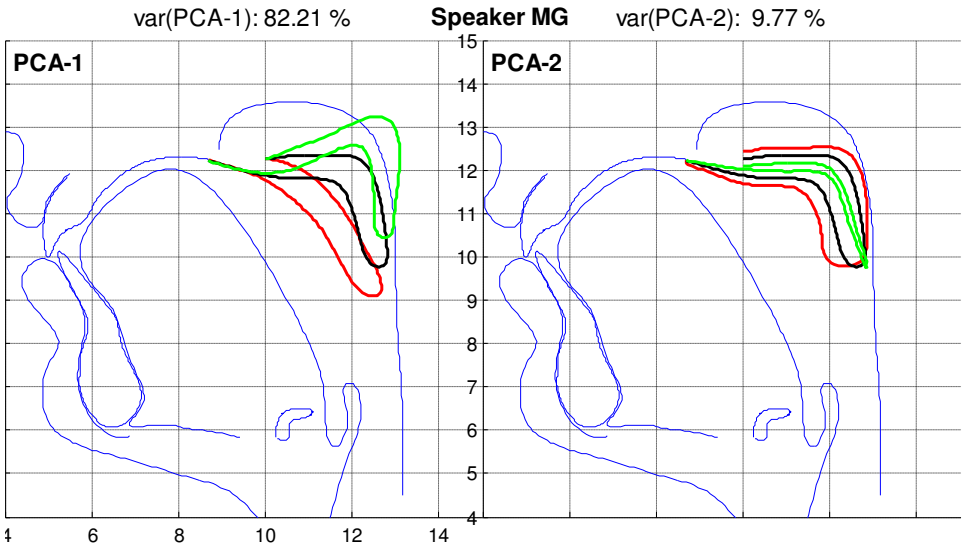

var(PCA-1): 67.30 %
Speaker RL
var(PCA-2): 25.65 %
PCA-1
PCA-2

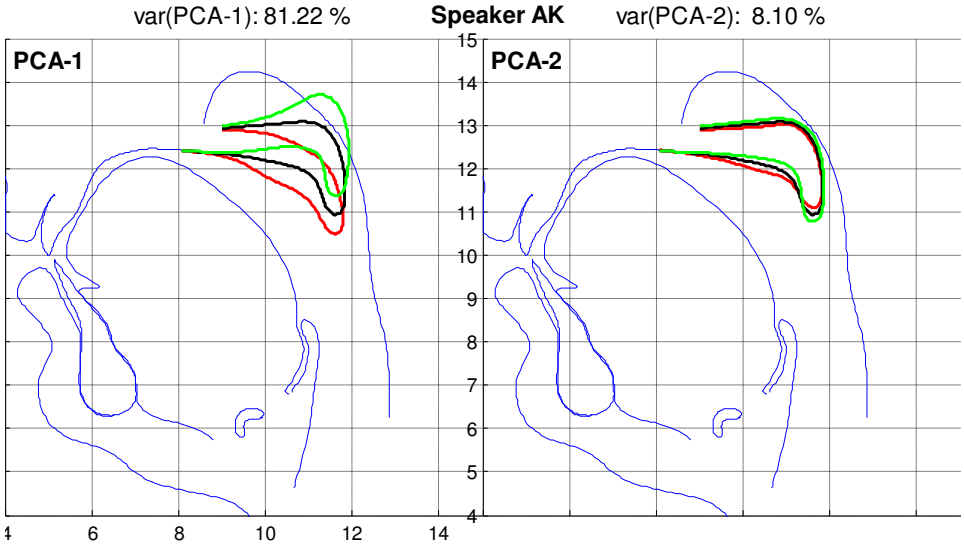

var(PCA-1): 68.69 %
Speaker LD
var(PCA-2): 16.41 %
PCA-1
PCA-2


var(PCA-1): 88.11 %
Speaker BR
var(PCA-2): 4.30 %
PCA-1
PCA-2


var(PCA-1): 69.17 %
Speaker HL
var(PCA-2): 12.66 %
PCA-1
PCA-2


var(PCA-1): 77.67 %
Speaker AA
var(PCA-2): 17.32 %
PCA-1
PCA-2


var(PCA-1): 82.21 %
Speaker MG
var(PCA-2): 9.77 %
PCA-1
PCA-2


var(PCA-1): 81.22 %
Speaker AK
var(PCA-2): 8.10 %
PCA-1
PCA-2

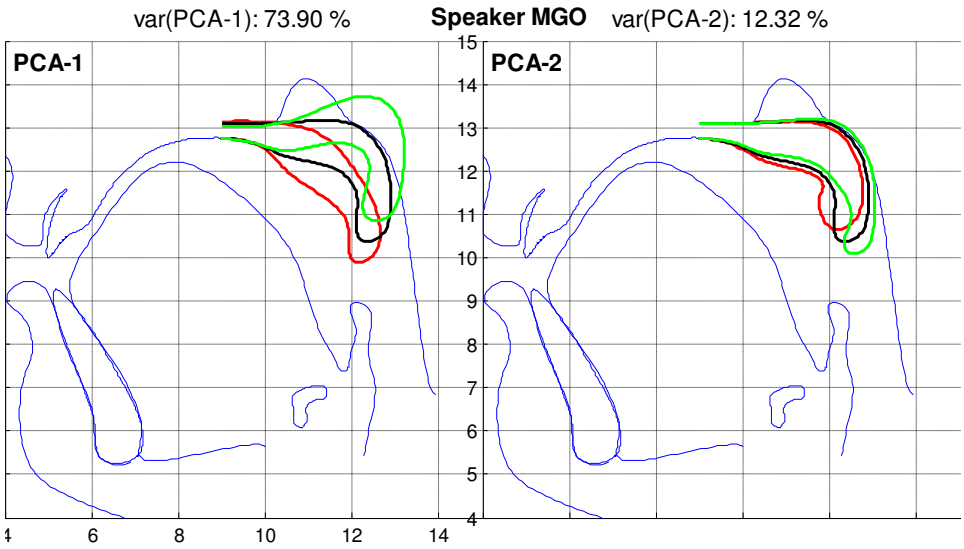


Figure 3-9 - Nomograms of velum components determined by PCA for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. Each predictor (PCA-1 and PCA-2) has values -2, 0 and 2 (red, black and green, respectively). The relative data variance explained by each component is displayed

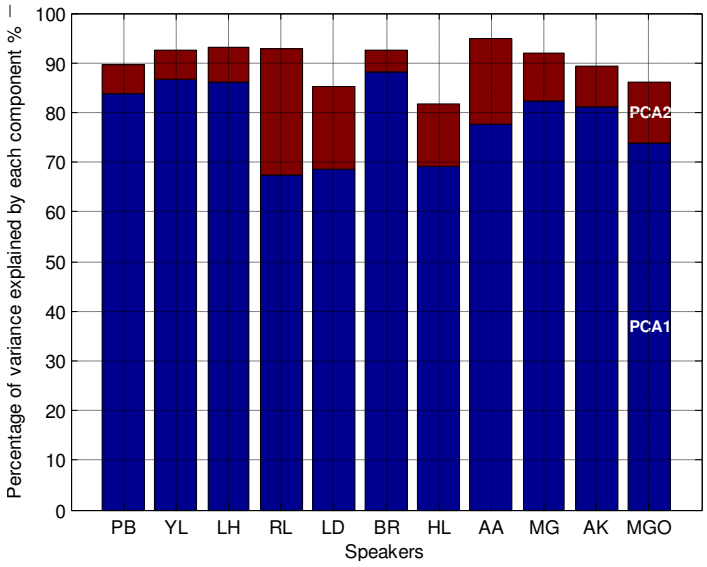


Figure 3-10 - Percentage of variance explained by each PCA component of the velum models, for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. PCA-1 contribution (blue section at the bottom), PCA-2 contribution (brown section at the top)

Furthermore, by looking at the MRI data we observed that speakers AA, BR, and LD have a special strategy to articulate the consonant /ʁ/ compared to other speakers. These speakers roll up their uvulas against the tongue to produce the consonant /ʁ/. This behaviour is repeated for almost all the vocalic contexts. The Figure 3-12 shows an example of a speaker who uses this velum strategy and a speaker who does not. Moreover, some examples of MRI images for the consonant /ʁ/ can be found in the annex A at the end of this manuscript. It was also observed that for speakers AA, BR, and LD, the consonant /ʁ/ is very well separated from the other two classes of nasal and non nasal articulations in the PCA1-PCA2 space, while for other speakers this

classification was not observed. Figure 3-11 shows an example of a speaker with this classification and a speaker for which this classification is not presented.

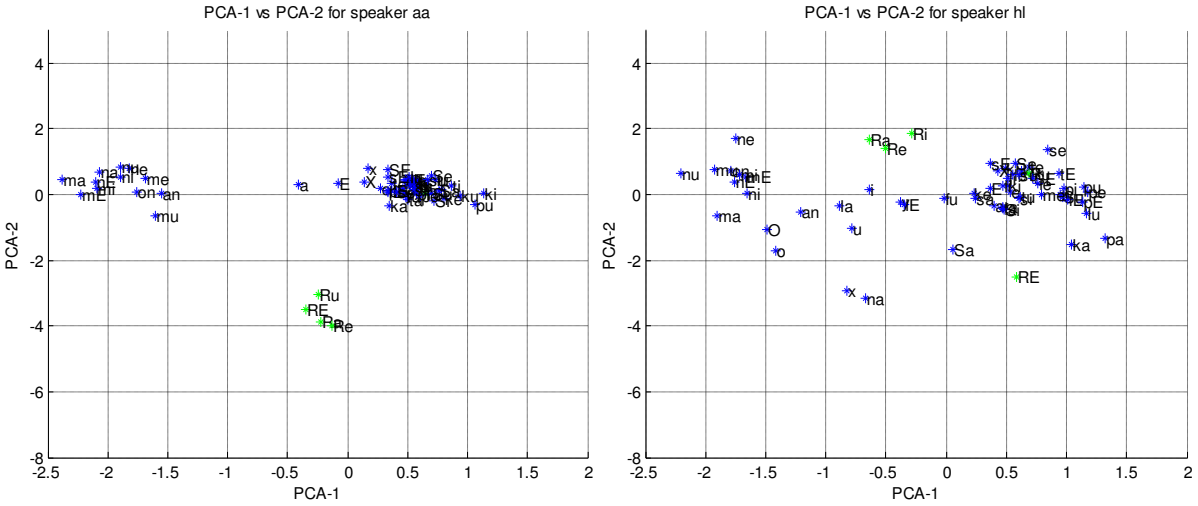


Figure 3-11 - PCA1-PCA2 space. Consonant /ʁ/ (in green dots . Left: speaker AA. Right: speaker HL.

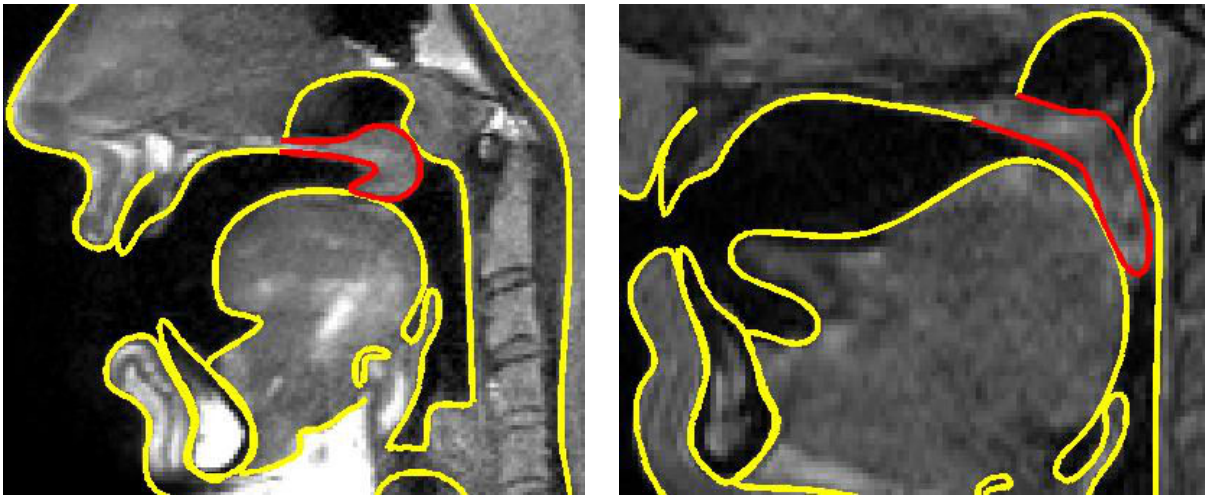


Figure 3-12 - Illustration of velum strategies producing the articulation /Ra/. Left: rolling of the uvula by speaker AA. Right: no rolling of the uvula by speaker HL.

3.7. Acoustics

In this section we describe and compare our speakers in terms of acoustic properties.
In order to compute the acoustic resonance of the vocal tract (Formants), the sagittal grid system explained in Chapter 2 was used. First, the grid was used to compute the midsagittal function. Second, the vocal tract area function was computed as the series of areas and lengths of each sagittal section. In order to compute the vocal tract area function, the α-β model was used (Beautemps et al.1995). In this type of model, the computation of the area depends on the parameters α and β. The value of these two parameters depends on the speaker and vocal tract location. In our specific case, α and β were calculated once from speaker PB and used for the other speakers.

Then, according to the theory of speech production (Fant, 1960; Badin & Fant, 1984), one can compute the vocal tract acoustic transfer function from the area function, and its maxima (formants). Figure 3-13 shows the triangles between vowels /i/, /a/ and /u/ in the F2-F1 space. Though it was expected that female speakers would have higher frequencies than male speakers, one can see some overlap between males and females.

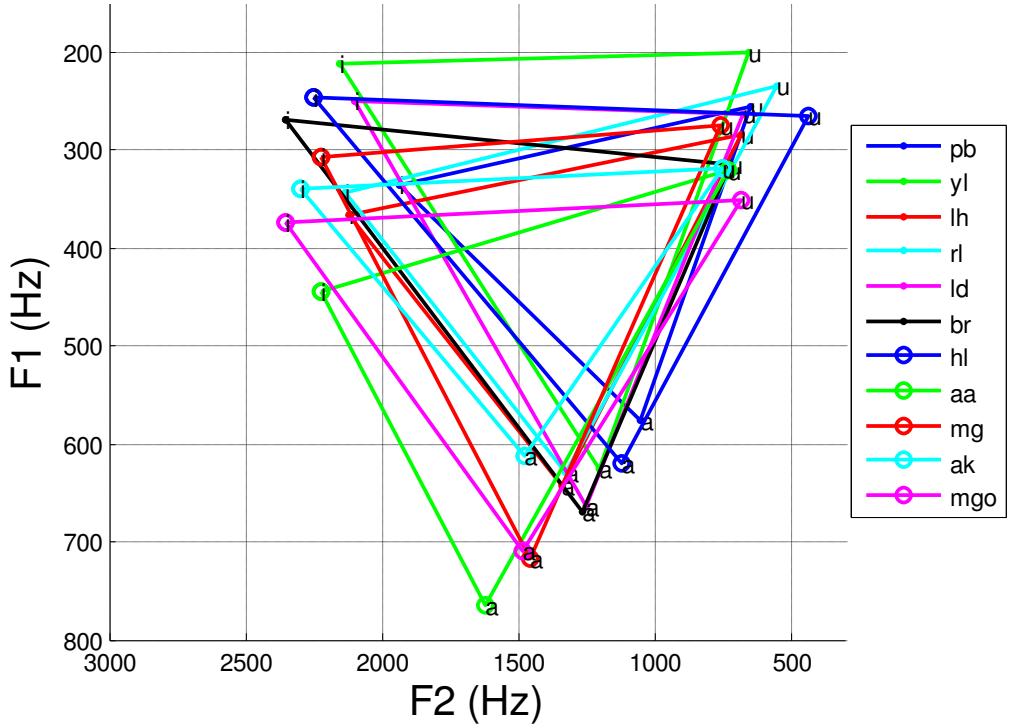
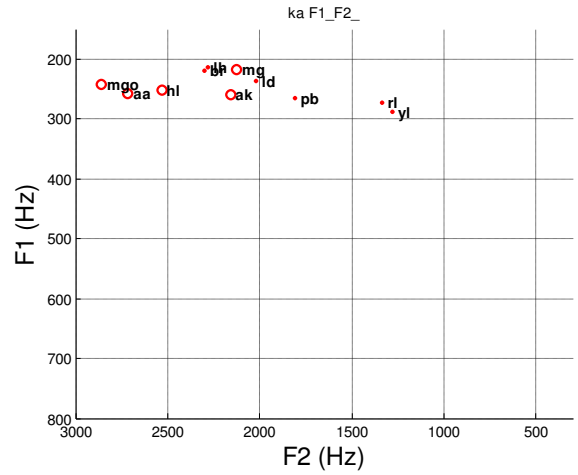
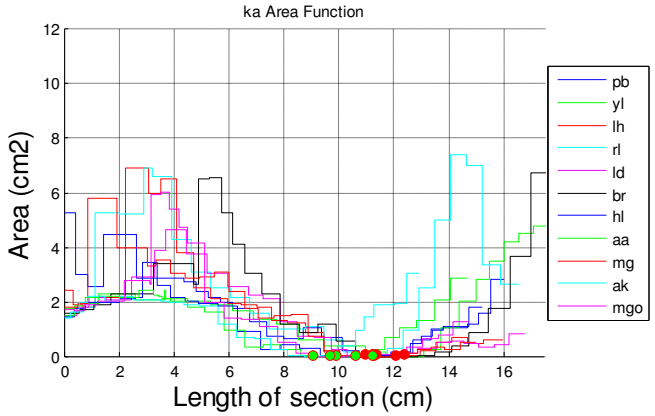
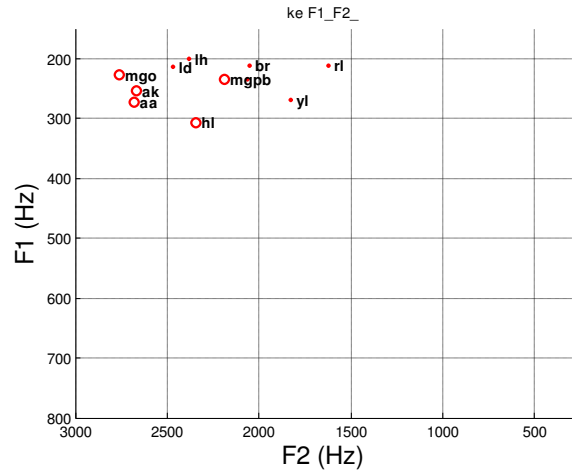
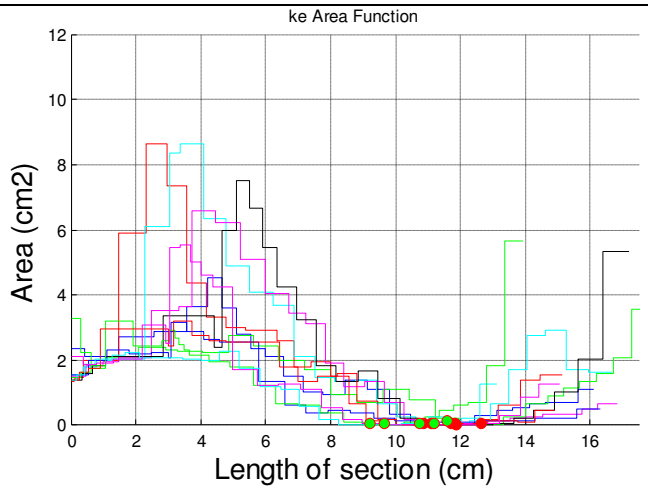
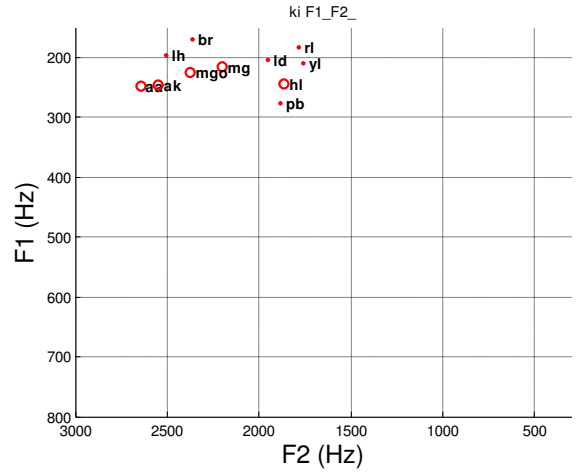
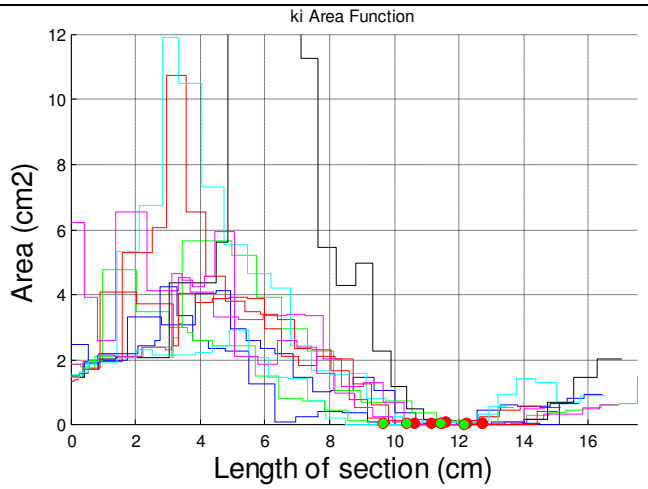


Figure 3-13 - Vocalic triangle of vowels /i/, /a/ and /u/ in the F2 - F1 space, for male speakers (with points) PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO (with circles)

The acoustic behaviour of different consonants in vocalic context was also studied. Two aspects were observed for the consonant /k/ in different vocalic contexts: first, the values for F1 ranged between 200 Hz - 300 Hz while the values for F2, along the axis x, were more spread for the whole set of speakers. Second, female speakers had higher F2 values than male speakers. These two aspects may be due to different vocal tract sizes and different speaker strategies to produce the constriction necessary between the tongue and palate to articulate the consonant /k/. According to the literature (Léon, 2012; Derivery, 1997), speakers can produce the consonant /k/ either in a palatal manner or in a velar manner. Figure 3-14 shows the formants (F2 vs. F1) and the area function of the consonant /k/ in different vocalic contexts, for each speaker. The area function shows that the constriction (Area ~= 0 cm^2) can be at different zones according to the speaker. Moreover, some examples of MRI images for the consonant /k/ can be found in the annex A at the end of this manuscript.

/k[a]/ formants in the F1 - F2 space	Area function of /K[a]/
	
/k[e]/ formants in the F1 - F2 space	**Area function of /K[e]/**
	
/k[i]/ formants in the F1 - F2 space	**Area function of /K[i]/**
	
/k[u]/ formants in the F1 - F2 space	**Area function of /K[u]/**

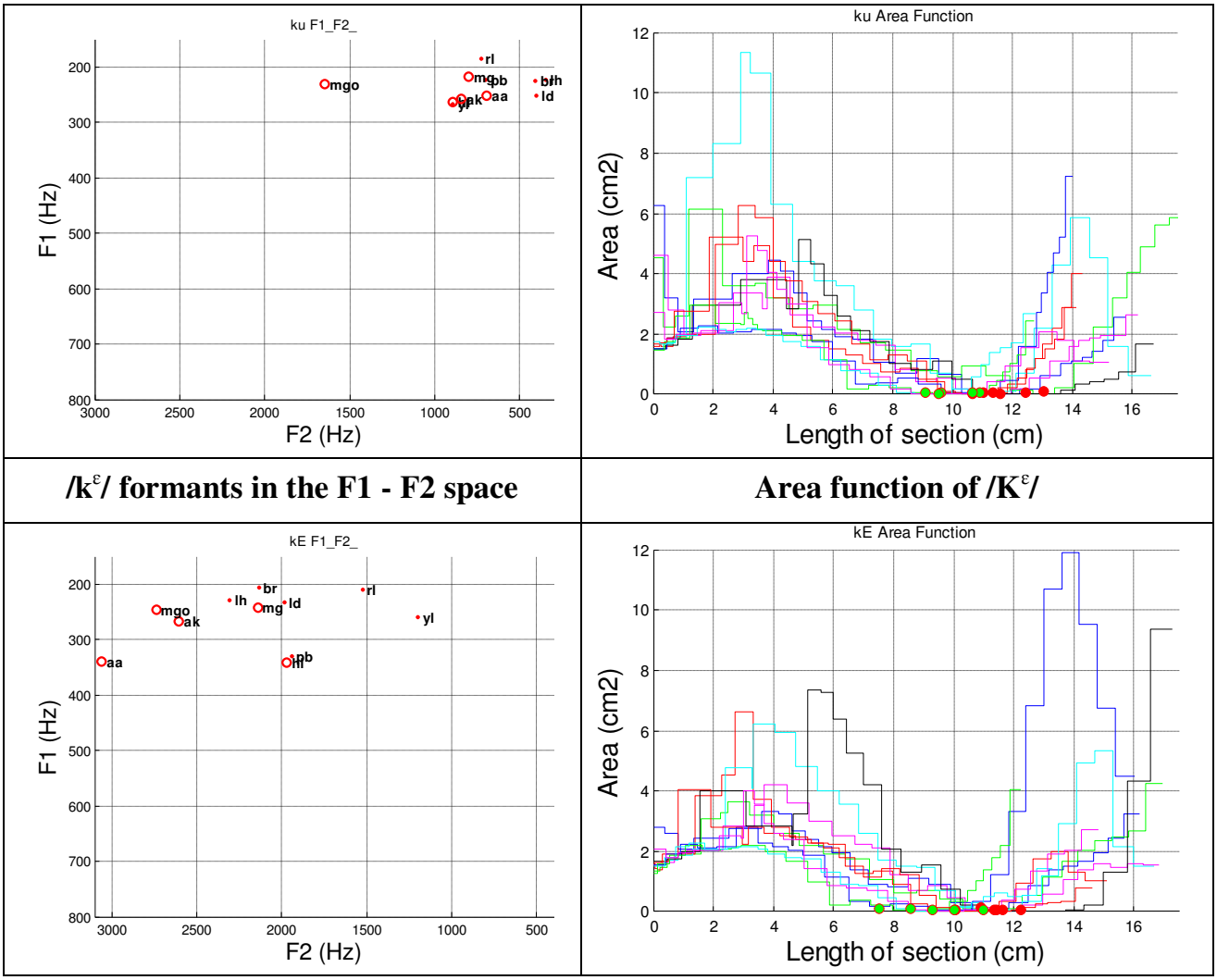
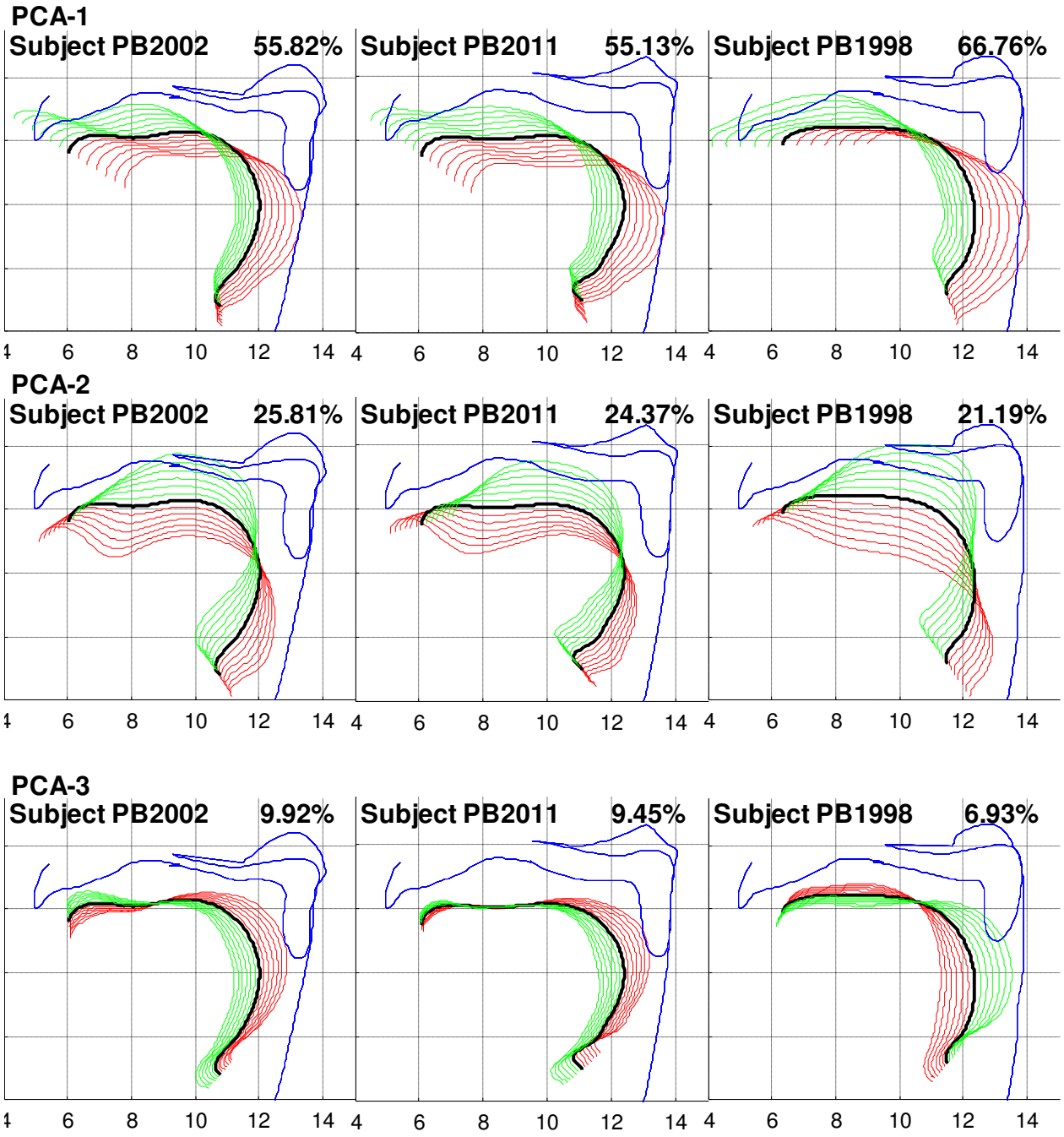


Figure 3-14 - Acoustic resonances (F2 vs. F1 space) and area function with constriction point (red point for male speakers and green point for female speakers) of the consonant /k/ in different vocalic contexts. The area function represents the area of each grid section from the epiglottis up to the tongue tip (from left to right respectively)

3.8. Intra-speaker variability

The purpose of this section is to illustrate the intra-speaker variability issue. A given speaker can use different strategies to produce the same phoneme at different moments. The Figure 3-15 compares the speaker PB, recorded in three different recording sessions (PB-1998, PB-2002 and PB-2011), as regards the first four components extracted by straight PCA. The PCA models, used for the nomograms on Figure 3-15, are based on a corpus composed of 42 articulations: the 10 French oral vowels /i e ɛ a y ø œ u o ɔ/, 2 nasal vowels /ã ɔ̃/ and the 10 consonants /p t k f s ʃ m n ʁ l/ articulated in symmetric vowel-consonant-vowel (VCV) context of the 3 vowels /a i u/. The 3 different MRI data sets were traced as explained in Chapter 2. Furthermore, Figure 3-16 shows an example of the articulations /i a u/ superposed for PB-1998, PB-2002 and PB-2011 to visualize the differences on the vocal tract

conformation. One can see some differences between the PCA components in terms of variance explained and the tongue movements explained by the different models. For PCA-1, we observe that PB-1998 represents a movement which is a bit dissimilar to PB-2002 and PB-2011. The same case is presented for PCA-2. Concerning PCA-3, it represents a tongue dorsum movement, but with a slightly different motion for the tongue tip of each version of speaker PB. The component PCA-4 illustrates a very similar tongue tip motion for the models of the three data sets.



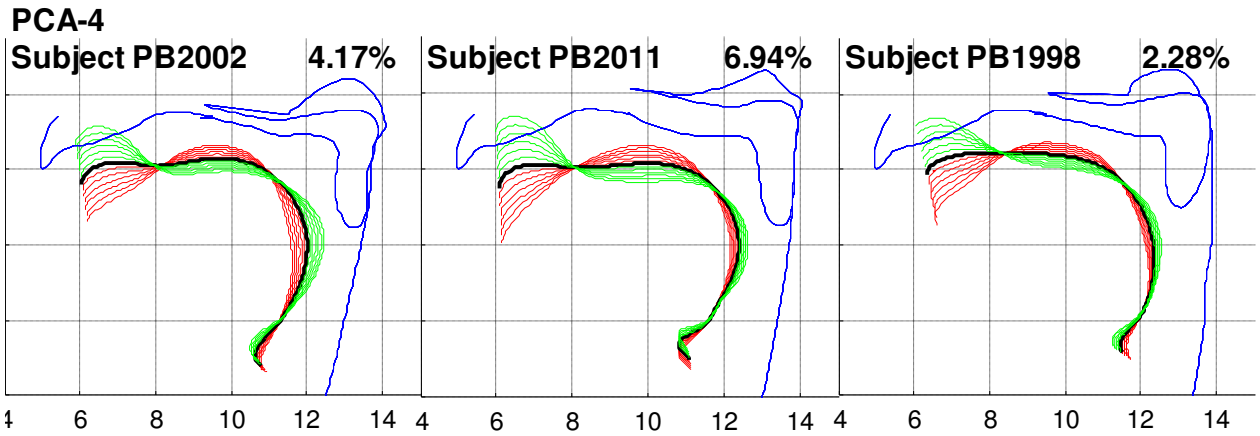


Figure 3-15 - Nomograms of the first four components determined by PCA for PB-1998, PB-2002 and PB-2011. Each predictor (PCA-1, PCA-2, PCA-3 and PCA-4) is varied from -3 to +3 with a 0.5 step. The reference wall (palate, velum and pharynx) is shown in blue. The relative data variance explained by each component is displayed

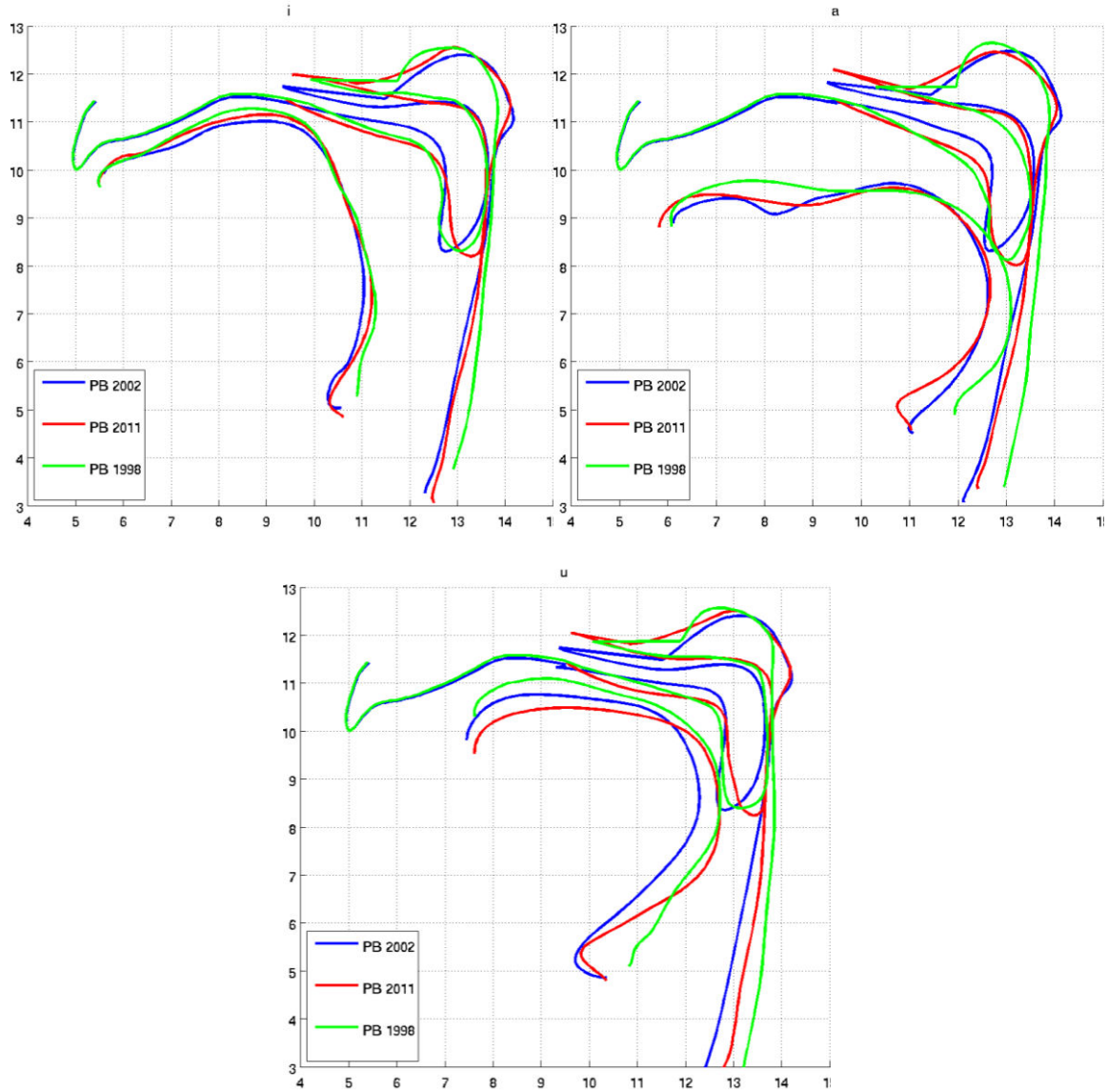


Figure 3-16 - Articulations superposed for PB-1998, PB-2002 and PB-2011. From top to bottom and from left to right: /i/, /a/ and /u/ respectively

3.9. Conclusion

In this chapter, the articulatory strategies employed by our speakers to control their tongues, jaws, and lips have been analysed and compared.

In order to compare the tongue control strategies, we used a procedure based on a guided PCA to extract four principal components (JH, TB, TD, and TT). We have seen that JH can be either associated with an independent movement of the front of the tongue or with a movement that carries the front and the back of the tongue together. We have also observed that TB controls front-back displacements of the tongue body. TB can be related to either a horizontal or a diagonal tongue movement. Besides, we have seen that TD controls the flattening-arching movements. It was noticed that our speakers use more front-back movements than flattening-arching movements, except for speakers YL and RL. Furthermore, TT controls the tongue tip motion. By looking at the nomograms, it was observed that apparently some speakers move their tongue tip more independently from the back than other speakers.

The synergy between tongue and jaw was also studied. It was observed that for a given displacement of the jaw, the tongue may globally move in a proportion that depends on the speaker. The proportion of tongue tip movement as regards the jaw movement ranges between 0.1 and 1.5. A particular compensation strategy was noticed for speaker MG: she uses a small range of vertical jaw movement, but compensates that by moving her tongue about 1.5 times compared to the jaw movement.

Since the lips constitute also an important part on speech articulation, the different strategies employed by our speakers to move their lips have been studied. Three components (JH, ULP, and ULH) were extracted by means of a guided PCA for the upper lip. Similarly, three components (JH, LLP, and LLH) were extracted for the lower lip. Overall, we observed that the jaw movement (JH) has usually more influence on the lower lip than on the upper lip. We also noticed that protrusion (ULP and LLP) can be stronger either on the upper lip or the lower lip depending on the speaker. Furthermore, the ULH parameter, which is related to the vertical movements of the upper lip, has usually more influence on the upper lip than LLH has on the lower lip, over all the speakers. Concerning the lip opening of different speakers, it was observed that the lip opening ranges between 0 cm and about 1.3 cm over all speakers. However, speaker AA performs a wider lip opening, up to about 2 cm, compared to other speakers.

The velum motion has also been studied. Two principal components were extracted by means of a PCA method. The first component represents an oblique movement. The second component corresponds mostly to a closure of the nasopharyngeal port by a back to front movement. By looking at the MRI data, we noticed that speakers AA,

MGO, BR, and LD usually roll up their uvulas against the tongue to produce the consonant /ʁ/ in all vocalic contexts.

Our speakers were also compared in terms of acoustics. The triangle of vowels /i/, /a/ and /u/ in the F2 - F1 space, for all speakers, was coherent with previous studies. Besides, it was observed that for the consonant /k/, in different vocalic contexts, the values for F1 ranged between 200 Hz - 300 Hz while the values for F2 were usually more spread along the axis x. This could be related to the fact that some speakers may produce the consonant /k/ either in a palatal manner or in a velar manner.

Finally, this chapter discussed the intra-speaker variability issue. PCA models of three different data bases, recorded for the same speaker, were built. We observed some differences between the PCA components in terms of variance explained and the tongue movements explained by the different models. The components PCA-1 and PCA-2 represented similar movements for the models of PB-2002 and PB-2011 but not for PB-1998. On the other hand, PCA-3 represented a tongue dorsum movement which was slightly different for each model. The component PCA-4 illustrated a very similar tongue tip motion for the models of the three data sets.

The next chapter will concentrate on modelling the tongue, lips and velum contours by means of multilinear decomposition methods.

Chapter 4. Individual and multilinear models of the tongue, lips and velum contours

4.1. Introduction

As explained in Chapter 1, one important issue when it comes to model vocal tract contours of several speakers is the variability as regards articulatory strategies. For instance, several speakers can employ different strategies to achieve sounds that are equivalent for communication purposes. Thus, the purpose of this chapter is to find common articulatory patterns among our speakers. Tongue, lips, and velum models, were built, by means of multilinear methods, to extract a set of common components that control the contours of several speakers.

Firstly, this chapter describes how the mean was subtracted from the data. The models are evaluated by means of two criteria: the relative explained variance and the Root Mean Square reconstruction Error (RMSE). The models were also built using a leave-one-out cross validation procedure (LOOCV) to ensure that there was not over-fitting. Second, the results of individual speaker models and various multilinear models are presented and compared. In order to have a reference starting point for the tongue models, our modelling is first limited to a repertoire of only French vowels and compared with the results quoted on the literature. Then, multilinear models of the tongue contour are built for a more extended corpus of vowels and consonants in vocalic context. Besides, this chapter presents a normalisation approach based on the mapping of articulatory spaces of the components extracted by the PCA tongue models. Finally, we also present models built for the lips and velum contours which are important organs during speech articulation. For instance, the role of the lips, in acoustic terms, is in particular to ensure a constriction for closed vowels and labio-dental fricatives (Fant, 1960). Besides, the velum contour controls the nasality in articulation (Serrurier & Badin, 2005).

4.2. Mean subtraction and orthogonality

Miranda et al. (2008) made a study about the mean centering issue for PCA. In that study, they discuss the necessity of subtracting the mean to find principal components that minimize the mean square error. The mean centering ensures that the first principal component corresponds to the direction of maximum variance. If the data are

not centered, the first PCA component might instead be related to the mean of the data. Therefore, in this study the models were built with centered data. The data of each speaker was centered by subtracting the mean of each articulatory measurement. Another important issue was the orthogonality between the components extracted. By definition in PCA, each component extracted is orthogonal to each other. The three-way linear decomposition methods PARAFAC, TUCKER and joint PCA were also set up to extract orthogonal components and the data were mean subtracted.

4.3. Assessment of models: variance explained and reconstruction error

The variance explained is defined by the ratio of the variance of the reconstructed data over the variance of the original measured data (as explained in Chapter 3). The Root Mean Square Error (RMSE) was also used to measure the precision of the data reconstructed from the models. The RMSE, for a given speaker X, is computed according to the following equation:

Root Mean Square Error: $\sqrt{\frac{\sum_{l}^{n}\sum_{l}^{m}(X_i - X_{i\_predicted})^2}{n.m}}$

Being n the number of observations and m the number of articulator measurements.

4.4. Leave-one-out cross validation procedure

According to Hawkins (2004), over-fitting might occur when a model fits the training data, but fails making predictions of new data since the model has not learned to generalize. In order to avoid over-fitting, it is necessary to use additional methods like the leave-one-out cross validation (LOOCV) procedure. The purpose of using such a method is to verify the capabilities of the model to generalize by evaluating its performance on data that were not used for training.

The models presented in this study are made and assessed by means of LOOCV. LOOCV is a process in which one observation of the data is left out; the model is built from the remaining data and used to predict the left-out observation; this process is then repeated for each observation on the set of data. LOOCV is 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).

4.5. Analysis and results

In this section the reader will be guided from linear models of individual speakers to multilinear models that take into account several speakers and various representations

of the tongue contour. First, in order to have a reference starting point, the modelling has been reduced to a corpus of vowels and compared to the results reported in the literature. The analyses presented in this section are limited to 150 equidistant points of the upper tongue contour (UpperTng), described in Chapter 2.

4.5.1. Linear tongue models with only 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 /i e ɛ a y ø œ u o ɔ/. Two versions of the tongue contour were used: the UpperTng and an under-sampled tongue contour of 3 points. In order to under-sample the tongue contour, three equidistant points were selected from the tongue tip to the back, for each speaker.

A PARAFAC model that extracted 2 components was able to reconstruct the tongue contour with an RMSE of 0.25 cm for UpperTng while the RMSE for tongue contours under-sampled to 3 points was 0.23 cm, accounting for a variance of 77.3 % and 84.6 %, respectively. Table 4-1 shows that, on the overall, our results are comparable with those reported in the literature. The challenge is then to extend this analysis to a corpus including consonants in vocalic contexts (63 articulations), as explained in following sections.

Type	Study	No. Speakers	Corpus	No. Points	Variance Exp
EMA	Hoole (1998)	7	15 vowels	4 sensors	80.0%
	Geng & Mooshammer (2000)	6	15 vowels	4 sensors	96.0%
	Hu (2006)	7	10 vowels	3 sensors	90.0%
X ray	Harshman et al. (1977)	5	10 vowels	13 points	92.7%
MRI	Hoole et al. (2000)	9	7 vowels	13 points	87.0%
	Zheng & Johnson (2003)	5	9 vowels	13 points	76.2%
	Ananthakrishnan et al. (2010)	3	13 vowels	150 points	71.0%
Our Results					
MRI	Valdés (2013)	11	10 vowels	3 points	84.6%
		11	10 vowels	150 points	77.3%

Table 4-1 - Comparison of our results with the literature using PARAFAC with 2 components

In order to have a reference to compare the models of following sections, models with 11 speakers, only the 10 French vowels /i e ɛ a y ø œ u o ɔ/, 150 measurements and several linear and multilinear methods were built. The results in terms of variance explained and RMSE are compared in **Figure 4-1**. Note that the performance of the methods from best to worst is given by PCA, joint PCA, TUCKER and PARAFAC. Further comparisons against these models are presented in the following sections.

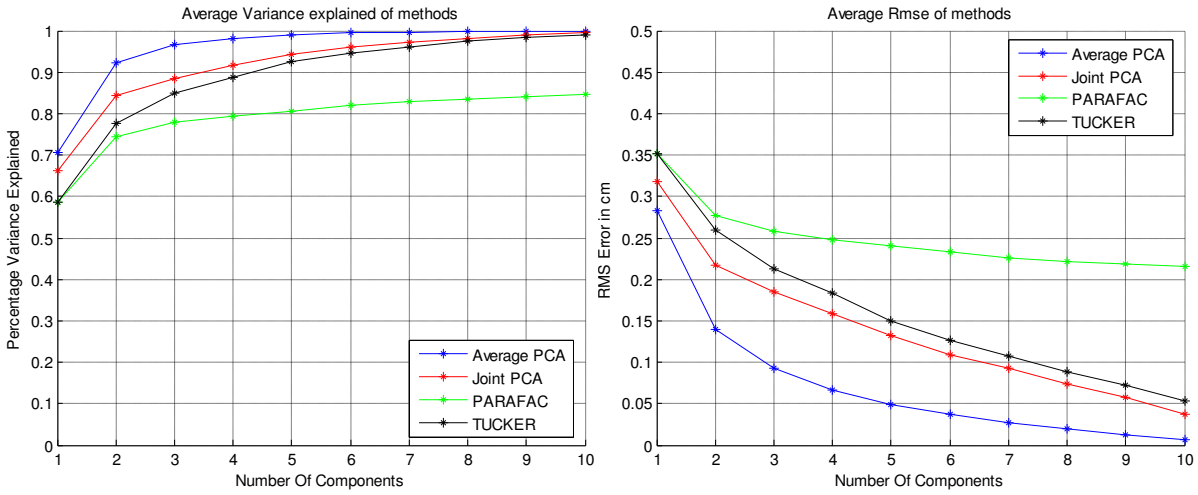


Figure 4-1 - Performance, established using LOOCV, of the PARAFAC, TUCKER and joint PCA as a function of number of components for the tongue contours for a corpus of only vowels. Left: variance explained. Right: RMSE in centimeters

4.5.2. Comparison of linear and multilinear methods as regards number of coefficients

In order to have a criterion of evaluation for the models of the following sections, the various linear and multilinear methods must be compared in terms of number of coefficients. The Table 4-2 shows the number of coefficients of each method as a function of number of components extracted. The computations are made using 11 speakers, 63 articulations and 150 measurements (x and y coordinate). The number of coefficients is computed as described by the following equations:

PCA = [(No. articulations × No. components) + (No. components × No. articulator measurements)] × No. Speakers

PARAFAC = (No. articulations × No. components) + (No. speakers × No. components) + (No. articulator measurements × No. components)

Joint PCA = (No. articulations × No. components) + (No. components × No. articulator measurements × No. speakers)

TUCKER = [(No. articulations × No. components) + (No. components × No. components × No. speakers) + [(No. speakers × No. components) × (No. speakers × No. articulator measurements)]]

According to the results on Table 4-2 the methods can be categorized in decreasing order of number of coefficients as: TUCKER, PCA, joint PCA and PARAFAC. Being

TUCKER the method which uses more coefficients and PARAFAC the one with less coefficients.

	No. of coefficients			
No. components	**PCA**	**PARAFAC**	**Joint PCA**	**TUCKER**
1	2343	224	1713	18224
2	4686	448	3426	36470
3	7029	672	5139	54738
4	9372	896	6852	73028
5	11715	1120	8565	91340
6	14058	1344	10278	109674
7	16401	1568	11991	128030
8	18744	1792	13704	146408
9	21087	2016	15417	164808
10	23430	2240	17130	183230
11	25773	2464	18843	201674
12	28116	2688	20556	220140
13	30459	2912	22269	238628
14	32802	3136	23982	257138
15	35145	3360	25695	275670
16	37488	3584	27408	294224
17	39831	3808	29121	312800
18	42174	4032	30834	331398
19	44517	4256	32547	350018
20	46860	4480	34260	368660
21	49203	4704	35973	387324

Table 4-2 – Number of coefficients of each method as a function of number of components extracted by PCA, PARAFAC, joint PCA and TUCKER

4.5.3. Linear tongue models extended to consonants

In this section linear models of individual speakers and multilinear models that take into account several speakers are presented and compared. A final model is selected and evaluated. The corpus of the models consisted of 63 articulations: 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 vowel-consonant-vowel (VCV) context of five vowels /a e ɛ i u/.

4.5.3.1. Individual tongue models (PCA)

Badin & Serrurier (2006) have shown that the first four components of a guided PCA model accounted for 78.4% of the tongue movement variance. Even though the individual speaker models considered in the present section are not guided, four components were chosen to have a reference of comparison. Figure 4-2 displays the variance explained and RMSE relative to the reconstruction of UpperTng, for the whole corpus of vowels and consonants, by means of PCA as a function of the number of components used. We have found that, when four components are used, the variance that individual speaker models explain from the tongue contour ranges, for the set of speakers, between 91% and 96.13% and the RMSE between 0.09 cm and 0.14 cm. On average, over our eleven speakers, the individual speaker models explain an amount of 93.23% of the data variance, with an RMSE of 0.12 cm, for an average PCA model of 4 components.

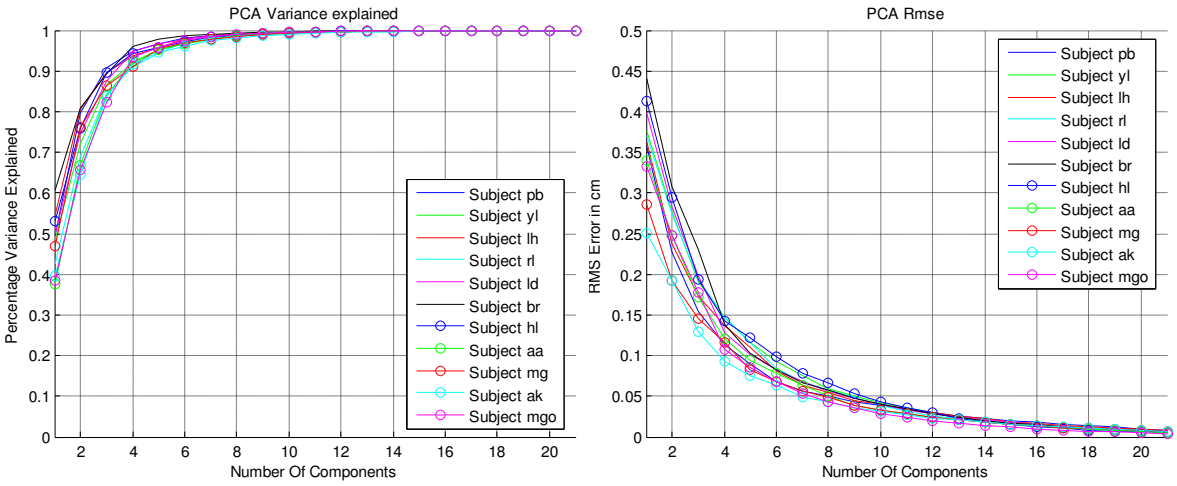


Figure 4-2 – Performance, established using LOOCV, of the PCA individual models as a function of number of components for the upper tongue contours of male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO for a corpus including vowels and consonants. Left: variance explained. Right: RMSE in centimeters

4.5.3.2. Multilinear tongue models

Figure 4-3 shows the performance of all the multilinear methods, as a function of the number of components, in terms of variance explanation and RMSE for UpperTng. As explained in section 1.2.3, the number of components for TUCKER can be fixed independently for each mode of variation (observations (Cmp_1), articulator measurements (Cmp_2) and speakers (Cmp_3)). In this study, Cmp_3 is fixed to be equal to the number of speakers and $Cmp_1 = Cmp_2 = Cmp$, being Cmp the number of components extracted for the model. This was done in order to simplify the comparison between different methods for a given number of components.

The variance explanation curve of TUCKER shows a very similar performance compared to joint PCA. As explained in section 1.2 and section 4.5.2, TUCKER is a method with a more complex structure and more parameters than joint PCA. Therefore, joint PCA was kept but TUCKER was not used anymore in the following sections. The results of Figure 4-3 are further analyzed in section 4.5.3.3.

We also compared the models with consonants to the models with only vowels, explained in section 4.5.1. Note that there is a decreasing of variance explanation and increasing of RMSE for the models extended to consonants. Overall with 4 components, The Average PCA model explains 98.16% for only vowels and 93.23% for vowels and consonants, accounting for a RMSE of 0.06 cm and 0.1 cm respectively. Joint PCA explains 91.66% for only vowels and 72.16% for vowels and consonants, accounting for a RMSE of 0.15 cm and 0.26 cm respectively. PARAFAC explains 79.42% for only vowels and 65.15% for vowels and consonants, accounting for a RMSE of 0.24 cm and 0.30 cm respectively. Finally, TUCKER explains 88.82% for only vowels and 70.17% for vowels and consonants, accounting for a RMSE of 0.18 cm and 0.27 cm respectively.

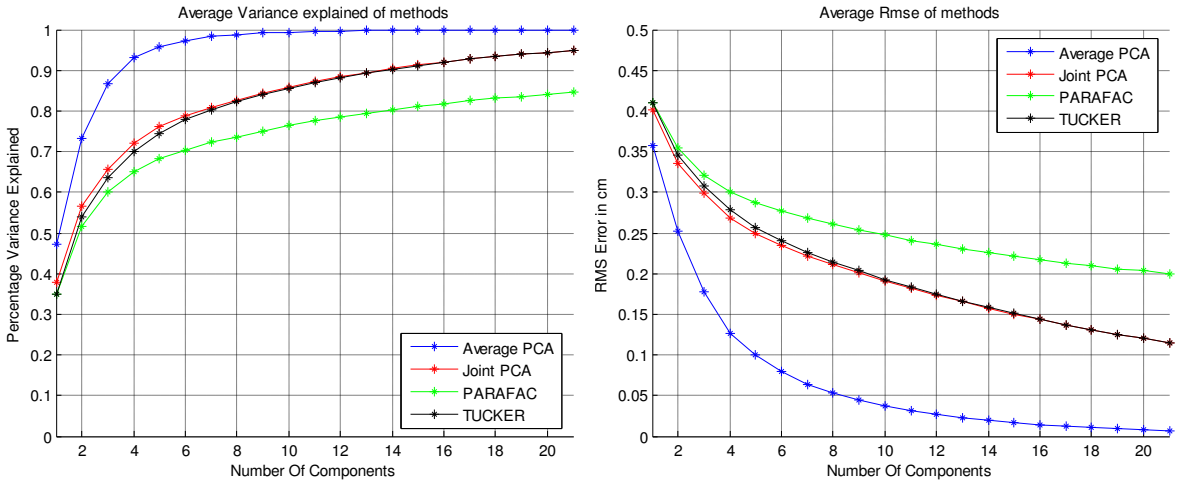


Figure 4-3 - Performance, established using LOOCV, of the PARAFAC, TUCKER and joint PCA as a function of number of components for the tongue contours for a corpus including vowels and consonants. Left: variance explained. Right: RMSE in centimetres

4.5.3.3. Models with different representations of the upper tongue contour

In this section, the multilinear tongue models explained in section 4.5.3.2 were built again for the two representations of tongue contour INT and INTRXY, based on the grid system explained in section 2.7.2. Figure 4-4 shows the performance of all methods in terms of variance explanation and RMSE.

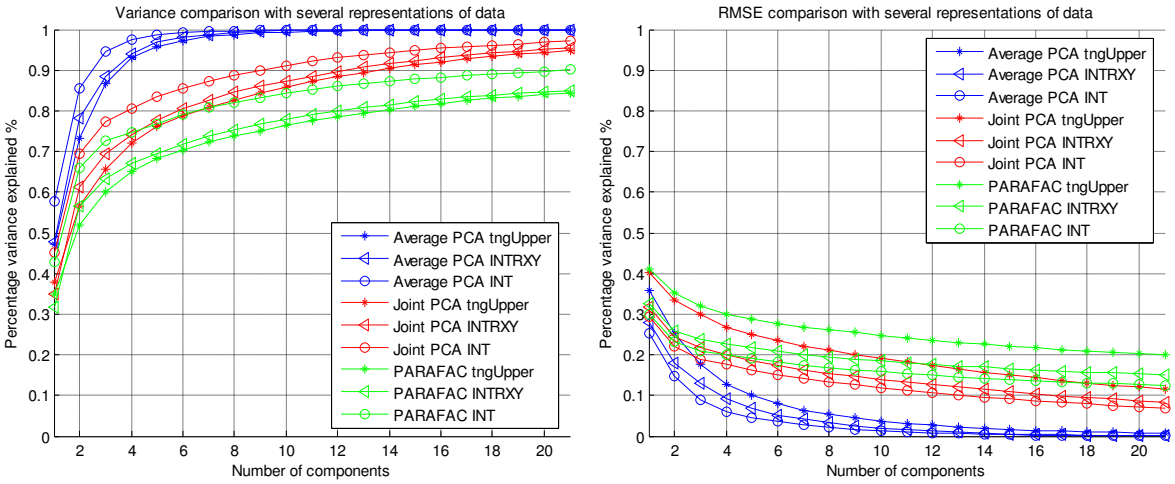


Figure 4-4- Performance, established using LOOCV, of the multilinear decomposition methods with several representations of data as a function of number of components for a corpus including vowels and consonants. Left: variance explained. Right: RMSE in centimetres.

PCA models have been used as baseline models to assess the performance of different multilinear decomposition methods. A Student's t-test at 5% significance level was used to determine the number of components for each method that gives an RMSE not statistically different from the one obtained by the reference individual PCA models. For PCA with UpperTng, four components were chosen as reference model to compute the t-test. On the other hand, for PCA with INTRXY and INT, three components were sufficient to explain about the same variance as the PCA with UpperTng. Thus, the PCA model, with 3 components, was chosen as reference model for INTRXY and INT to compute the t-test.

According to the Student's t-test for the models with UpperTng, PARAFAC needs more than 21 components to reach a variance explanation of 84.52 %. On the other hand, joint PCA needs between 14 and 21 components depending on the speaker. Figure 4-5 shows the statistics of numbers of components needed by joint PCA, for each speaker, to equal the performance of PCA according to a Student's t-test at 5% significance level. These results are also summarized and compared in Table 4-3. For instance, the performance of joint PCA with UpperTng is not statistically different to PCA between the 14th and the 21th component, accounting for a variance explanation between 90.33% and 94.88%. Joint PCA with UpperTng requires the minimum number of components for speaker YL (at least 14 components) and the maximum for speaker AK (21 components).

According to the Student's t-test for the models with INTRXY, PARAFAC needs more than 21 components accounting for a variance explanation of 88.35%. On the other hand, joint PCA needs between 12 components, accounting for a variance explanation of 89.63%, and 21 components, accounting for a variance explanation of 95.46%.

Joint PCA requires the minimum number of components for speaker LH (at least 12 components) and the maximum for speakers PB, AK and MGO (21 components).
According to the Student's t-test for the models with INT, PARAFAC needs more than 21 components accounting for a variance explanation of 90.07%. On the other hand, joint PCA needs between 12 components, accounting for a variance explanation of 93.03%, and 21 components, accounting for a variance explanation of 97.12%. Joint PCA requires the minimum number of components for speaker MG (at least 12 components) and the maximum for speaker AK (21 components). The Table 4-4 summarizes all the results explained above.
We could conclude that the re-sampling of the tongue contour in INTRXY and INT do not constitute any advantage for the modelling. By re-sampling the tongue contour we gain little extra variance explanation and we lose information. Thus, taking into account the conclusions of section 4.5.2 about the number of parameters used by each multilinear method, joint PCA with UpperTng appears to be the optimal solution.

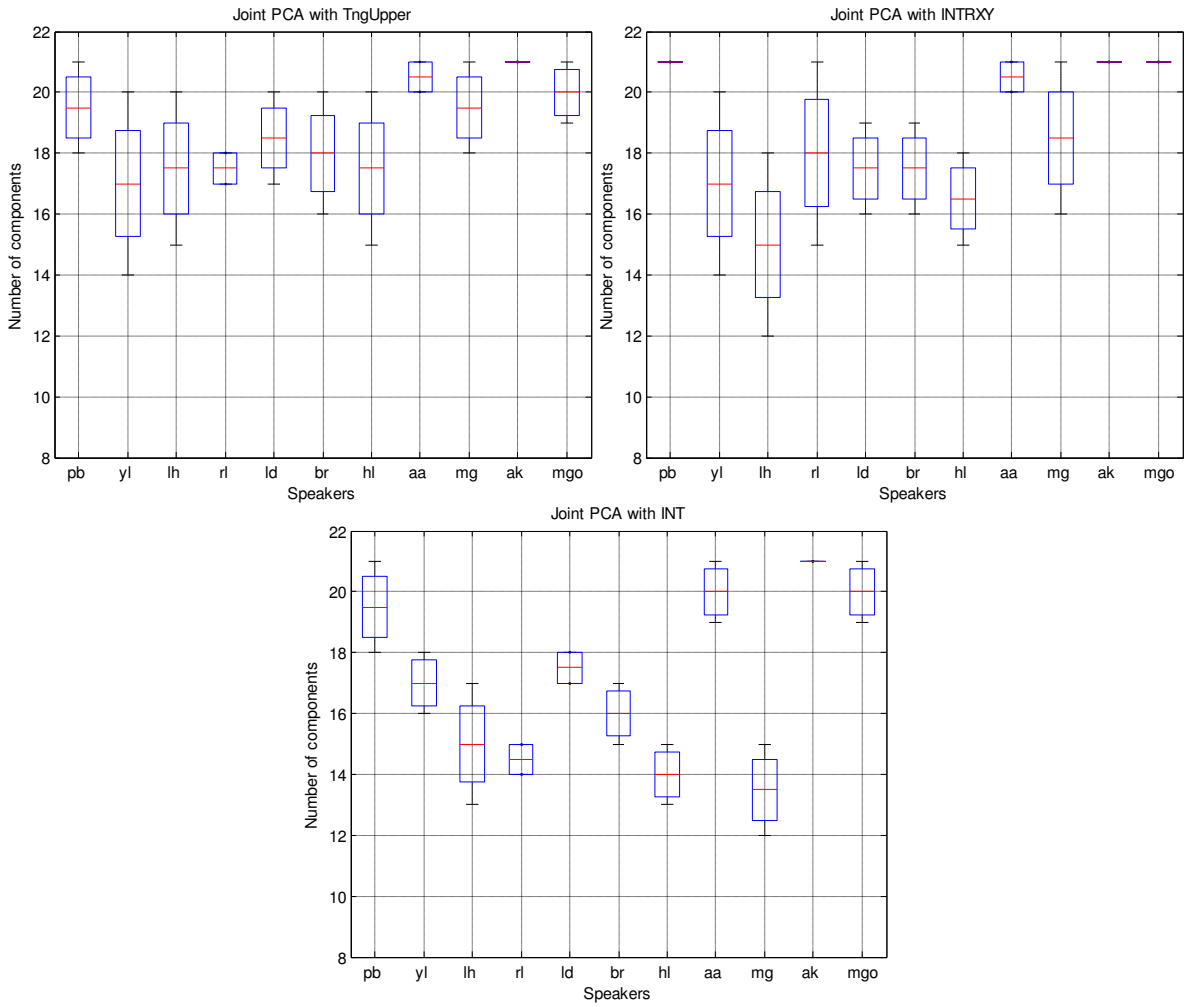


Figure 4-5 – Statistics of number of components needed for tongue models, with each representation of data (TngUpper, INTRXY, INT), according to a Student's t-test between the reference PCA and the multi-linear method joint PCA

Speaker	UpperTng	INTRXY	INT
PB	18	**21**	18
YL	**14**	14	16
LH	15	**12**	13
RL	17	15	14
LD	17	16	17
BR	16	16	15
HL	15	15	13
AA	20	20	19
MG	18	16	**12**
AK	**21**	**21**	**21**
MGO	19	**21**	19

Table 4-3 – Minimum number of components, for each speaker, needed for Joint PCA to reach the performance *of the reference PCA models according to a Student's t*-test at 5% significant level for each representation of data. The minimum and maximum numbers of components needed for each representation of data is highlighted in red

Representation of data	Average PCA		PARAFAC		Joint PCA	
	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.
UpperTng	4	93.23%	21	84.52%	14 - 21	90.33% - 94.88%
INTRXY	3	88.35%	21	85.07%	12 - 21	89.63% - 95.46%
INT	3	94.52%	21	90.07%	12 - 21	93.03% - 97.12%

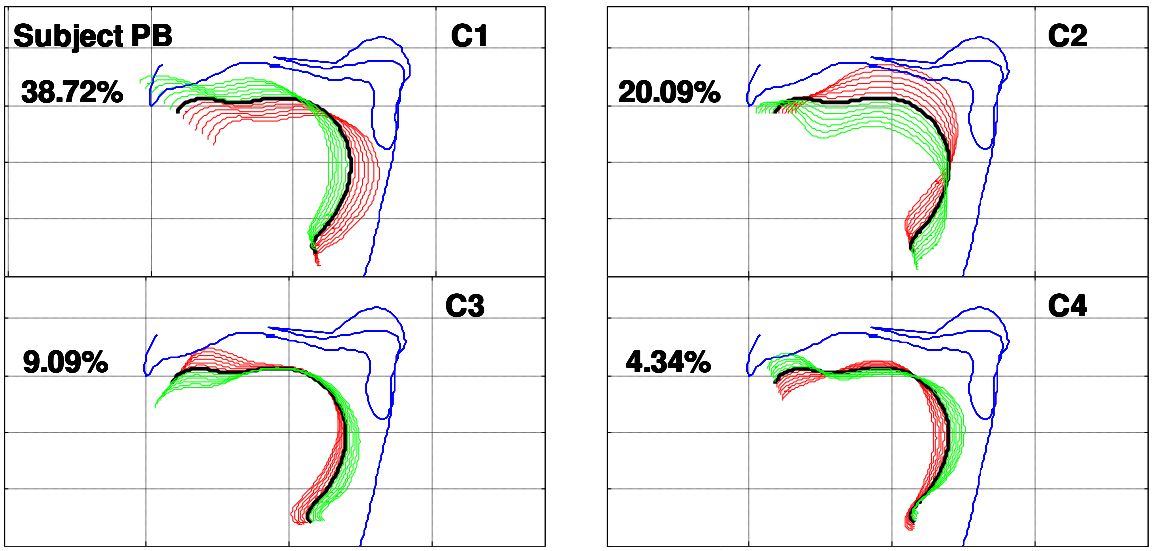
Table 4-4 – Summary of the number of components needed for the multilinear methods (PARAFAC and joint PCA) to reach the same performance of the reference PCA, according to a Student's t-test for each representation of data

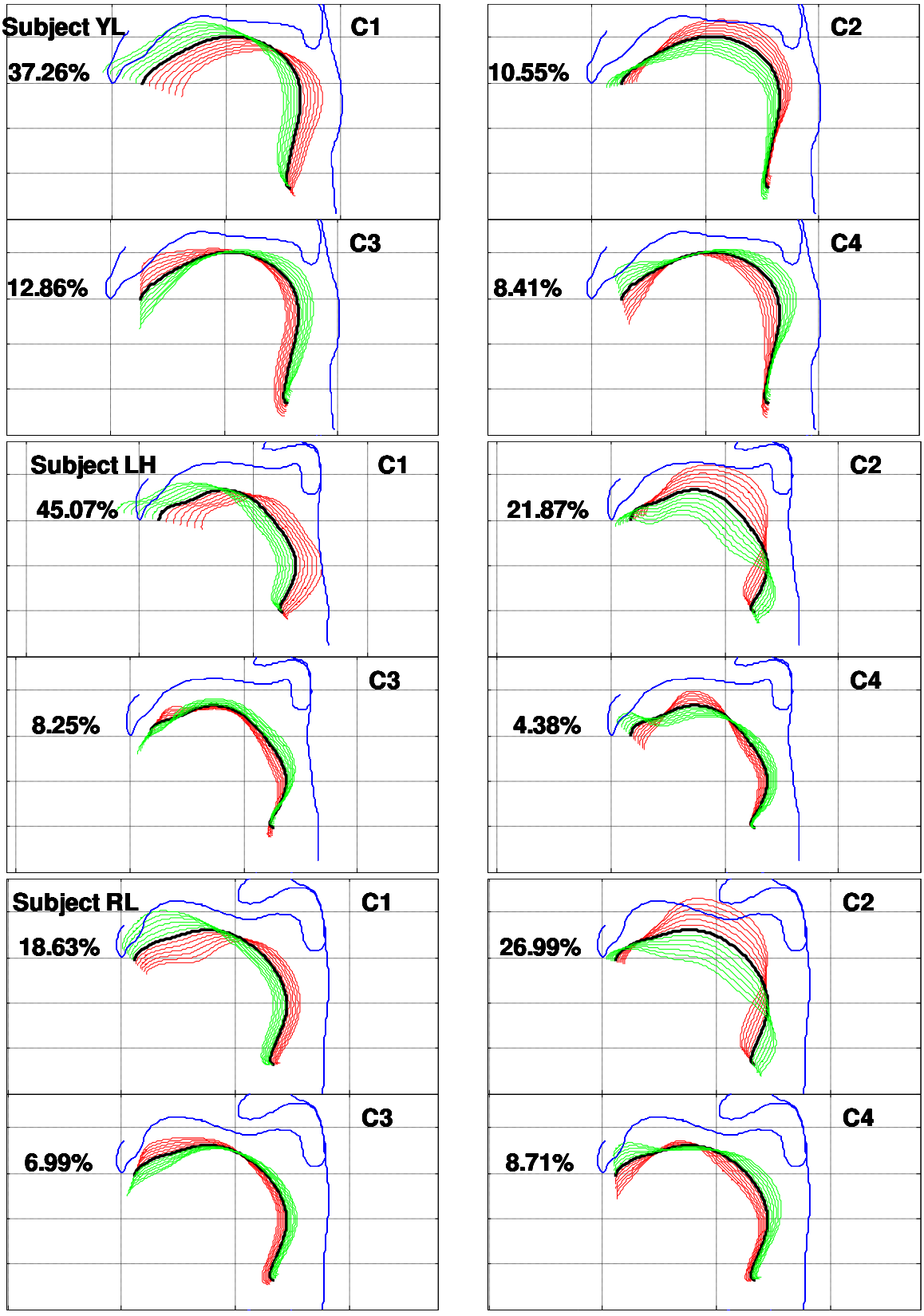
In Table 4-4 we have seen that the number of components needed by joint PCA is much higher than the number of components needed by the individual PCA models. We have thus analysed the meaning of the Joint PCA components by computing the correlations between the 4 guided PCA components, described in Chapter 3, and the 21 joint PCA components common to all speakers. The correlations for the first 6 components of joint PCA are displayed in Table 4-5. Correlations below 0.4 were not taken into account. The correlations from the 6th component on were also below 0.4 and thus not included in Table 4-5. The strongest correlations (yellow boxes in Table 4-5) indicate that the first 4 joint PCA components can be approximately interpreted in terms of jaw height (JH), tongue body (TB), tongue dorsum (TD) and tongue tip (TT). For instance, components 1 and 2 can be interpreted in terms of TB and TD for speaker PB. Similarly, the components from 1 to 3 are related to TB, TD and TT for speaker LD. One can also note a number of other correlations, though they are weaker, between 0.4 and 0.6 (green boxes in Table 4-5). Figure 4-6 illustrates the associated nomograms for the first 4 joint PCA components, for all speakers.

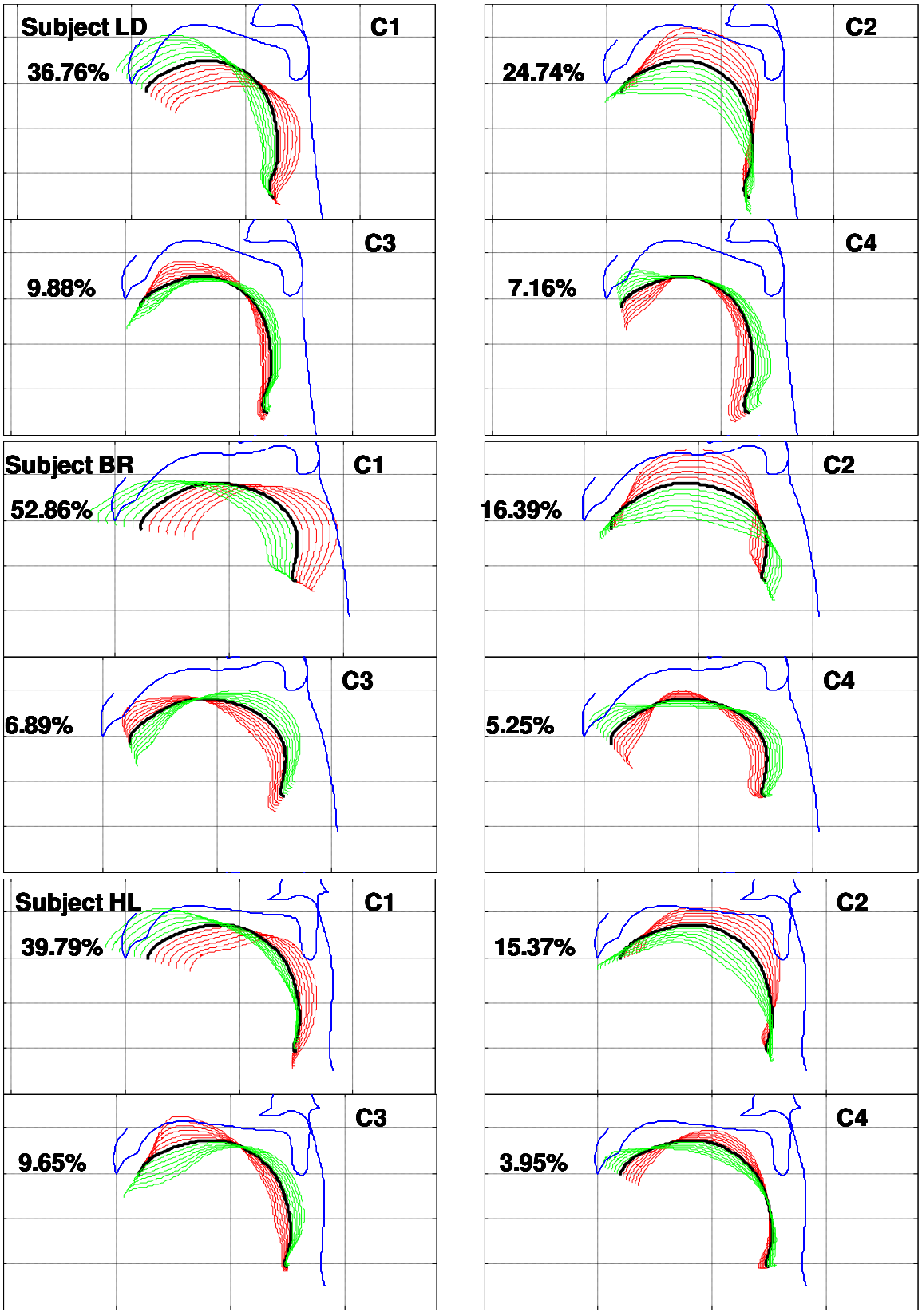
		1	2	3	4	5	6
pb	JH			-0,48			
	TB	**-0,78**					
	TD		**0,84**				
	TT						
yl	JH	0,53					
	TB		-0,54		0,57		
	TD	0,57		0,51			
	TT			0,49			
lh	JH	0,41					
	TB	**-0,82**					
	TD		**0,84**				
	TT			0,49		-0,40	
rl	JH					-0,41	
	TB		**0,75**				
	TD	0,50				0,42	
	TT			0,55	**-0,61**		
ld	JH	0,55			0,42		
	TB	**-0,62**					
	TD		**0,82**				
	TT			**0,71**			
br	JH	0,43			0,43		
	TB	**-0,82**					
	TD		**0,88**				
	TT			0,45	-0,50		

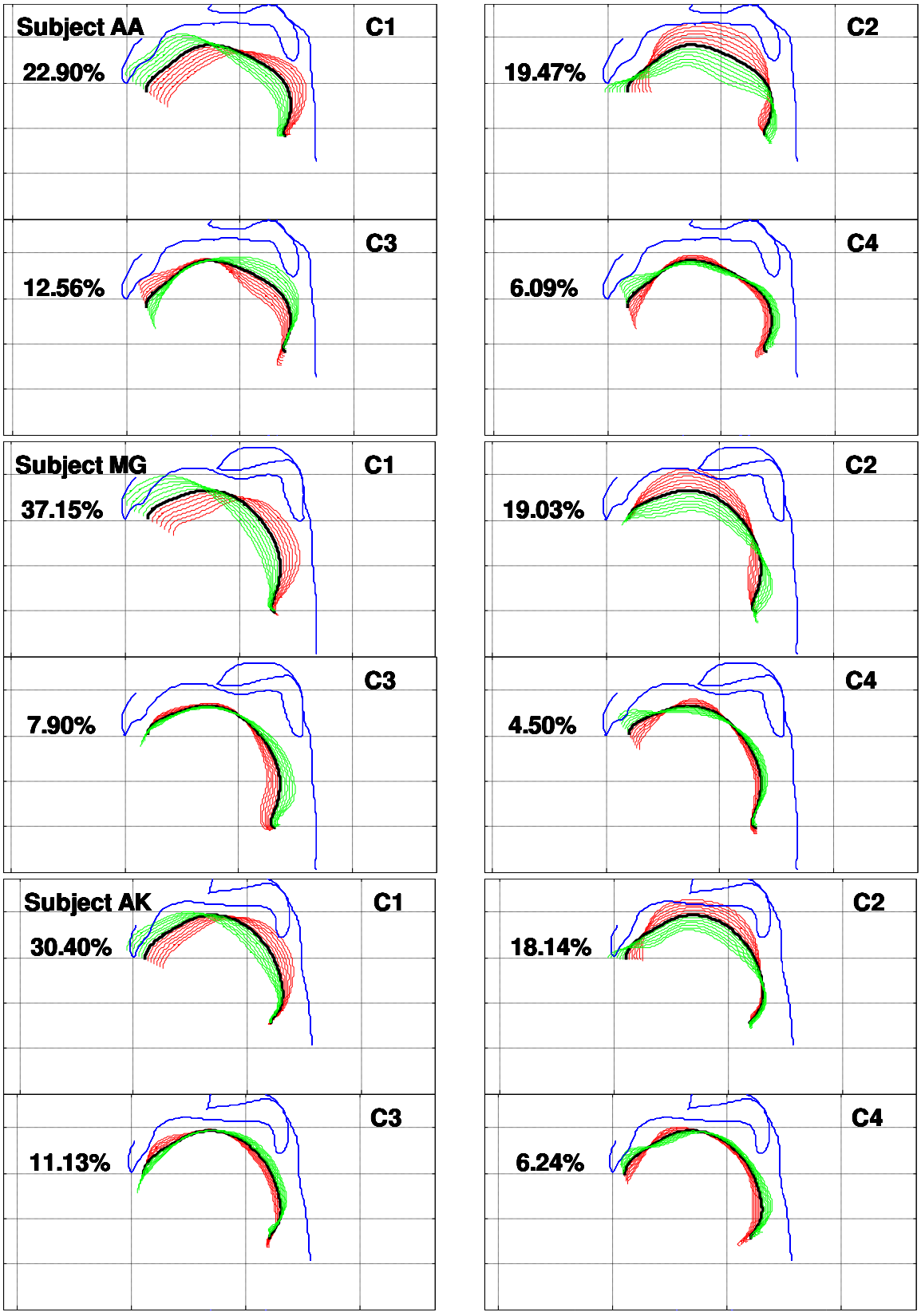
		1	2	3	4	5	6
hl	JH	0.52					
	TB	**-0.66**					
	TD		**0.73**				
	TT			**0.76**			
aa	JH			0.45	-0.50		
	TB						
	TD	0.52					
	TT	**-0.66**					
mg	JH			**0.76**			
	TB						
	TD					-0.42	
	TT	-0.42	-0.54	0.42			
ak	JH			0.48	-0.45		
	TB						
	TD						
	TT	**-0.68**	0.47				
mgo	JH				**-0.68**		
	TB						
	TD						-0.40
	TT	**-0.82**					

Table 4-5 - Correlations between the first 6 components of joint PCA and the 4 components of guided PCA (JH, TB, TD, and TT) for the tongue models. Only correlations higher or equal to 0.4 are shown. Correlations between 0.4 and 0.6 (green boxes) and correlations higher than 0.6 (yellow boxes). Rows: guided PCA components. Columns: joint PCA components




Subject YL
37.26%
C1
10.55%
C2
12.86%
C3
8.41%
C4
Subject LH
45.07%
C1
21.87%
C2
8.25%
C3
4.38%
C4
Subject RL
18.63%
C1
26.99%
C2
6.99%
C3
8.71%
C4


Subject LD
C1
36.76%
C2
24.74%
C3
9.88%
C4
7.16%
Subject BR
C1
52.86%
C2
16.39%
C3
6.89%
C4
5.25%
Subject HL
C1
39.79%
C2
15.37%
C3
9.65%
C4
3.95%


Subject AA
C1
22.90%
C2
19.47%
C3
12.56%
C4
6.09%
Subject MG
C1
37.15%
C2
19.03%
C3
7.90%
C4
4.50%
Subject AK
C1
30.40%
C2
18.14%
C3
11.13%
C4
6.24%

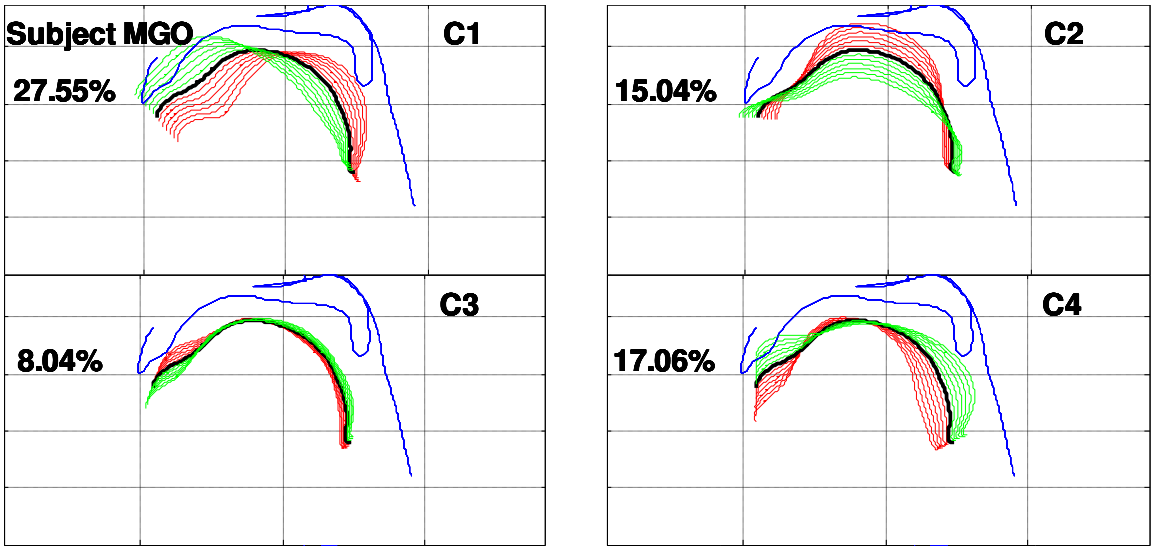


Figure 4-6 – Nomograms of the four upper tongue contour components determined by joint PCA for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO. Each predictor (C1, C2, C3 and C4) is varied from -3 to +3 with a 0.5 step. The reference wall (palate, velum and pharynx) is shown in blue. The relative data variance explained by each component is displayed

4.6. Multilinear regression between control parameters of couple of speakers

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 (TS) from those of a source speaker (SS). The goal is to predict PCA control parameters of a target speaker from PCA control parameters of a source speaker. Formally, a Multilinear Regression (MLR) model, with Cmp number of components, is expressed by:

$\prod_{TS} = \sum_{cmp=1}^{n} \beta_i \pi_{SSi} + \gamma$, for i = 1, 2, ..,Cmp. Where Υ is the residual error, β represents the coefficients of the linear regression; π_{TS} and π_{SSi} are respectively related to the set of control parameters predicted for speaker TS and the control parameters of speaker SS that is used as predictor.

We have built MLR models between each possible combination of couple of speakers. Figure 4-7 shows the evaluation for speaker PB. It appears that, the model gave strong signs of being over-fitted from the tenth component on. For instance, the cross-validated mean-square error decreases up to the tenth component, but increases from the tenth component. Indeed, increasing the number of predictors from the 10^{th} component might lead to an over-fitted or degenerated model. So, the meaningless components were discarded. Finally, with the first ten components, the MLR model was able to predict the tongue contour of speaker TS from speaker SS, on average

over all the SS speakers, accounting for about 64.32% of the variance and with a RMSE of 0.37 cm.

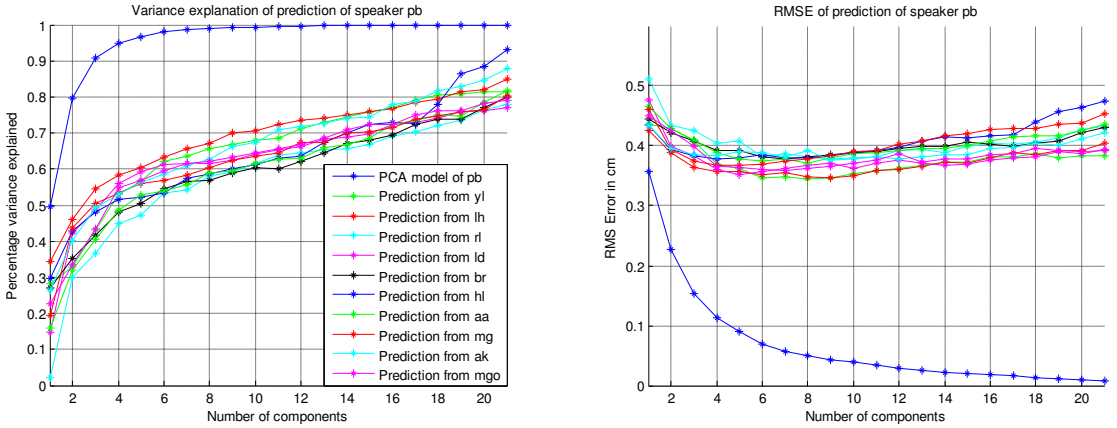


Figure 4-7 - Variance explained and RMSE , established using LOOCV, of the MLR models between control parameters of PB and the rest of speakers as a function of number of components.

4.7. Missing data PCA to model the tongue contour including the sublingual cavity

The models described in previous sections are built only for the upper tongue contour which contains the (x, y) coordinates from the tongue tip to the tongue root. However, the sublingual cavity which is not always visible and thus not possible to be traced, as explained in Chapter 2, was not included. In this section, a probabilistic PCA method, proposed by Verbeek (2009) that accepts missing data, is used to model the full tongue contour including the sublingual cavity. The sublingual cavity was thus included for the articulations in which it was visible and it was predicted by means of an expectation maximization (EM) algorithm for the articulations in which it was not visible. The expectation maximization algorithm is an iterative method that uses the model learnt from the known data to predict the missing data. The models of this section were built using the full tongue contour (FullTng). That is, the contour from the jawAttach to the base of the epiglottis.

The **Table 4-6** shows, for each speaker, the number of articulations with missing sublingual cavity data. Speaker HL has no missing data related to the sublingual cavity. Overall, the percentage of articulations with missing data was ~47% over a corpus of 63 articulations and 11 speakers. In order to build individual PCA models for the full tongue contour, the missing sublingual cavities were predicted by means of EM first. Then individual PCA models were built. Figure 4-8 shows the performance of PCA in terms of variance explanation and RMSE. As in section 4.5.3.1, four components were used to compare the performance of the individual models. Even

though the models were built for the full tongue contour, the variance explanation and RMSE were computed using only the upper tongue contour. The variance explained by individual speaker models ranges, for the set of speakers, between 89% and 94.5% and the RMSE between 0.10 cm and 0.16 cm. In average, over our eleven speakers, the individual speaker models explain an amount of 92.04% of the data variance, with an RMSE of 0.14 cm, using 4 components. Note that the performance of the missing data PCA models was a bit lower than the simple PCA models described in section 4.5.3.1.

	Missing sublingual cavities	**percentage of missing data (%)**
pb	32	50,79
yl	39	61,90
lh	27	42,86
rl	41	65,08
ld	43	68,25
br	16	25,40
hl	0	0
aa	8	12,70
mg	34	53,97
ak	37	58,73
mgo	49	77,78
MEAN	~30	~47

Table 4-6 – Number and percentage of articulations with missing data for each speaker

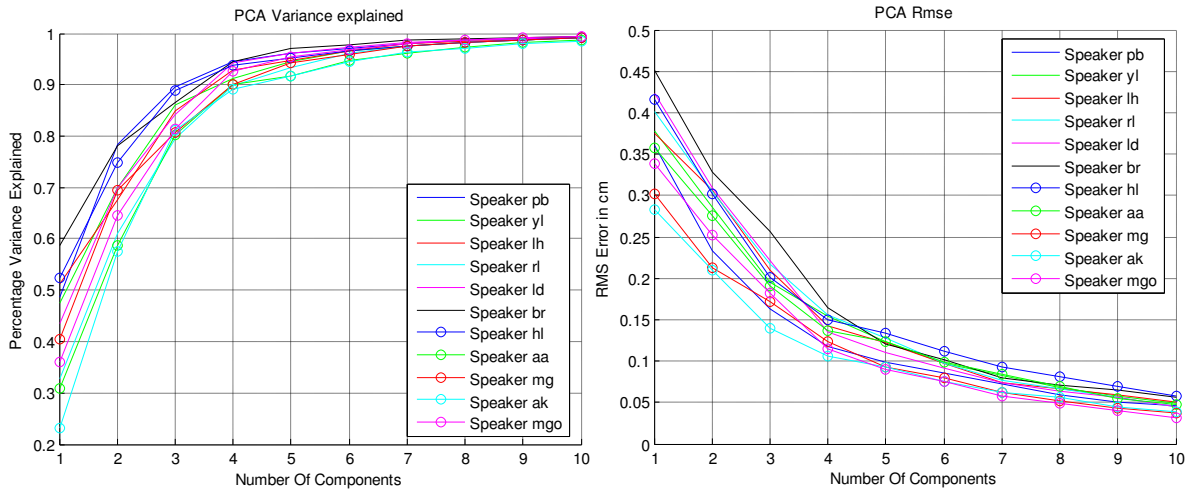


Figure 4-8 - Performance of the missing data PCA method as a function of number of components for the full tongue contour for male speakers PB, YL, LH, RL, LD, BR and female speakers HL, AA, MG, AK, MGO for a corpus including vowels and consonants. Left: variance explained. Right: RMSE in centimetres.

4.8. Non linear methods

The models explained in previous sections are built by means of linear methods. In this section, two non linear approaches are used to match a given (x, y) point in the articulatory space of a source speaker (SS) to its corresponding point in the articulatory space of a target speaker (TS). The following sections explain the methods and results of two techniques: neighbourhood average and center of gravity. The articulatory spaces are represented by the PCA components of the models described in sections 4.5.1 and 4.5.3.1. The studies in this section were made only for vowels. In order to match points in SS into TS, the articulatory space TB vs. JH was used because it corresponds roughly to the description of the vocalic space given in Chapter 2.

4.8.1. Neighbourhood Averaging between PCA control parameters

The goal of the neighbourhood average technique is to find the projection J of a given vowel I in a source speaker's space (SS) into a target speaker's space (TS) (Fontecave & Berthommier, 2009). The vowel I, contained in SS, is represented as a weighted sum of other vowels, called neighbours. The weight that a given neighbour (N), in SS, applies to I is computed as the inverse of the distance (see equation 1).Then, we assumed that the projection of I in TS corresponds to the weighted sum of its neighbours (Q) in TS. The projection is given by the equation 2.

$$W(I, N_n) = \frac{1}{d(I, N_n)} \qquad [1]$$

$$J = \frac{\sum_{n=1}^{K} \frac{Q_n}{d(I, N_n)}}{W(I, N_n)} \qquad [2]$$

N_n corresponds to the K neighbours in the source space (SS), Q_n is related to the K neighbours in the target space (TS), and d is the Euclidean distance between the point I and its K neighbours in SS. The Euclidean distance between two points (p and q) is computed as follows:

$$d(q, p) = \sum_{i=1}^{n} (q_i - p_i)^2$$

Figure 4-9 shows the projection of each vowel of SS into TS using different number of neighbours, for the source space of PB and the target space of YL. In order to project a given vowel of SS using a given number of neighbours, the vowel was first excluded from both spaces. Then, the equations explained above were used to project the given vowel into TS. This process was repeated for each vowel. The error of prediction

ranged between 0.86 and 1 using from 2 to 9 neighbours to predict each vowel. We have observed that increasing the number of neighbours used for predicting a given vowel does not necessarily improve the prediction. Note that the more neighbours were used, the more the prediction tended to centralise. This is also due to the use of only positive weights. These conclusions were also valid for all the possible combinations of source and target spaces using all the speakers. The following section introduces a method based on the same principle of neighbourhood average, but using negative and positive weights.

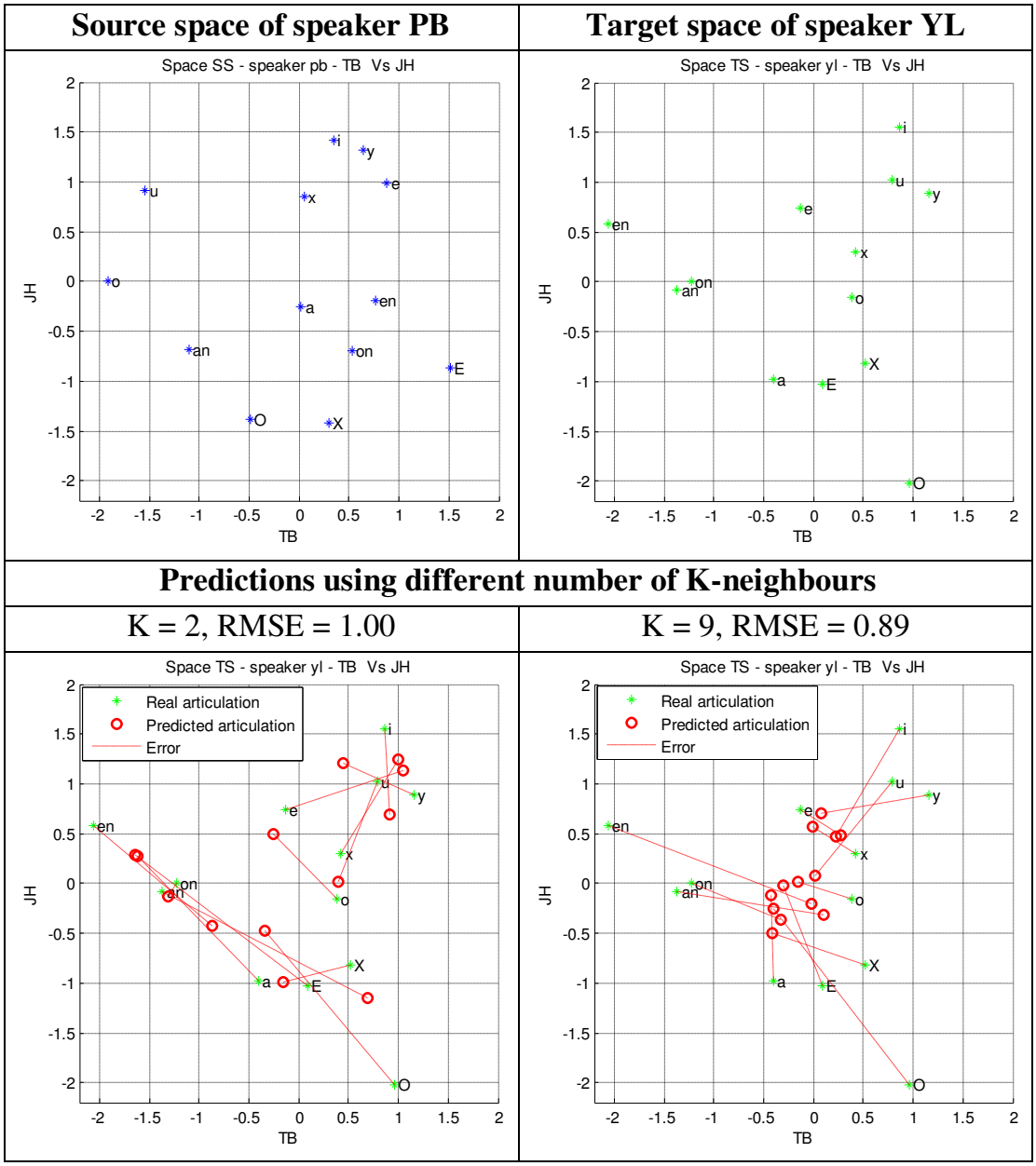


Figure 4-9 – Projection of a each vowel in the SS of speaker PB into the TS of speaker YL using different number of neighbours with the technique of K neighbourhood

4.8.2. Center of gravity

In the K neighbourhood technique, explained in previous section, the weights are always positive. Thus, the negative contribution that a neighbour may exert to a given point I is not taken into account. In this section, a local linear method that allows negative weights is used. The method is expressed by the following equations:

$$I_{SS} = W_{ss} N_{ss}$$

W represents the contribution (weights) of the neighbours N to the point I in the source space. This local lineal system can be decomposed as:

$$I_x = w_1 n_{1x} + w_2 n_{2x} + ... w_k n_{kx}$$
$$I_y = w_1 n_{1y} + w_2 n_{2y} + ... w_k n_{ky}$$
$$1 = w_1 + w_2 + ... w_k$$

Thus, the weights are obtained by inversing the equations above as:

$$W_{ss} = N_{SS}^{-1} I_{SS}$$

The projection J in the target space is finally computed as:

$$J = W_{ss} N_{Ts}$$

Figure 4-10 shows the projection of each vowel of SS into TS using different numbers of neighbours, for speaker PB as source space and speaker YL as target space. Each vowel was predicted using the same iterative procedure as the neighbourhood average. The error of prediction ranged between 1.03 and 1.86 using from 2 to 9 neighbours to predict each vowel. The source and target spaces of PB and YL seemed not to be homogenous enough to make good predictions using this method. These conclusions were also valid for all the possible combinations of source and target spaces using all the speakers.

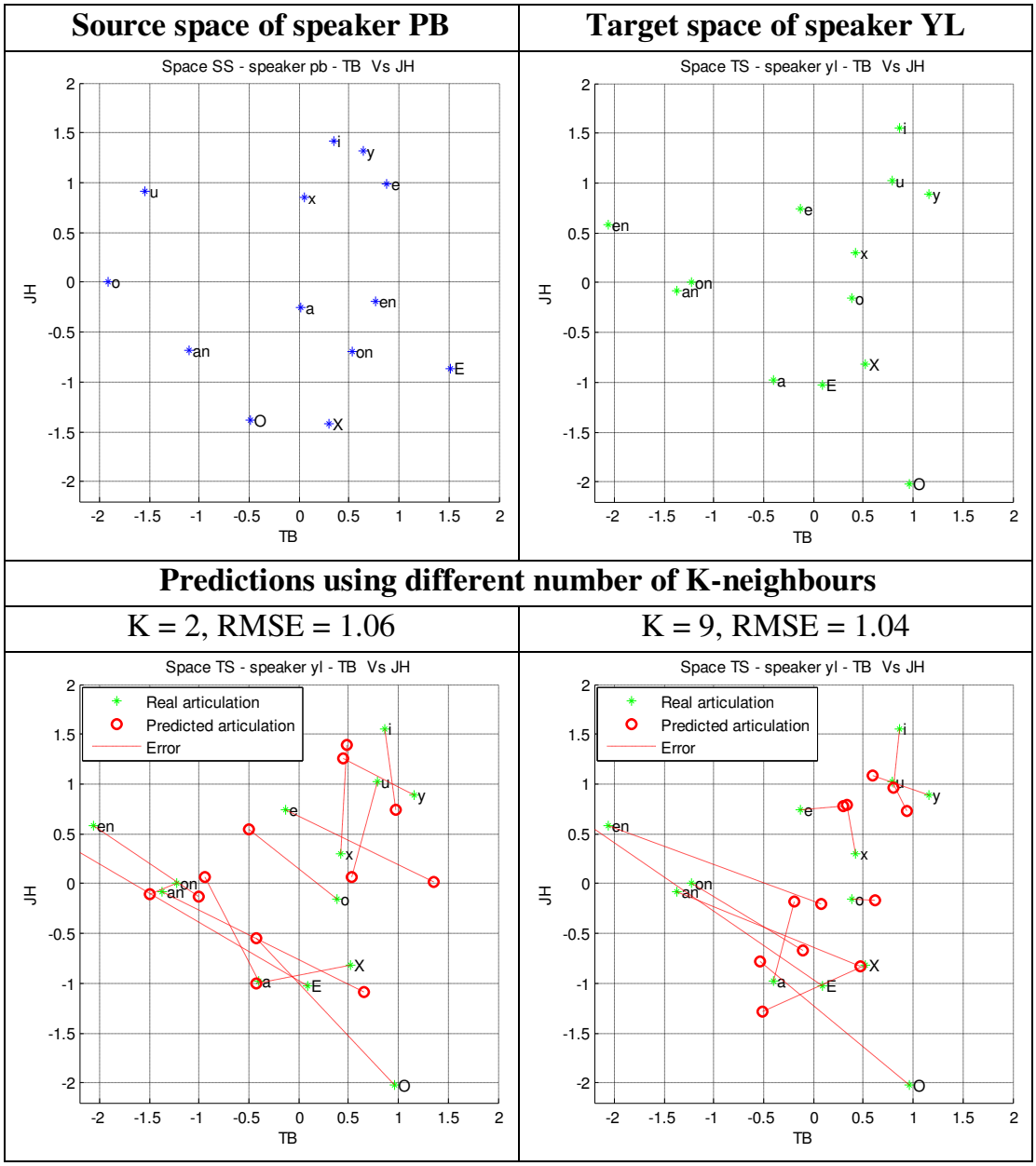


Figure 4-10 – Projection of a each vowel in the SS of speaker PB into the TS of speaker YL using different number of neighbours with the technique of center of gravity

4.9. Lip models

Figure 4-11 shows the performance of all the multilinear methods, described in Chapter 1, in terms of variance explanation and RMSE for the lip contours. As shown in Chapter 3, three components were extracted by means of a guided PCA applied to the lip contours. These three components represented the influence of the jaw on the lips, the protrusion and the lip height. The models described in the present section are not guided. However, three components were also taken as a reference of comparison. The variance explanation obtained for all speakers with individual PCA models with three components reached 94.9% and 94.5% for the upper and lower lip, respectively. The associated RMSE was 0.03 cm and 0.05 cm, respectively. We wanted now to

extract a set of articulatory components, common to all speakers. So the lips were also modelled by means of multilinear methods like: PARAFAC, TUCKER and joint PCA.

A Student's t-test at 5% significance level was used to determine the number of components that gives an RMSE not statistically different from the one obtained by the reference individual PCA with three components, for each multilinear method. The Table 4-7 summarizes the results of the Student's t-test. We first observed that the variance explanation curve of TUCKER showed a very similar performance compared to joint PCA. As explained in sections 1.2 and 4.5.2, TUCKER is a method with a more complex structure and more coefficients compared to joint PCA. Therefore, it was decided to keep joint PCA and not to use TUCKER for the Student's t-test. PARAFAC needs more than 21 components to be equivalent to the performance of PCA. PARAFAC models built with 21 components accounted for a variance explanation of 87.02% and 90.11% for the upper and lower lip respectively. Joint PCA needed between 15 and 21 components for the upper lip and between 11 and 21 components for the lower lip, depending on the speaker. Figure 4-12 shows the range of components needed by joint PCA, for each speaker, to equal the performance of PCA according to the Student's t-test. As regards the upper lip modelling, Joint PCA requires the minimum number of components for speakers PB and RL (at least 15 components) and the maximum for speaker MG (21 components), accounting for a variance explained between 93.4% and 96.7%. On the other hand, as regards the lower lip modelling, joint PCA needs the minimum number of components for speaker AA (at least 11 components) and the maximum for speaker MG (21 components), accounting for a variance explained between 91.2% - 96.9%.

According to the Student's t-test, joint PCA appears to be the optimal solution to model the upper and lower lips. It needed fewer components to normalise and model the lips contours compared to PARAFAC.

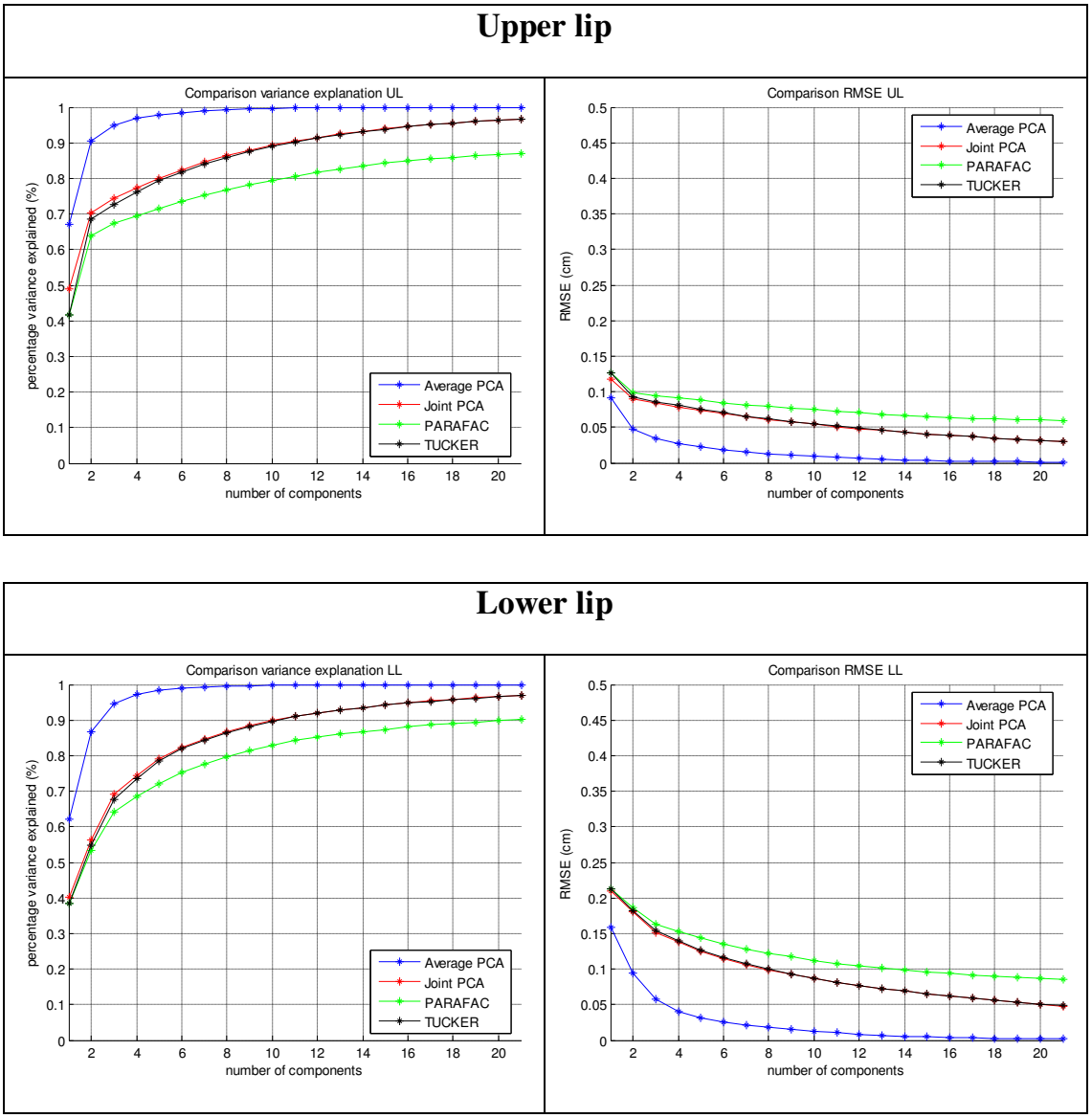


Figure 4-11 - Performance, established using LOOCV, of the average individual PCA, PARAFAC, TUCKER and joint PCA methods as a function of number of components for the lips contours for a corpus including vowels and consonants. Top: variance and RMSE of upper lip contour. Bottom: variance and RMSE of lower lip contour.

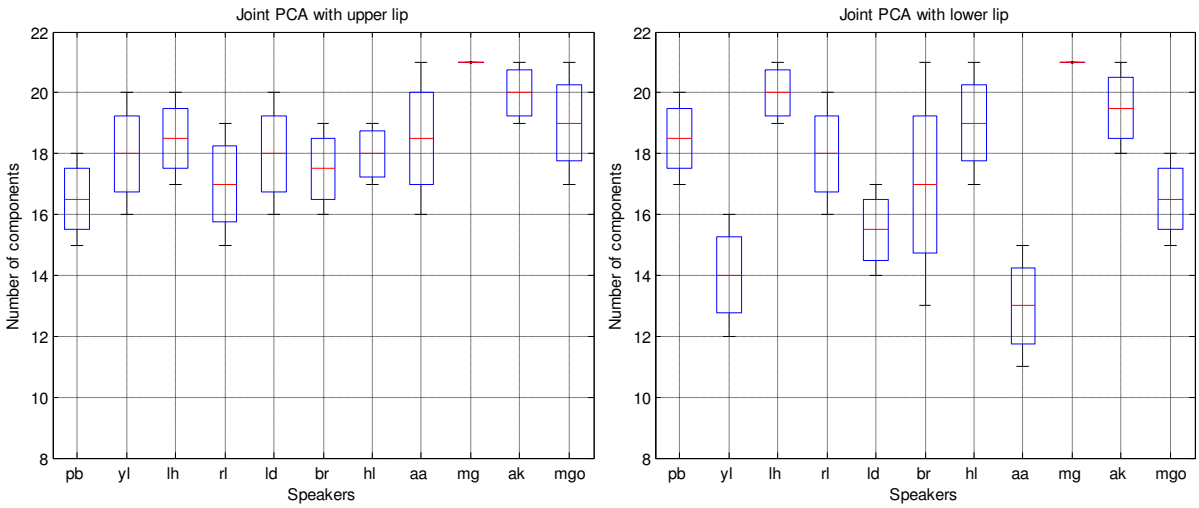


Figure 4-12 – Range of number of components needed for upper and lower lip models according to a Student's t-test between the reference PCA, with 3 components, and joint PCA. Left: number of components needed for upper lip model. Right: number of components needed for lower lip model

Representation of data	Average PCA		PARAFAC		Joint PCA	
	Ref. cmp	Var. Exp.	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.
Upper lip	3	94.9%	21	87.02%	15 - 21	93.4% - 96.7%
Lower lip	3	94.5%	21	90.11%	11 - 21	91.2% - 96.9%

Table 4-7 – Results of Student's t-test between reference PCA, with 3 components, and the multilinear methods (PARAFAC and joint PCA), for the upper and lower lip models

We have also analysed the meaning of the Joint PCA components, extracted from the lips, by computing the correlations between the 3 guided PCA components, described in Chapter 3, and the 21 joint PCA components common to all speakers (see Table 4-8).

The correlations from the 5th component on were below 0.4 and thus not included in Table 4-8. For the upper lip, the strongest correlations (yellow boxes) indicate that the first joint PCA components can be approximately interpreted in terms of jaw height (JH) and lip protrusion (ULP). However, in most of the cases the lip height component (ULH) has either weaker (green boxes) or zero correlations with the components extracted by joint PCA. For the lower lip, the strongest correlations indicate that the first components extracted by joint PCA can be approximately interpreted in terms of jaw height (JH), lip protrusion (LLP) and lip height (LLH). However, for speakers AA, MG, AK and MGO the component LLP has no correlation with the Joint PCA components.

Upper lip

		1	2	3	4	5
pb	JH	**0.77**	**0.58**			
	ULP	-0.51	**0.61**			0.41
	ULH					
yl	JH	**-0.84**				
	ULP		**0.67**			-0.43
	ULH					
lh	JH	**0.83**				
	ULP					
	ULH					
rl	JH	-0.51	0.58	0.44		
	ULP	**0.73**	0.43			
	ULH				0.41	
ld	JH	**0.92**				
	ULP		**0.86**			
	ULH					
br	JH	**0.91**				
	ULP		**0.83**			
	ULH					

		1	2	3	4	5
hl	JH	**0.71**			0.53	
	ULP		0.50			
	ULH		0.42			
aa	JH	**0.94**				
	ULP					
	ULH	**0.71**			0.53	
mg	JH	**0.84**				
	ULP					
	ULH	**0.94**				
ak	JH		0.59			
	ULP					
	ULH	**0.84**				
mgo	JH	**0.85**				
	ULP					
	ULH		0.59			

Lower lip

		1	2	3	4	5
pb	JH	**0.83**				
	LLP		**-0.79**			
	LLH			**-0.81**		
yl	JH	**0.81**				
	LLP		**-0.70**			0.58
	LLH			-0.53		
lh	JH	**0.65**				
	LLP		**-0.66**			
	LLH					
rl	JH	**0.75**		0.48		
	LLP		**-0.74**			
	LLH			-0.44		
ld	JH	**0.62**		-0.58		
	LLP		**0.78**			
	LLH	0.47	0.46	0.58		
br	JH	**0.87**				
	LLP		**-0.85**			
	LLH			**0.77**		

		1	2	3	4	5
hl	JH	**0.73**				
	LLP		0.50			
	LLH			**0.64**		
aa	JH	**0.89**				
	LLP					
	LLH	**0.73**				
mg	JH	**0.72**				
	LLP					
	LLH	**0.89**				
ak	JH	**0.67**			-0.47	
	LLP					
	LLH	**0.72**				
mgo	JH	**0.68**		-0.44	0.51	
	LLP					
	LLH	**0.67**			-0.47	

Table 4-8 - Correlations between the first 5 components of joint PCA and the 3 components of guided PCA for the upper and lower lip models. Correlations between 0.4 and 0.6 (green boxes) and correlations higher than 0.6 (yellow boxes). Columns: joint PCA components, Rows: Guided PCA components

4.10. Velum models

Figure 4-13 shows the performance of all the multilinear methods, described in Chapter 1, in terms of variance explanation and RMSE for the velum contour. As shown in Chapter 3, two components were extracted by means of a PCA applied to the velum contour. These two components represented an oblique movement related to the levator veli palatini muscle and a closure of the nasopharyngeal port by a back-front movement. In this section, two components were also taken as a reference of comparison. The variance explanation obtained for all speakers with individual PCA models with two components reached 90%. The associated RMSE was 0.08 cm. We wanted now to extract a common set of articulatory components, for all speakers. Thus, the velum contour was also modelled by means of multilinear methods like: PARAFAC, TUCKER and joint PCA.

A Student's t-test at 5% significance level was used to determine the number of components for each multilinear method that gives an RMSE not statistically different from the one obtained by the reference individual PCA models with two components. Table 4-9 summarizes the results of the Student's t-test. As well as we have seen before for the tongue and lip models, TUCKER showed a very similar performance compared to joint PCA. Thus, TUCKER was discarded and not taken into account for further analysis with the Student's t-test. PARAFAC needed between 4 and 21 components, depending on the speaker. Figure 4-14 shows the range of components needed by PARAFAC, for each speaker, to equal the performance of PCA according to the Student's t-test. PARAFAC requires the minimum number of components for speaker RL, LD and MG (at least 4, 10 and 10 components for each speaker respectively) and the maximum for the other speakers (21 components) accounting for a variance explained between 78.9% and 88.41%. Joint PCA needed between 1 and 14 components to equal the performance of the reference PCA models, depending on the speaker. The Figure 4-15 shows the range of components needed by joint PCA, for each speaker, to equal the performance of PCA according to the Student's t-test. Joint PCA requires the minimum number of components for speaker LD (at least 1 component) and the maximum for speaker AK (14 components) accounting for a variance explained between 60.02% and 94.2%.

According to the results of the Student's t-test, joint PCA appeared to be the optimal solution to model the velum contour. Joint PCA needed smaller ranges of components to normalise and model the velum contour, over all speakers, compared to PARAFAC.

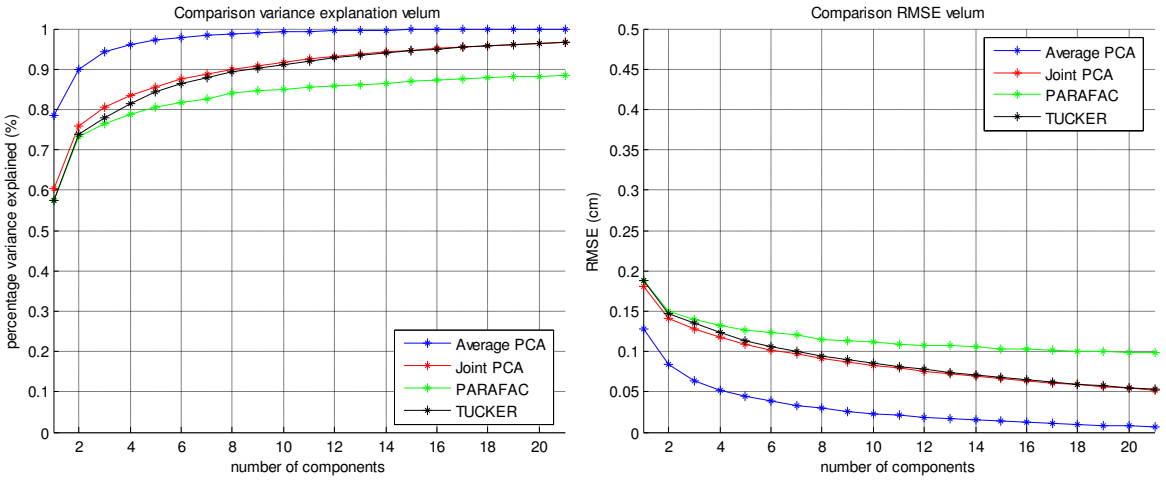


Figure 4-13 - Performance, established using LOOCV, of the Average individual PCA, PARAFAC, TUCKER and joint PCA as a function of number of components for the velum contour for a corpus including vowels and consonants. Left: variance explanation. Right: RMSE

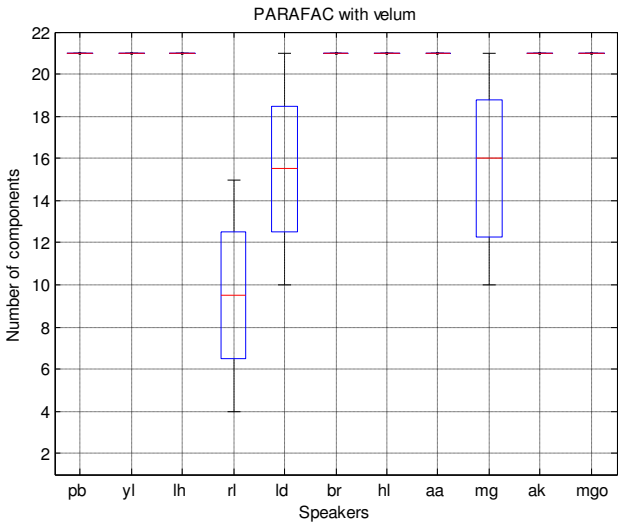


Figure 4-14 - Range of number of components needed by PARAFAC according to a Student's t-test between the reference PCA, with 2 components, and the multi-linear method PARAFAC for the velum contour

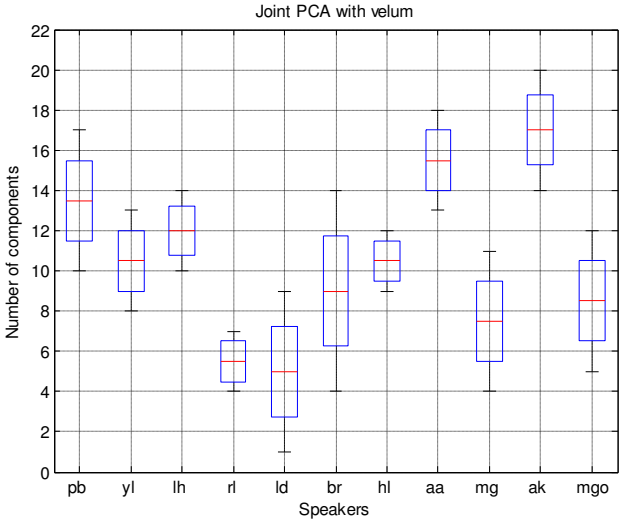


Figure 4-15 - Range of number of components needed by joint PCA according to a Student's t-test between the reference PCA, with 2 components, and the multi-linear method Joint PCA for the velum contour

Representation of data	Average PCA		PARAFAC		Joint PCA	
	Ref. cmp	Var. Exp.	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.
Velum	2	90%	4 - 21	78.9% - 88.41%	1 - 14	60.02% - 94.2%

Table 4-9 – Results of Student's t-test between reference PCA, with 2 components, and the multi-linear methods (PARAFAC and joint PCA), for the velum contour

We have also analysed the meaning of the Joint PCA components, extracted from the velum, by computing the correlations between the 2 PCA components, described in Chapter 3, and the 14 joint PCA components common to all speakers (see Table 4-10). The correlations from the 4th component on were below 0.4 and thus not included in Table 4-10. Overall, the strongest correlations (yellow boxes) indicate that the first components extracted by joint PCA can be approximately interpreted in terms of an oblique movement (PCA-1) and a back to front movement (PCA-2). However, for speakers AA, MG, AK and MGO the component PCA-2 has zero correlation with the joint PCA components.

Velum										
		1	2	3	4			1	2	3	4
pb	PCA-1	**0.83**				**hl**	PCA-1	**0.73**			
	PCA-2		**-0.79**				PCA-2		0.50		
yl	PCA-1	**0.81**				**aa**	PCA-1	**0.89**			
	PCA-2		**-0.70**				PCA-2				
lh	PCA-1	**0.65**				**mg**	PCA-1	**0.72**			
	PCA-2		**-0.66**				PCA-2				
rl	PCA-1	**0.75**		0.48		**ak**	PCA-1	**0.67**			-0.47
	PCA-2		**-0.74**				PCA-2				
ld	PCA-1	**0.62**		-0.58		**mgo**	PCA-1	**0.68**		-0.44	0.51
	PCA-2		**0.78**				PCA-2				
br	PCA-1	**0.87**									
	PCA-2			**0.77**							

Table 4-10 - Correlations between the first 4 components of joint PCA and the 2 components of PCA for the velum models. Only correlations higher or equal to 0.4 are shown. Correlations between 0.4 and 0.6 (green boxes) and correlations higher than 0.6 (yellow boxes). Columns: joint PCA components, Rows: PCA components

4.11. Conclusion

This chapter presented different linear and multilinear tongue models built from the data presented in Chapter 2.

The models are evaluated by means of two criteria: the relative explained variance and the Root Mean Square Error (RMSE). The models were also assessed using a leave-one-out cross validation procedure (LOOCV) to ensure that there was not over-fitting.

In order to have a reference starting point for the tongue models, our modelling was first limited to a repertoire of only French vowels and compared with the results quoted on the literature. Then, a PARAFAC model that extracted 2 components was built. The results in terms of variance explained of our PARAFAC model were coherent with the literature.

The results of individual speaker models and various multilinear models, using different representations of the tongue contour (UpperTng, INT, INTRXY), were compared. On average, over our eleven speakers, the individual PCA models explain an amount of 93.23% of the data variance, with an RMSE of 0.12 cm, using 4 components. The PCA models have been used as baseline models to assess the performance of the different multilinear methods. A Student's t-test was used to determine the number of components that gives an RMSE not statistically different from the one obtained by the reference individual PCA models. The results of the Student's t-test revealed that all multilinear methods needed at least 21 components to normalize the upper tongue contour of all the speakers together, for all the representations of data. Thus, the re-sampling of tongue contour with INTRXY and INT does not constitute an advantage for the modelling. By re-sampling the tongue contour we gain little extra variance explanation but lose information. Joint PCA with UpperTng appeared to be the optimal solution. This method needed between 14 and 21 components to model the tongue contour of all speakers, accounting for a variance explained between 90.33% and 94.88%. We have also analysed the semantics of the components extracted by joint PCA. This was done by computing the correlations between the 4 components of the individual guided PCA models and the 21 joint PCA components common to all speakers. The strongest correlations indicated that the first 4 joint PCA components can be approximately interpreted in terms of jaw height (JH), tongue body (TB), tongue dorsum (TD) and tongue tip (TT); accounting for a variance explained of 72.16% and an RMSE of 0.27 cm. Note that to model the tongue contour of all the eleven speakers we need 4 PCA components per speaker, accounting for a average variance explained of 93.23% and an average RMSE of 0.13 cm. Thus, we need 44 (4 components * 11 speakers) components in total to model the tongue contour of the whole set of speakers. On the other hand, joint PCA needs 21 components to model all the speakers, accounting for a variance explained of 94.88%

and an RMSE of 0.12 cm. Hence, there is a considerable reduction of number of components when using joint PCA.

Individual models for the full tongue contour (FullTng), including the sublingual cavity, were also built. The missing sublingual cavities were predicted by means of expectation maximization. The performance of the missing data PCA models was a bit lower than the simple PCA models. The variance explanation and RMSE were computed using only the upper tongue contour. In average, over our eleven speakers, the individual speaker models, for the missing and not missing data PCA using 4 components, explain an amount of 92.04% and 93.23% of the data variance, with an RMSE of 0.14 cm and 0.12 cm, respectively.

Besides, this chapter presents two normalisation approaches based on the mapping of articulatory spaces represented by the components extracted by the PCA tongue models. The goal of these techniques was to find the corresponding projection of a given point into a target space. The projection was computed by using the weights of the K closest neighbours. We concluded that the articulatory spaces of our speakers were not homogeneous enough to make good predictions.

We have also tested a technique called generalised procrustes analysis (explained in the annex B). This technique was used to align the data of our speakers to each other. Nevertheless, the alignment of data did not constitute a significant improvement of the modelling,

We have also modelled the upper and lower lip, and the velum contours. Joint PCA appeared to be the optimal solution to model these contours.

The individual PCA models for the upper and lower lip contours, with 3 components, accounted for an average variance explained of 94.89% and 94.5%, and an average RMSE of 0.03 cm and 0.05 cm, respectively. While the optimal joint PCA solution, with 21 components, accounted for a variance explained of 96.67% and 96.85% with an RMSE of 0.03 cm and 0.04 cm, respectively. A joint PCA model, using 3 components, accounted for a variance explained of 74.28% and 69.26% with an RMSE of 0.08 cm and 0.15 cm, respectively.

The individual PCA models for the velum contour, with 2 components, accounted for an average variance explained of 90% and an average RMSE of 0.08 cm. While the optimal joint PCA solution, with 14 components, accounted for a variance explained of 94.2% with an RMSE of 0.07 cm. A joint PCA model, using 2 components, accounted for a variance explained of 76.01% with an RMSE of 0.14 cm.

Note that there is a considerable reduction of number of components when using joint PCA, compared to the total number of components needed for the individual models of all speakers. For instance, on the one hand, for the upper and lower lip PCA models, we need 33 (3 components * 11 speakers) components in total to model all

the speakers. On the other hand, joint PCA needed 21 components to model the lips of all the speakers. The velum PCA models needed 22 (2 components * 11 speakers) to model all the speakers, while joint PCA needed 14 components for a model that includes all the speakers.

Chapter 5. Conclusions and perspectives

5.1. Conclusions

The twofold objective of this thesis work was to acquire knowledge about inter-speaker variability, and to propose models to adapt a given reference clone, composed of articulator's models (lips, tongue, velum, etc), to a variety of speakers using geometric and multilinear decomposition methods. This work has been conducted in the framework of the development of a visual articulatory feedback system, based on a given reference speaker, which automatically animates a 3D talking head from the speech sound. Thus, the main idea was to adapt this system to the morphology and articulatory strategies of several speakers. The applications lay on the domain of Computer Aided Pronunciation Training (CAPT) and speech rehabilitation.

In order to build articulatory models of various vocal tract contours, we have acquired data that cover the whole articulatory space in the French language. We collected a corpus of 11 French speakers (6 males and 5 females), and 63 articulations including vowels and consonants in vocalic context, recorded by means of magnetic resonance imaging (MRI). The data of 12 vocal tract contours were included (lips, palate, velum, pharynx, jaw, hyoid bone, tongue, epiglottis, glottis, backlarynx, trachea, and spinal cord). One of the main contributions of this study was the acquisition of data for the complete vocal tract contours, which allowed us to model and study the synergy between different organs involved in speech. Another important contribution is the inclusion of 10 consonants in vocalic contexts, rarely included in the literature, which also implied a challenge for modelling. Apart from the studies of Hoole (1998) which built models for the consonants (/p t k/), and Geng & Mooshammer (2000) which included models for the consonant /t/, only Ananthakrishnan et al. (2010) covered a more complete set of 10 consonants (/p t k f s ʃ m n ʁ l/). Besides, our database includes a larger set of speakers compared to all the studies in the literature. Our data were useful to characterise and compare our speakers by means of statistics of articulatory measurements, and to build models based on several multilinear decomposition methods.

The next step was the characterisation of our speakers concerning articulatory strategies. Individual guided PCA models were built for the tongue, lips and velum contours. The models of all speakers were analysed and compared. By looking at the

nomograms of the models, which are graphical representations of the components extracted, we observed that each speaker has his/her own strategy to achieve sounds that are considered equivalent for speech communication purposes. For instance, the tongue contour variability was decomposed in four principal movements: jaw height, tongue body, tongue dorsum and tongue tip. These movements are performed in different proportions according to the speaker. Besides, for a given displacement of the jaw, the tongue may globally move in a proportion that depends on the speaker. We also noticed that lip protrusion, lip opening, the influence of the jaw movement on the lips, and the velum's articulatory strategy can also vary according to the speaker. For example, some speakers roll up their uvulas against the tongue to produce the consonant /ʁ/ in vocalic contexts. The acoustic consequences of the different speakers' strategies were also compared by means of graphics in the F2-F1 space. The acoustic analysis for only vowels showed coherence with previous analysis reported in the literature. Besides, an acoustic and articulatory analysis of the consonant /k/ showed that this consonant can be articulated either in a palatal or a velar way, according to the vowel context and to the speaker. These findings constitute an important contribution to the knowledge of inter-speaker variability in speech production.

We have tested non-linear methods based on the mapping of a given point in the articulatory space of a source speaker to its corresponding point in the articulatory space of a target speaker. The articulatory spaces were represented by the components tongue body (x axis) and jaw height (y axis), extracted by the guided PCA models of each speaker. We noticed that the articulatory spaces of our speakers were not homogeneous enough to make good predictions.

In order to extract a set of common articulatory control parameters from all speakers, we built linear models of the tongue, lips and velum contours, based on several linear decomposition methods (PARAFAC, TUCKER and Joint PCA). For each decomposition method, a Student's t-test at 5% significance level was used to determine the number of components that gives an RMSE not statistically different from the one obtained by the individual PCA models of each contour: tongue, upper lip, lower lip and velum. The results showed that joint PCA was the optimal solution for modelling all the contours:

- For the tongue contour, the individual speakers' models needed 44 components (4 components * 11 speakers) in total to model all the speakers, accounting for an average variance explained of 93.23% and an average RMSE of 0.13 cm. On the other hand, joint PCA needed 21 components to model all the speakers,

accounting for a variance explained of 94.88% and an RMSE of 0.12 cm. A joint PCA model, using 4 components, accounted for a variance explained of 72.16% and an RMSE of 0.27 cm. Furthermore, the correlations between the joint PCA components and the guided PCA components showed that the first 4 joint PCA components can be approximately interpreted in terms of jaw height, tongue body, tongue dorsum and tongue tip.

- For the upper and lower lip contours, the individual speakers' models needed 33 components (3 components * 11 speakers) in total to model all the speakers, accounting for an average variance explained of 94.89% and 94.5%, and an average RMSE of 0.03 cm and 0.05 cm, respectively. On the other hand, joint PCA needed 21 components to model all the speakers, accounting for a variance explained of 96.67% and 96.85% with an RMSE of 0.03 cm and 0.04 cm, respectively. A joint PCA model, using 3 components, accounted for a variance explained of 74.28% and 69.26% with an RMSE of 0.08 cm and 0.15 cm, respectively. Besides, the correlations between the joint PCA components and the guided PCA components showed that the first joint PCA components can be globally interpreted in terms of jaw height, lip protrusion and lip height.

- For the velum contour, the individual speakers' models needed 22 components (2 components * 11 speakers) in total to model all the speakers, accounting for an average variance explained of 90% and an average RMSE of 0.08 cm. On the other hand, joint PCA needed 14 components to model all the speakers, accounting for a variance explained of 94.2% with an RMSE 0.07 cm. A joint PCA model, using 2 components, accounted for a variance explained of 76.01% with an RMSE 0.14 cm. Besides, the correlations between the joint PCA components and the PCA components indicated that the first components extracted by joint PCA can be approximately interpreted in terms of an oblique movement and a back to front movement of the velum.

Table 5-1 summarizes the results explained above. Note that there is a considerable reduction of number of components when using joint PCA compared to the individual PCA models. Furthermore, Table 5-2 compares our models built for vowels and consonants with the models reported in the literature for only vowels (see Table 4-1). Note that the PARAFAC models reported in the literature use 2 components to model between 7 and 15 vowels, for corpuses between 3 and 9 speakers, accounting for a variance explained between 71% and 96%. Thus, our two final joint PCA models with 4 and 21 components, built for a corpus of 63 articulations including vowels and consonants, which represent respectively 72.16% and 94.88% of the data variance, are comparable in terms of variance explanation with the models reported in the literature, built for only vowels.

Another important contribution of this study is the modelling for the upper lip, lower lip and the velum contour to extract a set of common articulatory patterns from several speakers. As far as I know, there are no studies in the literature about the normalisation of these important vocal tract contours.

Contour	Average PCA			Joint PCA according to Student's t-test			Joint PCA with reduced no. of components		
	No. Components	Variance Exp.	RMSE	No. Components	Variance Exp.	RMSE	No. Components	Variance Exp.	RMSE
Upper tongue	44 (4 *11)	93.23%	0.13 cm	21	94.88%	0.12 cm	4	72.16%	0.27 cm
Upper lip	33 (3*11)	94.89%	0.03 cm	21	96.67%	0.03 cm	3	74.28%	0.08 cm
Lower lip	33 (3*11)	94.50%	0.05 cm	21	96.85%	0.04 cm	3	69.26%	0.15 cm
Velum	22(2*11)	90%	0.08 cm	14	94.20%	0.07 cm	2	76.01%	0.14 cm

Table 5-1 - Comparison between average PCA models and joint PCA for the tongue, upper lip, lower lip and velum contour

	Models for Vowels	Models for a corpus of vowels and consonants	
Method	**PARAFAC**	**Joint PCA according to Student's t-test**	**Joint PCA with reduced no. of components**
No. components	2	21	4
Variance Exp.	71% - 96%	94.88%	72.16%
Corpus	7 - 15 vowels	63 articulations (vowels and consonants)	63 articulations (vowels and consonants)
No. speakers	3 - 9 speakers	11 speakers	11 speakers

Table 5-2 - Comparison between PARAFAC models built for vowels, reported in the literature, and our joint PCA models built for consonants and vowels, for the tongue contour

5.2. Perspectives

The results presented in this manuscript are promising but require further investigation. The first point to be considered is the increase of number of speakers. Indeed, one important issue is the inter-speaker variability which refers to how much speakers differ from each other. The inter-speaker variability should be large enough to extract articulatory patterns that are as general as possible. Thus, the data of more speakers should be included to the models presented in this study.

In the present work, the acoustic analysis was very limited, and no acoustic data were available. Future work should thus investigate the acoustic consequences of particular articulatory strategies of each speaker. One important aspect for vocal tract acoustics is the lip area. In this study, the lip area was not included. Only the area of the grid system sections between the tongue tip and the epiglottis was taken into account. Thus, it would be pertinent to extend our grid system up to the lips. We could then analyse and compare the F3 of all speakers, which is usually related to the lip rounding (Hardcastle & Marchal, 1990).

Another important matter is the modelling of the velum contour. On the one hand, the individual velum models needed 22 components (2 components * 11 speakers) in total to model all the speakers. On the other hand, joint PCA needed 14 components to model the velum of all speakers. However, a better modelling performance could be expected. Future work could focus on cross-speaker velum variability. A deeper understanding about the velum behaviour is needed to propose new modelling solutions. For instance, the velum contour may be somehow pushed backwards by the tongue action for certain articulations. Moreover, some speakers could use a given degree of nasality even though they are not producing nasal articulations. All this facts introduce variability in the data, which makes modelling more complex.

Most of the methods used in this study were linear. Nevertheless, one can suppose that linear methods may not offer the best solution to model the variability among different speakers, especially in the presence of consonants. Thus, future work is to be directed at using non linear methods. A first step towards that purpose is to model the individual speakers by means of a non-linear method like Kernel PCA (Schölkopf et al., 1998; Mika et al., 1999).

Data could be acquired by means of a different technique like EMA or real time MRI (Narayanan, 2004; 2011), which allow obtaining much larger quantities of data because of a faster acquisition speed compared to MRI. These data could be directed at building stochastic models. Statistical learning methods like Hidden Markov models (HMMs) and Gaussian mixture models (GMMs) could be used to estimate articulatory features directly from the data (Toda et al., 2008; Zen et al. 2009; Ling et al., 2010; Ben Youssef et al., 2011b).

Bibliography

Aalto, D., Malinen, J., Vainio, M., Saunavaara, J., & Palo, P. (**2011**). Estimates for the measurement and articulatory error in MRI data from sustained vowel production. 17th International Congress of Phonetic Sciences (ICPhS XVII), (pp. 180-183). Hong Kong, China.

Ananthakrishnan, G., Badin, P., Valdés Vargas, J. A., & Engwall, O. (**2010**). Predicting unseen articulations from multi-speaker articulatory models. 11th Annual Conference of the International Speech Communication Association, (pp. 1588-1591). Makuhari, Japan.

Apostol, L., Perrier, P., & Bailly, G. (**2004**). A model of acoustic interspeaker variability based on the concept of formant--cavity affiliation. The Journal of the Acoustical Society of America , 115 (1), 337-351.

Badin, P., & Fant, G. (**1989**). Fricative production modelling: aerodynamic and acoustic data. In 1st EuroSpeech Conference , 2, 23-26.

Badin, P., & Serrurier, A. (**2006**). Three-dimensional modeling of speech organs: Articulatory data and models. IEICE Technical Report. 106, pp. 29-34. Kanazawa, Japan: The Institute of Electronics, Information, and Communication Engineers.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., & Savariaux, C. (**2002**). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics , 30 (3), 533-553.

Badin, P., Koncki, A., Vargas, J. A., Lamalle, L., & Savariaux, C. (**2012**). Développement et mise en œuvre de marqueurs fiduciaires pour l'imagerie IRM du conduit vocal en vue de la modélisation articulatoire de la parole. In ATALA-AFCP (Ed.), *Actes des 29èmes Journées d'Etude de la Parole*, (pp. 81-88). Grenoble, France.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (**2010a**). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Communication , 52 (6), 493-503.

Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., & Hueber, T. (**2010b**). Visual articulatory feedback for phonetic correction in second language learning. L2SW, Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, (pp. P1-10). Tokyo, Japan.

Bailly, G. (**1997**). Learning to speak. Sensori-motor control of speech movements. Speech Communication , 22, 251-267.

Bailly, G., Badin, P., & Vilain, A. (**1998**). Synergy between jaw and lips/tongue movements: Consequences in articulatory modelling. In R. H. Mannell, & J. Robert-Ribes (Ed.), 5th International Conference on Spoken Language Processing, 5, pp. 1859-1862. Sydney, Australia.

Bailly, G., Bégault, A., Elisei, F., & Badin, P. (**2008**). Speaking with smile or disgust: data and models. Auditory-Visual Speech Processing Workshop, AVSP 2008, (pp. 111-114). Moreton Island, Australia.

Beautemps, D., Badin, P., & Bailly, G. (**2001**). Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. Journal of the Acoustical Society of America , 109 (5), 2165-2180.

Beautemps, D., Badin, P. & Laboissière, R. (**1995**). Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. Speech Communication, 16, 27-47.

Berger, M. (**1987**). Geometry I. Springer.

Ben Youssef, A., Hueber, T., Badin, P., Bailly, G., & Elisei, F. (**2011a**). Toward a speaker-independent visual articulatory feedback system. 9th International Seminar on Speech Production, ISSP9.

Ben Youssef, A., Hueber, T., Badin, P., & Bailly, G. (**2011b**). Toward a multi-speaker visual articulatory feedback system. In Interspeech 2011 , 589-592.

Boë, L.-J., Ménard, L., & Maeda, S. (**2000**). Adaptation of control strategies during vocal tract growth inferred from simulation studies with an articulatory model. 5th Seminar on Speech Production: Models and Data \& CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, (pp. 277-280). Kloster Seeon, Germany.

Derivery, N. (**1997**). La phonétique du Français. (Seuil, Ed.) mémo.

Edwards, J., & Harris, K. S. (**1990**). Rotation and translation of the jaw during speech. Journal of Speech and Hearing Research, 33, 550-562.

Engwall, O. (**2003**). A revisit to the application of MRI to the analysis of speech production – testing our assumptions. Sixth International Seminar on Speech Production, (pp. 1-6). Sydney, Australia.

Engwall, O. (**2006**). Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation. Speech production: Models, Phonetic Processes and Techniques. (J. H. Tabain, Ed.) New York: Psychology Press.

Engwall, O. (**2004**). Speaker adaptation of a three-dimensional tongue mode. 8th International Conference on Spoken Language Processing - Interspeech 2004, (pp. 465-468). Jeju Island, Korea.

Engwall, O., & Beskow, J. (**2003**). Effects of corpus choice on statistical articulatory modeling. Sixth International Seminar on Speech Production. Sydney, Australia.

Erber, N. P. (**1975**). Auditory-visual perception of speech. Journal of Speech and Hearing Disorders , XL, 481-492.

Fant, G. (**1960**). Acoustic Theory of Speech Production. The Hague: Mouton.

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

Geng, C., & Mooshammer, C. (**2009**). How to stretch and shrink vowel systems: Results from a vowel normalization procedure. Journal of the Acoustical Society of America , 125 (5), 3278-3288.

Geng, C., & Mooshammer, C. (**2000**). Modeling the German stress distinction. 5th Seminar on Speech Production: Models and Data \& Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, (pp. 161-164). Kloster Seeon, Germany.

Goldstein, U. (**1980**). An articulatory model for the vocal tracts of growing children. Ph.D. dissertation, MIT, Cambridge, MA.

Hardcastle, W. J., & Marchal, A. (**1990**). Speech production and speech modelling. NATO Advanced Study Institute on Speech Production and Speech Modelling Bonas, France: Springer.

Harshman, R. (**1970**). Foundations of the PARAFAC procedure: Models and conditions for an "exploratory" multimodal factor analysis. UCLA Working Papers in Phonetics , 16, 1-84.

Harshman, R., & Lundy, M. (**1994**). PARAFAC: Parallel factor analysis. Computational Statistics and Data Analysis , 18, 39-72.

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

Hashi, M., Westbury, J., & Honda, K. (**1998**). Vowel posture normalization. Journal of the Acoustical Society of America , 104 (4), 2426-2437.

Hawkins, D. M. (**2004**). The Problem of Overfitting. J. Chem. Inf. Comput. , 44 (1), 1-12.

Hoole, P. (**1998**). Modelling tongue configuration in German vowel production. In R. Mannell, & J. Robert-Ribes (Ed.), 5th International Conference on Spoken Language Processing, (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. 5th Seminar on Speech Production: Models and Data \& CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, (pp. 157-160). Kloster Seeon, Germany.

Hu, F. (**2006**). On the lingual articulation in vowel production: case study from Ningbo Chinese. In H. C. Yehia, D. Demolin, & R. Laboissière (Ed.), 7th International Seminar on Speech Production, ISSP7. Ubatuba, SP, Brazil: UFMG, Belo Horizonte, Brazil.

Johnson, K., Ladefoged, P., & Lindau, M. (**1993**). Individual differences in vowel production. The Journal of the Acoustical Society of America , 94 (2), 701-714.

Léon, P. (**2012**). Phonétisme et prononciations du Français. (a. Colin, Ed.)

Ling, Z.-H., Richmond, K., & Yamagishi, J. (**2010**). An Analysis of HMM-based prediction of articulatory movements. Speech Communication , 52 (10), 834-846.

Legou, T., Marchal, A., Meynadier, Y., & André, C. (**2008**). 3D Palatography. International Seminar on Speech Production, (pp. 369-372). Strasbourg, France.

Maeda, S. (**1988**). Improved articulatory models. Journal of the Acoustical Society of America , 84, S146.

Mathieu, B., & Laprie, Y. (**1997**). Adaptation of Maeda's model for acoustic to articulatory inversion. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Ed.),

EUROSPEECH '97 - 5th European Conference on Speech Communication and Technology, (pp. 2015–2018). Rhodes, Greece.

Matthies, M. L., Svirsky, M., Perkell, J., & Lane, H. (**1996**). Acoustic and articulatory measures of sibilant production with and without auditory feedback from a cochlear implant. Journal of Speech and Hearing Research , 39 (5), 936-946.

McFarland, D. H. (**2009**). L'anatomie en orthophonie: Parole, déglutition et audition. (E. Masson, Ed.)

Meyer, E. A. (**1907**). First X-ray photographs of vowel position. Medizinisch-pädagogischen Monatschrift, f. d. gesamte Sprachheilkunde , 17, 8, 9.

Mika, S., Schölkopf, B., Smola, A., robert Müller, K., Scholz, M., & Rätsch, G. (**1999**). Kernel pca and de-noising in feature spaces. Advances in Neural Information Processing Systems 11 (pp. 536-542). MIT Press.

Miranda, A. A., Borgne, Y.-A. L., & Bontempi, G. (**2008**). New Routes from Minimal Approximation Error to Principal Components. Neural Processing Letters, Springer , 27 (3).

Mosher, H. P. (**1927**). X-ray study of movements of the tongue, epiglottis and hyoid bone in swallowing, followed by a discussion of difficulty in swallowing caused by retropharyngeal diverticulum, post-cricoid webs and exostoses. The Laryngoscope .

Narayanan, S. S., Nayak, K., Lee, S., Sethy, A., & Byrd, D. (**2004**). An approach to real-time magnetic resonance imaging for speech production. Journal of the Acoustical Society of America , 115 (4), 1771–1776.

Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y.-C., et al. (**2011**). A multimodal real-time MRI articulatory corpus for speech research. Interspeech 2011 (12th Annual Conference of the International Speech Communication Association), (pp. 837-840). Florence, Italy.

Pearson, K. (**1901**). On lines and planes of closest fit to systems of points in space. Philosophical Magazine , 2 (6), 559-572.

Peterson, G. E., & Barney, H. L. (**1952**). Control methods used in a study of the vowels. Journal of the Acoustical Society of America , 24 (175-184).

Ridouane, R. (**2006**). Investigating speech production A review of some techniques.

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.

Rokkaku, M., Hashimoto, K., Imaizumi, S., Niimi, S., & Kiritani, S. (**1986**). Measurement of the three-dimensional shape of the vocal tract based on the Magnetic Resonance Imaging technique. Annual Bulletin of the Research Institute for Logopedics and Phoniatrics , 20, 47-54.

Ross, A. Procrustes Analysis. Procrustes Analysis .

Schölkopf, B., Smola, A., & Müller, K.-R. (**1998**). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation , 10, 1299-1319.

Serrurier, A., & Badin, P. (**2008**). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. Journal of the Acoustical Society of America , 123 (4), 2335-2355.

Serrurier, A., & Badin, P. (**2005**). A three-dimensional linear articulatory model of velum based on MRI data. Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, (pp. 2161-2164). Lisbon, Portugal.

Serrurier, A., & Badin, P. (**2005**). Towards a 3D articulatory model of velum based on MRI and CT images. ZAS Papers in Linguistics, Speech production and perception: Experimental analyses and models (Susanne Fuchs, Pascal Perrier and Bernd Pompino-Marschall, eds) , 40, 195-211.

Stegmann, M. B., & Gomez, D. D. (**2002**). A Brief Introduction to Statistical Shape Analysis. A Brief Introduction to Statistical Shape Analysis .

Stone, M. S. (**2007**). Comparison of speech production in upright and supine position. The Journal of the Acoustical Society of America , 122.

Tiede, M. K., Masaki, S., & Vatikiotis-Bateson, E. (**2000**). Contrasts in speech articulation observed in sitting and supine conditions. 5th Seminar on Speech Production: Models and Data \& CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, (pp. 25-28). Kloster Seeon, Germany.

Toda, T., Black, A. W., & Tokuda, K. (**2008**). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication , 50 (3), 215-227.

Tucker, L. R. (**1966**). Some mathematical notes on three-mode factor analysis. Psychometrika , 31 (3), 279-311.

Verbeek, J. (2009). Notes on Probabilistic PCA with Missing Values.

Zen, H., Tokuda, K., & Black, A. W. (**2009**). Statistical parametric speech synthesis. Speech Communication , 51 (11), 1039-1064.

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.

Julián Valdés' publications

Badin, P., **Valdés Vargas, J. A.**, Koncki, A., Lamalle, L. & Savariaux, C. (**2013**). Development and implementation of fiduciary markers for vocal tract MRI imaging and speech articulatory modeling. Interspeech 2013 (14th Annual Conference of the International Speech Communication Association). Lyon, France, USA.

Valdés Vargas, J. A., Badin, P., & Lamalle, L. (**2012**). Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods. Interspeech 2012 (13th Annual Conference of the International Speech Communication Association). Portland, Oregon, USA.

Valdés Vargas, J. A., Badin, P., Lamalle, L. & Ananthakrishnan, G. (**2012**). Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods (Normalisation articulatoire du locuteur par méthodes de décomposition tri-linéaire basées sur des données IRM). In 29èmes Journées d'Etude de la Parole (L. Besacier, B. Lecouteux & G. Sérasset, Eds.), 1, pp. 529-536. Grenoble, France.

Badin, P., Koncki, A., **Valdés Vargas, J. A.**, Lamalle, L., & Savariaux, C. (**2012**). Développement et mise en oeuvre de marqueurs fiduciaires pour l'imagerie IRM du conduit vocal en vue de la modélisation articulatoire de la parole. 29èmes Journées d'Etude de la Parole (L. Besacier, B. Lecouteux & G. Sérasset, Eds.), 1, pp. 81-88. Grenoble, France.

Ananthakrishnan, G., Badin, P., **Valdés Vargas, J. A.**, & Engwall, O. (**2010**). Predicting unseen articulations from multi-speaker articulatory models. 11th Annual Conference of the International Speech Communication Association, (pp. 1588-1591). Makuhari, Japan.

Annex A: MRI examples

This annex includes MRI examples for speakers LD, RL and AK. These speakers have been chosen to illustrate in a general way individual articulatory strategies. On the one hand, speakers LD and RL are male speakers from which RL has an atypical tongue behaviour compared to other speakers, as explained in Chapter 4. On the other hand, speaker AK, who is a female speaker, was included to elucidate vocal tract differences compared to male speakers. MRI examples of the vowels /i, a, u/ were included to illustrate backness and roundness of French vowels. Furthermore, the consonants /k, ʁ, l/ were included to illustrate different constriction places, particular velum behaviours and tongue shapes.

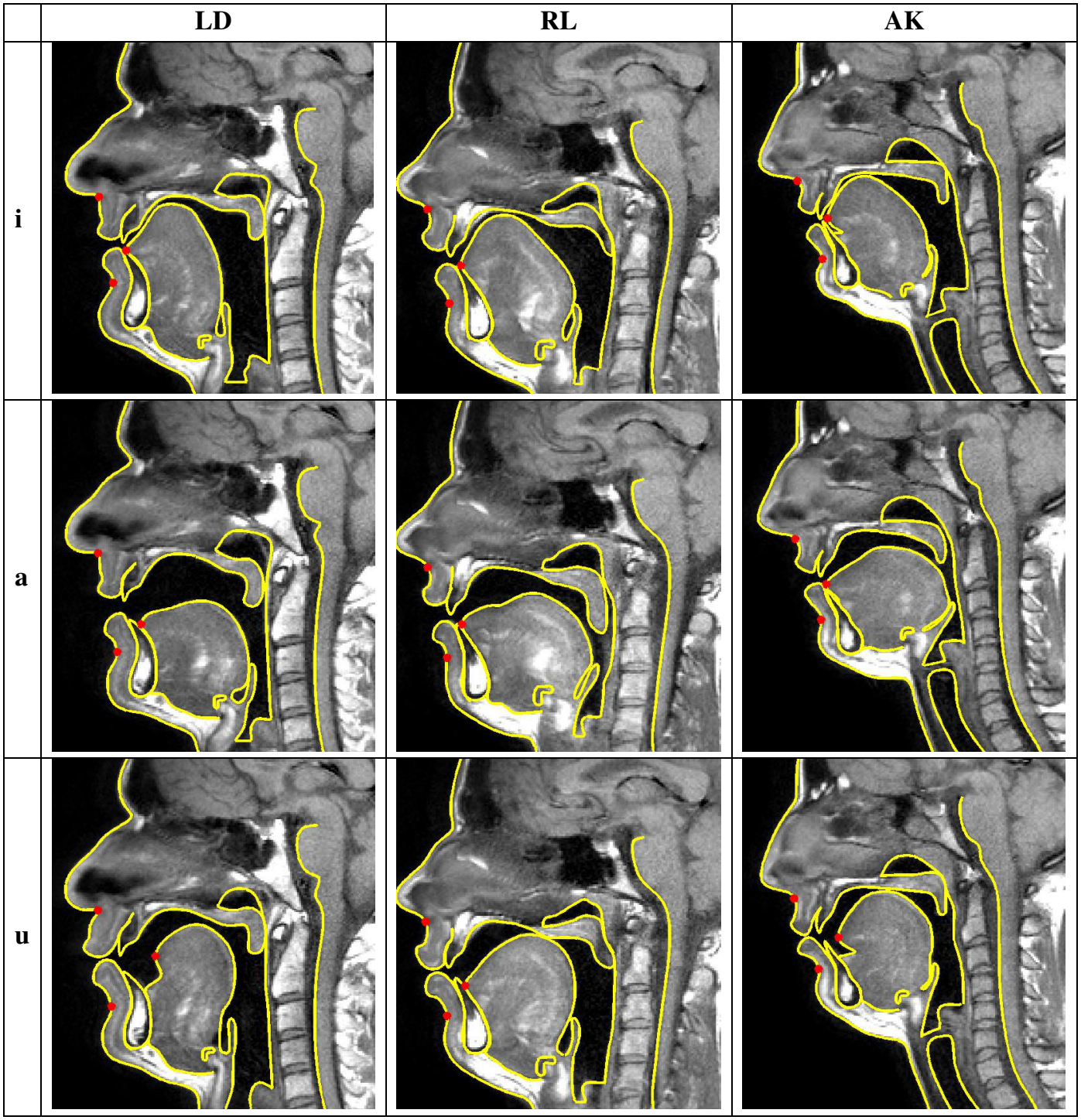


Fig 1 - MRI of articulations /i, a, u/ for male speakers LD and RL and female speaker AK

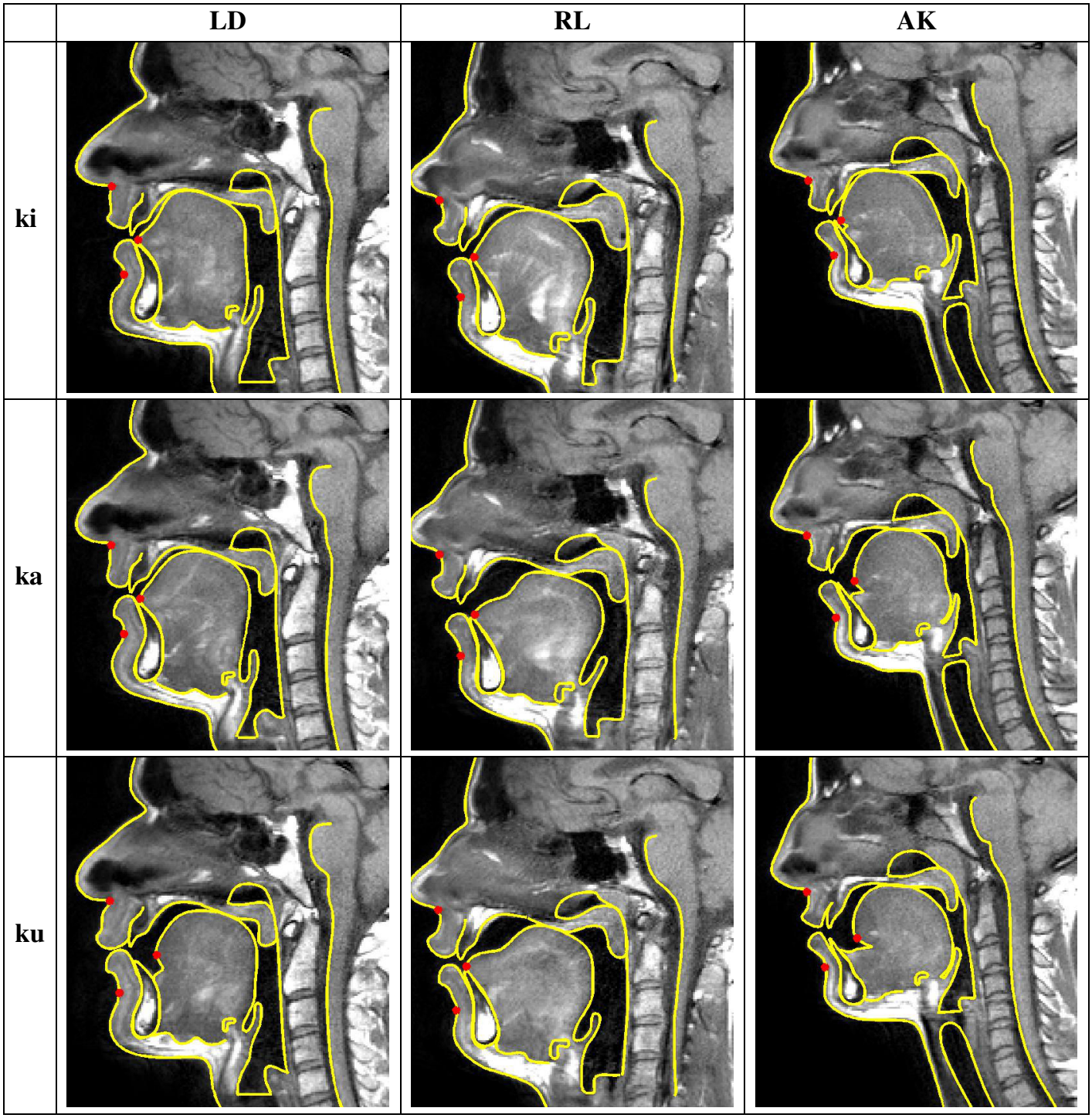


Fig 2 - MRI of articulations /ki, ka, ku/ for male speakers LD and RL and female speaker AK

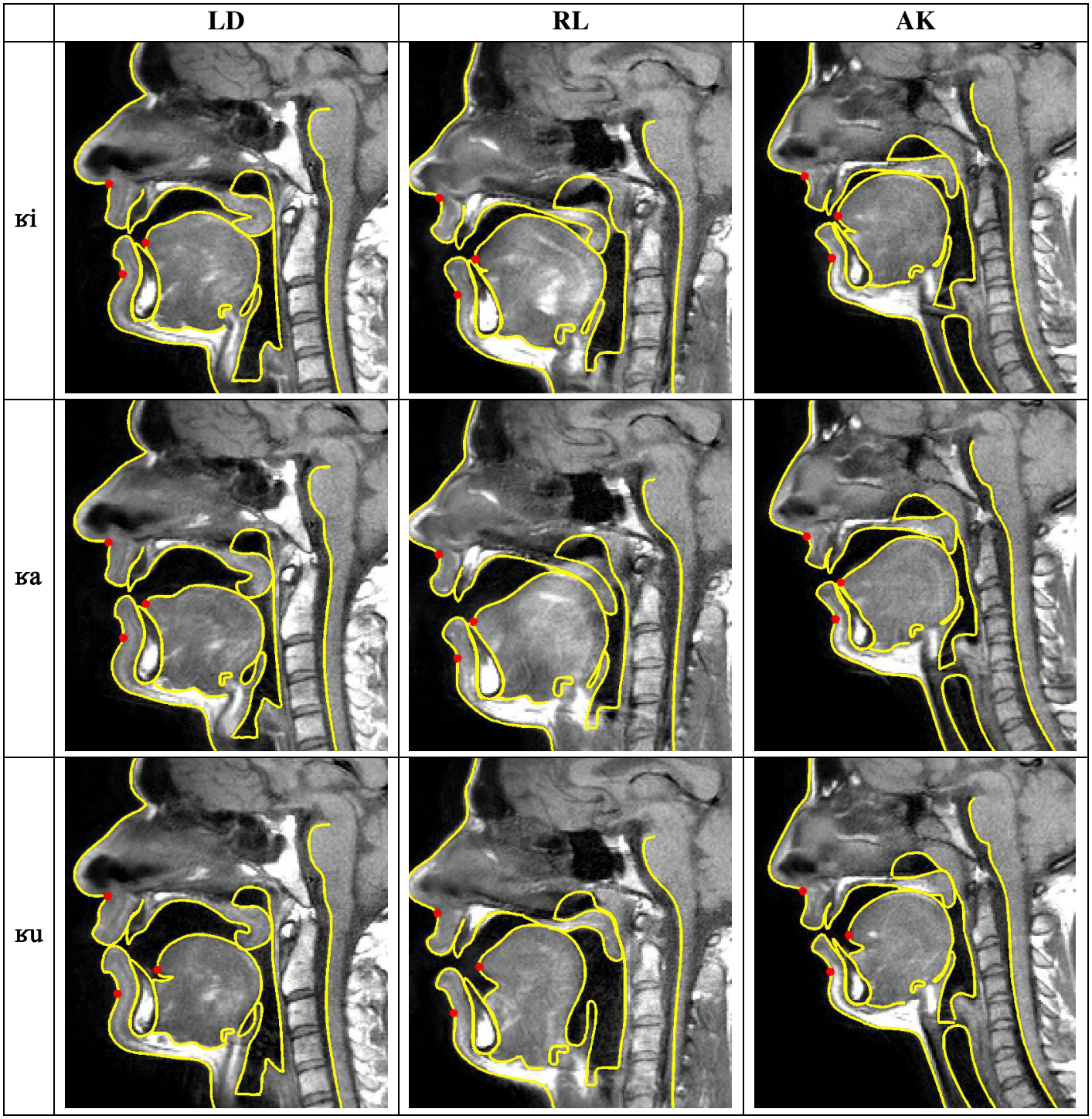


Fig 3 - MRI of articulations /*ʁi, ʁa, ʁu*/ for male speakers LD and RL and female speaker AK

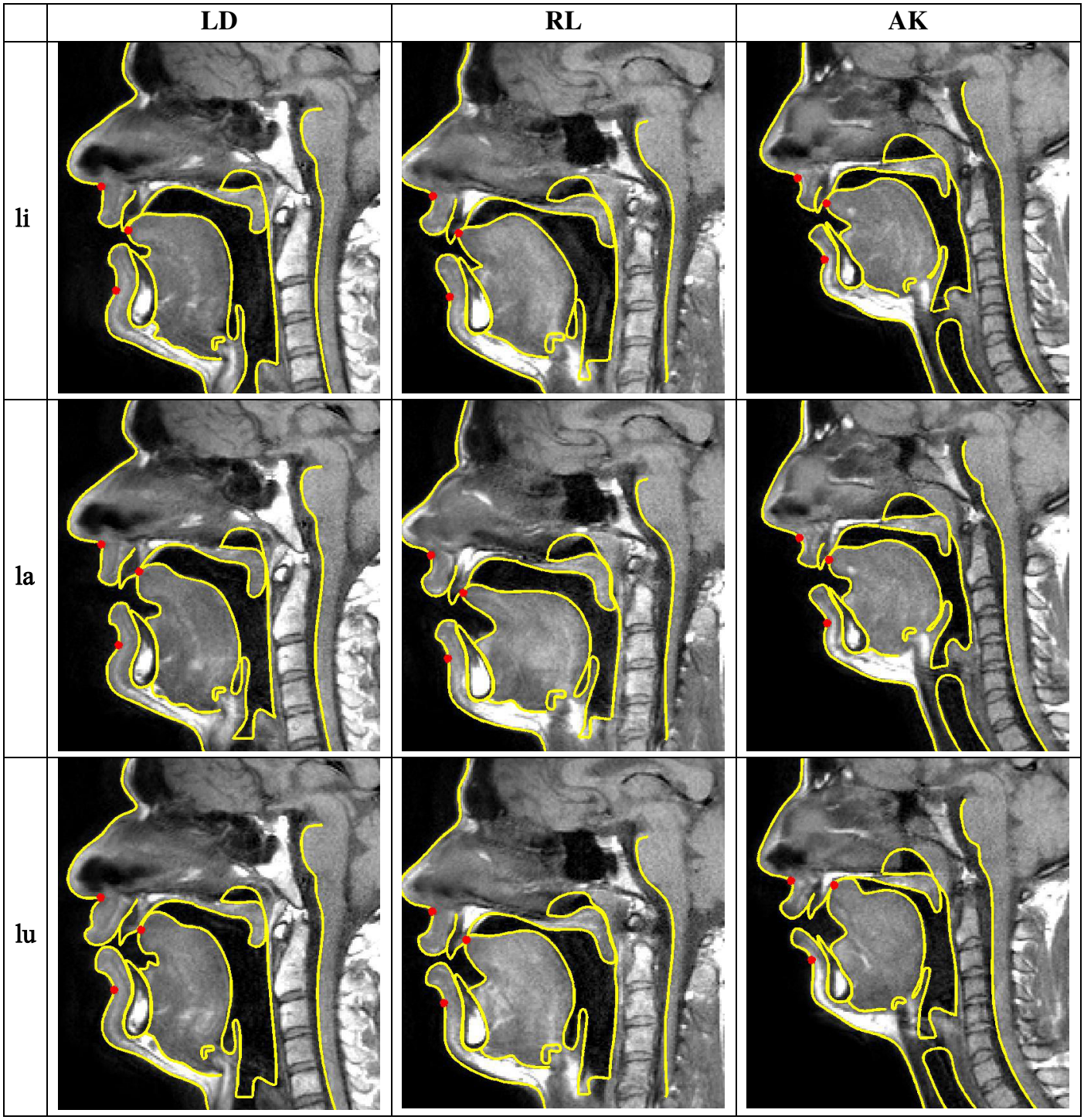


Fig 4 - MRI of articulations /*li, la, lu*/ for male speakers LD and RL and female speaker AK

Annex B: Geometric normalisation of tongue contours

8.1. Introduction

As explained in Chapter 1, one of the main difficulties to model vocal tract contours is the inter-speaker variability as regards morphology. In seek of normalising vocal tract shapes among several speakers; the literature in the domain has proposed methods based on scaling transformations (Hashi et al., 1998; Engwall, 2004; Geng & Mooshammer, 2009; Apostol et al., 2004). The geometric normalisation procedures presented in this chapter are based on several affine transformations: translation, rotation and scaling.

In this chapter we first describe the set of affine transformations applied to the tongue contour. Then, the explanation of an iterative method called Procrustes, based in affine transformations, is given. Finally, the performance of multilinear models, built from the data of aligned tongue contours, is described.

8.2. Affine transformations

The main characteristic of affine transformations is that they keep straight lines and relative distances between the lines (Berger, 1987). For instance, a set of parallel lines will remain parallel after applying an affine transformation. Following sections explain the basis of some affine transforms like translation, rotation and scaling applied to 2-dimensional data.

8.2.1. Translation

Given a set of points, the translation transformation moves every point by a constant distance in certain direction (Berger, 1987). The matrix form of a translation $[Z_x, Z_y]$ applied to a set of points $[W_x, W_y]$ can be written as follows:

$$\begin{bmatrix} 1 & 0 & Z_X \\ 0 & 1 & Z_Y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} W_X \\ W_Y \\ 1 \end{bmatrix} = \begin{bmatrix} W_X + Z_X \\ W_Y + Z_Y \\ 1 \end{bmatrix} = W + Z$$

Fig 5 – Matrix notation of the translation transformation of a point W by a vector Z

8.2.2. Rotation

The rotation transformation is described as the motion around a fixed point (Berger, 1987). The rotation of a point (x, y) by an angle θ clockwise about the origin can be described as:

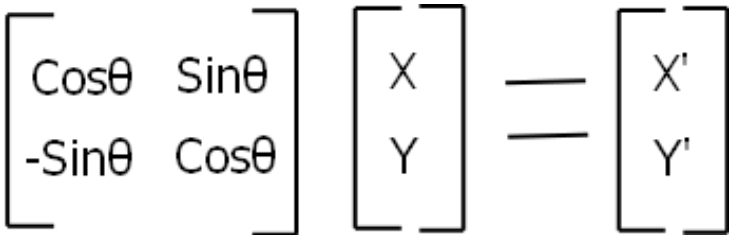


Fig 6 – *Matrix notation of the rotation transformation of a point (x, y) by an angle θ clockwise about the origin*

8.2.3. Scaling

The scaling transformation enlarges or shrinks the shape formed by a set of points (Berger, 1987). The scaling of a point (x, y), by a scaling factor in x (S_x) and a scaling factor in y (S_y) can be described as:

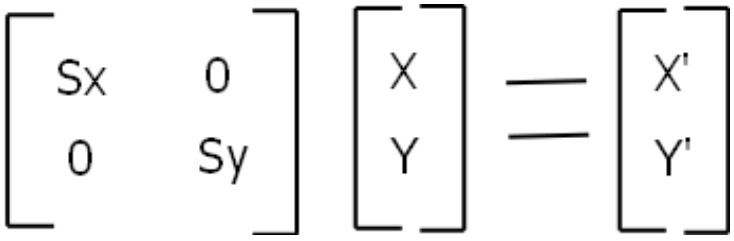


Fig 7 – Matrix notation of the scaling transformation of a point (x, y) by the scaling factor Sx and Sy

8.3. Procrustes and Generalised Procrustes Analysis (GPA)

Procrustes is a method that determines a linear transformation of a set of points Y to best conform the set of points X (Ross). The final linear transformation applied by Procrustes is composed by an iterative combination of translations, orthogonal rotations, and scaling transformations. The goodness-of-fit criterion is defined by the sum of squared errors. However, Procrustes analysis is limited to the alignment of two sets of data. A more advanced technique called Generalized Procrustes Analysis (GPA), allows the alignment of several sets of data (Stegmann & Gomez, 2002). The alignment by GPA involves an iterative process of four steps: first, an initial data set is chosen. Second, all the remaining data sets are aligned to the initial data set. Third, the estimated mean is calculated from the aligned shapes. Finally, if the estimated mean has changed, then the step 2 is repeated. When the mean shape has not changed significantly within the last iteration it is considered that the algorithm has converged. Fig 8 shows how the articulation /i/ was aligned for speakers PB, YL and HL by means of GPA. Only the coordinates of the tongue contour were aligned. All the eleven

speakers, including 63 articulations, were aligned between them by means of GPA. Following section describes the results of models built from the aligned data.

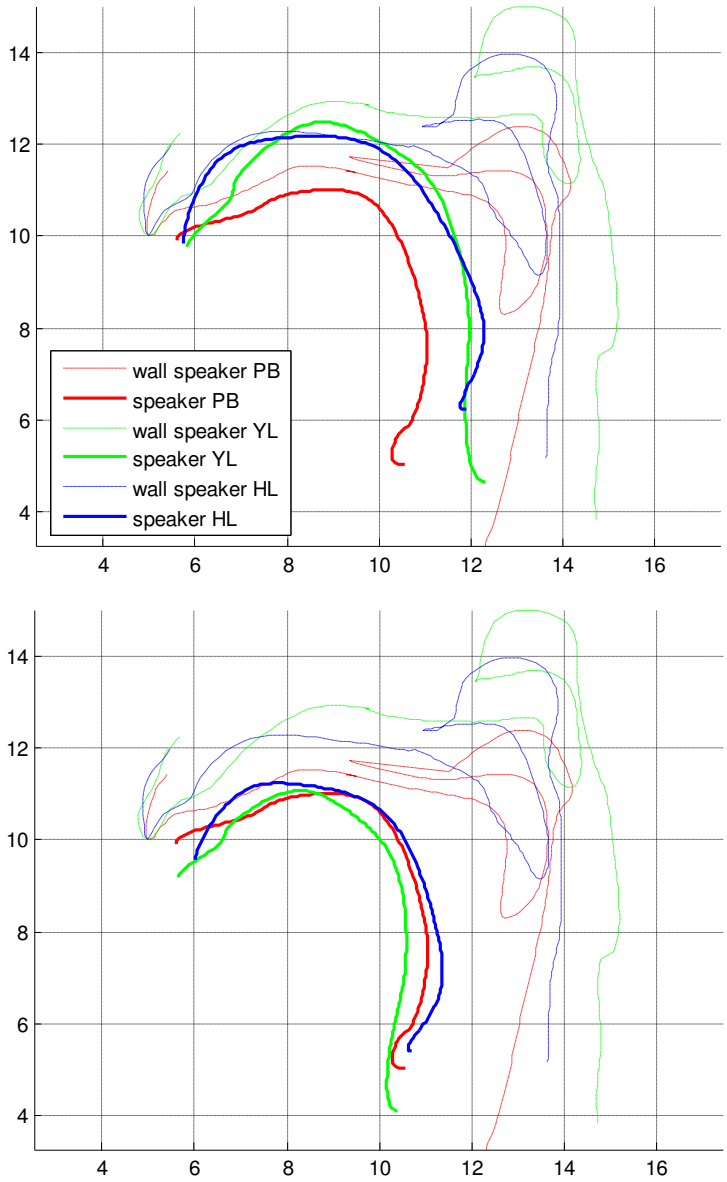


Fig 8 – Alignment of upper tongue contour by means of GPA for articulation /i/ of speakers PB, YL and HL. Tongue in solid lines and reference wall (palate, velum and pharynx) in dot dashed lines. Top: initial data. Bottom: Data aligned by means of GPA

8.4. Effect of affine transformations on linear models

Fig 9 shows the performance of all the multilinear methods, described in section 1.2, in terms of variance explanation and RMSE for the aligned UpperTng of all speakers. For each linear method, the steps to compute the variance explanation and RMSE were: (1) building the model, (2) Reconstruction of the data from the model, (3) the predicted data was transformed back to the original (x, y) coordinate space by applying the inverse of the GPA (GPA^{-1}), finally the variance explained and RMSE

were computed as explained in Chapter 3 and Chapter 4. Fig 9 reveals that aligning the tongue contours, between all the speakers, constitutes a very little improvement to the linear models in terms of variance increasing and RMSE decreasing.

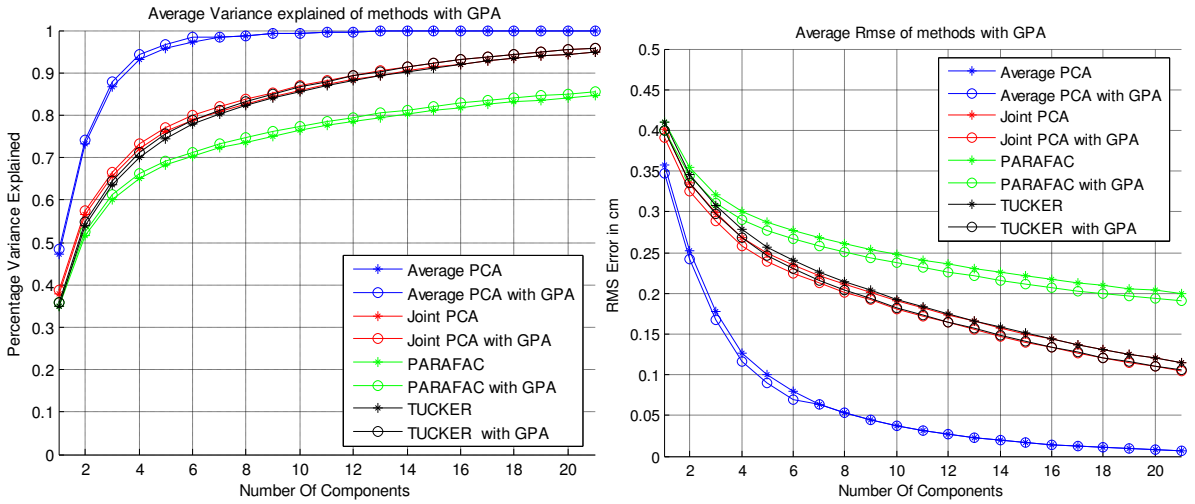


Fig 9 - Performance, established using LOOCV, of the PARAFAC, TUCKER and joint PCA as a function of number of components for the tongue contours aligned by means of GPA. Left: variance explained. Right: RMSE in centimetres

8.5. Conclusion

In this chapter we first described a set of affine transformations. The matrix notation of the rotation, translation and scaling transformations were given. Then, the Procrustes Analysis, an iterative method based on affine transformations, was presented. This method is limited to the alignment of only two sets of data. A more advanced method, the Generalised Procrustes Analysis, which can align more than two sets of data, was described. Finally, the performance of several multilinear methods, applied to the aligned tongue contours, was described in terms of variance explanation and RMSE. We can conclude that even though by applying GPA we were able to reduce the distance between the coordinates of all speakers, this alignment did not constitute a significant improvement to the modelling. The reason of this is that the variability due to different articulatory strategies was still present after the GPA alignment. For example, if some speakers usually positioned their tongues in an anterior position, across all the articulations, while other speakers habitually positioned their tongues in a posterior position, this divergence would be still present after the GPA alignment.

The next chapter will concentrate on modelling the lips and velum contours by means of multilinear decomposition methods.

Annex C: Résumé en Français de la thèse

Cette annexe contient un résumé détaillé en français du travail effectué dans cette thèse.

9.1. Introduction

La production de la parole nécessite une maîtrise précise des différents articulateurs du conduit vocal (lèvres, mâchoire, langue, voile du palais, épiglotte, etc.) Cette habileté est apprise et maintenue au moyen d'une boucle de perception-action qui permet au locuteur de corriger sa production en fonction du retour perceptif (Matthies et al., 1996; Bailly, 1997). Le retour que le locuteur reçoit de sa propre production est auditif et proprioceptif, mais pas vraiment visuel. D'autre part, Erber (1975) a démontré la contribution de la vision des lèvres à la perception de la parole, tandis que Badin et al. (2010b) ont récemment mis en évidence la contribution de la vision de la langue à la reconnaissance des consonnes dans des signaux audio bruités. Par ailleurs, un certain nombre d'études ont exploré l'importance du retour perceptif dans des domaines tels que l'orthophonie, la correction phonétique ou l'acquisition du langage (Badin et al., 2010b). Ainsi, les systèmes de retour articulatoire visuel, qui visent à fournir un retour visuel de l'articulation prononcée au locuteur, semblent appropriés pour améliorer l'intelligibilité de la parole (Badin et al., 2010a)

Le département de parole et cognition à GIPSA-lab a donc développé un système d'inversion acoustique-articulatoire. Ce système est capable de créer un retour articulatoire visuel à partir du signal acoustique (Ben Youssef et al., 2011). Ce système est basé sur un clone orofacial constitué de modèles articulatoires comme la mâchoire, les lèvres, la langue, le voile, etc.

Cependant, le clone de notre système de retour articulatoire visuel développé au GIPSA-lab est basé sur des données articulatoires acquises sur un seul locuteur (Badin et Serrurier, 2006). Par conséquent, le clone représente fidèlement les caractéristiques de ce locuteur spécifique, mais pas nécessairement celles d'autres locuteurs qui peuvent avoir des morphologies et stratégies de contrôle articulatoire différent. Ainsi, le double objectif de cette thèse était d'acquérir des connaissances sur la variabilité interlocuteur et intra-locuteur, et de proposer des modèles du conduit vocal pour adapter un clone de référence à plusieurs locuteurs.

La principale difficulté de modéliser les contours du conduit vocal est la variabilité au niveau morphologique et celle des stratégies articulatoires des différents locuteurs.

Une question importante est ce que nous appelons le problème de la normalisation: comment les modèles d'un clone orofacial pour un locuteur spécifique peuvent être adaptés à d'autres locuteurs? Cette tâche est particulièrement difficile car elle implique de découvrir comment différents locuteurs avec des morphologies différentes peuvent produire des sons articulés considérés comme équivalents à des fins de communication de la parole.
Ce qui suit est un résumé de nos contributions et résultats.

Note : projet ARTIS. Le travail présenté dans cette thèse a constitué une contribution importante au projet ANR-08-EMER-001-02 ARTIS en collaboration entre le GIPSA-lab, le LORIA, IRIT et TSI-Télécom ParisTech. L'objectif principal de ce projet de recherche est de fournir de la parole augmentée au moyen d'une tête parlante.

9.2. Corpus articulatoire

Les données articulatoires constituent un élément crucial pour les modèles statistiques. Afin de décider quel type d'information doit être inclus dans le corpus, il faut d'abord clarifier le but des modèles. Les méthodes de décomposition linéaires, présentées dans les sections suivantes, ont pour but l'extraction des patrons articulatoires communs à plusieurs locuteurs francophones. Il y a trois questions principales à prendre en compte dans la construction d'un corpus pour la modélisation articulatoire: (1) la variabilité interlocuteur, qui se réfère à combien les locuteurs sont différents les uns des autres, doit être suffisamment grande pour extraire des patrons articulatoires qui soient aussi généraux que possible, (2) la couverture phonétique articulatoire, qui est liée à l'ensemble des sons de la parole produits par les locuteurs, devrait couvrir autant que possible la gamme des mouvements articulatoires présents dans la langue française, et (3) la taille de la base de données, qui est lié aux données articulatoires disponibles, devrait être assez grande pour que les modèles articulatoires puissent estimer des paramètres statistiques fiables. Cependant, en pratique, une session d'enregistrement de données articulatoires est limitée à environ deux heures. Ainsi, afin de remplir la condition de temps, la taille du corpus doit être limitée à certaines voyelles et consonnes en contexte vocalique. Les propriétés de ce corpus seront détaillées dans les sections suivantes.

9.2.1. Méthodes d'acquisition de données articulatoires

Les études antérieures sur la normalisation articulatoire ont été principalement basées sur trois méthodes d'enregistrement: rayons X, imagerie par résonance magnétique (IRM) et Articulographie Électromagnétique (EMA). Les rayons X ont été utilisés pour la première fois par Meyer (1907) et Mosher (1927). Les données recueillies au

moyen de cette méthode ont été utiles pour obtenir des représentations picturales du contour langue. Toutefois, les informations contenues dans ces images, même en étant relativement complètes, ne sont pas suffisantes pour identifier avec précision les contours du conduit vocal dans les images. L'IRM a été utilisée dans de très nombreuses études d'articulation de la parole depuis 1986 (Rokkaku et al., 1986). Cette méthode fournit des informations détaillées du conduit vocal. Néanmoins, le locuteur enregistré doit maintenir l'articulation pendant plusieurs secondes en raison de la vitesse d'acquisition relativement lente qui caractérise les systèmes d'IRM. Un autre inconvénient de l'IRM est que les locuteurs doivent être couchés sur le dos pour l'enregistrement. Les effets gravitationnels de cette posture peut avoir une certaine influence sur l'articulation (Tiede et al., 2000; Stone et al., 2007). L'articulographie électromagnétique offre une bonne solution pour suivre les mouvements articulatoires, mais son principal inconvénient est que des petites bobines réceptrices électromagnétiques doivent être collées sur les articulateurs d'intérêt. Ainsi, il est difficile de maintenir les capteurs fixés lors de sessions d'enregistrement de longue durée. En outre, les capteurs et fils dans la bouche peuvent perturber l'articulation naturelle. Par ailleurs, certaines régions du conduit vocal ne sont pas facilement accessibles pour coller les bobines réceptrices (i.e. le voile du palais, la partie arrière de la langue, etc.) Le Tableau 1 montre une comparaison des méthodes d'enregistrement mentionnées ci-dessus:

	EMA	**IRM**	**Rayons X**
Conduit vocal complet	Non	Oui	Oui
Image de la langue	Pellets	longueur totale	longueur totale
Résolution temporelle	500 Hz	0-24 Hz	30-60 Hz
Danger pour la santé	Non	Non	Oui
Qualité des signaux	Bonne	Bonne	Bonne
Mouvement de la tête	Limité	Limité	Libre
Portabilité	Non	Non	Non
coût	Oui	Oui	Oui

Tableau 1 - Comparaison des 3 méthodes d'enregistrement (modifiée à partir de(Ridouane, 2006))

Malgré ses inconvénients, l'IRM offre une information plus complète du conduit vocal comparé à EMA et fournit des données plus lisibles par rapport aux rayons X. Cela a motivé le choix de l'IRM comme méthode d'acquisition de données articulatoires dans cette étude.

9.2.2. Dispositif expérimental et protocole

Au cours d'une session d'enregistrement le locuteur passe par trois étapes différentes. Premièrement, le locuteur est installé dans les conditions les plus confortables possible sur un lit qui peut coulisser et transporter le locuteur dans l'aimant de l'image IRM. Comme l'imageur produit des sons qui pourraient être dommageables pour les oreilles humaines, le locuteur est protégé par des bouchons d'oreilles. Lors d'une session d'enregistrement, le locuteur peut appeler l'opérateur IRM à tout moment à l'aide d'un interphone. Deuxièmement, la position de la tête du locuteur est déterminée à l'aide de clichés d'exporation, et l'opérateur peut ainsi positionner les plans d'images souhaités pour le sujet, à savoir le plan médiosagittal dans notre cas. Pendant ce temps, le locuteur est demandé de ne pas bouger car les propriétés d'alignement de la machine sont en cours de définition. Enfin, le locuteur est chargé de prononcer et maintenir la forme du conduit vocal de certaines articulations pendant 8 secondes chacune. Les articulations incluses dans le corpus sont décrites dans la section suivante.

9.2.3. Base de données IRM

Dans cette étude, des données IRM ont été recueillies pour onze locuteurs francophones (six hommes: PB, YL, LH, RL, LD, BR et cinq femmes: HL, AA, MG, AK, et MGO). Trois bases de données contenant les mêmes informations ont été enregistrées pour le locuteur PB (PB-1998, PB-2002 et PB-2011). Idéalement, on aurait aimé enregistrer d'autres locuteurs pour rendre cette étude plus générale. Cependant, dans la pratique, trouver des locuteurs n'est pas toujours facile et la préparation d'une session d'enregistrement prend un certain temps. De même, on aurait aimé enregistrer toutes les combinaisons possibles de consonnes dans tous les contextes vocaliques existant en français. Mais, lors d'une session d'enregistrement, le locuteur en cours d'enregistrement ne doit pas ressentir trop de fatigue. Par conséquent, une séance d'enregistrement est limitée à deux heures pour les locuteurs les plus résistants. Ces faits ont forcé la réduction de la taille du corpus. Par exemple, il a été décidé de ne conserver que 13 voyelles parmi les 16 voyelles orales et nasales en français. La raison de cette diminution est que la plupart des locuteurs n'étaient pas capables de distinguer la prononciation des certaines voyelles. Par exemple: les voyelles nasales / ɛ / et / œ /, les voyelles orales / ə / et / œ / et les voyelles orales / a / et / ɑ /. Ainsi, il a été décidé de ne garder que / ɛ /, / œ / et / a /. Engwall et al. (2003) ont montré qu'un corpus limité, qui couvre l'espace articulatoire peut capturer les mêmes caractéristiques articulatoires qu'un corpus plus complet. Beautemps et al. (2001) ont montré qu'un modèle basé sur un corpus réduit pouvait reconstruire les données avec une précision proche de celle obtenue avec le modèle basé sur l'ensemble du corpus, à condition d'être choisi de manière optimale. Ainsi, également

dans cette étude, le corpus a été réduit à un seul représentant pour chaque type de consonne. Le corpus commun à tous les locuteurs dans notre étude est composé de 63 articulations: les 10 voyelles orales françaises /i e ɛ a y ø œ u o ɔ/, les 3 voyelles nasales /ã ɛ̃ ɔ̃/ et les 10 consonnes /p t k f s ʃ m n ʁ l/ articulées en contexte vocalique symétrique (voyelle-consonne-voyelle) pour les cinq voyelles /a e ɛ i u/.

9.2.4. Comparaison de notre corpus avec les corpus de la littérature

Le **Tableau 2** suivant montre la comparaison entre notre corpus et ceux dans la littérature. Cette comparaison est faite en termes de nombre de locuteurs, de taille des corpus et de nombre de mesures articulatoires. Comme on peut le voir, la présente étude comprend plus de locuteurs, avec des corpus plus grands que ceux de la littérature, comprenant des voyelles et des consonnes. Un apport important de notre étude est l'acquisition de données pour les contours complets du conduit vocal, ce qui nous permet de modéliser et d'étudier la synergie entre les différents organes impliqués dans la parole. Une autre contribution importante est l'inclusion de consonnes dans des contextes vocaliques, ce qui implique également un défi pour la modélisation.

Méthode d'enregistrement	Etude	No. Locuteurs	Taille du corpus	No. Points
EMA	Hoole (1998)	7	15 voyelles	4 capteurs
	Geng & Mooshammer (2000)	6	15 voyelles	4 capteurs
	Hu (2006)	7	10 voyelles	3 capteurs
Rayons X	Harshman et al. (1977)	5	10 voyelles	13 points
IRM	Hoole et al. (2000)	9	7 voyelles	13 points
	Zheng & Johnson (2003)	5	9 voyelles	13 points
	Ananthakrishnan et al. (2010)	3	13 voyelles	150 points
Nos Résultats				
MRI	Valdes (2013)	11	13 voyelles et 10 consonnes dans 5 contextes vocaliques	150 points

Tableau 2 - Comparaison entre notre corpus et ceux de la littérature

9.3. Le système de grille

Dans les sections suivantes, nous présentons les modèles articulatoires du contour de la langue. Afin de représenter les points du contour langue de différentes manières et calculer différentes mesures articulatoires, le système de grille proposé par Beautemps et al. (2001) a été utilisé (voir Fig 10).

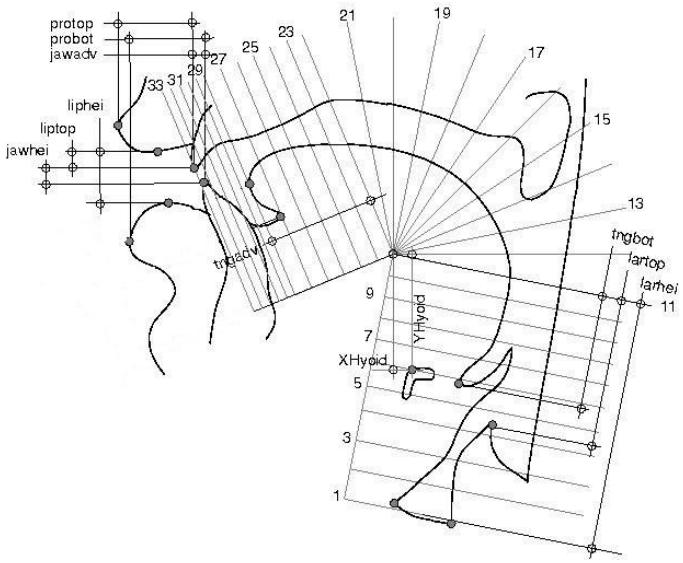


Fig 10 - illustration du système de grille pour représenter le contour de lalangue (à partir Beautemps et al., 2001)

Le système de grille a été mis en place une fois pour chaque locuteur. Tout d'abord, le centre de la grille a été défini comme la moyenne du centre de gravité de l'ensemble des contours de la langue. D'autre part, l'une des lignes centrales de la grille est orientée pour être parallèle à la position moyenne de la paroi pharyngée, alors que l'autre ligne centrale est orientée pour être parallèle au palais. Fig 11 montre un exemple du centre et de l'orientation de la grille de la locutrice AA. Notons que la ligne de grille n° 28 correspond à la pointe de la langue et la suit, et que la ligne n° 6 est fixée à la jonction entre la racine de la langue et de l'épiglotte.

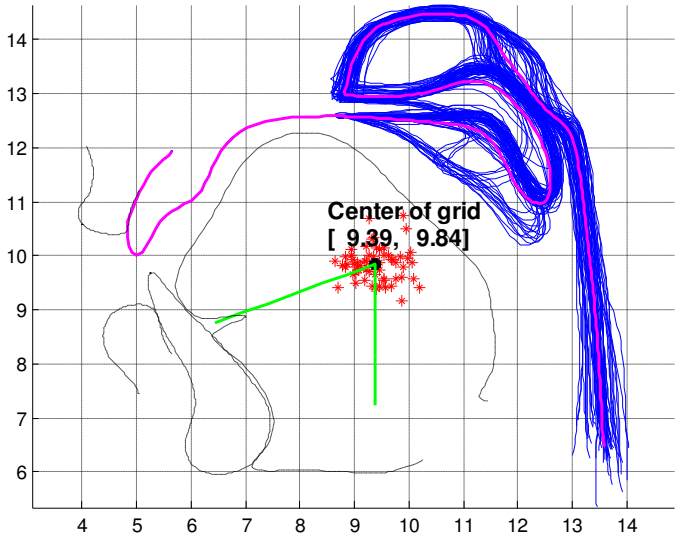


Fig 11 - Centre et orientation des lignes de la grille de la locutrice AA. Centre de gravité de chacune des articulations (étoiles rouges), centre de la grille (point noir), lignes centrales de la grille (vert), superposition de la paroi pharyngée de chaque articulation (bleu), moyenne de la paroi pharyngée (magenta)

9.3.1. Représentation du contour de la langue et d'échantillonnage

Le contour de la langue a été représenté de quatre manières. La première est une représentation de 200 points (x, y) appelée ‘full tongue contour’ (FullTng). FullTng est le contour de la mâchoire jusqu’à la jonction entre la racine de la langue et de l'épiglotte. Les 50 premiers points équidistants représentent la cavité sublinguale située entre le plancher de la bouche et la pointe de la langue. Les 150 points restants sont liés au contour de la pointe de la langue jusqu'à la jonction entre la racine de la langue et de l'épiglotte. La deuxième représentation du contour de la langue est appelée ‘Upper tongue contour’ (UpperTng). Cette représentation est définie par 150 points. UpperTng est le contour de la langue de la pointe jusqu'à la jonction entre la racine de la langue et de l'épiglotte. Les deux autres représentations sont basées sur un ré-échantillonnage d’UpperTng en utilisant le système de grille. La représentation INTRXY fait référence aux 23 points d'intersection entre les lignes de grille et le contour de la langue, de la sixième ligne à la 28ème ligne. Les coordonnées INT sont liées aux 23 distances entre les lignes centrales de la grille et le contour de la langue, de la sixième ligne à la 28ème ligne. Cette dernière représentation de la langue comprend des paramètres supplémentaires (*‘tongue bottom’* et *‘tongue advance’)* nécessaire pour définir la grille et modéliser les mouvements haut-bas et avant-arrière de la langue. Le paramètre tngbot est défini comme la distance entre la sixième ligne de grille, fixée au début de l'épiglotte, et la 11ème ligne de grille. L'avancement de la langue (tngadv) a été mesuré comme la distance entre la 22ème ligne de grille et la 28ème ligne de grille qui est fixé à la pointe de la langue.

9.4. Modèles linéaires individuels et multiples du contour de la langue

En vue de normaliser le contour de la langue de notre ensemble de locuteurs pour trouver des paramètres articulatoires communs, cette section présente les résultats de différents modèles linéaires construites à partir des données présentées dans la section 9.2.3. Tout d'abord, les méthodes utilisées sont expliquées. Ensuite, les méthodes sont comparées par rapport au nombre de coefficients qu’elles utilisent. En plus, cette section décrit comment la moyenne des données a était soustraite. Les modèles sont évalués au moyen de deux critères: la variance expliquée et l'erreur quadratique moyenne (RMSE). Les modèles ont également été évalués en utilisant une procédure de validation croisée leave-one-out (LOOCV) pour s'assurer qu'il n'y avait pas de sur-apprentissage. Finalement, les résultats des modèles linéaires pour les voyelles uniquement d’une part et pour les voyelles et consonnes ensemble d’autre part sont présentés.

9.4.1. Méthodes de décomposition linéaire

9.4.1.1. ACP

L'ACP est une méthode d'analyse souvent utilisée pour la réduction de dimensionnalité et l'analyse d'ensembles de données pour extraire des régularités (Pearson, 1901). Des mesures articulatoires $X_s = [x_1, x_2, ... , x_A]$ pour un locuteur donné s, qui se compose de N mesures pour chaque articulation de 1 à A, est décomposé dans un ensemble de paramètres de commande $\pi_s^{[A \times Cmp]}$ (ensemble de composantes Cmp qui expliquent les variations des articulations) et un modèle articulatoire $C_s^{[N \times Cmp]}$ (coefficients qui expliquent la contribution de chaque mesure articulatoire sur les composantes) par l'équation suivante:

$X_s = \pi_s * C_s^T + \xi_s$, où ξ est l'erreur résiduelle .

9.4.1.2. PARAFAC

La PARAFAC est une approche d'analyse en facteurs qui est souvent utilisée pour décomposer des données en 3 modes (Harshman, 1970; Harshman & Lundy, 1994). Dans notre cas spécifique, les trois modes sont liées aux articulations, mesures articulatoires et locuteurs, respectivement. La différence entre PARAFAC et ACP est que la PARAFAC extrait des patrons à partir de plusieurs locuteurs, alors que l'ACP ne décompose que les données d'un locuteur individuel. La décomposition des données d'un locuteur donné X_s, par PARAFAC, est représentée par l'équation suivante:

$X_s = \pi * \Phi_s * C^T + \xi_s$ où ξ est l'erreur résiduelle.

π est la matrice des paramètres de commande universels, qui représente les caractéristiques communes extraites pour tous les locuteurs. La matrice Φ est une matrice diagonale avec des poids spécifiques au locuteur par rapport à la contribution de chaque composante. La matrice des coefficients C, aussi appelée modèle articulatoire universel, représente la contribution de chaque mesure articulatoire aux composantes universelles extraites pour tous les locuteurs.

9.4.1.3. Décomposition de Tucker

La décomposition de Tucker est une extension de la PARAFAC (Tucker, 1966). Les matrices de paramètres de contrôle universels (Π), les poids spécifiques au locuteur (Φ) et les coefficients (C) représentent les mêmes modes que pour la PARAFAC. Par contre, ces matrices peuvent être décomposées avec un nombre différent de composantes chacune. En d'autres termes, la décomposition de Tucker permet l'extraction de nombres différents de composantes pour chaque dimension. Dans la PARAFAC toutes les dimensions sont décomposées avec le même nombre de

composantes. Les données d'un locuteur X_s donné, décomposé selon Tucker, peuvent être représentées comme suit:

$\sum_{i=1}^{wl}\sum_{m=1}^{w2}\sum_{n=1}^{w3} \pi * \Phi s * C * G + \xi_s$ où ξ est l'erreur résiduelle.

La matrice supplémentaire G contient des poids pour les trois modes de variation. Cette matrice représente l'interaction entre les composantes extraites pour chaque mode de variation.

9.4.1.4. ACP conjointe

Cette méthode a été proposée par Ananthakrishnan et al. (2010) et appelée ACP à deux niveaux. Nous utiliserons la dénomination d'ACP conjointe, qui reflète mieux la réalité. L'ACP conjointe est une extension de l'ACP pour décomposer les données de plusieurs locuteurs au lieu d'un locuteur individuel. Dans cette technique, les données sont décomposées en utilisant l'APC régulière, mais en imposant d'extraire un ensemble de paramètres de commande communs à tous les locuteurs. Les données de plusieurs locuteurs sont regroupées dans une matrice $X = [X_{s1}; X_{s2};. . . ; X_{sy}]$, qui contient les mesures articulatoires $X_s = [x_1, x_2, ..., x_A]$ pour les articulations de 1 à A de chacun des locuteurs.

9.4.2. Comparaison des méthodes linéaires par rapport au nombre de coefficients

Afin de disposer d'un critère d'évaluation pour les modèles des sections suivantes, les différentes méthodes linéaires doivent être comparées en termes de nombre de coefficients à déterminer. Le Tableau 3 montre le nombre de coefficients de chaque méthode en fonction du nombre de composantes extraites. Les calculs sont effectués en utilisant 11 locuteurs, 63 articulations et 2×150 mesures articulatoires (coordonnées x et y). Le nombre de coefficients est calculé comme décrit par les équations suivantes:

ACP = [(No. articulations × No. composantes) + (No. composantes × No. mesures articulatoires)] × No. Locuteurs

PARAFAC = (No. articulations × No. composantes) + (No. locuteurs × No. composantes) + (No. mesures articulatoires × No. composantes)

ACP conjointe = (No. articulations × No. composantes) + (No. composantes × No. mesures articulatoires × No. locuteurs)

TUCKER = [(No. articulations × No. composantes) + (No. composantes × No. composantes × No. locuteurs) + [(No. locuteurs × No. composantes) × (No. locuteurs × No. mesures articulatoires)]]

Selon les résultats dans le tableau suivant, les méthodes peuvent être classées en ordre descendant en fonction du nombre de coefficients comme: TUCKER, ACP, ACP

conjointe et PARAFAC. La décomposition de Tucker est la méthode qui utilise le plus de coefficients et la PARAFAC le moins de coefficients.

	Nombre de coefficients			
Nombre de composantes	**ACP**	**PARAFAC**	**ACP conjoint**	**TUCKER**
1	2343	224	1713	18224
2	4686	448	3426	36470
3	7029	672	5139	54738
4	9372	896	6852	73028
5	11715	1120	8565	91340
6	14058	1344	10278	109674
7	16401	1568	11991	128030
8	18744	1792	13704	146408
9	21087	2016	15417	164808
10	23430	2240	17130	183230
11	25773	2464	18843	201674
12	28116	2688	20556	220140
13	30459	2912	22269	238628
14	32802	3136	23982	257138
15	35145	3360	25695	275670
16	37488	3584	27408	294224
17	39831	3808	29121	312800
18	42174	4032	30834	331398
19	44517	4256	32547	350018
20	46860	4480	34260	368660
21	49203	4704	35973	387324

Tableau 3 - Nombre de coefficients de chaque méthode en fonction du nombre de composantes extraites par ACP, PARAFAC, ACP conjointe et TUCKER

9.4.3. Soustraction de la moyenne et orthogonalité

Miranda et al. (2008) ont réalisé une étude sur la question de soustraction de la moyenne des données préalable à l'ACP. Dans cette étude, ils discutent de la nécessité de soustraire la moyenne pour trouver les composantes principales qui minimisent l'erreur quadratique moyenne. La soustraction de la moyenne assure que la première composante principale correspond à la direction de variance maximale. Si les données ne sont pas centrées, la première composante ACP pourrait plutôt être liée à la moyenne des données. Par conséquent, dans notre étude, les modèles ont été construits avec des données centrées. Les données de chaque locuteur ont été centrées en soustrayant la moyenne de chaque mesure articulatoire. Une autre contrainte

importante est la condition d'orthogonalité. Par définition en ACP, les composantes extraites sont orthogonales entre eux. PARAFAC, Tucker et l'ACP conjointe ont également été mis en place pour extraire des composantes orthogonales et les données ont été centrées.

9.4.4. Évaluation des modèles: variance expliquée et erreur de reconstruction

Les différents modèles présentés dans cette étude ont été validés en fonction de la variance relative expliquée. La variance expliquée est donnée par le ratio de la variance des données prédites (X_p) sur la variance des données originales (X), comme montré dans les équations suivantes:

$$\mathrm{VARIANCE}(X) = \frac{\sum_1^n \sum_1^m (X_i - \overline{X_i})^2}{n.m}$$

$$\mathrm{VARIANCE\_EXPLAINED}(X, X_p) = \frac{\mathrm{VARIANCE}(X_p)}{\mathrm{VARIANCE}(X)}$$

Par ailleurs, l'erreur quadratique moyenne (RMSE) a également été utilisée pour mesurer la précision des données reconstruites à partir des modèles. L'erreur quadratique moyenne, pour un locuteur X donné, est calculée selon l'équation suivante:

Erreur quadratique moyenne : $\sqrt{\frac{\sum_1^n \sum_1^m (X_i - X_{i\_predicted})^2}{n.m}}$

n étant le nombre d'observations et m le nombre de mesures d'articulatoires.

9.4.5. Validation croisée

Selon Hawkins (2004), du sur-apprentissage pourrait se produire si un modèle représente très précisément les données d'apprentissage, mais ne parvient pas à faire des prédictions pour de nouvelles données car il n'a pas appris à généraliser. Afin d'éviter le sur-apprentissage, il est nécessaire d'utiliser des méthodes d'évaluation comme la validation croisée (LOOCV). Le but de l'utilisation d'une telle méthode est de vérifier les capacités du modèle à généraliser en évaluant ses performances sur des données qui n'ont pas été utilisés pour la construction du modèle.

Les modèles présentés dans cette thèse sont réalisés et évalués au moyen de la méthode de validation croisée « leave-one-out cross validation » (LOOCV). La LOOCV est une méthode dans lequelle une observation de la base de données est enlevée, le modèle est construit à partir des données restantes et est utilisé pour prédire l'observation enlevée ; ce processus est ensuite répété pour chaque observation

de l'ensemble de données. La LOOCV est utile pour déterminer le nombre de prédicteurs à utiliser. Par exemple, l'erreur quadratique moyenne aura la tendance à diminuer si des prédicteurs valables sont ajoutés, mais elle augmentera si des prédicteurs inutiles sont ajoutés. En effet, l'augmentation du nombre de prédicteurs pourrait conduire à un modèle dégénéré (Riu & Bro, 2003).

9.4.6. Modèles linéaires du contour de la langue pour les voyelles

Pour faire une comparaison équitable de nos résultats avec ceux de la littérature, nous avons établi un modèle limité aux 10 voyelles orales français /i e ɛ a y ø œ u o ɔ/. Deux versions du contour de la langue ont été utilisées: l'UpperTng et un contour de la langue sous-échantillonné à 3 points. Afin de sous-échantillonner le contour de la langue, trois points équidistants ont été sélectionnés à partir de la pointe de la langue vers l'arrière, pour chaque locuteur. Un modèle PARAFAC avec 2 composantes a pu reconstruire le contour de la langue avec une erreur quadratique moyenne de 0,25 cm pour UpperTng, alors que l'erreur quadratique moyenne pour le contour de la langue sous-échantillonné à 3 points était de 0,23 cm, ce qui représente une variance expliquée de 77,3% et 84,6%, respectivement. Le Tableau 4 montre que, dans l'ensemble, nos résultats sont comparables à ceux rapportés dans la littérature. Le défi est alors d'étendre cette analyse à un corpus comprenant des consonnes dans différents contextes vocaliques (63 articulations), comme expliqué dans les sections suivantes.

Type	Etude	No. locuteurs	Corpus	No. Points	Variance Exp
EMA	Hoole (1998)	7	15 voyelles	4 capteurs	80.0%
	Geng & Mooshammer (2000)	6	15 voyelles	4 capteurs	96.0%
	Hu (2006)	7	10 voyelles	3 capteurs	90.0%
Rayons-X	Harshman et al. (1977)	5	10 voyelles	13 points	92.7%
IRM	Hoole et al. (2000)	9	7 voyelles	13 points	87.0%
	Zheng & Johnson (2003)	5	9 voyelles	13 points	76.2%
	Ananthakrishnan et al. (2010)	3	13 voyelles	150 points	71.0%
Nos résultats					
IRM	Valdés (2013)	11	10 voyelles	3 points	84.6%
		11	10 voyelles	150 points	77.3%

Tableau 4 - Comparaison de nos résultats avec la littérature en utilisant PARAFAC avec 2 composantes

Afin d'avoir une référence pour comparer les modèles présentés dans les sections qui suivent, nous avons construit des modèles avec 11 locuteurs, seulement les 10 voyelles françaises / i e ɛ a y ø œ u o ɔ/, 150 mesures articulatoires, et testé plusieurs

méthodes linéaires. Les modèles ont été construits avec la représentation des données UpperTng. Les résultats en termes de variance expliquée et de RMSE sont comparés dans Fig 12. On voit qu'il est possible de classer les méthodes selon leur performance de la meilleure à la moins bonne : APC, ACP conjointe, Tucker et PARAFAC. D'autres comparaisons par rapport à ces modèles sont présentées dans les sections suivantes.

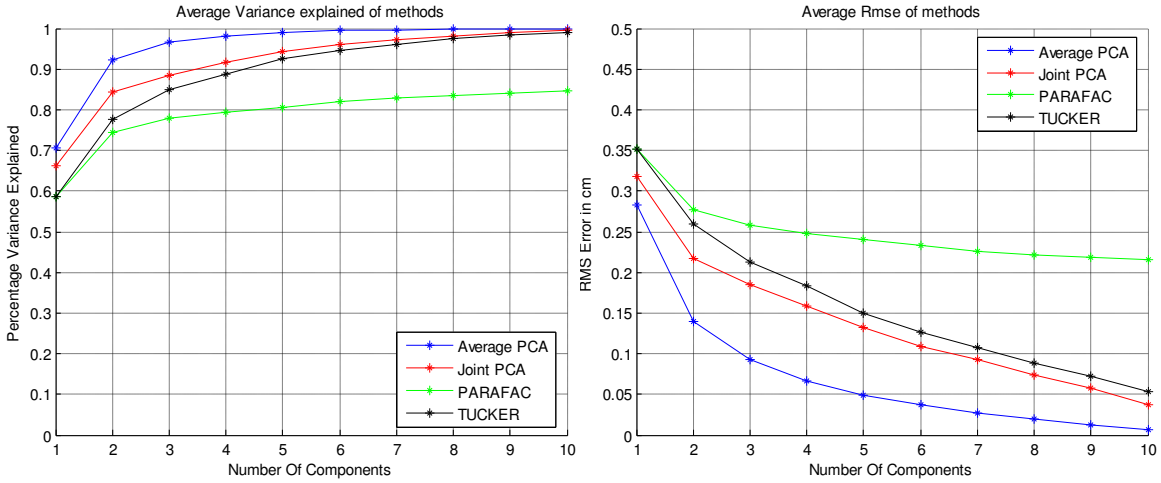


Fig 12 - *Performance, établie à l'aide de LOOCV, pour la PARAFAC, Tucker et l'ACP conjointe en fonction du* nombre de composantes pour les contours de la langue avec un corpus de seulement voyelles. A gauche: la variance expliquée. A droite: RMSE en centimètres

9.4.7. Modèles linéaires du contour de la langue étendus aux consonnes

Dans cette section, les modèles linéaires des locuteurs individuels et des modèles multiples, qui prennent en compte plusieurs locuteurs, sont présentés et comparés. Un modèle final est sélectionné et évalué. Le corpus est composé de 63 articulations du français: 10 voyelles orales /i e ɛ a y ø œ u o ɔ/, 3 voyelles nasales /ã ɛ̃ ɔ̃/ et 10 consonnes /p t k f s ʃ m n ʁ l/ articulés en contextes symétriques voyelle-consonne-voyelle (VCV) avec les cinq voyelles /a e ɛ i u/.

Les modèles présentés dans cette section ont été construits à partir de plusieurs représentations du contour de la langue (UpperTng, INTRXY et INT), basé sur le système de grille expliqué dans la section 9.3. Fig 13 illustre la performance de toutes les méthodes en termes d'explication de la variance et d'erreur quadratique moyenne de reconstruction des données.

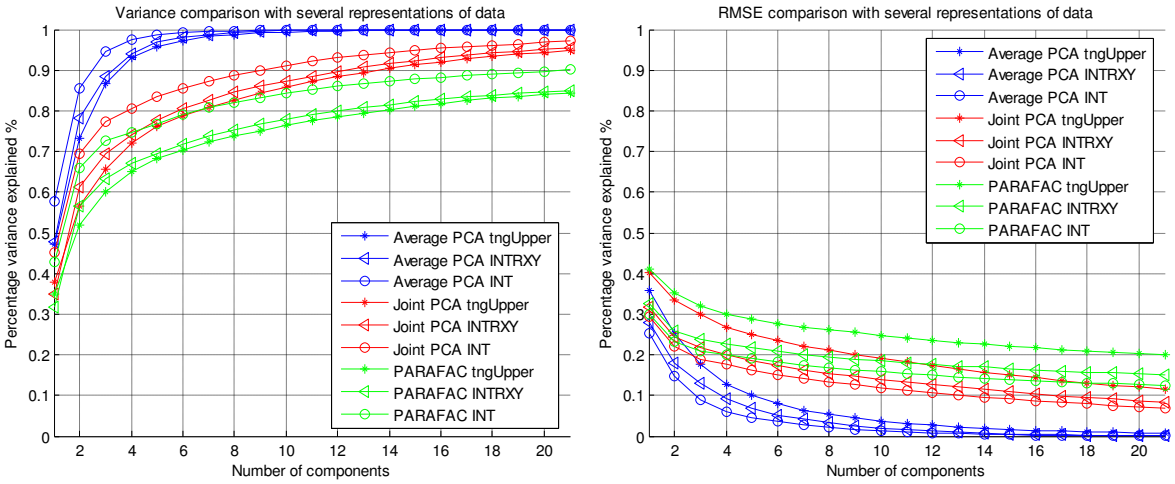


Fig 13 - Performances, établie à l'aide de LOOCV, des multiples méthodes de décomposition linéaires avec plusieurs représentations de données en fonction du nombre de composantes pour un corpus comprenant voyelles et consonnes. A gauche: la variance expliquée. A droite: RMSE en centimètres.

Nous avons également comparé les modèles restreints aux voyelles (section 9.4.6) avec les modèles étendus aux consonnes en contexte vocalique, en utilisant seulement la représentation UpperTng. Notez qu'il y a une diminution de l'explication de la variance et une augmentation de l'erreur de reconstruction pour les modèles étendus aux consonnes. En utilisant 4 composantes, le modèle ACP moyen explique 98,16% pour les voyelles et 93,23% pour les modèles étendus aux consonnes, avec une RMSE de 0,06 cm et 0,1 cm respectivement. L'ACP conjointe explique 91,66% pour les voyelles et 72,16% pour les modèles étendus aux consonnes, ce qui représente une RMSE de 0,15 cm et 0,26 cm respectivement. PARAFAC explique 79,42% pour les voyelles et 65,15% pour les modèles étendues aux consonnes, avec une RMSE 0,24 cm et 0,30 cm respectivement. Enfin, la méthode de Tucker explique 88,82% pour les voyelles et 70,17% pour les modèles étendues aux consonnes, avec une RMSE 0,18 cm et 0,27 cm respectivement.

Vu que la méthode de Tucker obtient à-peu-près la même performance que l'ACP conjointe, mais est plus complexe et utilise beaucoup plus de coefficients comparé aux autre méthodes, nous avons décidé de ne plus l'inclure dans les analyses suivantes.

Les modèles ACP individuels ont été utilisés comme modèles de référence pour évaluer la performance de différents modèles multiples, comme on peut le voir sur la Fig 13. Le test de Student a été utilisé pour déterminer, pour chaque méthode, le nombre de composantes qui donnent une RMSE non statistiquement différente de celle obtenue par les modèles de référence ACP. Pour l'ACP avec UpperTng, quatre composantes ont été choisies comme modèle de référence pour calculer le test de Student. Par ailleurs, pour les modèles ACP avec INTRXY et INT, trois composantes étaient suffisantes pour expliquer la même variance que l'ACP avec UpperTng. Ainsi,

le modèle ACP, avec 3 composantes, a été choisi comme le modèle de référence pour INTRXY et INT pour calculer le test de Student.
Selon le test de Student, la PARAFAC nécessite plus de 21 composantes pour atteindre une explication de la variance de UpperTng de 84,52%. L'ACP conjointe nécessite entre 14 et 21 composantes, en fonction du locuteur, ce qui représente une explication de la variance entre 90,33% et 94,88%. L'ACP conjointe avec UpperTng nécessite le nombre minimum de composantes pour le locuteur YL (au moins 14 composantes) et le maximum pour la locutrice AK (21 éléments).
Selon le test de Student pour les modèles avec INTRXY, PARAFAC nécessite plus de 21 composantes en représentant une explication de la variance de 88,35%. L'ACP conjointe nécessite entre 12 composantes et 21 composantes, ce qui représente une explication de la variance de 89,63% et 95,46%, respectivement. L'ACP conjointe nécessite le nombre minimum de composantes pour le locuteur LH (au moins 12 composantes) et le maximum des composantes pour les locuteurs PB, AK et MGO (21 composantes).
Selon le test de Student pour les modèles avec INT, PARAFAC nécessite plus de 21 composantes en représentant une explication de la variance de 90,07%. D'autre part, L'ACP conjointe nécessite entre 12 et 21 composantes, ce qui représente une explication de la variance de 93,03% et 97,12%, respectivement. L'ACP conjointe nécessite le nombre minimum de composantes pour la locutrice MG (au moins 12 composantes) et le maximum pour la locutrice AK (21 éléments). Le tableau suivant résume l'ensemble des résultats expliqués ci-dessus pour le test de Student.
Nous pourrions conclure que le ré-échantillonnage du contour de la langue avec INTRXY et INT ne constitue pas un avantage pour la modélisation. En ré-échantillonnant le contour de la langue nous gagnons peu d'extra explication de la variance et nous perdons d'informations. Ainsi, en tenant compte des conclusions de la section 9.4.2 sur le nombre de coefficients utilisés par chaque méthode linéaire, l'ACP conjointe avec UpperTng semble être la solution optimale.

Types des données	ACP moyenne		PARAFAC		ACP conjoint	
	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.	Nb. cmp	Var. Exp.
UpperTng	4	93.23%	21	84.52%	14 - 21	90.33% - 94.88%
INTRXY	3	88.35%	21	85.07%	12 - 21	89.63% - 95.46%
INT	3	94.52%	21	90.07%	12 - 21	93.03% - 97.12%

Tableau 5 - Résumé du nombre de composantes nécessaires pour les méthodes linéaires multiples (PARAFAC et ACP conjoint) pour atteindre la même performance de l'APC de référence, selon un test de Student pour chaque représentation de données

Nous voyons dans le Tableau 5 que le nombre de composantes nécessaires pour l'ACP conjointe est beaucoup plus élevé que le nombre de composantes nécessaires pour un modèle ACP individuel. Nous avons donc essayé d'analyser la signification

des composantes communes à tous les locuteurs, à partir de l'ACP conjointe, en calculant les corrélations entre les 4 composantes obtenues par ACP et les 21 composantes obtenues par l'ACP conjointe. Les corrélations ont indiquées que les 4 premières composantes de l'ACP conjointe peuvent être interprétées en termes de hauteur de la mâchoire (JH), mouvement du corps de la langue (TB), mouvement du dos de la langue (TD) et mouvement de la pointe de la langue (TT).

Enfin, on a testé des approches de normalisation basées sur la projection des points dans des espaces articulatoires crées a partir des composantes extraits par ACP. L'objectif de ces techniques était de trouver la projection correspondante d'un point donné dans un espace cible. La projection a été trouvée en utilisant les pondérations des k voisins les plus proches. Nous avons conclu que les espaces vocaliques JH vs. TB de nos locuteurs n'étaient pas suffisamment homogènes pour faire de bonnes prédictions.

9.5. Normalisation géométrique

Dans cette section, nous décrivons une méthode de normalisation géométrique appelée analyse procustéenne. Ensuite, la performance des modèles linéaires construits à partir des données alignées au moyen de Procruste, est comparée à celle des modèles basés sur les données originales.

L'analyse procustéenne est une méthode qui détermine une transformation linéaire d'un ensemble de points Y pour mieux se conformer à l'ensemble des points X (Ross). La transformation linéaire appliquée par Procuste est composé par une combinaison itérative de translations, rotations, et écaillage. Le critère d'ajustement est défini par la somme carrée des erreurs. Toutefois, l'analyse Procuste est limitée à l'alignement des deux ensembles de données. Une technique plus avancée, appelée analyse procustéenne généralisée (Generalised Procrustes Analysis, GPA), permet l'alignement de plusieurs ensembles de données (Stegmann & Gomez, 2002). L'alignement par GPA implique un processus itératif en quatre étapes. Tout d'abord, une première série de données est choisie. Ensuite, tous les ensembles de données restants sont alignés par rapport à la série de données initiales. Troisièmement, la moyenne estimée est calculée à partir des formes alignées. Enfin, si la moyenne estimée a changé, l'étape 2 est répétée. Lorsque la moyenne n'a pas changé de manière significative dans la dernière itération, on considère que l'algorithme a convergé. La Fig 14 montre comment l'articulation /i/ a été alignée pour les locuteurs PB, YL et HL au moyen de la GPA. Seules les coordonnées du contour de la langue ont été alignées. Les 63 articulations des nos onze locuteurs ont été alignés entre eux.

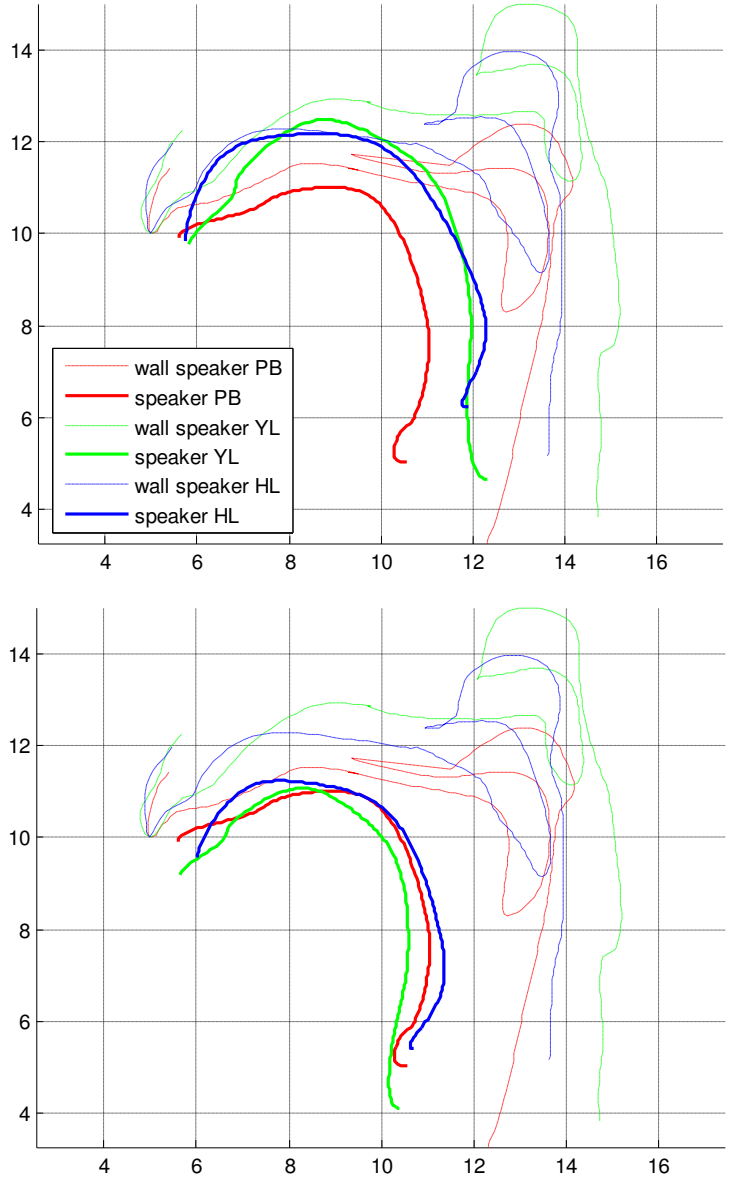


Fig 14 - Alignement du contour de la langue au moyen de GPA pour l'articulation / i / des locuteurs PB, YL et HL. En haut: données initiales. En bas: données alignées à l'aide de GPA

Fig 15 illustre la performance de toutes les méthodes linéaires, décrites dans la section 9.4.1, en termes de variance expliquée et d'erreur quadratique moyenne pour *l'UpperTng* alignés de tous les locuteurs. Pour chaque méthode linéaire, les étapes pour calculer l'explication de la variance et RMSE étaient les suivantes: (1) construction du modèle, (2) reconstruction des données à partir du modèle, (3) transformation des données prédites vers l'espace de coordonnées originales en appliquant l'inverse de GPA (GPA^{-1}), enfin calcul de la variance expliquée et de la RMSE comme expliqué dans la section 9.4.4. Fig 15 montre que l'alignement des contours de la langue, entre tous les locuteurs, apporte très peu d'amélioration à la performance des modèles linéaires en termes de la variance expliquée et de RMSE.

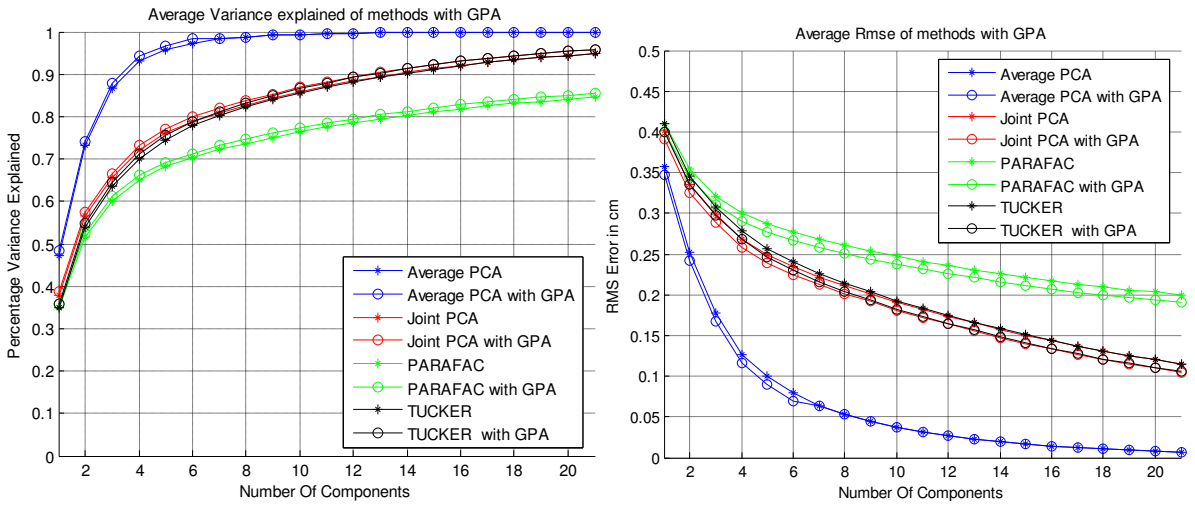


Fig 15 - Performance, établie à l'aide LOOCV, de la PARAFAC, Tucker et ACP conjointe en fonction du nombre de composants pour les contours de la langue aligné au moyen de GPA. A gauche: la variance expliquée. A droite: RMSE en centimètres

9.6. Modèles linéaires individuels et multiples des contours des lèvres et voile du palais

Comme les sections précédentes présentent des modèles linéaires du contour de la langue, il est également intéressant de modéliser d'autres contours du conduit vocal qui sont importants pour l'articulation en parole. Par exemple, les lèvres, en termes acoustiques, visent en particulier à assurer une constriction pour les voyelles fermées et les consonnes fricatives labio-dentales (Fant, 1960). Un autre organe important est le contour de voile du palais qui contrôle la nasalité dans l'articulation (Serrurier & Badin, 2005). Dans cette section, nous décrivons les modèles construits pour les lèvres et le contour du voile du palais.

9.6.1. Modèles linéaires pour les lèvres

Fig 16 montre la performance de toutes les méthodes linéaires, décrites dans la section 9.4.1, en termes d'explication de la variance et l'erreur quadratique moyenne pour les contours des lèvres. Un modèle à trois composantes a été pris comme référence de comparaison. L'explication de la variance obtenue pour les modèles individuels ACP des lèvres supérieure et inférieure, avec 3 composantes, atteint 94,9% et 94,5%, respectivement. L'erreur quadratique moyenne associée est 0,03 cm et 0,05 cm, respectivement. Nous voulions maintenant extraire un ensemble de composants articulatoires communs à tous les locuteurs. Ainsi, les lèvres ont également été modélisées à l'aide de plusieurs méthodes linéaires tels que: PARAFAC, Tucker et ACP conjointe.

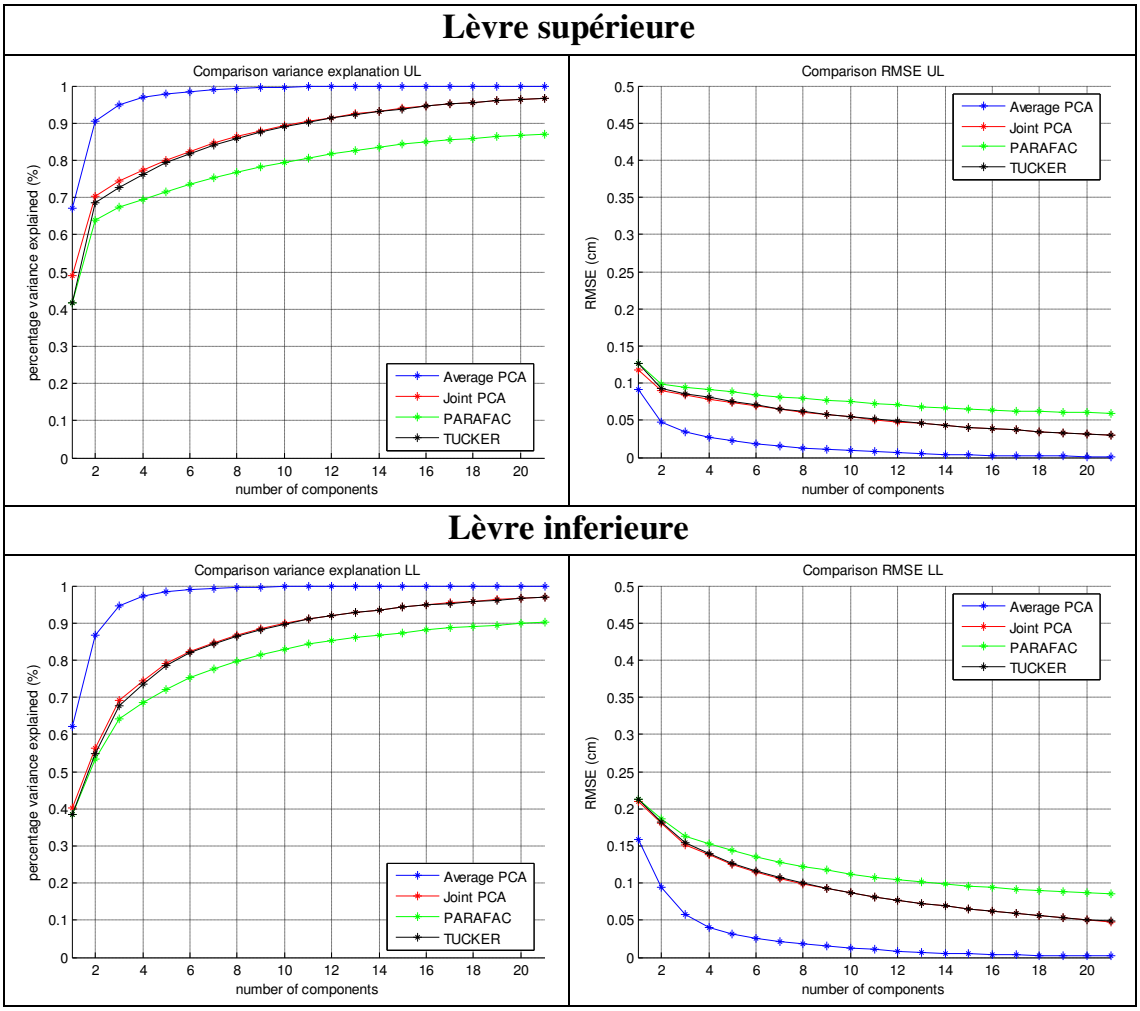


Fig 16 - *Performance, établie à l'aide de LOOCV, de l'ACP , PARAFAC, Tucker et ACP conjointe* en fonction du nombre de composants pour les contours des lèvres pour un corpus comprenant des voyelles et consonnes. *En haut: la variance expliquée et l'erreur quadratique moyenne du contour de lèvre supérieure. En bas: la variance expliquée et l'erreur* quadratique moyenne du contour de lèvre inférieure.

Un test de Student a été utilisé pour déterminer le nombre de composants qui donne un RMSE pas statistiquement différent de celui obtenu par le L'ACP de référence individuelle avec trois composants, pour chaque méthode linéaire. Le **Tableau 6** résume les résultats du test de Student. Nous avons d'abord observé que la courbe d'explication de la variance de TUCKER affiche une performance très similaire par rapport à l'ACP conjoint. Par conséquent, il a été décidé de garder l'ACP conjointe et de ne plus utiliser TUCKER pour le test de Student. PARAFAC nécessite plus de 21 composants pour être équivalente à la performance de l'ACP. Les modèles PARAFAC construits avec 21 composantes ont représenté une explication de la variance de 87,02% et 90,11% pour la lèvre supérieure et inférieure, respectivement. L'ACP conjointe nécessite entre 15 et 21 composants pour la lèvre supérieure et entre 11 et 21 composants pour la lèvre inférieure, selon le locuteur. Fig 17 montre la gamme de composants nécessaires par l'ACP conjoint, pour chaque locuteur, pour égaler la performance de l'ACP selon le test de Student. En ce qui concerne la

modélisation de la lèvre supérieure, l'ACP conjointe nécessite le minimum nombre de composants pour les locuteurs PB et RL (au moins 15 composantes) et le maximum pour la locutrice MG (21 composant), ce qui représente une variance expliquée entre 93,4% et 96,7%. D'ailleurs, en ce qui concerne la modélisation de la lèvre inférieure, l'ACP conjointe nécessite le nombre minimum de composantes pour la locutrice AA (au moins 11 composantes) et le maximum pour la locutrice MG (21 composantes), ce qui représente une variance expliquée entre 91,2% et 96,9 %. D'après le test de Student, l'ACP conjointe est la solution optimale pour modéliser les lèvres supérieure et inférieure car Il a fallu moins de composantes par rapport à PARAFAC.

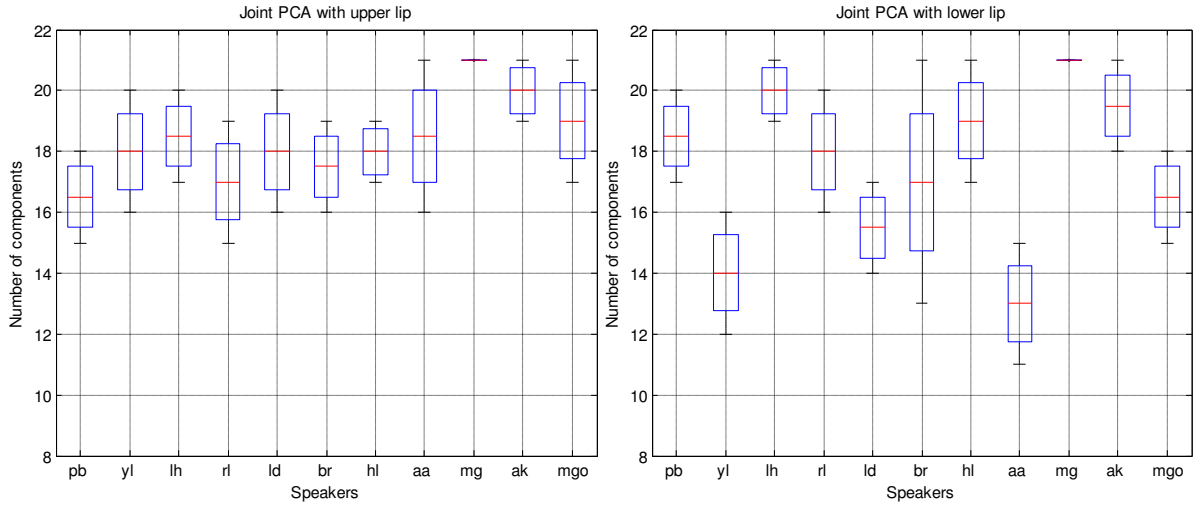


Fig 17 - Variation de nombre de composants nécessaires pour modèles de lèvre inférieure et supérieure , selon un test de Student entre le ACP de référence, avec 3 composants, et conjointe ACP. À gauche: nombre de composants nécessaires pour le modèle lèvre supérieure. Droite: nombre de composants nécessaires pour le modèle lèvre inférieure

données	**ACP moyen**		**PARAFAC**		**ACP conjointe**	
	Ref. cmp	**Var. Exp.**	**Nb. cmp**	**Var. Exp.**	**Nb. cmp**	**Var. Exp.**
Lèvre sup.	3	94.9%	21	87.02%	15 - 21	93.4% - 96.7%
Lèvre inf.	3	94.5%	21	90.11%	11 - 21	91.2% - 96.9%

Tableau 6 - Résultats de test de Student entre l'ACP de référence, avec 3 composants, et les méthodes linéaires multiples (PARAFAC et ACP conjointe), pour les lèvres supérieure et inférieure

9.6.2. Modèles linéaires pour le voile du palais

Fig 18 montre la performance de toutes les méthodes linéaires multiples, décrites dans sa section 9.4.1, en termes d'explication de la variance et d'erreur quadratique moyenne pour le contour du voile du palais. Deux composantes ont été prises comme référence de comparaison. Ces deux composantes représentent un mouvement oblique lié au muscle levator veli palatini et à la fermeture du port nasopharyngeal par un mouvement horizontal, comme décrit par Serrurier et Badin (2005; 2008). L'explication de la variance obtenue pour tous les locuteurs avec des modèles ACP individuels avec deux composantes atteint 90%. La RMSE associé était de 0,08 cm. Nous voulions extraire un ensemble de composantes articulatoires commun à tous les

locuteurs. Donc, le contour du voile du palais a également été modélisée au moyen de plusieurs méthodes linéaires telles que: PARAFAC, Tucker et ACP conjointe.

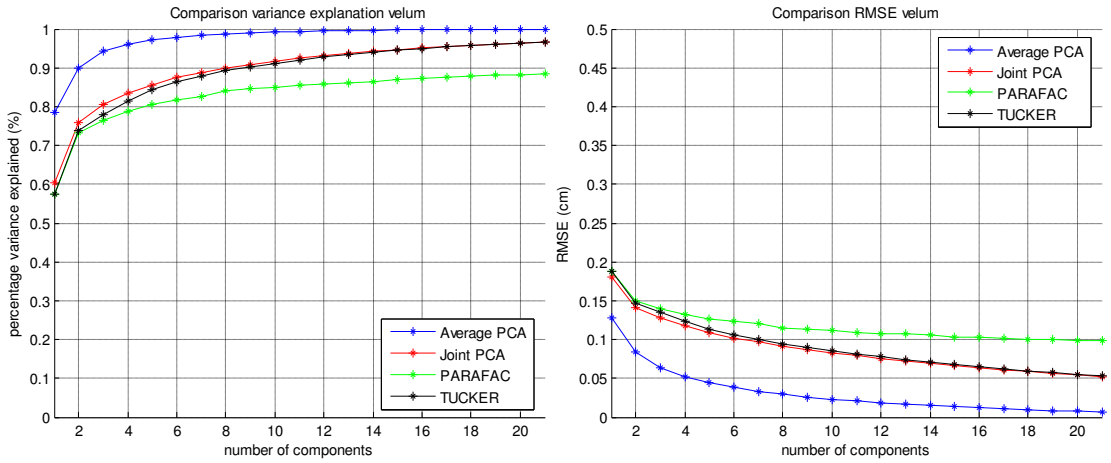


Fig 18 - Performance, établie à l'aide *de LOOCV, pour l'ACP moyen, PARAFAC, TUCKER et ACP conjointe* en fonction du nombre de composantes pour le contour de la voile du palais pour un corpus avec des voyelles et consonnes. Gauche: explication de la variance. A droite: RMSE

Un test de Student a été utilisé pour déterminer le nombre de composantes pour chaque méthode linéaire qui donne un RMSE pas statistiquement différent de celui obtenu par les modèles de référence ACP individuels. Le **Tableau 7** résume les résultats du test de Student. TUCKER a affiché une performance très similaire par rapport à l'ACP conjoint. Ainsi, TUCKER a été écartée et non pris en compte pour une analyse plus approfondie avec le test de Student. PARAFAC nécessite entre 4 et 21 composantes, en fonction du locuteur. Fig 19 montre la gamme de composantes nécessaires par PARAFAC, par chaque locuteur, pour égaler la performance de l'ACP en fonction du test de Student. PARAFAC nécessite le nombre minimum de composants pour les locuteurs RL, LD et MG (au moins 4, 10 et 10 composantes pour chaque locuteur, respectivement) et le maximum pour les autres locuteurs (21 composantes) représentant une variance expliquée entre 78,9% et 88,41%. L'ACP conjointe nécessite entre 1 et 14 composantes pour égaler la performance des modèles de référence ACP, selon le locuteur. Fig 19 montre la gamme de composants nécessaires par l'ACP conjoint, pour chaque locuteur, pour égaler la performance de l'ACP. L'ACP conjointe nécessite le minimum nombre de composantes pour le locuteur LD (au moins 1 composante) et le maximum pour la locutrice AK (14 composantes) représentant une variance expliquée entre 60,02% et 94,2%. Selon les résultats du test de Student, l'ACP conjointe est la solution optimale pour modéliser le contour de la voile du palais car il nécessite moins de composantes que PARAFAC.

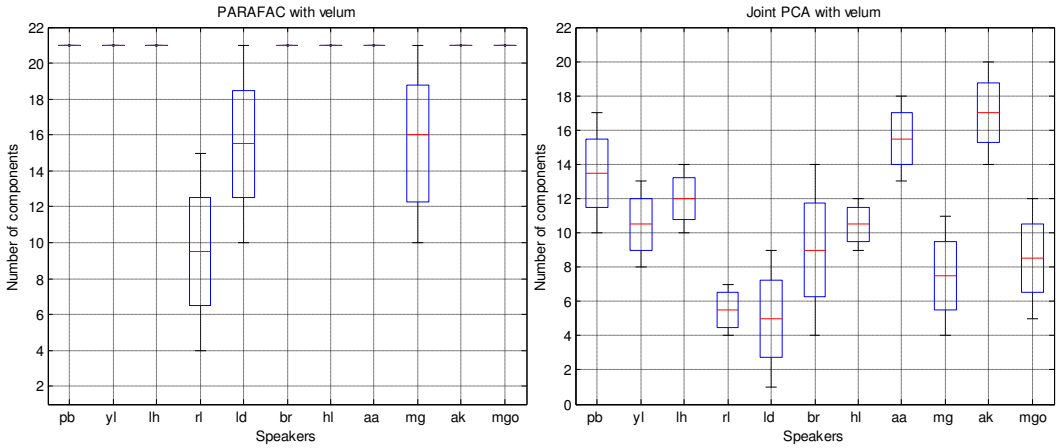


Fig 19 - *Variation du nombre de composantes nécessaires pour la PARAFAC et l'ACP conj*ointe selon un de test de Student entre la ACP de référence, avec 2 composantes, et les méthodes PARAFAC et ACP conjointe pour le contour de la voile du palais

données	**ACP moyen**		**PARAFAC**		**ACP conjointe**	
	Ref. cmp	**Var. Exp.**	**Nb. cmp**	**Var. Exp.**	**Nb. cmp**	**Var. Exp.**
Voile	2	90%	4 - 21	78.9% - 88.41%	1 - 14	60.02% - 94.2%

Tableau 7 - Résultats de test de Student entre l'ACP de référence, à 2 composants, et les méthodes (PARAFAC et ACP conjoint), pour le contour de la voile du palais

9.7. Conclusions et perspectives

9.7.1. Conclusions

Le double objectif de cette thèse était d'acquérir des connaissances sur la variabilité inter-locuteur, et de proposer des modèles pour adapter un clone de référence, composé des modèles articulatoires (lèvres, langue, voile du palais, etc.), à une variété de locuteurs en utilisant des méthodes de décomposition linéaire. Ce travail a été mené dans le cadre du développement d'un système de retour articulatoire visuel, basé sur un locuteur donné, qui anime automatiquement une tête parlante 3D à partir du son de la parole. Ainsi, l'idée principale était d'adapter ce système à la morphologie et aux stratégies articulatoires de plusieurs locuteurs. Les applications envisagées se situent dans le domaine de la prononciation assistée par ordinateur et la réhabilitation de la parole.

Afin de construire des modèles de contours pour les différents articulateurs du conduit vocal, nous avons acquis des données qui couvrent l'espace articulatoire de la langue française. Nous avons recueilli un corpus de 11 locuteurs français (6 hommes et 5 femmes), prononçant 63 articulations, incluant les voyelles et les consonnes en contexte vocalique, enregistrées au moyen d'un système d'imagerie par résonance magnétique (IRM). Les contours de 12 articulateurs du conduit vocal ont été inclus (lèvres, palais, voile du palais, pharynx, mâchoire, os hyoïde, langue, épiglotte, glotte, larynx, trachée et moelle épinière). Ces données constituent l'un des apports important

de ce travail puisqu'elles nous ont permis de modéliser et d'étudier la synergie entre les différents organes impliqués dans la production de la parole. Une autre contribution importante est l'inclusion de 10 consonnes dans des contextes vocaliques, ce qui figure rarement dans la littérature, et implique également un défi pour la modélisation. Outre les études de Hoole (1998) qui a construit des modèles pour les consonnes (/p t k/), et l'étude de Geng & Mooshammer (2000) qui comprenait des modèles pour la consonne /t/, seuls Ananthakrishnan et al. (2010) ont couvert un ensemble plus complet de 10 consonnes (/p t k f s ʃ m n ʁ l/). En outre, notre base de données comprend un vaste ensemble de locuteurs par rapport à toutes les études de la littérature. Nos données ont été utiles pour caractériser et comparer nos locuteurs au moyen des statistiques sur les mesures articulatoires, et pour construire des modèles basés sur plusieurs méthodes de décomposition linéaire multiple.

L'étape suivante a été la caractérisation des stratégies articulatoires de nos locuteurs. Des modèles individuels d'ACP guidée ont été construits pour la langue, les lèvres et le voile du palais. Les modèles de tous les locuteurs ont été analysés et comparés. En observant les nomogrammes des modèles, qui sont des représentations graphiques des composantes extraites, nous avons observé que chaque locuteur a sa propre stratégie pour atteindre des articulations qui sont considérées comme équivalentes du point de vue de la communication parlée. Par exemple, la variabilité du contour de la langue a été décomposée en quatre composantes principales: hauteur de la mâchoire, corps, dos et pointe de la langue. Les mouvements associés sont effectués dans des proportions différentes selon le locuteur. Ainsi, pour un déplacement donné de la mâchoire, la langue peut globalement se déplacer dans une proportion qui dépend du locuteur. Nous avons également remarqué que la protrusion des lèvres, l'ouverture des lèvres, l'influence du mouvement de la mâchoire sur les lèvres, et la stratégie articulatoire de la voile du palais peuvent également varier en fonction du locuteur. Par exemple, certains locuteurs replient leur voile du palais contre la langue pour produire la consonne /ʁ/. Les conséquences acoustiques des stratégies des différents locuteurs ont également été comparées au moyen des graphiques dans l'espace F2-F1. Une analyse acoustique pour les voyelles a montré une cohérence avec les analyses précédentes rapportées dans la littérature. Par ailleurs, une analyse acoustique et articulatoire de la consonne /k/ a montré que cette consonne peut être articulée d'une manière palatale ou vélaire, selon le contexte de la voyelle et le locuteur. Ces résultats constituent une contribution importante à la connaissance de la variabilité interlocuteur dans la production de la parole.

Nous avons testé des méthodes non linéaires basées sur la projection d'un point donné dans l'espace articulatoire d'un locuteur source sur son point correspondant dans

l'espace articulatoire d'un locuteur cible. Les espaces articulatoires étaient représentés par les composantes corp de la langue (axe X) et la hauteur de la mâchoire (axe Y), extraites par les modèles ACP guidées de chaque locuteur. Nous avons remarqué que les espaces articulatoires de nos locuteurs ne sont pas suffisamment homogènes pour faire de bonnes prédictions.

Afin d'extraire un ensemble de paramètres de contrôle articulatoire communs à tous les locuteurs, nous avons construit des modèles linéaires pour la langue, les lèvres et le voile du palais, basés sur diverses méthodes de décomposition linéaires (PARAFAC, Tucker et ACP conjointe). Pour chaque méthode de décomposition, un test de Student avec un seuil de signification de 5% a été utilisé pour déterminer le nombre de composantes qui donne une erreur quadratique non statistiquement différente de celle obtenue par les modèles d'ACP individuels de chaque articulateur: langue, lèvre supérieure, lèvre inférieure et voile du palais. Les résultats ont montré que l'ACP conjointe était la solution optimale pour la modélisation de ces contours:

- Pour le contour de la langue, les modèles des différents locuteurs nécessitent 44 composantes (4 composantes * 11 locuteurs) au total pour modéliser tous les locuteurs, ce qui représente une variance expliquée moyenne de 93,23% et une erreur quadratique moyenne de 0,13 cm. L'ACP conjointe nécessite de son côté 21 composantes pour modéliser tous les locuteurs, pour une variance expliquée de 94,88% et une erreur de 0,12 cm, tandis qu'un modèle avec seulement 4 composantes aboutit à une variance expliquée de 72,16% et une erreur de 0,27 cm. Par ailleurs, les corrélations entre les composantes de l'ACP conjointe et les composantes de l'ACP guidée ont montré que les 4 premières composantes de l'ACP conjointe peuvent être interprétées en termes de hauteur de la mâchoire, corps, dos et pointe de la langue.

- Pour les contours des lèvres supérieure et inférieure, les modèles des différents locuteurs nécessitent 33 éléments (3 éléments * 11 locuteurs) au total pour modéliser tous les locuteurs, ce qui conduit à des variances expliquées moyennes de 94,89% et 94,5%, et des erreurs de reconstruction moyennes de 0,03 cm et 0,05 cm, respectivement. L'ACP conjointe de son côté nécessite 21 composantes pour modéliser tous les locuteurs, pour une variance expliquée de 96,67% et 96,85%, avec une erreur quadratique moyenne de 0,03 cm et 0,04 cm, respectivement, tandis qu'un modèle, avec seulement 3 composantes, aboutit à une variance expliquée de 74,28% et 69,26%, avec une erreur quadratique moyenne de 0,08 cm et 0,15 cm, respectivement. Par ailleurs, les corrélations entre les composantes de l'ACP conjointe et les composantes d'ACP guidée ont révélé que les premières composantes de l'ACP conjointe

peuvent être globalement interprétés en termes de hauteur de la mâchoire, protrusion et hauteur des lèvres.

- Pour le contour du voile de palais, les modèles des locuteurs individuels nécessitent 22 composantes (2 composantes * 11 locuteurs) au total pour modéliser tous les locuteurs, ce qui représente une variance expliquée moyenne de 90% et une erreur moyenne de 0,08 cm. L'ACP conjointe nécessite de son côté 14 composantes pour modéliser tous les locuteurs, pour une variance expliquée de 94,2% et une erreur de 0,07 cm, tandis qu'un modèle, avec seulement 2 composantes aboutit à une variance expliquée de 76,01% avec une erreur de 0,14 cm. Par ailleurs, les corrélations entre les composantes de l'ACP conjointe et les composantes des modèles ACP individuels ont indiqué que les premières composantes extraites par l'ACP conjointe peuvent être approximativement interprétés en termes d'un mouvement oblique et un mouvement horizontal du voile du palais.

Le **Tableau 8** résume les résultats expliqués ci-dessus. Notez qu'il y a une réduction considérable du nombre de composantes lors de l'utilisation d'ACP conjointe par rapport aux modèles d'ACP individuels. En outre, le **Tableau 9** compare nos modèles construits pour voyelles et consonnes avec les modèles décrits dans la littérature pour les voyelles seulement (voir Tableau 4). Notez que les modèles PARAFAC décrits dans la littérature utilisent 2 composantes pour modéliser entre 7 et 15 voyelles, pour des ensembles de 3 à 9 locuteurs, ce qui conduit à une variance expliquée entre 71% et 96%. Ainsi, nos deux modèles d'ACP conjointe avec 4 et 21 composantes, construit pour un corpus de 63 articulations, incluant les voyelles et les consonnes, qui représentent respectivement 72,16% et 94,88% de la variance des données, sont comparables en termes d'explication de la variance avec les modèles décrits dans la littérature, construits pour les voyelles seulement.

Une autre contribution importante de ce travail de thèse est la modélisation des lèvres inférieure et supérieure ainsi que du voile de palais pour extraire un ensemble de patrons articulatoires communs à plusieurs locuteurs. A notre connaissance, il n'existe pas dans la littérature sur la normalisation d'autres études de ces contours importants du conduit vocal.

Contour	Average PCA			Joint PCA according to Student's t-test			Joint PCA with reduced no. of components		
	No. Components	Variance Exp.	RMSE	No. Components	Variance Exp.	RMSE	No. Components	Variance Exp.	RMSE
Upper tongue	44 (4 *11)	93.23%	0.13 cm	21	94.88%	0.12 cm	4	72.16%	0.27 cm
Upper lip	33 (3*11)	94.89%	0.03 cm	21	96.67%	0.03 cm	3	74.28%	0.08 cm
Lower lip	33 (3*11)	94.50%	0.05 cm	21	96.85%	0.04 cm	3	69.26%	0.15 cm
Velum	22(2*11)	90%	0.08 cm	14	94.20%	0.07 cm	2	76.01%	0.14 cm

Tableau 8 - *Comparaison entre les modèles d'ACP moyen et l'ACP conjointe pour la langue, la lèvre* supérieure, la lèvre inférieure et le voile du palais

	Modèles for voyelles	**Modèles pour un corpus de voyelles et consonnes**	
Méthode	**PARAFAC**	**ACP conjointe selon test de Student**	**ACP conjointe avec composantes réduites**
No. composantes	2	21	4
Variance Exp.	71% - 96%	94.88%	72.16%
Corpus	7 - 15 vowels	63 articulations (vowels and consonants)	63 articulations (vowels and consonants)
No. locuteurs	3 - 9 speakers	11 speakers	11 speakers

Tableau 9 - Comparaison entre les modèles PARAFAC construites pour les voyelles, rapportées dans la *littérature, et nos modèles d'ACP conjointe construites pour consonnes et voyelles, pour le contour langue*

9.7.2. Perspectives

Les résultats présentés dans ce manuscrit sont prometteurs mais nécessitent une recherche plus approfondie. Le premier point à considérer est l'augmentation du nombre de locuteurs. En effet, une question importante est la variabilité inter-locuteur qui fait référence à combien les locuteurs sont différents les uns des autres. La variabilité inter-locuteur doit être suffisamment grande pour permettre d'extraire des modèles articulatoires aussi généraux que possible. Ainsi, un nombre plus grand de locuteurs devrait être intégré aux modèles présentés dans cette étude.

Dans le présent travail, l'analyse acoustique était très limitée; aucunes données acoustiques n'étaient disponibles. Les travaux futurs devraient donc étudier les conséquences acoustiques des stratégies articulatoires particulières de chaque locuteur. Un aspect important de l'acoustique du conduit vocal est lié aux contours de lèvres. Dans cette étude, les lèvres n'ont pas été incluses. Uniquement les sections du système de grille entre la pointe de la langue et l'épiglotte ont été prises en compte. Ainsi, il serait pertinent d'étendre notre système de grille jusqu'aux lèvres. Nous pourrions alors analyser et comparer le formant F3 de tous les locuteurs, qui est généralement lié à l'arrondissement des lèvres (Hardcastle & Marchal, 1990).

Une autre question importante est la modélisation du voile de palais. D'une part, les modèles individuels du voile de palais nécessitent 22 composantes (2 composantes * 11 locuteurs) au total pour modéliser tous les locuteurs. D'autre part, l'ACP conjointe nécessite 14 composantes pour modéliser le voile de palais de tous les locuteurs. Toutefois, une meilleure performance de modélisation pourrait être prévue. Les travaux futurs pourraient se concentrer sur la variabilité inter-locuteurs du voile de palais. Une meilleure compréhension du comportement du voile est nécessaire pour proposer de nouvelles solutions de modélisation. Par exemple, le voile du palais peut être en quelque sorte poussé vers l'arrière par l'action de la langue pour certaines articulations. Par ailleurs, il n'est pas exclu que certains locuteurs utilisent un certain degré de nasalité, même pour des articulations non nasales. Tous ces faits introduisent une variabilité dans les données qui rend la modélisation plus complexe.

La plupart des méthodes utilisées dans cette étude sont linéaires. Néanmoins, on peut supposer que les méthodes linéaires n'offrent pas la meilleure solution pour modéliser la variabilité entre les différents locuteurs, en particulier en présence de consonnes. Ainsi, de futurs travaux devraient être réalisés à l'aide des méthodes non linéaires. Une première étape vers ce but est de modéliser les locuteurs individuels au moyen d'une méthode non-linéaire comme l'ACP à noyaux (Schölkopf et al., 1998;.. Mika et al., 1999).

Enfin, des données pourraient être obtenues au moyen d'autres techniques comme l'articulographie électromagnétique ou l'IRM en temps réel (Narayanan, 2004; 2011), qui permettent d'obtenir de beaucoup plus grandes quantités de données grâce à des vitesses d'acquisition plus rapides que l'IRM statique. Ces données pourraient être utilisées pour construire des modèles stochastiques. Des méthodes d'apprentissage statistique comme les modèles de Markov cachés (HMMs) et les modèles de mélange gaussien (GMMs) pourraient être utilisées pour estimer des caractéristiques articulatoires directement à partir des données (Toda et al., 2008; Zen et al., 2009; Ling et al., 2010; Ben Youssef et al., 2011b).