Journal of Phonetics 107 (2024) 101374

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Formant-based articulatory strategies: Characterisation and inter-speaker variability analysis

Antoine Serrurier *, Christiane Neuschaefer-Rube

Clinic for Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Aachen 52057, Germany

ARTICLE INFO

Article history:
Received 13 October 2023
Received in revised form 15 October 2024
Accepted 18 October 2024

Keywords:
Vocal tract
Acoustics
Formant
Articulation
Articulatory-acoustic relationship
Vowel

ABSTRACT

Vowels are articulatorily characterised by the shape of the vocal tract and acoustically by their three lowest formants. The relationship between formant variations and articulatory variations is well documented. This study addresses the opposite problem: describing the main articulatory variations associated with the variations of single formants. A data-driven modelling-based approach was chosen for this purpose. Midsagittal vocal tract contours from the glottis to the lips for 532 vowels from 41 speakers of three different languages were obtained from MRI data. Corresponding formant values were obtained by acoustic modelling. For each speaker, linear regressions of the contours on the formant values were performed. It led to five articulatory components, characterising the vocal tract variations associated with variations of the first three formants and their differences. Inter-speaker variability was analysed by applying principal components analysis on the components in a second level of modelling. A correlation analysis of the resulting inter-speaker components with morphological features was performed to determine whether a speaker's strategy could be driven by the morphology. Results show that the palate shape and the vertical pharyngeal height, related to the male–female difference, have a small influence on the speaker's strategy. Associated Matlab code is publicly available.

1. Introduction

In acoustics, vowels are characterised by their formants (Ladefoged & Johnson, 2011). As emphasised by Ladefoged (Ladefoged & Johnson, 2011), the notion of formant had already been suggested by Isaac Newton around 1665. Since then, many studies have formalised the concept of formants (Vilain, Berthommier, & Boë, 2015), and it is well-known that vowels are identified by their two to three lowest formants (Peterson & Barney, 1952; Ladefoged & Johnson, 2011). In articulatory phonetics, vowels are characterised by the shape of the vocal tract (Ladefoged & Maddieson, 1996). Since early studies in phonetics (Bell, 1867) and advanced articulatory analyses (Russell, 1929), vowels can clearly be classified according to the shape and position of the speech articulators (Ladefoged & Maddieson, 1996; Ladefoged & Johnson, 2011).

The understanding of the relationships between articulation and acoustics relies on the pioneering works of Chiba and Kajiyama (1941) and Fant (1960) on the perturbation and source-filter theories. In articulatory acoustics, the formants are associated with the acoustic resonances of the various cavities of the vocal tract (Fant, 1960; Stevens, 2000). Further, the movements of the vocal tract articulators can be associated with formant changes. Nowadays, the impacts of the vocal tract movements on the few lowest formats are well understood (Stevens, 2000).

The first formant, F1, corresponding to the lowest resonance of the vocal tract, varies for adults between 270 Hz and 850 Hz in average in the reference study of Peterson and Barney (1952). As the lowest formant, it can be considered as the most important, carrying on average 80% of the energy (Ladefoged & Johnson, 2011). Irrespective of the vowel's type, an increase of F1 is achieved by a lowering of the jaw and the tongue, corresponding to an increase of the size of the mouth cavity (Ladefoged & Johnson, 2011). This articulatory property

* Corresponding author.
E-mail address: aserrurier@ukaachen.de

https://doi.org/10.1016/j.wocn.2024.101374
0095-4470/

is referred to as the height of the vowels and F1 is inversely correlated with the vowel height (Ladefoged & Johnson, 2011). The second formant F2, corresponding to the second lowest resonance of the vocal tract, varies for adults between 840 Hz and 2790 Hz in average according to Peterson and Barney (1952). A decrease of F2 is achieved by a backward movement of the tongue, referred to as the backness property (Ladefoged & Johnson, 2011). F2 and the vowel backness are therefore positively correlated. This correlation is however not as good as the correlation between F1 and the height (Ladefoged & Johnson, 2011) as F2 is also affected by the lip rounding, which contributes to its lowering (as well as all the other formants to various extents (Stevens, 2000)). Strictly speaking, vowel backness is more closely related to the difference between F1 and F2, where the effect of lip rounding is eliminated (Ladefoged & Johnson, 2011). The third formant F3, corresponding to the third lowest resonance of the vocal tract, varies for adults between 1690 Hz and 3310 Hz in average (Peterson & Barney, 1952). F3 is less unequivocally related to an articulatory feature than F1 and F2. F3 conveys little information regarding height and backness (Ladefoged & Johnson, 2011). It is usually associated with the third features of the vowels, roundedness, in an inverse correlation (Ladefoged & Johnson, 2011). Its impact is however more dependant on the type of vowels than F1 and F2. In addition, it is often related to retroflex sounds (Ladefoged & Johnson, 2011) with an inverse correlation between F3 and the r-colour of a vowel. As for F2 with F1, F3 is also sometimes considered in interaction with the previous formant F2, coming for instance in proximity of F2 to make more impact on the F2-F3 spectral peak (Aaltonen, 1985; Stevens, 2000).

The representation of the vowels on the trapeze chart (Jones, 1917; Pfitzinger & Niebuhr, 2011) summarises the height and backness properties of the vowels. Remarkably, the representation of the vowels in the F1-F2 formant chart (Peterson & Barney, 1952) tends to superpose itself with the articulatory trapeze chart, despite some differences (Ladefoged & Johnson, 2011). The chart is sometimes represented in 3D to take into account the roundedness in the third dimension (Ladefoged & Johnson, 2011). Complementarily, 3D charts in the F1-F2-F3 plane have already been attempted (Boë, Perrier, Guérin, & Schwartz, 1989), but remain rarely used in phonetics due to the smaller amount of cross-vowels acoustic and articulatory features in F3.

In the acoustic theory of speech production, the vocal tract from the glottis to the lips is approximated by the concatenation of cylindrical tubes of various cross-sectional areas. The variation of the cross-section area along the cylinder midline from the glottis to the lips is the area function. As little as two to four tubes are enough to acoustically reproduce the various sounds generated by a vocal tract (Fant, 1960). In this configuration, the resonance cavities of the vocal tract are represented by single tubes. The frequency resonances, i.e. the formants, can be theoretically recovered from the tubes' size by solving simplified forms of the sound wave propagation equations (Fant, 1960; Stevens, 2000). In complement, the wave propagation equations can be solved for the concatenation of tubes, providing the acoustic transfer function of the whole vocal tract and in consequence its resonances. This makes the link between the articulation (through the area function) and the formants (the cavity resonances).

In this framework, exploring the articulation-formant relationship becomes straightforward: The effects of the modification of the area function can be directly measured on the formants. This relies on the calculation of so-called *sensitivity functions* (Fant, 1980). This is the current approach for establishing the articulation-formant relationships (Fant, 1960; Stevens, 2000). This modelling-based approach mirrors the speech production process, where the speaker sets the vocal tract and generates the sound corresponding the articulation. While this approach has proven to be powerful, simple articulatory variations can however have complex acoustic consequences. In addition, simple formant variations can be the result of several concomitant articulatory variations. As a consequence, finding the optimal multiple articulatory variations resulting in the increase or decrease of a single formant is the consequence of a complex trial and error process. Various simple articulatory variations have to be explored as well as their combination, without guarantee to capture at the end the most effective articulatory variations associated with a formant. In other words, the articulation-formant relationship problem is traditionally tackled from one perspective only: Observing formant variations from simple articulatory variations. Addressing the opposite problem, i.e. determining the articulatory variations associated with the variations of a formant, is handled through a complex and uncertain trial and error process. As a consequence, while an effect on the formants of an articulatory change is well-known, the full characterisation of all articulatory changes involved in the modification of one formant is surprisingly not explicitly addressed in the literature.

We intend in this study to address exactly this opposite problem: Given the increase or decrease of a single formant, deriving the associated articulatory variations. The associated articulatory variations are referred to in this paper as the *articulatory attributes* of the formant. Although it is naturally expected to have a broad overlap with existing common knowledge, such as obtaining primarily the tongue lowering while F1 increases, the current study aims at uncovering the articulatory attributes of the few lowest formants and characterising them accurately from the glottis to the lips. This is the complementary approach than what has been done in previous studies, and aims at determining how speakers organise the different articulatory variations associated with the variations of a single formant. For instance, while a decrease of F2 is obtained by bringing the tongue backward, it can also be partly achieved by rounding the lips. How do speakers combine these two strategies to effectively achieve a decrease of F2? Are other secondary articulatory strategies involved? Surprisingly, it seems that this has not been addressed so far in the literature and appears to us important, as the implemented strategy might differ between speakers. In other words, the articulatory impacts on the formants are well known but the articulatory attributes of the formants remain largely unexplored. In addition, as emphasised earlier, the difference between the first and second and between the second and third formants may play a role as important as the formants themselves. Thus, this study intends to characterise also the articulatory attributes of

the differences between F1 and F2 and between F2 and F3. In this paper, these differences are referred to as $\Delta F1F2$ and $\Delta F2F3$. In analogy with the formants they are named Δ*formants* and are treated in the same way as formants. Evidently, during production, speakers do not process formants independently but rather as a whole to achieve specific targets (Fant, 1960). Regardless of the speaker intention, there exists however an optimal articulatory variation resulting in the variation of a formant or a Δformant. Recovering these articulatory variations constitute the first target of the paper. It must be noted that the recovered articulatory variations associated with the variations of a formant might also be associated with variations of other formants. Indeed, articulatory variations, even very simple ones, have usually impact on several formants, if not all. This is not in contradiction with the objectives of the paper, which aims at determining the main articulatory variations associated with the variation of a formant, regardless of their impacts on the other formants. At first, this approach can look similar to the acoustic-to-articulatory inversion problem, which consists in recovering the vocal tract shape, or the area function, from speech acoustic inputs (Wakita, 1973). Traditional approaches rely on the same principles of acoustic modelling presented above (Wakita, 1973; Fant, 1980). However, acoustic-to-articulatory inversion aims at recovering an unknown vocal tract shape from an acoustic input (possibly formant values), facing the one-to-many problem: Several different vocal tract shapes can lead to similar acoustic inputs (Atal, Chang, Mathews, & Tukey, 1978). In the current study, the articulatory-acoustic mapping is *a priori* known, the question being to recover the optimal variations of the vocal tract shape that correspond to a variation of the acoustic input.

Speakers vary significantly from one another both in articulation (e.g. Johnson, Ladefoged, & Lindau, 1993; Serrurier, Badin, Lamalle, & Neuschaefer-Rube, 2019) and acoustics (e.g. Peterson & Barney, 1952) while realising a similar speech task. As a consequence, the articulatory attributes of the formants and Δformants are expected to be to a certain extent speaker-dependant. In addition to determining the articulatory attributes of the formants and Δformants, our second objective will thus be to explore their inter-speaker variability. Several speakers will be considered and the inter-speaker variability will be analysed qualitatively and quantitatively.

Inter-speaker variability can be attributed to the social background of the speaker or to her/his anatomic morphology, thereafter the *morphology* (Ladefoged & Broadbent, 1957). It has been shown that the morphology influences the articulations produced by the speakers (see Serrurier & Neuschaefer-Rube, 2023 for a recent review). It is therefore expected that the articulatory attributes of the formants and Δformants differ between speakers according to the morphology, to a certain extent. The current study intends to relate the inter-speaker variability of the articulatory attributes of the formants and Δformants to the inter-speaker morphological variability. It will take advantage of our recent work on the morphology of the vocal tract (Serrurier & Neuschaefer-Rube, 2023), which proposes a morphological model of the vocal tract and relates the components to morphological features. The current study will investigate whether these features, that characterise the morphology of a vocal tract, are also related to the inter-speaker variability of the articulatory attributes. This intends to reveal which morphological feature influences the articulatory attributes of the formants and Δformants, to which extent, and to characterise this influence. In comparison to previous studies that aim to determine whether the morphology influences the formant values, this on the contrary aims to determine whether the morphology influences the strategy implemented by a speaker to control her/his formant values.

In summary, the aims of the study are: (1) to characterise precisely the cross-vowel articulatory attributes of the F1, F2, F3 formants and the ΔF1F2, ΔF2F3 Δformants for the whole vocal tract from the glottis to the lips, (2) to analyse qualitatively and quantitatively the inter-speaker variability of the attributes of these formants /Δformants, and (3) to uncover the potential relationships between this inter-speaker variability and the morphological inter-speaker variability. These three objectives enhances therefore common knowledge of phonetics. First, rather than determining the way variations of the vocal tract impact the formants, it determines the strategy implemented by the speakers to control in practise their formants. Second, it analyses how this strategy varies between speakers. And third, it investigates whether the morphology of a speaker influences this strategy.

In addition, the study adopts a data-driven approach based on real vocal tract contours, as opposed to artificial area functions, which are simplified representations of the vocal tract for the purpose of acoustic analysis, commonly targeted for this purpose. Technically, full contours of the vocal tract for several speakers sustaining oral vowels were obtained. The corresponding F1, F2, F3, ΔF1F2 and ΔF2F3 formants and Δformants were estimated by means of sound wave propagation simulation. The articulatory attributes of each formant and Δformant for each speaker were obtained by linear regressions of the vocal tract contours on the formant and Δformant values. For the increase or decrease of a formant/Δformant, the linear regression provides the variations of the articulatory contours it is correlated with, i.e. the articulatory attributes of the formants/Δformants. The articulatory attributes of each formant/Δformant remain however speaker-dependant. To deal with this issue, an analysis of variability of the linear regressions were performed, inspired by the concept of *model of models* proposed in Serrurier et al. (2019). Briefly, the linear regressions represent the first level models, which predict the variations of the vocal tract contours from variations of formant/Δformant frequency values. The parameters of these models, i.e. the coefficients of the linear regressions, are themselves analysed statistically per formant/Δformant by means of Principal Component Analysis (PCA), the number of observations corresponding to the number of speakers. This constitutes the second level of modelling. The components correspond to the main modes of variation of the linear regressions between speakers. It characterises the inter-speaker variability of the articulatory attributes of each formant and Δformant. Finally, based on the morphological parameters described in Serrurier and Neuschaefer-Rube (2023), a correlation analysis were carried out to uncover the relationships between the morphological characteristics of the speakers and the inter-speaker components of the second-level models. This reveals how the inter-speaker variability of the articulatory attributes of the formants and Δformants are driven by the morphology.

For reproducibility and transparence, the Matlab code associated with the results is publicly available, together with the articulation contours, on the following link: https://github.com/tonioser/FormantComponents.

A preliminary version of this work has been published in Serrurier and Neuschaefer-Rube (2022). This study constitutes a very large extension of this initial work.

2. Material and methods

2.1. Data

The data considered for the study are static midsagittal Magnetic Resonance Imaging (MRI) data of 41 speakers articulating the oral vowels of their native language. The data form a subset of the data presented in Serrurier and Neuschaefer-Rube (2023). Three different native languages are considered: French (5 females, 7 males, [i, e, ɛ, a, y, ø, œ, u, o, ɔ]), German (4 females, 8 males, [aː, eː, iː, oː, uː, ɛː, øː, yː, a, ɛ, ɪ, ɔ, ʊ, ʏ, œ, ə]) and English (9 females, 8 males, [iː, ɪ, ei, ɛ, æ, ɑː, ʌ, ɔː, oʊ, uː, ʊ, ɹ]), for a total of 18 females and 23 males. The speakers sustain each of the oral vowels of their native language for about 5 to 10 s. The midsagittal images encompass the entire vocal tract from the glottis to the lips and have a resolution between 1 mm and 1.56 mm depending on the dataset. An example is visible in Fig. 1. Altogether, 532 images are considered for this study, ranging from 10 to 16 per speakers, with an average of 12.98 images per speaker. For each image, the contours of the vocal tract articulators from the glottis to the lips are available from a previous study (Serrurier & Neuschaefer-Rube, 2023). In the current study, only a restriction of these contours to the sections surrounding the vocal tract have been considered, excluding for instance the inferior part of the tongue or the lower part of the mandible. The extended contours are considered for illustration purposes but analyses and measurements are performed on the vocal tract restriction only. An illustration of these contours is visible in Fig. 1. Altogether, the data consist of p = 41 speakers $\times$ 10–16 articulations $\times$ n = 1036 vocal tract contour points $\times$ 2 x-y coordinates. For conciseness, the set of articulation point coordinates are referred to as *articulation* in the paper.

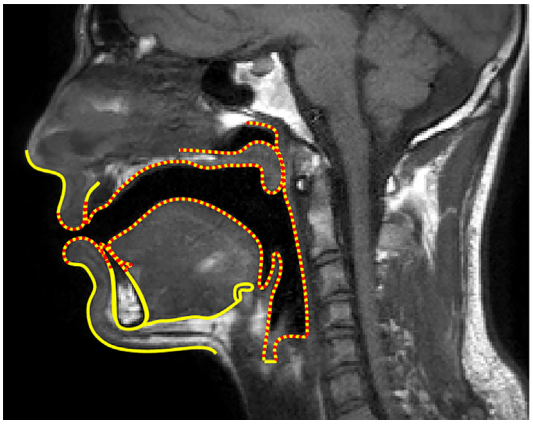

Fig. 1. MRI image of a French male speaker articulating [œ] superimposed with the contours (solid yellow lines) and their restriction to the vocal tract points (dashed red lines).

Further details regarding the data and the manual annotations are provided in Serrurier and Neuschaefer-Rube (2023).

2.2. Method

The method consists in determining the linear articulatory components corresponding to F1, F2, F3, ΔF1F2 and ΔF2F3 variations for each speaker and to analyse qualitatively and quantitatively their common ground and the inter-speaker variability.

2.2.1. Formant articulatory components

The articulatory components were obtained by a linear regression of the articulations on the formant frequencies. The steps involved in the process are described in the following.

- The formant frequencies were obtained by simulating the propagation of a plane acoustic wave in the vocal tract and by calculating the acoustic transfer function between the glottis and the lips. Such acoustic simulations have proven to be powerful to recover the formant frequencies up to 5000 Hz (Fant, 1960; Stevens, 2000). For this purpose, for each articulation of each speaker, a sagittal function was calculated. It represents the variations along the vocal tract midline from the glottis to the lips of the transverse distances between the lower and upper vocal tract contours. The transverse vocal tract areas were derived from the transverse vocal tract distances using an α-β model (Soquet, Lecuit, Metens, & Demolin, 2002), leading to the area function. The acoustic wave propagation was simulated by means of an electrical equivalent (Fant, 1960; Stevens, 2000). The output is an acoustic transfer function from which the first three formant frequencies simulated for the considered articulation were extracted. An illustration of this process is provided in Fig. 2.
- For each speaker and each formant and Δformants, a linear regression of the articulations on the formant and Δformant frequencies was carried out. It captures the linear articulatory variations associated with the variations of the formants and Δformants. They form the articulatory components related to the formant and Δformant variations and are referred to as *F1 articulatory component*, *F2 articulatory component*, *F3 articulatory component*, *ΔF1F2 articulatory component* and *ΔF2F3 articulatory component* in the following. For conciseness, unless specified, the formants and Δformants are referred to in the following as *formants* and all the components as *formant components*. Articulatory nomograms for the F2 component for four different speakers, i.e. variation of the speaker articulation around her/his mean articulation for the formant varying at regular steps between the minimal and maximal values found in the data, are presented for illustration in Fig. 3.

Formally, an articulation is a matrix A_{ik} of size $n \times 2, i$ representing the articulation number and k the speaker, that can be decomposed as follows:

$$A_{ik} = F^j_{ik} R^j_k + \epsilon_1$$

where F^j_{ik} represents the frequency of the formant j for the articulation i and the speaker k, R^j_k, of size $n \times 2$, represents results of the multiple linear regressions for the formant j of the speaker k, and ϵ_1 the residue. Technically, for each speaker and each formant, there exist 10 to 16 articulations with their corresponding formant values. The articulations are regressed on the formant values, so that a simple linear regression occurs between each set of coordinates of the n contour points and

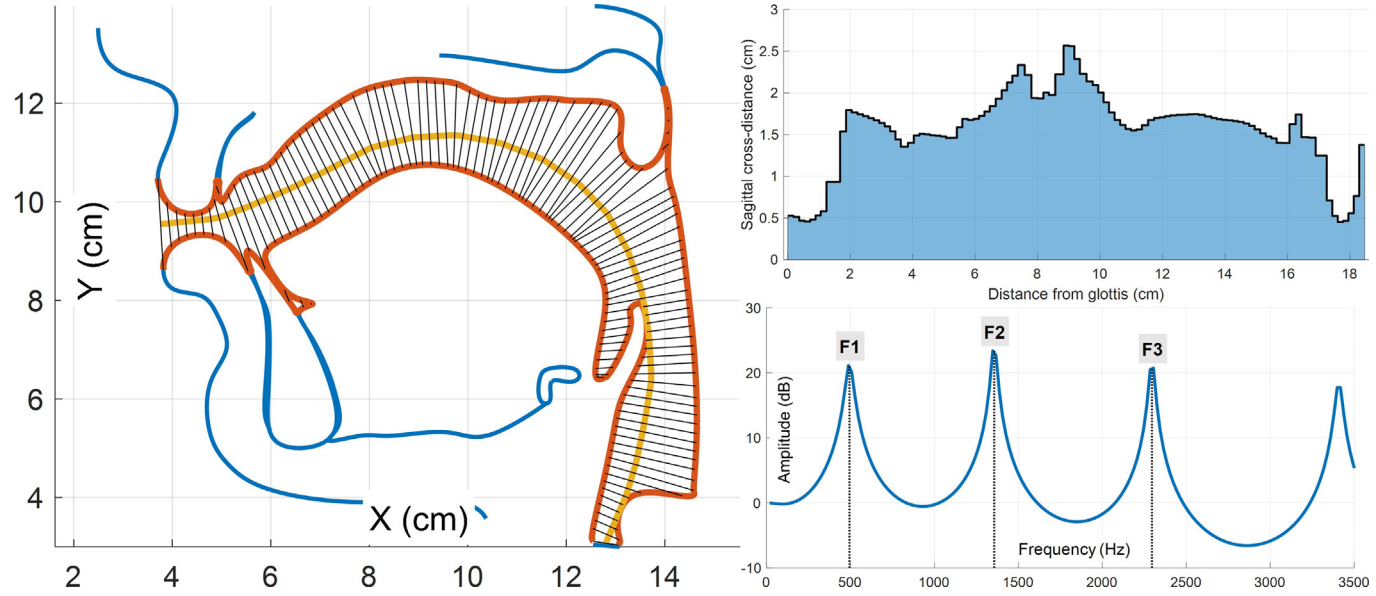

Fig. 2. Steps of the process for the calculation of the acoustic transfer function for the articulation presented in Fig. 1. Left: articulation (blue), vocal tract contours (orange), vocal tract midline (yellow) and vocal tract transverse lines (black). Top right: sagittal function. Bottom right: acoustic transfer function with identification of the first three formants.

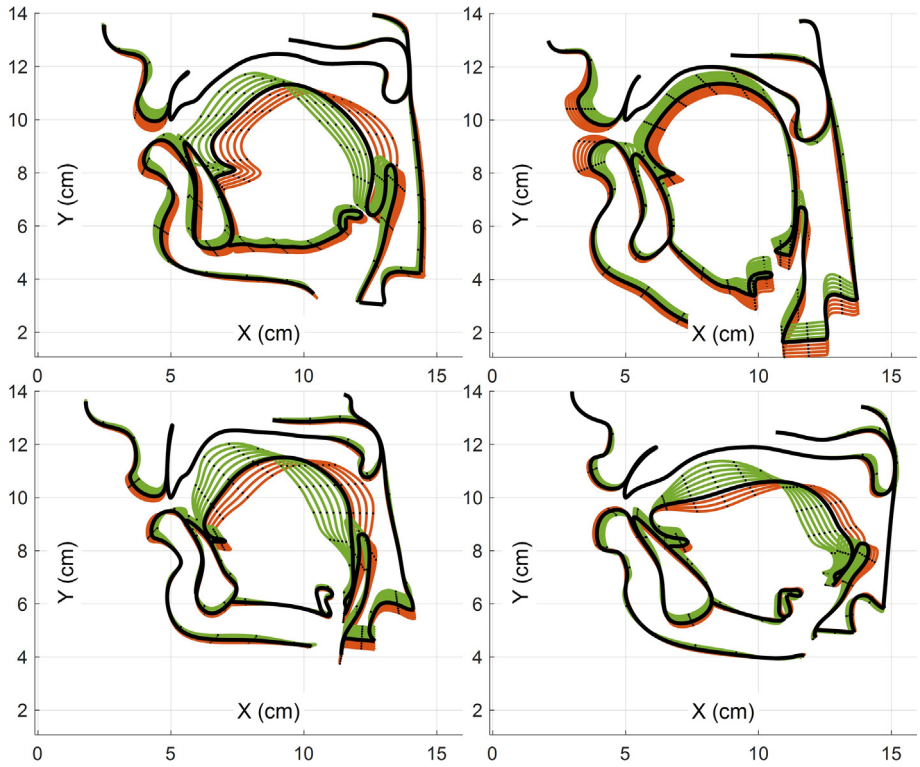

Fig. 3. Articulatory nomograms of the F2 formant component for four different speakers. The green and red contours represent the contours obtained for formant frequencies above and below the averaged formant values, respectively. The black contours represent the mean articulation of the mean speakers' articulations. One every 20 point is represented by a dot to emphasize the directions of deformation.

the formant values. The results of these multiple regressions are stored in the R_k^j matrix, which provides therefore the linear relationship between the value of a formant and the displacement of the contour points.

The matrix R_k^j represents the formant articulatory component. It is the equivalent of the eigenvector in an articulatory model driven by PCA (Serrurier et al., 2019). Note that unlike traditional articulatory modelling, our objective here is not to maximise the variance explanation nor to minimise the residual error. Our objective is to capture the linear relationships between formant values and articulations, including weak relationships. Weaker relationships will simply results in formant components with smaller amplitudes, but remain fully valid for further analyses.

2.2.2. Morphological parameters

In order to uncover possible links between the formant articulatory components and the morphology of the speakers, parameters characterising the vocal tract morphology have been extracted. They were inspired by the morphological analysis of the vocal tract carried out in a previous study (Serrurier & Neuschaefer-Rube, 2023). To summarise, there are five parameters:

- The *morphology X*, thereafter MX, measuring the horizontal length of the vocal tract, calculated as the projection on the horizontal axis of the distance between the upper teeth and the pharyngeal wall.
- The *morphology Y*, thereafter MY, measuring the vertical height of the vocal tract, calculated as the projection on the vertical axis of the distance between the glottis and the upper teeth.
- The *morphology angle*, thereafter MA, representing the angle between horizontal buccal part and the vertical pharyngeal part of the vocal tract, calculated as the angle between the vertical axis and the pharyngeal wall line.
- The *morphology palate anteriority*, thereafter MPA, representing the variation of the palate shape in the anterior-posterior direction, similar to the anteriority mode described by Lammert, Proctor, and Narayanan (2013), calculated as the first component scores of a PCA applied on the horizontal coordinates of the hard palate contours of the 41 speakers.
- The *morphology palate concavity*, thereafter MPC, representing the variation of the concavity of the palate, similar to the concavity mode described by Lammert et al. (2013), calculated as the radius of the lest square circle fitting the rounded part of the hard palate.

The morphological parameters were calculated for all speakers from the so-called morphological average-articulation (Serrurier & Neuschaefer-Rube, 2023), an artificial articulation representative of a speaker morphology and free from the speaking task. They have been calculated in a recursive way, where the contribution of each parameter is removed from the morphological average-articulation before calculating the next parameter. This method is analogous with the *guided PCA* method in articulatory modelling (Maeda, 1990; Beautemps, Badin, & Bailly, 2001) and ensures that the resulting parameters are decorrelated. The same variability in the data is therefore not captured multiple times by several parameters. As an illustration, one can imagine that the length of the palate, partly captured by MPA, is correlated with the length of the mouth, captured by MX. As a consequence, the parameter MPA captures variability also captured by MX, bringing confusion in its interpretation. Using the present method, MX captures the variability related to the length and MPA only the remaining horizontal (palate) variability not related to the overall length variability already captured by MX. More generally, each parameter is ensured not to capture some variablity already captured by the one or several parameters calculated before. Note that in this approach the order in which the parameters are calculated plays a role. Please see (Serrurier & Neuschaefer-Rube, 2023) for further details on this aspect. Formally, a parameter is a vector of length p (the number of speakers, 41) and all parameters can be grouped in a matrix M of size $p \times 5$ (41 speakers, 5 morphological parameters).

2.2.3. Direct analysis

As illustrated for a few examples in Fig. 3, different speakers present different components. In order to visualise the general trends, the formant articulatory components were averaged over speakers to obtain average formant articulatory components:

$$\forall j \in [1,..,5] \quad \overline{R^j} = \frac{1}{p}\sum_{k=1}^{p} R_k^j$$

where R_k^j stands for the formant articulatory component of the formant j, varying from 1 to 5, and of the speaker k, varying from 1 to $p = 41$.

Articulatory nomograms were then calculated and displayed for analysis.

Averaging by sex was initially performed but did not reveal any major difference at this stage (see for instance Serrurier & Neuschaefer-Rube, 2022 for a display of preliminary results). It is therefore not reported in the paper. Note that the further inter-speaker analysis revealed however slight differences between males and females, as emphasised later in the paper.

As F1 variations are known to be related to the opening-closing of the front part of the vocal tract, a special attention was brought to determine the weight of the opening-closing of the jaw in the F1 variations. More specifically, the Pearson correlation coefficient between F1 and the vertical coordinate of the lower teeth, referred to as JY, was calculated. In a second step, another correlation study between this correlation coefficient and the morphological parameters was performed to analyse whether the F1-jaw relationship could be ascribed to some extent to the speaker morphology.

2.2.4. Analysis by modelling

Despite common trends, large inter-speaker variability was observed in the formant components, as visualised in Fig. 3 for the F2 component for four different speakers. This section describes the approach chosen to characterise this inter-speaker variability.

Principles of second-level modelling. The method relies on the principles of second-level modelling, also referred to as *model of models* (Serrurier et al., 2019). In this approach, the first level is represented by the modelling approach having led to the formant components: The formant components R_k^j form the first level of model. For each formant, there are 41 observations, one per speaker. For a formant j, they can be grouped together in another matrix of size -$p \times n \times 2$-, on which a PCA can be further applied, leading to a model of models. This uncovers the main linear variations of the formant components over the speakers, also referred to as the *inter-speaker components*. In other words, a PCA is applied to the formant components to uncover their main linear modes of variation.

Formally, the p matrices R_k^j can be grouped in the single matrix Q^j of size $p \times n \times 2$, that can be decomposed as follows:

$$\forall j \in [1,..,5] \quad Q^j = \overline{Q^j} + S^j E^j + \epsilon_2$$

where S^j and E^j, of respective size $p \times m$ and $m \times n \times 2$ (m represents the numbers of components of the PCA), represent respectively the scores and eigenvectors of the PCA for the formant j, and ϵ_2 the residue.

Second-level modelling. This principle was applied to each formant component to build a model of models per formant.

The performance are evaluated in terms of percentage of variance explained. It relates to the variance of the data inputted to the PCA, i.e. the inter-speaker variance of the formant components. These percentage are provided in Section 3. In complement, nomograms are provided in order to interpret the inter-speaker components. As inter-speaker components explain variations of formant components, i.e. represent components of components, they are hardly interpretable as such. Instead, nomograms of the underlying formant components are provided. For this purpose, an inter-speaker component is used to generate a formant component, for which articulatory nomograms can be plotted. This technique was used to generate, for one inter-speaker component, two formant components corresponding to the most extreme possible predictions of the inter-speaker component. The nomograms of the obtained formant components are then plotted in Section 3. In summary, two extreme formant component nomograms are provided for one inter-speaker component.

Relationship with the morphology. The objective of this section is to determine whether the large inter-speaker variability in the formant components can be ascribed to some extent to morphological differences. In other words, to know if the strategy implemented by the speakers to modify their formants can be driven by the morphology. For instance, one could wonder whether a more domed palate (Brunner, Fuchs, & Perrier, 2009) would force a speaker to focus primarily on vertical movements of the tongue to make valuable changes of F1.

For this purpose, a first approach was to compute Pearson's correlation coefficients between the PCA scores of the second-level decomposition, i.e. the matrices S^j for $j = [1, .., 5]$, and the morphological parameters M described previously. The results are provided in Section 3.

In a second approach, for each formant, recursive linear regressions of the formant component matrix Q^j embedding all speakers' formant components was applied on the morphological parameters. This method consists in applying a guided PCA (Serrurier & Neuschaefer-Rube, 2023) to the formant components rather than a PCA as in the previous section. A guided PCA decomposes data into two matrices: the matrix of the independent variables, in our case the morphology parameters M, analogous to the score matrix in a PCA, and the regression matrix G^j, analogous to the eigenvector matrix in a PCA:

$$\forall j \in [1, .., 5] \quad Q^j = \overline{Q^j} + MG^j + \epsilon_3$$

where G^j is of size $5 \times n \times 2$ and ϵ_3 represents the residue. The decomposition is carried out recursively so that the data regressed on a new morphological parameter are the data Q^j where the contribution of all morphological parameters calculated so far, i.e. MG^j, is removed. This approach leads to a linear decomposition for which the control parameters, equivalent to the scores in the PCA, are the morphological parameters. The performance are measured in terms of percentage of data variance explanation like for the PCA. Further information and description of the guided PCA can be found in previous publications (Maeda, 1990; Beautemps et al., 2001; Serrurier & Neuschaefer-Rube, 2023). In this context, it allows to determine to which extent the morphology can explain the inter-speaker variability of the formant articulatory components. The performance are provided in Section 3.

3. Results

3.1. Averaged formant articulatory components

The nomograms describing the articulatory attributes of the three first formants and the two Δformants averaged over all speakers are visible on Figs. 4 and 5. The F1 component is mainly associated with an opening-closing of the jaw supplemented with a vertical variation of the tongue shape and position ranging from a bunched shape in a high position to a flat shape in a low position. The two effects combine themselves together to amplify the effect on the front cavity. The F2 component is mainly associated with a frontward-backward variation of the tongue position, slightly oblique so that the anterior position is higher than the posterior position. The variation in the backward direction is associated with a slight protrusion of the lips and a slight lowering of the larynx. The F3 component is associated with a frontward-backward variation of the tongue position combined with an opposite protrusion-retraction of the lips of similar scale. The ΔF1F2 component appears as a composition of the F1 and F2 components: A large tongue variation in an oblique direction, from a high front bunched position to a low back flat position, without much variations of the jaw. The slight co-variations of the lips and larynx as for the F2 component is also observed. This component

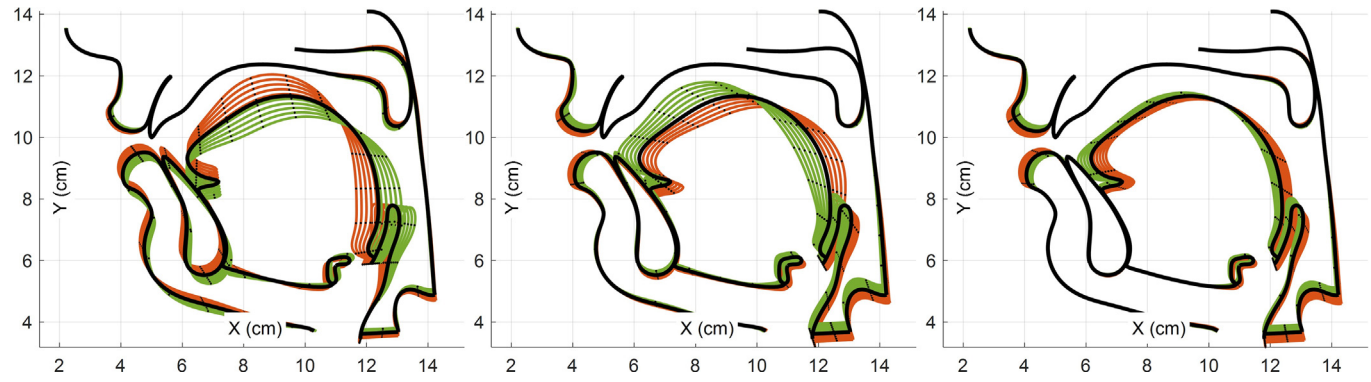

Fig. 4. Articulatory nomograms of the F1 (left), F2 (middle) and F3 (right) formant components averaged over the speakers. The green and red contours represent the contours obtained for formant frequencies above and below the averaged formant values, respectively. The black contours represent the mean articulation of the mean speakers' articulations. One every 20 point is represented by a dot to emphasize the directions of deformation.

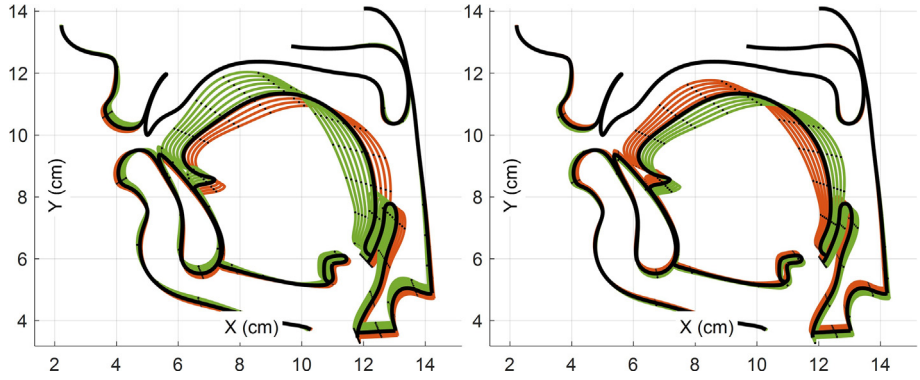

Fig. 5. Similar to Fig. 4 for the ΔF1F2 (left) and ΔF2F3 (right) components.

associate large variations of the front and back cavities. The ΔF2F3 appears rather similar to the F2 component, with slightly smaller variations on the lips and the larynx.

The distribution of the speakers according to the Pearson correlation coefficient between F1 and JY is visible in Fig. 6. There is a good correlation between F1 and JY with an average correlation coefficient of 0.69. While the jaw opening-closing is largely involved in the variation of F1 for most speakers, it is however noticeable that for some speakers, changes in F1 are achieved almost without opening or closing the jaw. Pearson correlation coefficients were calculated between this Pearson correlation coefficient and the morphological parameters. It indicates whether the degree of opening-closing of the jaw during variations of F1 could be related to one or several morphological features. The highest coefficient was obtained for the MPA parameter with a value of 0.48. This parameter is associated with the variation of the anterior-posterior position of the palatal dome once the contributions of the vocal tract size in the horizontal and vertical dimensions and of the vocal tract angle, as defined in Section 2.2.2, have been removed. In other words, the degree of opening-closing of the jaw during variations of F1 is associated to a certain extent with the anteriority of the palatal dome, as illustrated on Fig. 7: Speakers with a palatal dome flatter and more posterior tend to associate variations of F1 with the opening-closing of the jaw more than speakers with a palatal dome more pronounced and more anterior.

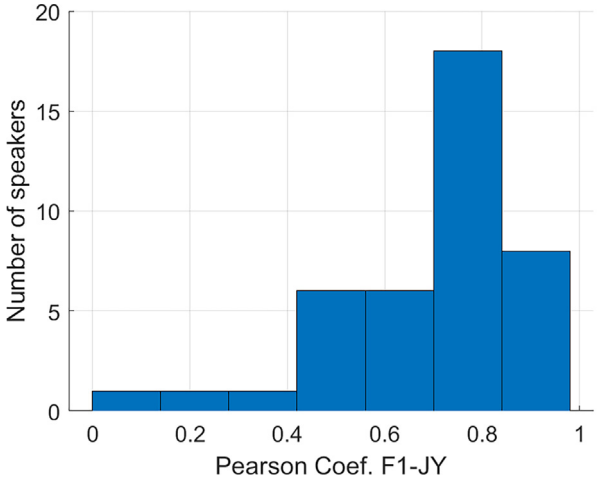

Fig. 6. Distribution of the speakers according to their Pearson correlation coefficient between F1 and JY.

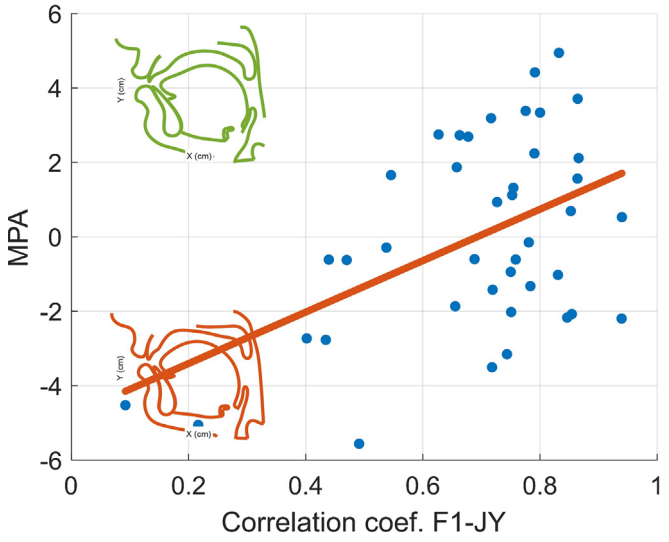

Fig. 7. MPA morphological parameter vs. Pearson correlation coefficient between F1 and JY for all speakers (blue dots) with the associated linear regression (orange line). The predicted articulations for the two extreme values of MPA found in the data are represented in green and orange.

3.2. Inter-speaker analysis

The averaged articulatory attributes characterised in the previous section are in general agreement with the knowledge of which acoustic effects are driven by which articulatory variation. The current section aims at investigating deeper the inter-speaker variabilities.

3.2.1. Second-level modelling

Only the first inter-speaker component was retained as the main mode of variation easily interpretable. This component explains respectively 35%, 37%, 33%, 36% and 45% of the inter-speaker variance of the F1, F2, F3, ΔF1F2 and ΔF2F3 articulatory components. The corresponding nomograms are visible in Figs. 8 and 9.

The main linear variation of the F1 component ranges from a F1 component associated with an oblique variation of the tongue, from a bunched front high position of the tongue to a flat back low position of tongue on the one hand, to a F1 com-

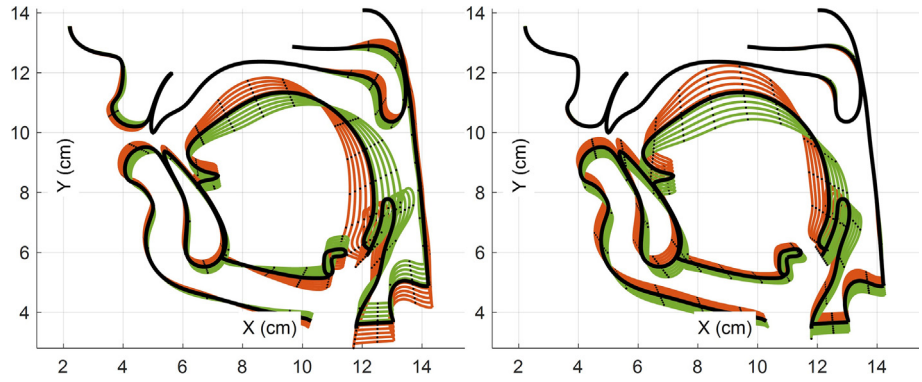

Fig. 8. Articulatory nomograms of the F1 formant component obtained from the first inter-speaker component for the minimum score (left) and maximum score (right) found in the data. The green and red contours represent the contours obtained for formant frequencies above and below the averaged formant values, respectively. The black contours represent the mean articulation of the mean speakers' articulations. One every 20 point is represented by a dot to emphasize the directions of deformation.

ponent associated with a larger vertical movement of the tongue body combined with a larger range of jaw opening-closing on the other hand. In other words, speakers vary primarily from focusing strongly on the volume of the front cavity to combining a variation of the front cavity size of smaller range together with a variation of the back cavity.

The main linear variation of the F2 component ranges from a F2 component associated with a large frontward-backward variation of the tongue position combined with an opposite protrusion-retraction of the lips on the one hand, to a F2 component associated with a smaller oblique variation of the tongue, from a bunched front high position to a flat back low position on the other hand. For the F2 component, speakers vary therefore primarily from a strong focus on the frontward-backward position of the tongue to a more limited range of variation that modifies concomitantly the front and back cavities of the vocal tract.

The main linear variation of the F3 component ranges from a F3 component associated with a large and slightly oblique frontward-backward variation of the tongue position combined with an opposite protrusion-retraction of the lips on the one hand, to a F3 component associated with a smaller range of variation of the tongue from a higher and more bunched to a lower and flatter tongue shape on the other hand. For the F3 component, speakers vary therefore primarily from a strong focus on the frontward-backward position of the tongue and lips protrusion-retraction to a slighter variation of the shape of the tongue without lip protrusion.

The main linear variation of the ΔF1F2 component ranges from a ΔF1F2 component associated with a very large and slightly oblique frontward-backward variation of the tongue position, combined with an upward-downward variation of the larynx position and an opposite protrusion-retraction of the lips on the one hand, to a ΔF1F2 component associated with a smaller range of variation of the tongue only from a high bunched to a lower flat tongue position on the other hand. For the ΔF1F2 component, speakers vary therefore primarily from a strong focus on the most anterior part of the front cavity and of the back cavities to a slighter focus on the middle part of the front cavity.

Finally, the main linear variation of the ΔF2F3 component ranges from a ΔF2F3 component associated with a large variation of the tongue in the horizontal direction, combined with a slight upward-downward variation of the larynx position and a slight opposite protrusion-retraction of the lips on the one hand, to a ΔF2F3 component associated with smaller range of variation of the tongue only in the vertical direction on the other hand. For the ΔF2F3 component, speakers vary therefore primarily from an horizontal to a vertical variation of the tongue.

3.2.2. Relationship with the morphology

The Pearson correlation coefficients between the scores of the first inter-speaker component of each formant component and the morphological parameters are provided in Table 1. The plot of the scores vs. the morphological parameter values for correlation coefficients above 0.3 are provided in Fig. 10.

A very slight correlation between the first inter-speaker component of the F1 component with the palate shape morphological parameters, both anteriority and concavity, is observed. Speakers with a more anterior and domed palate tend to present a F1 component rather similar to the left part of Fig. 8 whereas speakers with a more posterior and flatter palate tend to present a F1 component rather similar to the right part.

A stronger correlation is observed between the first inter-speaker component of the F2, the ΔF1F2 and the ΔF2F3 components and the vertical height of the vocal tract. Speakers with a longer vertical height of the vocal tract tend to present a F2, ΔF1F2 and ΔF2F3 components rather similar to the left part of Fig. 9 (top, third and bottom rows), whereas speakers with a shorter vertical height tend to present components rather similar to the right part. In summary, speakers with a shorter vocal tract height tend to focus on more vertical movements of the tongue while speakers with a longer vocal tract height tend to focus on larger horizontal movements of the tongue. Because MY is related to male–female differences (Serrurier & Neuschaefer-Rube, 2023), females tend to focus on vertical movements and males on horizontal movements.

The results of the guided PCA of the formant articulatory components on the five morphological parameters led to an overall variance explanation of respectively 21%, 30%, 24%, 30% and 29% for the F1, F2, F3, ΔF1F2 and ΔF2F3 articulatory components. For comparison, a raw PCA of the formant articulatory components with only three components led to variance explanations between 65% and 78%. Only limited inter-speaker variability of the formant articulatory components can therefore be explained linearly by the morphology param-

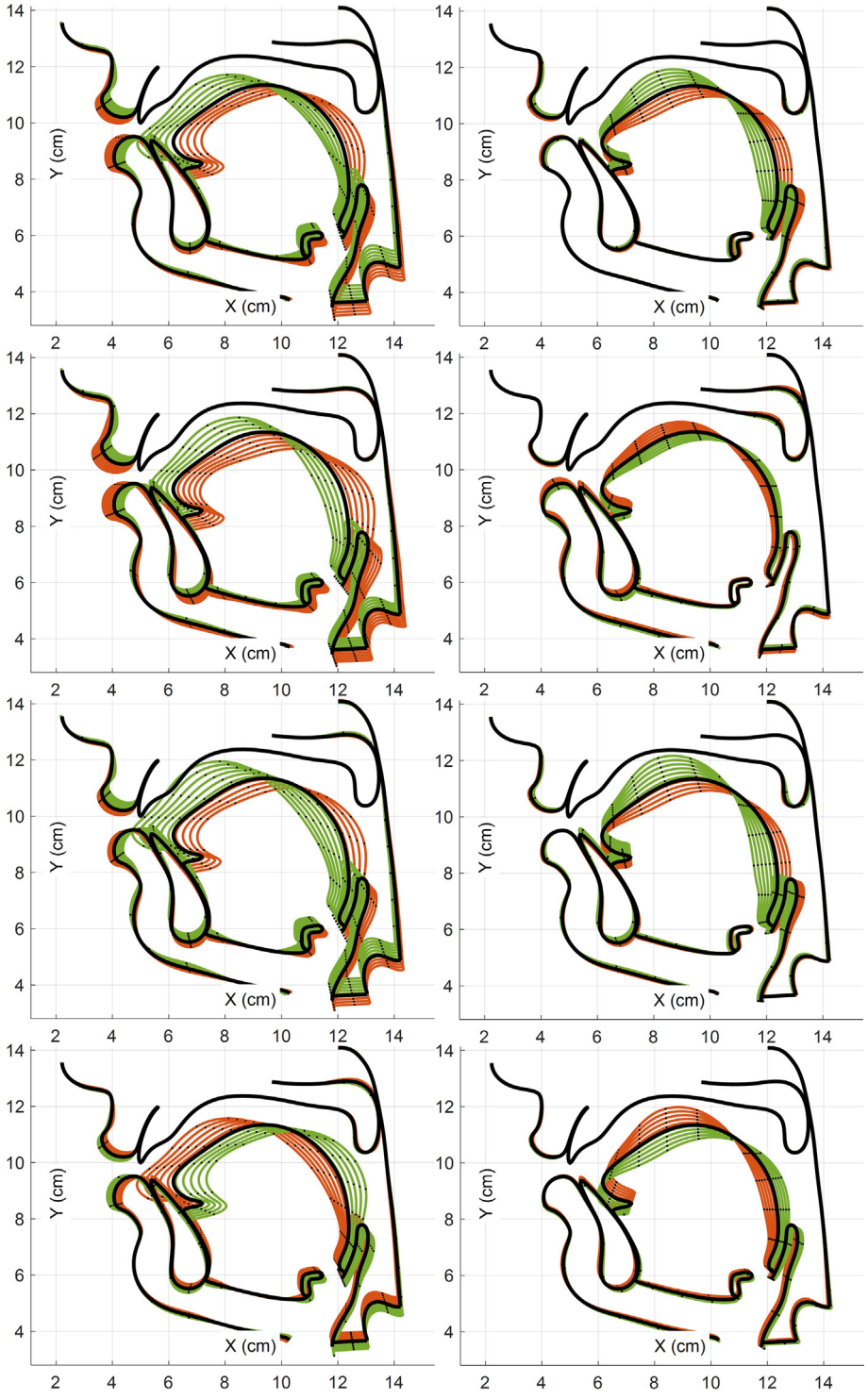

Fig. 9. Similar to Fig. 8 but for the F2 (top row), F3 (second row), ΔF1F2 (third row) and ΔF2F3 (bottom row) formant components.

Table 1
Absolute values of the Pearson correlation coefficients between the scores of the first inter-speaker component of each formant component and the morphological parameters. In bold the values above 0.3.

	MX	MY	MA	MPA	MPC
1st inter-speaker component of F1 comp.	0.08	0.23	0.14	**0.35**	**0.33**
1st inter-speaker component of F2 comp.	0.06	**0.64**	0.19	0.16	0.04
1st inter-speaker component of F3 comp.	0.02	0.20	0.04	0.13	0.24
1st inter-speaker component of ΔF1F2 comp.	0.09	**0.61**	0.18	0.15	0.11
1st inter-speaker component of ΔF2F3 comp.	0.22	**0.52**	0.24	0.28	0.08

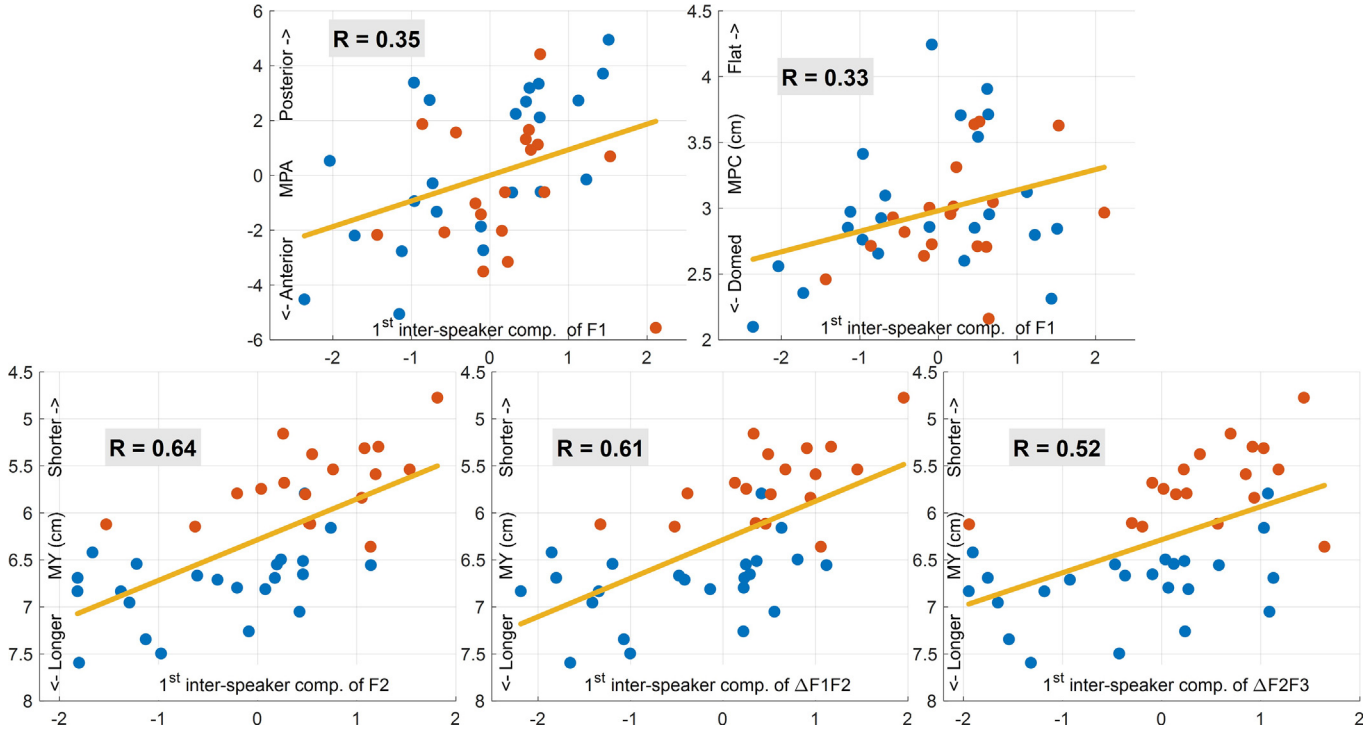

Fig. 10. Morphological parameters vs. scores of the inter-speaker components for all speakers for the correlations of Table 1 above 0.3. Blue and orange dots represent the male and female speakers, respectively. The solid lines represent the associated linear regressions. Note the reversed axis orientation of the MY parameter to mirror the height of the larynx.

eters. The variance explanation of individual morphological parameters were almost always equal or below 10%, except for the MY parameter explaining 15% to 17% of the variability of the F2, ΔF1F2 and ΔF2F3 articulatory components. This confirms the link between the inter-speaker variability of these three components and the MY morphological parameter exhibited earlier (see Table 1 and Fig. 10).

In summary, some links between the formant articulatory components and the morphology of the vocal tract have been exhibited, emphasising the role played by the shape of the palate on the one side and by the height of the pharyngeal cavity on the other side. However, these influences remain limited and a large part of the inter-speaker variability of the formant components cannot be explained by the morphology.

4. Discussion and conclusion

4.1. Summary

41 speakers articulating the oral vowels of their native language and recorded by means of static midsagittal MRI were considered for this study. The images were manually annotated to segment the contours of the vocal tract from the glottis to the lips. The acoustic transfer function for these articulations were calculated by means of acoustic modelling and the three first formant frequencies identified. They were complemented by the differences between the F1 and F2 formants and between the F2 and F3 formants, leading to 5 formant related frequencies.

For each formant and Δformant and each speaker, multiple linear regressions of the articulation contour points on the frequencies were applied to derive the corresponding formant/Δformant articulatory components. They capture the linear articulatory variations associated with the formant/Δformant variations, i.e. their articulatory attributes. An averaging of the formant/Δformant components over the speakers revealed the average articulatory components: An upward-downward variation of the tongue, from a high bunched to a low flat position, for the F1 component, an oblique anterior-posterior variation of the tongue, associated with slight larynx height variations and slight opposite lip protrusion-retraction, for the F2 component, and a small anterior-posterior variation of the tongue, also associated with slight larynx height variations and slight opposite lip protrusion-retraction, for the F3 component. These are in general agreement with the phonetic knowledge derived so far from the direct analyses, i.e. when the formant variations are derived from simple vocal tract changes. For the ΔF1F2 component, an oblique variation of the tongue from a high front to low back position, and for the ΔF2F3 component, a similar oblique variation of the tongue of smaller amplitude. F1 was found to be correlated with the jaw opening-closing with an average Pearson coefficient of 0.69, although some speakers, especially those with a more pronounced and more anterior palatal dome, appeared not to relate F1 variations with the jaw movement.

A large inter-speaker variability in the formant/Δformant articulatory components was observed and analysed by means of second-level modelling. It consists in applying a PCA on the formant/Δformant components to uncover the main inter-speaker modes of variation. The first inter-speaker component explains 33% to 45% of the formant/Δformant component variance. It reveals in general that the main modes of variation tend to relate to the orientation and amplitude of the tongue variations, from speakers presenting larger and rather

anterior-posterior variations to speakers presenting smaller and more upward-downward variations.

Some of these inter-speaker components appeared to be related to the speakers' morphology. For the F1 component, speakers with a more anterior and domed palate have a slight tendency to present more oblique variations of the tongue as opposed to speakers with a more posterior and flat palate who have a slight tendency to present more vertical variations of the tongue, with Pearson coefficients between 0.33 and 0.35. For the F2, ΔF1F2 and ΔF2F3 components, speakers with a shorter vocal tract height, including females, have a tendency to present smaller vertical variations of the tongue position while speakers with a longer vocal tract height, including males, have a tendency to present larger horizontal variations, with Pearson coefficients between 0.52 and 0.64. In general, most of the inter-speaker component variance cannot be explained by the morphology, with a maximum of 15% to 17% of the F2, ΔF1F2 and ΔF2F3 component variance for the vertical height of the vocal tract.

4.2. Discussion

Unlike many studies related to the analysis of the articulation-formant relationship and based on artificial area functions, the current approach relies on real articulations: 10 to 16 articulations per speaker for 41 speakers. This represents on the one side a relatively limited set of data to perform statistics. On the other side, the data, a subset of the data presented in Serrurier and Neuschaefer-Rube (2023), represent more than 500 medical images where the full vocal tract contours from the glottis to the lips have been manually outlined. It represents very valuable data that can hardly be extended to another scale with such reliability. Real-time MRI (Ramanarayanan et al., 2018), where images are recorded at a temporal resolution that allows natural speech, constitute naturally an interesting alternative. This would provide much more articulations per speaker. The technology is however currently emerging and not widely spread as for the static MRI. In addition, the lower image quality (see e.g. Lingala, Sutton, Miquel, & Nayak, 2016) and the large number of articulations available per speaker call for challenging automatic methods to segment the vocal tract that are currently under development (Labrunie et al., 2018; Ribeiro, Isaieva, Leclere, Vuissoz, & Laprie, 2022; Belyk, Carignan, & McGettigan, 2023).

The considered articulations are static articulations that may differ from real speech articulations. However, as the articulations are carefully sustained during several seconds by the speakers, one can assume that they are well articulated and represent the phoneme. They may have the tendency to be hyper-articulated. As they are used to perform a linear regression to obtain the formant components, their role is however to sample the articulatory space. As well-formed or hyper-articulated articulations, they ensure to cover up to the extremes of the articulatory space. Although a larger sample of articulations may bring robustness and should be considered in the future, this limited set appears therefore appropriate for the current usage. Such approach with a limited set of well-chosen articulations is commonly used in articulatory modelling studies (e.g. Harshman, Ladefoged, & Goldstein, 1977; Serrurier & Badin, 2008; Lammert et al., 2013).

The acoustic data are obtained by means of modelling as acoustic recordings were not performed with the image recordings. This is to be ascribed to the heavier experimental setup to integrate microphones compatible with MRI scanners and to the high level of noise during image acquisition (Bresch, Nielsen, Nayak, & Narayanan, 2006). This noise requires post-processing to remove the scanner component (Bresch et al., 2006; Echternach, Burk, Burdumy, Traser, & Richter, 2016) and may produce an acoustic signal with relatively low signal-to-noise ratio, limiting its use for accurate measurement purposes. Alternatively, plane wave simulations by means of electrical equivalent have proven to be reliable to estimate formants below 5 kHz (Fant, 1960; Stevens, 2000). One limitation in this approach comes from the calculation of the area function for the sagittal distances. This is not a straightforward calculation and traditional approaches, like the one used in our study (Soquet et al., 2002), rely on α-β models (Heinz & Stevens, 1965; Mermelstein, 1973). Further studies emphasise however the complexity and potential limitations of these calculations (McGowan, Jackson, & Berger, 2012). In the future, three-dimensional MRI data or recorded acoustics could solve this issue.

The models are built by means of linear analysis: The first-level models (regression to obtain the formant components) and the second-level models (PCA to obtain the inter-speaker components, guided PCA on the morphological parameters). This is due to the relatively limited number of observations: 10 to 16 for the first level, 41 for the second level. Non-linear modelling, such as those based on deep learning (Ribeiro & Laprie, 2022), may provide complementary results, possibly more accurate. Such approaches should be considered in the future, but require a very large amount of data that does not appear reachable at this stage, although temporary solution could be achieved by data augmentation (e.g. Serrurier, 2023). The linear modelling may constitute the bottleneck of the method. It can only capture the major relationships but may miss some more subtle effects. However, the current study provides a benchmark in the inter-speaker analysis of the articulatory attributes of the formants and could be refined in the future when more data and non-linear models become available.

The formant and Δformant components provide a full characterisation of the vocal tract from the glottis to the lips and uncover for the first time the articulatory attributes. A global characterisation has been carried out by averaging the components over the speakers. It could however be the case that the averaged components remain far from the effective components of the speakers, for instance in the case of a bimodal distribution of the speakers' components. Until further analysis of the distribution of the speaker components, these results should therefore be considered with care. The average F1 and F2 components are in general agreement with knowledge from the literature (Stevens, 2000; Ladefoged & Johnson, 2011), where analyses are based on a direct approach, i.e. when the vocal tract shape varies first and the effects on the formants are observed afterwards. An increase of F1 is achieved by a lowering of the tongue to open the buccal cavity and a backward movement to close the pharyngeal cavity. This is partly carried by the opening of the jaw. An increase of F2 is achieved by a frontward movement of the tongue that lengthens the back cavity of the vocal tract. It is associated with a

raising of the larynx, known to rise the F2 formant (Sundberg & Nordström, 1976), and with a retraction of the lips. In the current analysis, the frontward tongue movement is shown to be slightly oblique, probably to have a larger impact on the most anterior part of the front cavity, where the narrowing should take place to increase F2 (Stevens, 2000). This phenomenon was already observed in previous studies (Lee, Shaiman, & Weismer, 2016). F3 is known to present less cross-vowel articulatory characteristics than F1 and F2. In the literature, it is usually associated with lip rounding (Ladefoged & Johnson, 2011), and in relation with F2, supporting the principle of the ΔF2F3 formant of the current study. In the current study, an increase of F3 is associated with a slight forward movement of the tongue, which could be related to the F3-tongue blade correlation already noted in the literature (Lindblom & Sundberg, 2007; Pennington, 2022). The protrusion is logically associated with a decrease of F3. The slight larynx lowering has also been reported as decreasing F3 (Sundberg & Nordström, 1976). This can be considered as the counterpart of the lip protrusion at the other end of the vocal tract to lengthen it, which tends to decrease all formants (Fant, 1960; Stevens, 2000), as observed also for F2. The smaller amplitude for the F3 component than for the F1 and F2 components could be ascribed to the various antagonistic vowel-specific strategies leading to a weak correlation between the articulations and the formant variations. This is captured by small slopes in the linear regression performed to calculate the formant component. Indeed, the F3 component appeared to explain in average 23% of the articulations' variance, as opposed to 34% to 46% for the other formant and Δformant components. Antagonistic inter-speaker strategies for F3 could also partly cancel each other in the averaging process and participate in a smaller average F3 component.

As emphasised above, a raising and frontward movement of the tongue decreases F1 and increases F2. Logically, an increase of the ΔF1F2 Δformant is therefore achieved by a clear frontward and upward movement of the tongue. This is consistent with acoustic considerations which represent the F1-F2 distance as being rather oblique in the F1-F2 plane (e.g. Fig. 1 in Auracher, Menninghaus, & Scharinger (2020)). In average, ΔF1F2 appears equally correlated with F1 and F2 (average Pearson coefficients of 0.26 and 0.27) in our data. Since the lip protrusion and larynx lowering seem to be consistently related to an F2 decrease but not much related to F1, these movements are also logically involved in the decrease of ΔF1F2. However, in agreement with Ladefoged and Johnson (2011), ΔF1F2 variations seem more related to the sole tongue movements than for F2, more affected by the lip rounding and the larynx lowering. Any effect decreasing F2 and ΔF1F2 may in mirror tend to increase ΔF2F3. Similarly, any effect increasing F3 will tend to increase ΔF2F3. As a result, a decrease of ΔF2F3 is achieved by an oblique forward movement of the tongue, somewhat similar to the one corresponding to the increase of F2 but of slightly lower amplitude. The effect of lip rounding on ΔF2F3 seems however rather marginal in our data, unlike reported in the literature (Stevens, 2000). As noted earlier for F3, the less salient articulatory effects for ΔF2F3 could be partly ascribed to our attempt to emphasise general cross-vowels cross-speaker effects while F3 may be more sensitive to the vowel and the speaker.

In general, an increase of F1 is partly achieved by jaw opening. The degree of use of the jaw is however unequal between speakers, with speakers relying almost entirely on it, a tendency for those with a rather more posterior palatal dome, and others not at all, a tendency for those with a rather more anterior palatal dome. This could be explained by the necessity to fill the gap between the tongue and the palate to decrease F1. Speakers with a more posterior palatal dome might achieve it by simply rising the jaw, which leads to a global rise of the tongue that may be enough to significantly reduce the gap between the tongue and the palate. On the contrary, the same action of the jaw for speakers with a more anterior palatal dome might leave a larger gap between the tongue and the palate, simply due to the more rounded shape of the palate in the front. As a consequence, these speakers might be required to actively rise the tongue, in addition or instead the rise of the jaw, to close the gap with the palate. The effect remains however a slight tendency and should be confirmed on more speakers in the future.

The first PCA inter-speaker components explain a good third of the inter-speaker variance of the formant components. For the F1 component, the variability relates to the relative impact of the tongue variations on the buccal and pharyngeal cavities: It varies from large variations of the tongue impacting mostly the buccal cavity to smaller variations of the tongue impacting equally the buccal and pharyngeal cavities, so as to multiply the effects on F1 with smaller tongue variations. This variability can be partly ascribed to the palate morphology, as for the degree of use of the jaw: more oblique movements of the tongue body tend to be associated with more anterior positions of the palatal dome, as opposed to more vertical movements associated with more posterior positions of the dome, so as to target with the tongue the palatal dome. The larger amplitude of the tongue body movement for the more posterior palatal dome positions is less intuitive, as this is also associated with flatter palates. This might be explained by the smaller amplitude of the tongue root movement in the back cavity in this case, compensated by a stronger effect on the front cavity. Nonetheless, the correlation with the morphology remains very slight and requires further analysis on a broader set of speakers.

For the F2 component, the variability relates also to the orientation of the tongue body movement. Interestingly, this variability is correlated with the vertical height of the vocal tract, itself related to the male–female difference. The relation with F2 does not seem straightforward. It could be related to the global size of the vocal tract, longer for males than females. As a consequence, males might need to stretch their tongue further forward and backward in the mouth to achieve substantial variations in F2, while smaller articulatory variations towards the palate might be enough for females to generate even larger acoustic variations (Simpson, 2002). One would however expect also a correlation with the MX parameter, capturing the horizontal length variability, which is not the case in our data. This could also be ascribed to biomechanical reasons, as for instance for jaw opening (Weirich, Fuchs, Simpson, Winkler, & Perrier, 2016). Finally, this could be related to the socio-linguistic background, leading males and females to implement different acoustic strategies not related to their morphological differences (van Bezooijen, 1995; Weirich & Simpson, 2018). This latter reason seems however

unlikely, as the correlation remains valid, although lower, within the male and female groups, with Pearson coefficients of 0.44 and 0.5, respectively. In any case, further insights might be provided by a finer speaker-specific or sex-specific formant-cavity affiliation analysis for various categories of vowels, as for instance performed by Fant (1966) and Badin, Perrier, Boe, and Abry (1990), beyond the scope of the current study.

The same inter-speaker component as for the F2 component is observed for the ΔF1F2 and ΔF2F3 components: Their scores are correlated with the scores of the inter-speaker component of the F2 component with Pearson coefficients of 0.98 and 0.89, respectively. Similar variations of nomograms and similar correlations with the vocal tract vertical height are therefore observed for these components than for the F2 component mentioned above.

For the F3 component, the variability captured by the first PCA inter-speaker component also relates to the orientation of the tongue body movement, as for the F1 and F2 components, and to the amplitude. No substantial link with the morphology was found for this component.

For each formant component, only the first inter-speaker component of the second-level modelling was considered. This is due to the complexity of interpretation of second-level components, which goes through the illustration of their effects on the first-level components. The other inter-speaker components were initially analysed, but revealed only subtle differences with the first PCA inter-speaker components that were hardly interpretable, all that for a lower percentage of variance explanation, varying between 17% and 27% for the second component. For this reason, the focus was brought on the first inter-speaker component only, exhibiting the most salient and interpretable inter-speaker variability.

4.3. Conclusion

The current study determines the articulatory attributes of the first three formants for the complete vocal tract. Unlike standard approaches in phonetics where the acoustic correlates of specific articulatory pattern are investigated, the present study addresses the opposite problem: Uncovering the articulatory movements associated with specific acoustic variations. In a second step, models of models were used to identify the main inter-speaker modes of variation, so as to provide a qualitative and quantitative inter-speaker analysis. Finally, the relationship between these main modes of variations and morphological features, such as the palate shape or the vertical height of the vocal tract, is exhibited.

The study is entirely reproducible thanks to the code provided publicly: https://github.com/tonioser/FormantComponents.

The current study emphasises the cross-vowel articulatory behavior associated with formant variations. Finding cross-vowels general trends has always been a motivation in phonetic studies, but constitutes only a starting point. In complement, front and back vowels as well as high and low vowels present intrinsic characteristics that deserve to be analysed separately, for example using the original methods of the current study. This calls for much more data and should be considered with the recent development in medical image acquisition and medical image processing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are very grateful to P. Badin and T. Sawallis for providing some data and to L. Lamalle and S. Romanzetti for participating in recording some others. The corpus of the data acquired in Aachen has been designed with the help of B. Kröger and C. Busch. Some initial tracings were performed by J.-A. Valdés Vargas, G. Ananthakrishnan, M. Eslami, A. Hülsdünker, and F. Stepp. This research project is supported by the START-Program of the Faculty of Medicine, RWTH Aachen University. This work was also supported by the Brain Imaging Facility of the Interdisciplinary Center for Clinical Research (IZKF) Aachen within the Faculty of Medicine at RWTH Aachen University. The recording of the Grenoble data has been partially funded by the French ANR (Grant No. ANR-08-EMER-001–02 "ARTIS"). The IRMaGe MRI facility was partly funded by program "Investissement d'Avenir" run by the "Agence Nationale de la Recherche"—Grant "Infrastructure d'avenir en Biologie et Santé"—ANR-11-INBS-0006.

References

Aaltonen, O. (1985). The effect of relative amplitude levels of f2 and f3 on the categorization of synthetic vowels. *Journal of Phonetics, 13*(1), 1–9. https://doi.org/10.1016/s0095-4470(19)30721-1.

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America, 63*(5), 1535–1555. https://doi.org/10.1121/1.381848.

Auracher, J., Menninghaus, W., Scharinger, M. (2020). Sound predicts meaning: Cross-modal associations between formant frequency and emotional tone in stanzas. *Cognitive Science* 44 (10) (oct 2020). doi:10.1111/cogs.12906.

Badin, P., Perrier, P., Boe, L.-J., & Abry, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *The Journal of the Acoustical Society of America, 87*, 1292–1300.

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. *The Journal of the Acoustical Society of America, 109*, 2165–2180.

Bell, A. M. (1867). *Visible speech*. London: Simpkin, Marshall and Co..

Belyk, M., Carignan, C., & McGettigan, C. (2023). An open-source toolbox for measuring vocal tract shape from real-time magnetic resonance images. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02171-9. jul.

Boë, L. -J., Perrier, P., Guérin, B., & Schwartz, J.-L. (1989). Maximal vowel space. In: *Proc. First European Conference on Speech Communication and Technology (Eurospeech 1989)*, 1989, (pp. 2281–2284). doi:10.21437/Eurospeech.1989-238.

Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S. (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America* 120 (4) (2006) 1791–1794. arXiv: https://doi.org/10.1121/1.2335423, doi:10.1121/1.2335423. doi: 10.1121/1.2335423.

Brunner, J., Fuchs, S., Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *The Journal of the Acoustical Society of America* 125 (6) (2009) 3936–3949. arXiv:http://asa.scitation.org/doi/pdf/10.1121/1.3125313, doi:10.1121/1.3125313. http://asa.scitation.org/doi/abs/10.1121/1.3125313.

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure*. Tokyo, Japan: Tokyo-Kaiseikan.

Echternach, M., Burk, F., Burdumy, M., Traser, L., & Richter, B. (2016). Morphometric differences of vocal tract articulators in different loudness conditions in singing. *PLOS ONE, 11*(4), 1–17. https://doi.org/10.1371/journal.pone.0153792.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1966). A note on vocal tract size factors and non-uniform f-pattern scalings. *STL-QPSR, 7*(4), 22–30.

Fant, G. (1980). The relations between area functions and the acoustic signal. *Phonetica, 37*(1–2), 55–86. https://doi.org/10.1159/000259983.

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

Heinz, J. M., & Stevens, K. N. (1965). On the relations between lateral cineradiographs area functions, and acoustic spectra of speech. In *Proceedings of 5th International Congress on Acoustics* (pp. A44).

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America, 94*(2), 701–714. https://doi.org/10.1121/1.406887. arXiv:http://asa.scitation.org/doi/pdf/10.1121/1.406887. URL http://asa.scitation.org/doi/abs/10.1121/1.406887.

Jones, D. (1917). *An English Pronouncing Dictionary*. London: J.M. Dent.

Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., & Boë, L.-J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication, 99*, 27–46. https://doi.org/10.1016/j.specom.2018.02.004.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America 29(1)*, 98–104. https://doi.org/10.1121/1.1908694. arXiv:http://asa.scitation.org/doi/pdf/10.1121/1.1908694, URL: http://asa.scitation.org/doi/abs/10.1121/1.1908694.

Ladefoged, P., & Johnson, K. (2011). *A course in phonetics*. Cengage Learning. https://books.google.de/books?id=FjLc1XtqJUUC.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford, UK: Blackwell Publishers.

Lammert, A., Proctor, M., & Narayanan, S. (2013). Morphological variation in the adult hard palate and posterior pharyngeal wall. *Journal of Speech, Language, and Hearing Research, 56*(2), 521–530. https://doi.org/10.1044/1092-4388(2012/12-0059).

Lee, J., Shaiman, S., Weismer, G. (2016). Relationship between tongue positions and formant frequencies in female speakers. *The Journal of the Acoustical Society of America* 139 (1) (pp. 426–440). arXiv:https://doi.org/10.1121/1.4939894, doi:10.1121/1.4939894. doi: 10.1121/1.4939894.

Lindblom, B. & Sundberg, J. (2007). Springer Handbook of Acoustics, Springer, New York, 2007, Ch. The Human Voice in Speech and Singing (pp. 669–712).

Lingala, S. G., Sutton, B. P., Miquel, M. E., & Nayak, K. S. (2016). Recommendations for real-time speech MRI. *Journal of Magnetic Resonance Imaging, 43*(1), 28–44. https://doi.org/10.1002/jmri.24997. URL https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.24997.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: *Speech Production and Speech Modelling* (pp. 131–149).

McGowan, R. S., Jackson, M. T.-T., & Berger, M. A. (2012). Analyses of vocal tract cross-distance to area mapping: An investigation of a set of vowel images. *The Journal of the Acoustical Society of America, 131*(1), 424–434. https://doi.org/10.1121/1.3665988.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America, 53*, 1070–1082.

Pennington, M. (2022). Acoustic-articulatory correlations in a four-region model of the vocal tract: Theoretical bases and a comparison of two data sets. In: IULC Working Papers, Vol. 22.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184. https://doi.org/10.1121/1.1906875. URL http://link.aip.org/link/?JAS/24/175/1.

Pfitzinger, H. R. & Niebuhr, O. (2011). Historical development of phonetic vowel systems - the last 400 years. In: *Proc. of the 17th International Congress of Phonetic Sciences (ICPhS), Hong-Kong, China* (pp. 160–163). URL https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/SpecialSession/Session7/Pfitzinger/Pfitzinger.pdf.

Ramanarayanan, V., Tilsen, S., Proctor, M., Töger, J., Goldstein, L., Nayak, K. S., & Narayanan, S. (2018). Analysis of speech production real-time mri. *Computer Speech & Language, 52*, 1–22. https://doi.org/10.1016/j.csl.2018.04.002. URL http://www.sciencedirect.com/science/article/pii/S0885230817301894.

Ribeiro, V., Laprie, Y. (2022). Autoencoder-based tongue shape estimation during continuous speech. In *Proc. Interspeech 2022*, (pp. 86–90). doi:10.21437/Interspeech.2022-10272.

Ribeiro, V., Isaieva, K., Leclere, J., Vuissoz, P.-A., & Laprie, Y. (2022). Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated. *Speech Communication, 141*, 1–13. https://doi.org/10.1016/j.specom.2022.04.004.

Russell, G. O. (1929). The mechanism of speech. *The Journal of the Acoustical Society of America, 1*(1), 83–109. https://doi.org/10.1121/1.1901471.

Serrurier, A. (2023). Can Deep Learning help to understand speech production mechanisms? In *Proceedings of the 34th Conference on Electronic Speech Signal Processing ESSV, Munich, Germany* (pp. 1–8).

Serrurier, A., & Badin, P. (2008). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *The Journal of the Acoustical Society of America, 123*(4), 2335–2355. https://doi.org/10.1121/1.2875111. URL http://link.aip.org/link/?JAS/123/2335/1.

Serrurier, A., & Neuschaefer-Rube, C. (2022). F1 and F2 formant variations and inter-speaker articulatory variability: a preliminary analysis. In *Proceedings of the 33rd Conference on Electronic Speech Signal Processing ESSV, Online* (pp. 172–179).

Serrurier, A., & Neuschaefer-Rube, C. (2023). Morphological and acoustic modeling of the vocal tract. *The Journal of the Acoustical Society of America, 153*(3), 1867–1886. https://doi.org/10.1121/10.0017356.

Serrurier, A., Badin, P., Lamalle, L., & Neuschaefer-Rube, C. (2019). Characterization of inter-speaker articulatory variability: a two-level multi-speaker modelling approach based on MRI data. *The Journal of the Acoustical Society of America, 145*(4), 2149–2170. https://doi.org/10.1121/1.5096631. URL https://asa.scitation.org/doi/10.1121/1.5096631.

Simpson, A. P. (2002). Gender-specific articulatory–acoustic relations in vowel sequences. *Journal of Phonetics, 30*(3), 417–435. https://doi.org/10.1006/jpho.2002.0171.

Soquet, A., Lecuit, V., Metens, T., & Demolin, D. (2002). Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication, 36*(3–4), 169–180.

Stevens, K. N. (2000). *Acoustic phonetics, current studies in linguistics*. MIT Press.

Sundberg, J., & Nordström, P.-E. (1976). Raised and lowered larynx - the effect on vowel formant frequencies. *Speech Transmission Laboratory - Quarterly Progress and Status Report - Stockholm, Sweden, 17*(2–3), 35–39.

van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between japanese and dutch women. *Language and Speech, 38*(3), 253–265. https://doi.org/10.1177/002383099503800303.

Vilain, C., Berthommier, F., & Boë, L.-J. (2015). A brief history of the articulatory-acoustic representation of vowels. In *International Workshop on the History of Speech Communication Research*. .

Wakita, H. (1973). Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics, 21* (5), 417–427. https://doi.org/10.1109/tau.1973.1162506.

Weirich, M., & Simpson, A. P. (2018). Individual differences in acoustic and articulatory undershoot in a german diphthong – variation between male and female speakers. *Journal of Phonetics, 71*, 35–50. https://doi.org/10.1016/j.wocn.2018.07.007.

Weirich, M., Fuchs, S., Simpson, A., Winkler, R. & Perrier, P. (2016). Mumbling: Macho or morphology? *Journal of Speech, Language, and Hearing Research* 59 (6) (dec 2016). doi:10.1044/2016_jslhr-s-15-0040.