Characterization of inter-speaker articulatory variability: A two-level multi-speaker modelling approach based on MRI data

Antoine Serrurier, Pierre Badin, Laurent Lamalle, Christiane Neuschaefer-Rube

Characterization of inter-speaker articulatory variability: A two-level multi-speaker modelling approach based on MRI data

Antoine Serrurier,[1,a)] Pierre Badin,[2] Laurent Lamalle,[3] and Christiane Neuschaefer-Rube[1]
[1]*Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Aachen, Germany*
[2]*Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France*
[3]*Inserm US 17—CNRS UMS 3552— Université Grenoble Alpes & CHU Grenoble Alpes, UMS IRMaGe, Grenoble, France*

(Received 4 May 2018; revised 12 March 2019; accepted 14 March 2019; published online 19 April 2019)

Speech communication relies on articulatory and acoustic codes shared between speakers and listeners despite inter-individual differences in morphology and idiosyncratic articulatory strategies. This study addresses the long-standing problem of characterizing and modelling speaker-independent articulatory strategies and inter-speaker articulatory variability. It explores a multi-speaker modelling approach based on two levels: statistically-based linear articulatory models, which capture the speaker-specific articulatory variability on the one hand, are in turn controlled by a speaker model, which captures the inter-speaker variability on the other hand. A low dimensionality speaker model is obtained by taking advantage of the inter-speaker correlations between morphology and strategy. To validate this approach, contours of the vocal tract articulators were manually segmented on midsagittal MRI data recorded from 11 French speakers uttering 62 vowels and consonants. Using these contours, multi-speaker models with 14 articulatory components and two morphology and strategy components led to overall variance explanations of 66%–69% and root-mean-square errors of 0.36–0.38 cm obtained in leave-one-out procedure over the speakers. Results suggest that inter-speaker variability is more related to the morphology than to the idiosyncratic strategies and illustrate the adaptation of the articulatory components to the morphology.
https://doi.org/10.1121/1.5096631

[ZZ] Pages: 2149–2170

I. INTRODUCTION

In order to be effective, speech communication relies on articulatory and acoustic codes shared between speakers and listeners (cf., e.g., Lindblom, 1990). These codes are obviously language-dependent but are expected to be speaker-independent, though "*most studies of speech production find some differences between speakers*" (Johnson *et al.*, 1993). Indeed, as emphasized by Ladefoged and Broadbent (1957), the "*socio-linguistic features*," related to the "*general background of the speaker*" and the "*idiosyncratic features*" of the speaker, are conveyed by speech on top of the linguistic information. These latter features may partly be "*due to anatomical and physiological considerations such as the particular shape of the vocal cavities*" (Ladefoged and Broadbent, 1957). In other words, a speaker is characterized by a specific *morphology*, i.e., the intrinsic size and shape of the speech articulators irrespective of the articulatory tasks, and adapts her/his *articulatory strategy*, i.e., the displacement and deformation of the speech articulators, to perform the speech task and achieve the articulatory-acoustic goals common to all speakers in a language. Disentangling the morphology variability from the articulatory strategy variability constitutes a challenging problem, which is tackled in the present study by means of a multi-speaker modelling approach.

Inter-speaker articulatory variability in speech production has been analyzed in a number of studies. Based on statistical analysis of the cross-section areas of the vocal tract, Mokhtari *et al.* (2007) illustrated the inter-speaker variability per phoneme while Story (2005, 2007) reported similar modes of variation around speaker-specific morphologies across speakers. By means of scaling transformations, Hashi *et al.* (1998), using X-ray microbeam data, showed that the variability of the palate morphology is partly compensated for by the speaker articulation. Using a similar technique, Geng and Mooshammer (2009) achieved a global reduction of the cross-speaker variability for electromagnetic articulography (EMA) data. Many studies, using various measurement methods such as X-ray microbeam, EMA, electropalatography or magnetic resonance imaging (MRI), showed that the shape and size of the palate have an influence on the articulations (Hashi *et al.*, 1998; Brunner *et al.*, 2005, 2009; Fuchs *et al.*, 2008; Geng and Mooshammer, 2009; Yunusova *et al.*, 2012; Rudy and Yunusova, 2013; Weirich and Fuchs, 2013). They emphasized, however, that the palate variability could not explain all the inter-speaker variability. Using X-ray scans, Honda *et al.* (1996) observed relationships between geometry and articulatory variation and suggested that speakers' vowel articulations adapt to the shape of their respective articulatory

a)Electronic mail: aserrurier@ukaachen.de

space, while consonant articulations seem to be independent of this space. More recently, Sorensen *et al.* (2016), using real-time MRI, highlighted the various strategies used by the speakers to achieve a target constriction in the vocal tract.

Linear statistically-based articulatory models based on principal component analysis (PCA), referred to as *linear articulatory models* or more simply as *linear models* in this article, have proved in the last decades to be powerful to extract and characterize the basic articulatory components of a speaker (Lindblom and Sundberg, 1971; Maeda, 1979, 1990; Hoole, 1999; Engwall, 2000; Beautemps *et al.*, 2001; Badin *et al.*, 2002). In such approaches, the correlations between the various shapes of the speech organs over the set of considered tasks are exploited to reduce the dimensionality of the models. A variant of this method, referred to as the *guided PCA* and detailed later in the article, aims at exploiting the sole correlations related to biomechanisms while excluding the correlations clearly related to pure control strategies.[1] The limited set of resulting components corresponds to a set of simple gestures considered as *independent degrees of freedom*, i.e., that are linearly uncorrelated and can be executed independently of each other by the articulators of the vocal tract (cf., e.g., Beautemps *et al.*, 2001).

Articulatory models, largely explored for single speakers, both in the two-dimensional (2D) midsagittal and the three-dimensional volume spaces, could be extended to multi-speaker models. The multi-speaker approach consists in modelling the vocal tract articulators simultaneously for a set of speakers performing the same speech tasks, while taking into account both the speaker-specific characteristics related to morphology and the idiosyncratic articulatory control strategies. The multi-speaker modelling approach has shown to be powerful in the context of variability analysis as it brings out the articulatory background common to speakers and determines to what extent each speaker complies with this common background. The Parallel Factor Analysis (PARAFAC), introduced by Harshman *et al.* (1977) in the field of articulation studies, appears by far to be the most popular approach in multi-speaker speech articulation studies. It aims at extracting a set of articulatory components considered as common to all the speakers. Similarly, the set of parameters controlling these components, i.e., the relative contribution of each component to achieve the articulations corresponding to given phonemes, is also common to all the speakers. Together, they represent what might be considered as the common articulatory background. Finally, a set of speaker-specific weights is determined to provide the contribution of each articulatory component for each speaker. The PARAFAC method assumes that all the speakers share the same sets of articulatory components and associated control parameters, and that they only differ in the amount of use of each component. The amount of data not explained by the model is therefore considered as purely speaker-specific. This simple clear-cut separation between universal and speaker-specific components could explain the popularity of this method. So far, it has mainly been used to study the inter-speaker articulatory variability, in most cases for the tongue, within one given language (Harshman *et al.*, 1977; Johnson *et al.*, 1993; Hoole, 1998, 1999; Geng and Mooshammer, 2000; Hoole *et al.*, 2000; Zheng *et al.*, 2003; Hu, 2006; Ananthakrishnan *et al.*, 2010; Valdés Vargas *et al.*, 2012), or sometimes across different languages (Lindau, 1986; Jackson, 1988; Nix *et al.*, 1996). In a cross-language framework, Linker (1982) applied this decomposition to the lips, articulators that were also succinctly considered together with the tongue by Johnson *et al.* (1993). PARAFAC has therefore proved to be a powerful tool to characterize both common articulatory background and speaker strategies. However, it suffers also from important limitations. Unlike the guided PCA approach mentioned earlier, the decomposition underlying PARAFAC appears less conducive to obtaining articulatory components related to plausible underlying biomechanisms. The components are not ensured to be uncorrelated, which makes the model design sometimes very challenging and relying strongly on the modeler's expertise (cf., e.g., the re-analysis and re-interpretation of Jackson's (1988) data by Nix *et al.*, 1996). Mathematically, the algorithm might not converge, or might converge to local minima, and must thus be explicitly monitored by an expert modeler. In terms of pre-requisites, it assumes that "*the ratio of any two speakers' usage of a given [articulatory component] must be the same for all [considered articulations]*" (Harshman *et al.*, 1977). In other words, "*if speaker A uses more of [component] 1 than does speaker B for a particular [articulation], then speaker A must use more of [component] 1 than speaker B in all other [articulations]*" (Harshman *et al.*, 1977). This assumption may, however, not systematically hold. As reported by Hoole (1999), PARAFAC showed moreover some limitation in modelling various consonant contexts. To overcome some of these limitations, extensions have been proposed in the literature (Hoole, 1999; Geng and Mooshammer, 2000), usually at the cost of a higher number of parameters to estimate. The issue of the number of components and parameters to estimates has also been discussed by Valdés Vargas *et al.* (2012) and Valdés Vargas (2013). Alternative approaches have also been explored and compared with PARAFAC by Ananthakrishnan *et al.* (2010), Valdés Vargas *et al.* (2012), and Valdés Vargas (2013). Their joint PCA approach, based on the two-way decomposition of the speakers' data assembled per articulation, seems to provide better results than the PARAFAC and the direct three-way decomposition proposed by Tucker (1966), though at a "*much higher*" cost in the number of parameters to estimate than PARAFAC (Ananthakrishnan *et al.*, 2010). The Tucker and PARAFAC decompositions require also setting in advance the number of the various modes of variation. Besides, the Tucker decomposition is a "*method with a more complex structure and more parameters than joint PCA*" (Valdés Vargas, 2013, p. 73). In addition, all the decompositions are usually performed on centered data, i.e., the averaged data are first subtracted per speaker separately, which prevents taking into account the inter-speaker variability carried by the averages.

Regarding the speech material analyzed, almost all the studies mentioned above considered vowels only, sometimes in different consonant contexts, on a few speakers. As far as we know, only Hoole *et al.* (2000) with nine speakers, Ananthakrishnan *et al.* (2010) with three speakers, Valdés

Vargas *et al.* (2012) with seven speakers, and Valdés Vargas (2013) with 11 speakers considered also consonants in their corpora.

The ambition of the present study was to explore alternative approaches overcoming the previously mentioned limitations of multi-speaker articulatory modelling. The first issue is to ensure a large coverage of the speech repertoire of the studied language by relying on a substantial set of vowels and consonants acquired on a substantial set of speakers. The present work makes use of the multi-speaker set of French MRI data initially collected by Ananthakrishnan *et al.* (2010), Valdés Vargas *et al.* (2012), and completed by Valdés Vargas (2013). The second aim is to build an organ-based multi-speaker articulatory model of the whole vocal tract, from the larynx to the lips, based on the guided PCA so that the articulatory components can be related to plausible movements in terms of biomechanisms. The present approach explores a modelling in two levels to solve this problem: individual articulatory model parameters obtained for each speaker by a classical two-way decomposition, representing the first level models, are grouped together and further decomposed themselves by a two-way decomposition, representing the second level model. This *two-level* architecture leads to what might be called a *model of models (MoM)*. The third focus of the study is to characterize the variability of the morphology and of the strategy among speakers and to evaluate their relative contribution to the articulatory realizations. As the proposed approach aims to maintain a low dimensionality for the models, so that every speaker can be represented by a very limited number of parameters, the correlations between the morphology and the strategy are exploited as much as possible in the modelling design. A complementary analysis takes advantage of the two-level modelling approach presented here to attempt to disentangle them. For these purposes, in order to ensure capturing all the inter-speaker variability, the speakers' average articulations, usually discarded from statistical analyses in the literature, are also considered.

The manuscript is organized as follows: Sec. II presents the modelling approach and the data, Sec. III details the whole vocal tract articulatory model, Sec. IV the multi-speaker modelling, while Sec. V characterizes the morphology and strategy variability, and Sec. VI discusses the results and draws conclusions.

II. APPROACH AND MATERIAL

A. Linear statistically-based articulatory modelling: Formulation and terminology

In a linear articulatory model obtained by PCA from a set of observations, the shape of an articulator for a given phone—or articulation—is expressed as a linear combination of *eigenvectors* forming an orthogonal basis of the organ shape space, weighted by the set of *control parameters* corresponding to the articulation. Mathematically, a column vector of articulatory measures $\boldsymbol{x}_i$ of an articulator contour for an articulation i is decomposed as follows (to simplify, the equation is presented in a vector mode and the index j corresponding to the points has been omitted):

$$\boldsymbol{x}_i = \sum_{k=1}^{n_q} p_{ik}\boldsymbol{e}_k + \bar{\boldsymbol{x}} + \boldsymbol{\varepsilon}_i, \tag{1}$$

where: (1) $X = (x_{ij})$, $i = 1\ldots n_a$, $j = 1\ldots 2n_c$ are the coordinates of the n_c contour points of the considered articulator for the set of the n_a observed articulations, (2) $E = (e_{kj})$, $k = 1\ldots n_q$, $j = 1\ldots 2n_c$ the n_q (orthogonal) eigenvectors of the articulator contours, (3) $P = (p_{ik})$, $i = 1\ldots n_a$, $k = 1\ldots n_q$ the weights for the n_q eigenvectors of the n_a articulations, (4) $\bar{\boldsymbol{x}} = (x_j), j = 1\ldots 2n_c$ the average of the articulator contour coordinates over the n_a articulations and (5) $R = (\varepsilon_{ij})$, $i = 1\ldots n_a$, $j = 1\ldots 2n_c$ the residue not explained by the former parameters for each coordinate value. R represents the modelling error.

The *eigenvectors* in the matrix E are also referred to as *factors* or *modes* in previous publications (e.g., Harshman *et al.*, 1977). These vectors are controlled by the weighting matrix P, thereafter the *predictors*, also referred to in the literature as *control parameters* (Beautemps *et al.*, 2001; Badin *et al.*, 2002), *factor loadings* (e.g., Harshman *et al.*, 1977) or *[weighting] coefficients* (e.g., Story, 2005, 2007). In the formulation adopted in the present article, the eigenvectors of the matrix E are associated with the notion of *articulatory components*. They correspond to the degrees of freedom of the articulators described in the introduction. The *articulatory model* is finally defined by the association of the matrix of *eigenvectors* E with the *mean articulation* vector $\bar{\boldsymbol{x}}$.

A schematic representation of a linear articulatory model is proposed in Fig. 1. Note that the term *articulation* will sometimes be used in the manuscript instead of "the coordinates of the contour points of an articulation" to simplify the formulation.

B. Speakers

At present, medical imaging techniques such as MRI can easily generate large amounts of data (e.g., Narayanan *et al.*, 2014), but processing them to obtain segmented articulators is still an arduous problem (e.g., Labrunie *et al.*, 2018). As a trade-off between the necessity to elicit as much as possible inter-speaker variability and keeping a manageable amount of data to process, 11 speakers have been recorded by means of static MRI in a previous study (Valdés Vargas, 2013): six male and five female, with ages spanning from 18 to 48 at the time of the recordings (cf. Table I). All the speakers were native French speakers and uttered the same corpus of artificially sustained French articulations.

C. Corpus

The corpus, designed to be balanced and representative of the French phonemic repertoire, consisted of the ten French oral vowels [i e ɛ a y ø œ u o ɔ], two of the four nasal vowels [ɑ̃ ɔ̃], and of ten consonants [p t k f s ʃ m n ʁ l] in five symmetric vowel consonant vowel (VCV) contexts [i e ɛ a u], leading to a total of 62 articulations. For the consonants, the speakers were instructed to repeat the VCV sequences a few times in a natural manner, and then to freeze the

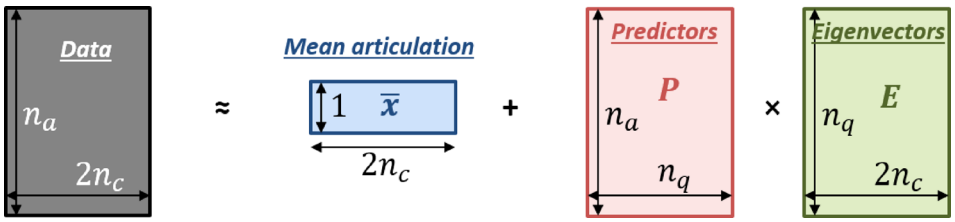

FIG. 1. (Color online) Schematic representation of a single speaker linear articulatory model. See the text for the definitions of the variables. The replication of the line vector $\bar{x}$ in n_a rows was omitted to enhance the comprehension. The color code will be followed in the other figures to facilitate the reading: grey for the data, blue for the mean articulation, red for the predictors, and green for the eigenvectors.

consonant for the last repetition; the scan was started as soon as the operator heard that the consonant was hold. This protocol ensures that the consonant is truly coarticulated with the vowel. Due to the difficulty to sustain the articulations for several seconds during data collection, the voiced consonant counterparts have been discarded. They are, however, expected to bring little additional variability. Similar corpora have been extensively used in the past for such articulatory studies (cf., e.g., Serrurier and Badin, 2008), following the validation proposed by Beautemps *et al.* (2001).

D. Data

Midsagittal static MRI data were recorded at IRMaGe MRI facility (Grenoble, France) with Achieva 1.5T and Achieva 3.0T TX scanners (Philips, Best, The Netherlands) between 2010 and 2012 for 10 out of 11 speakers. One speaker was recorded at ATR (Japan) in 2002 with a Marconi Eclipse 1.5 T scanner. The speakers, in the supine position, were instructed to sustain each of the 62 articulations between 8 to 24 s without movement. During this time, at least one midsagittal image was acquired, spanning at least from below the glottis to above the nasal tract in the vertical direction and from before the external nose to behind the neck in the horizontal direction (cf., e.g., Fig. 2). The characteristics of the images and of their acquisition sequences are displayed in Table I.

In order to outline the contours of the teeth, which cannot be distinguished from the air on MRI, additional articulations have been recorded where the soft tissues (lips, tongue) are in contact with the teeth to materialize their contours by contrast. The boundaries of the upper teeth and hard palate as well as those of the lower teeth and jaw bone have been manually outlined on such reference images for each speaker, using B-splines and control points. These two rigid contours are referred to as the *palate* and *jaw* in the rest of the article. Similarly, the outlines of the hyoid bone in the midsagittal plane have been manually segmented on one of the 62 articulations. The contours of these three rigid structures have then been manually aligned by means of a rigid 2D transformation (translation and rotation) for each of the 62 images for each speaker.

The following deformable contours of the vocal tract relevant for speech have also been manually traced on each of the 62 images for all the speakers using B-splines and control points: the upper and lower lips (extended along the face contours respectively up to the nose and down to the larynx prominence), the tongue, velum, pharyngeal wall (referred to as *pharynx* in the rest of the article), the epiglottis and posterior supraglottic region (referred to as *posterior supraglottis* in the rest of the article[2]). An example of MRI with manually superimposed contours is visible in Fig. 2 (yellow lines). These structures will be referred to as speech *articulators* in the following, although this term might be in question for the structures such as the pharynx and posterior supraglottis.

A series of landmarks attached to these contours or to anatomical features has been determined, as visible in Fig. 2 (green points). They correspond to geometrical and anatomical singularities such the lower tip of the velum, apex of the tongue, vermilion border, etc. While some landmarks correspond to obvious characteristics, such as the velum tip,

TABLE I. List of speakers and MRI data characteristics. *F* = Female; *M* = Male; 1.5T = Philips Achieva; 1.5T-ME = Marconi Eclipse; 3T = Philips Achieva 3.0T TX; n/a = not available. TE: echo time; FA: flip angle; TR: repetition time; FOV: field of view.

Speaker	Age(years)	Imaging System	TE (ms)	FA	TR (ms)	Slice Thickness (mm)	FOV (mm^2)	Resolution (mm/px)	Acquisition Time (s)
f1	26	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
f2	22	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
f3	42	*1.5T*	3.5	12°	7.7–8.6	1.25	230 × 160	0.958	21.8–24.1
f4	31	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
f5	18	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
m1	29	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
m2	27	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
m3	29	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
m4	47	*1.5T-ME*	n/a	n/a	n/a	5	256 × 256	1	n/a
m5	24	*3T*	10.74	80°	4.26	4	256 × 256	1	8.1
m6	48	*1.5T*	3.5	12°	5.6	1.25	230 × 160	0.958	15.7
Mean	31								

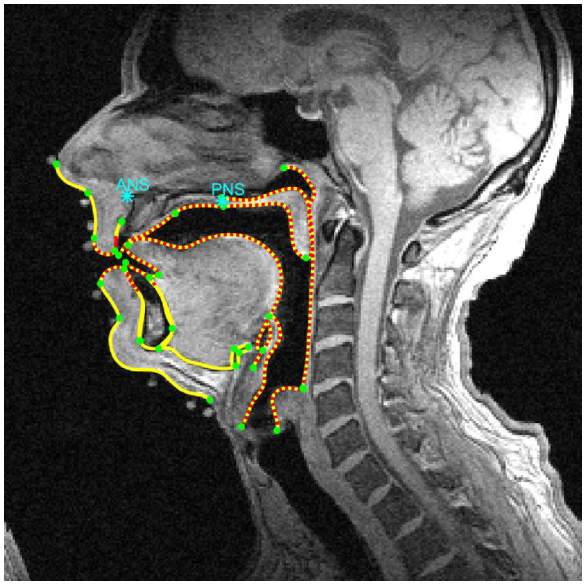

FIG. 2. (Color online) Example of MRI image superimposed with the manually segmented contours (solid yellow lines), their restriction to the vocal tract region (dotted red lines), the contour landmarks (green points), and the two ANS and PNS anatomical landmarks (cyan stars). Note that the small grey spots that can be seen along the external profile correspond to fiducial markers not exploited in the present study.

others are more uncertain and have been determined at best on the images based on the expert's experience. This is for instance the case of the landmark of the sublingual cavity, not visible for all articulations (cf. Ananthakrishnan *et al.*, 2010) but nevertheless estimated for all of them.

All the contours and landmarks have been aligned for each speaker on his/her common palate shape. Two further anatomical landmarks have been identified: the *Anterior Nasal Spine* (ANS) and the *Posterior Nasal Spine* (PNS). In order to enhance the reliability of the determination of these points, they have been manually located on all of the 62 images with a mouse click and their averages have been calculated for each speaker after alignment on the speaker common palate shape. This has resulted for each speaker in two single points ANS and PNS rigidly connected with the palate (cf. cyan stars in Fig. 2). Using these points, the contours of each of the $n_s \times n_a = 11 \times 62$ articulations have been aligned in a common anatomical space: after transformation into centimeters, they have been rotated and translated so that the ANS-PNS line is horizontal and the origin of the coordinate system is set so that the upper incisors lower edge coordinates are arbitrarily at $X = 5$ cm and $Y = 10$ cm.

Finally, the contours have been resampled between each landmark with a fixed number of points in order to build a dataset consistent across articulations and speakers. The fixed number of points for each section has been set so that the average distance between two consecutive contour points remains close to an arbitrary value, although the length of each sub-contour may vary depending on the articulation and speaker. On this occasion, the contours of the tongue, upper lip and lower lip have been restricted to the more relevant regions that influence the vocal tract shape and acoustics: the tongue from its root junction with the epiglottis up to its attachment to the jaw, and similarly the upper lip (respectively, lower lip) from the contact with the palate (respectively, jaw) until the external border of the vermilion (cf. Fig. 2).

III. ARTICULATORY MODEL OF THE MEAN SPEAKER

In an attempt to develop an articulatory model which retains general but not speaker-specific characteristics, a (virtual) mean speaker with 62 articulations has been created by averaging the contours of the 11 speakers for each articulation. An articulatory model of this speaker, which represents the average articulatory space of the set of speakers, is expected to smooth out speaker-specific articulatory strategies while retaining the speakers' common language background, and thus concentrates on the components common to the speakers. This kind of general speaker will thereafter be referred to as the *mean speaker*. Note that this speaker can be considered as an intermediate between male and female due to the approximately balanced number of males (six) and females (five) in the dataset.

Following the general principles described in Secs. I and II A, the guided PCA has been applied to the contour coordinates of the 62 articulations for each articulator of the mean speaker, in a hierarchical order, as detailed in Table II. The number and nature of components retained for each articulator depend on the desired percentage of explained variance, residual root-mean-square (rms) error and likelihood in terms of biomechanical interpretation. The percentage of variance explanation and the rms reconstruction error are provided for each articulator in Table III. The nomograms of the contours, i.e., the variation of the contours resulting from the variations of the articulatory predictors by regular steps between the minimal and maximal values found in the data are displayed in Fig. 3 for all the articulatory components. Note that the tongue component TT, which appears to be the fifth component in terms of variance explanation in the PCA, is presented in this study as the fourth component, before the component TR, to maintain the consistency of interpretation with previous publications (Badin and Serrurier, 2006; Valdés Vargas, 2013).

The overall model of the vocal tract articulators is made of 14 articulatory components, leading to a global rms reconstruction error of 0.05 cm and a variance explanation of 96%. The variance of each articulator is explained up to a level varying from 83%–86% (pharynx, epiglottis, posterior supraglottis) to 93%–98% (tongue, lips, velum). This model provides therefore a very accurate reconstruction of the 62 mean articulations of the corpus.

The JH component, a nearly vertical displacement of the jaw, partly controls the lips, and also the tongue, with a higher amplitude in the front region. TB corresponds to an oblique backward-frontward movement of the tongue and of the epiglottis, TD to an oblique flattening-arching movement of the tongue, and TT to an oblique upward-downward movement of the tongue tip. These observations are in agreement with previous results (e.g., Beautemps *et al.*, 2001), though they concern here an average speaker rather than individual speakers, for whom these components may vary significantly (cf. Valdés Vargas, 2013). Interestingly, an additional component TR accounting for 13% of the

TABLE II. Pseudo-code description of the guided PCA process for the mean speaker. All contours were initially centered, i.e., their means over the 62 articulations subtracted. Predictors and eigenvectors described in Sec. II A and in Fig. 1 are postfixed, respectively, with "*p*" and "*e*" for each component. LR stands for *Linear Regression*. All predictors are *z*-scored and the associated eigenvectors accordingly adjusted when required. Associated articulatory nomograms are displayed in Fig. 3.

Articulator	Component	Predictor and eigenvector
Jaw	JH (Jaw Height)	- JHp = vertical coordinate of the upper point of the lower teeth (LT)
		- JHe_jaw = LR of LT on JHp
		- Jaw contour follows the rigid displacement of LT according to its JH component
Upper lip	JH	- JHe_upperlip = LR of upper lip contour on JHp
	ULH (Upper Lip Height)	- Residual contour = upper lip contour - JHp × JHe_upperlip
	ULP (Upper Lip Protrusion)	- ULHp = vertical coordinate of the outer point of the residual contour
		- ULHe = LR of residual contour on ULHp
		- Residual contour = residual contour - ULHp × ULHe
		- ULPp = horizontal coordinate of the outer point of the residual contour
		- ULPe = LR of residual contour on ULPp
Lower lip	JH	- JHe_lowerlip, LLHp and LLHe, LLPp and LLPe are determined in the same way as for the upper lip, but on the lower lip contour
	LLH (Lower Lip Height)	
	LLP (Lower Lip Protrusion)	
Tongue	JH	- JHe_tongue = LR of the tongue contour on JHp
	TB (Tongue Body)	- Residual contour = tongue contour - JHp × JHe_tongue
	TD (Tongue Dorsum)	- TBp and TDp = predictors obtained by PCA on the back region of the residual contour
	TT (Tongue Tip)	- TBe_tongue and TDe = LR of the entire residual contour on TBp and TDp
	TR (Tongue Rounding)	- Residual contour = residual contour – TBp × TBe_tongue - TDp × TDe
		- TTp and TTe and TRp and TRe = eigenvectors and predictors obtained by PCA on the entire residual contour
Velum	VL (Velum Levator)	- VLp and VLe and VSp and VSe = eigenvectors and predictors obtained by PCA on the velum contour
	VS (Velum Shape)	
Pharynx	LH (Larynx Height)	- LHp (Larynx Height) = vertical coordinate of the center of gravity of the glottis contour
	PH (Pharynx Horizontal)	- LHe_pharynx = LR of the pharynx contour on LHp
		- Residual contour = pharynx contour - LHp × LHe_pharynx
		- PHp and PHe_pharynx = eigenvectors and predictors obtained by PCA on the residual contour
Epiglottis	LH	- LHe_epiglottis = LR of the epiglottis contour on LHp
	TB	- Residual contour = epiglottis contour - LHp × LHe_epiglottis
	EH (Epiglottis Horizontal)	- TBe_epiglottis = LR of the residual contour on TBp
		- Residual contour = residual contour - TBp × TBe_epiglottis
		- EHp and EHc = eigenvectors and predictors obtained by PCA on the residual contour
Posterior supraglottis	LH	- LHe_posteriorsupraglottis = LR of the posterior supraglottis contour on LHp
	PH	- Residual contour = posterior supraglottis contour - LHp × LHe_ posteriorsupraglottis
		- PHe_posteriorsupraglottis = LR of the residual contour on PHp

variance of the tongue has been highlighted. It corresponds to a bunching-flattening movement complementary to those controlled by TB and TD: [a] (flat) and [u] (round) are associated with opposed extreme values of TR, while [i] is associated with an intermediate value. ULH, LLH, ULP, and LLP correspond to the height and protrusion of the upper and lower lips (cf. Badin *et al.*, 2013). VL corresponds to an oblique movement of the velum from a low flat position to a high position folded in its middle, while VS corresponds mainly to a small frontward-backward movement with the greater amplitude on the uvula, which is in general agreement with earlier observations on a single speaker (Serrurier and Badin, 2008). Note that the VS movement may partly correspond to a shift of the velum mechanically pushed by the tongue observed on several articulations for several speakers (Valdés Vargas, 2013) and may not result from muscular activity. Interestingly, the movement of the velum for VL is not associated with a movement of the upper region of the pharyngeal wall to form a nasopharyngeal sphincter, referred to as Passavant's pad (Passavant, 1869) in the literature and already observed on a single speaker (Serrurier and Badin, 2008). LH, representing a vertical displacement of the larynx, is related to the vertical movements of the organs in the neighborhood, namely, the epiglottis and posterior supraglottis and by extension the pharynx. The remaining pharynx component PH corresponds to a frontward-backward movement, associated with a similar frontward-backward movement of the posterior supraglottis. Finally, the remaining component of the epiglottis EH corresponds to a frontward-backward tilting movement.

It is worth mentioning significant correlations between predictors. The upper and lower lip protrusions have a correlation coefficient of 0.74, in line with earlier findings (Beautemps *et al.*, 2001, found 0.91, while Abry and Boë, 1986, found 0.98 on a corpus without labiodentals). A correlation of 0.63 has also been observed between lower lip

TABLE III. Percentage of variance explanation per articulatory component (column "%") and rms reconstruction error cumulated over the articulatory components in cm (column "cm") for each articulator for the model of the mean speaker.

	Jaw		Tongue		Upper lip		Lower lip		Velum		Pharynx		Epiglottis		Posterior supraglottis	
	%	cm	%	cm	%	cm	%	cm	%	cm	%	cm	%	cm	%	cm
JH	91	0.06	21	0.35	9	0.15	18	0.20								
LH											57	0.06	47	0.15	66	0.10
TB			33	0.27									13	0.13		
TD			24	0.18												
TT			7	0.15												
TR			13	0.05												
ULP					49	0.10										
ULH					38	0.03										
LLP							48	0.13								
LLH							27	0.06								
VL									92	0.05						
VS									6	0.02						
PH											25	0.04			20	0.06
EH													24	0.08		
Cum.	**91**	**0.06**	**98**	**0.05**	**96**	**0.03**	**93**	**0.06**	**98**	**0.02**	**83**	**0.04**	**83**	**0.08**	**86**	**0.06**

protrusion and larynx height: this correlation has already been documented by Beautemps *et al.* (2001), who found a correlation of 0.66, and by Hoole and Kroos (1998); they ascribed this covariation to a control strategy aiming at maximizing overall vocal tract length variations to enhance acoustic differences between /i/ and /u/.

Although Johnson (1991) already performed analyses on data averaged over several speakers, this model constitutes as far as we know the only organ-based articulatory model of the full vocal tract representing an average speaker. It highlights the common articulatory background between speakers, but does not take into account the speaker-specific strategies. Section IV focuses on this aspect.

IV. MULTI-SPEAKER MODELLING

As explained in Sec. I, multi-speaker articulatory modelling aims at modelling the vocal tract articulators simultaneously for a set of speakers producing the same speech tasks. Multi-speaker modelling can be handled in various ways. Four different approaches are explored and compared in this section. Section IV A describes the design, while Sec. IV B presents an evaluation.

A. Model design

The multi-speaker modelling presented in this study relies on a two-level linear modelling approach. The first

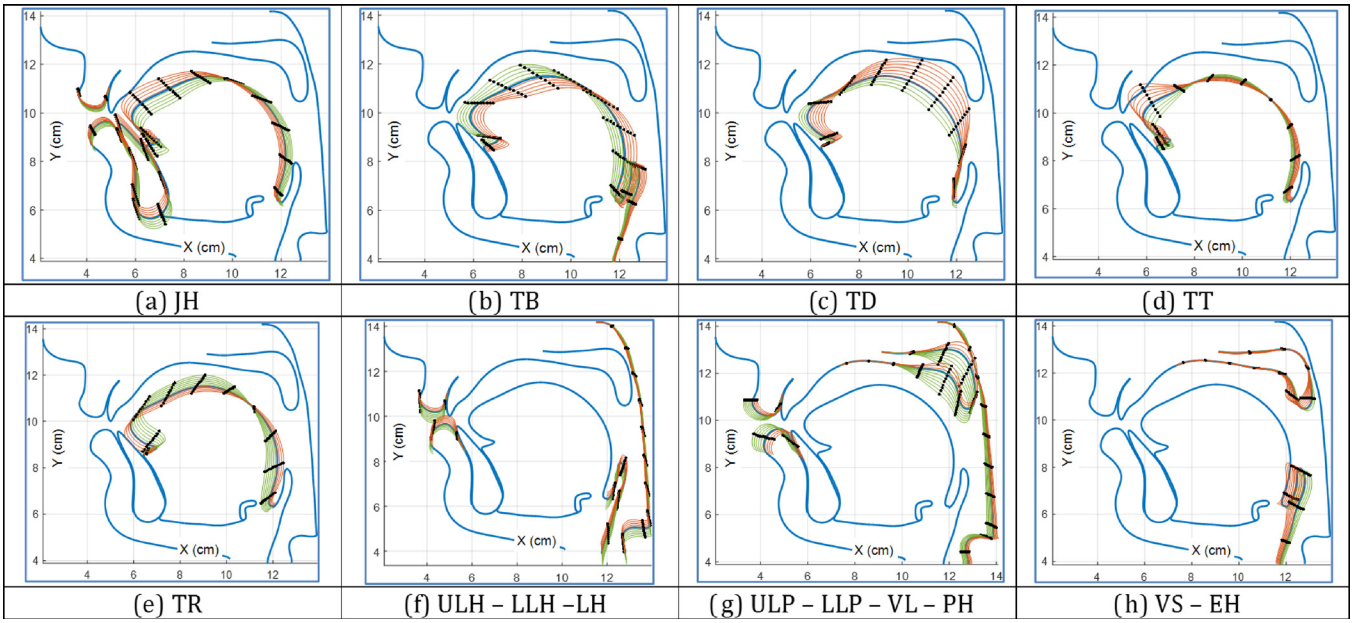

FIG. 3. (Color online) Nomograms of the contours for the 14 articulatory components of the mean speaker for predictors varying at regular steps between the minimal and maximal values found in the data. Contours with negative (respectively, positive) predictor values are plotted in green (respectively, orange); one every 25 points is plotted as a black dot to emphasize deformation directions. The full contour of the average articulation is displayed (in blue) for better comprehension. Some components are grouped on the same picture (*f*, *g*, and *h*) and some components apply to several articulators (JH on *a*, TB on *b*, LH on *f* and PH on *g*).

level model is a speaker-specific model of articulation, which provides speaker-specific articulations from *articulatory predictors*. The second level model is a model of speaker, which controls the parameters of the articulatory model of a given speaker from *speaker predictors*. This latter model is therefore a model of models. An overview is displayed in Fig. 4. Being speaker-specific, the first level articulatory model represents the intra-speaker variability, while the associated predictors control the realization of the articulation produced with the specific morphology and strategy of the speaker. The second level model represents the inter-speaker variability: the associated predictors control the parameters of the speakers' articulatory models, including their specific morphology and strategy, and possibly the articulatory predictors. Together, the first level articulatory model and the second level speaker model constitute a *two-level multi-speaker model*.

The multi-speaker models are derived from the data. The individual articulatory model parameters are obtained for each speaker by a standard two-way decomposition [as in Eq. (1) and Fig. 1]; the parameters of all the speakers are then concatenated and further modelled using another two-way decomposition to obtain the parameters of the second level speaker model. The articulatory components identified in the first level models are expected to be similar for different speakers, as they aim at the same phonetic goals, but may also underlie slightly different speaker-specific strategies. This has been emphasized by Valdés Vargas *et al.* (2012) who carried out qualitative inter-speaker comparisons of homologous components with similar interpretations or functions. The underlying idea of the second level model is to take advantage of these similarities and of the possible correlations among speakers between the morphology and the strategy parameters to uncover the principal strategy and morphology components common to the speakers, namely, the speaker components.

In the current formulation, the parameters of the first level models are the n_s sets of articulatory models, i.e., the eigenvectors E and the mean articulations $\bar{\boldsymbol{x}}$, representing, respectively, the speakers' strategies and morphologies, together with the n_s sets of articulatory predictors P. The four approaches explored are detailed in the following.

1. General model

In this model, the first level corresponds to a speaker-specific articulatory model together with the associated articulatory predictors, i.e., the model and articulatory predictors

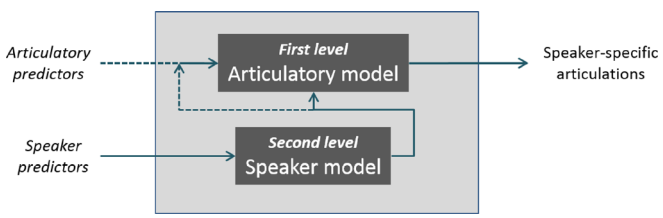

FIG. 4. (Color online) Schematic general overview of the two-level multi-speaker model. Note that, in some cases, the speaker model may also control the articulatory predictors, as illustrated by the dashed lines.

that represent the articulations in the space of the speaker. The second level model aims to represent the speaker-specific articulatory model parameters, i.e., the parameters P, E, and $\bar{\boldsymbol{x}}$ related to the first level. The second level model is controlled by a small number of speaker components n_g representing the morphology and the articulatory strategy of the speaker. As schematized in Fig. 5, the model is built as follows. First, n_s individual articulatory models together with their predictors are built using the methodology described in Sec. III and form the first level models. The homologous articulatory components in these models are assumed to lead to contour deformations approximately in the same directions for all the speakers.[3] Then, for each speaker, the P, E, and $\bar{\boldsymbol{x}}$ parameters are recast in a single line vector and the resulting n_s line vectors are finally stacked in a single matrix (cf. Fig. 5). The second level model is obtained by submitting this matrix directly to PCA so as to exploit the covariations between speakers and to reduce the dimensionality of the model parameters. Note that the P, E, and $\bar{\boldsymbol{x}}$ parameters composing this matrix are of different nature, with different units, and may have different orders of magnitude; a simple PCA might therefore generate components mainly explaining the variance of the parameters with the largest order of magnitude. The other parameters that would be poorly taken into account by this PCA might, however, play an important role in the first level models despite their lower order of magnitude. This potential problem has been dealt with by attributing different weights to the three sets of parameters before applying the PCA. E has been arbitrarily chosen as reference, while empirical weightings for P and $\bar{\boldsymbol{x}}$ have been determined automatically so as to minimize the global reconstruction error of the multi-speaker model, leading, respectively, to values of 0.4 and 1.2.

In the second level, the first principal components determined with the PCA are expected to represent simultaneously the morphology and strategy variations of the speakers. They will be referred to in the following as SPg, standing for *SPeaker* components of the *general model*. This MoM simultaneously reconstructs speaker-specific articulatory models and speaker-specific articulatory predictors with a limited number n_g of second-level speaker predictors.

2. Universal predictor model

The universal predictor model approach is similar to the general model approach, but aims at ensuring universal, i.e., speaker-independent, articulatory predictors. As schematized in Fig. 6, the model is built in two steps as follows.

In the first step, n_a sets of articulations pooled over speakers are formed, i.e., made of the concatenation of the articulations of each speaker, and referred to as *pooled articulations* in the following. This approach is inspired by the work of Ananthakrishnan *et al.* (2010), Valdés Vargas *et al.* (2012), and Valdés Vargas (2013), where it is referred to as *two-level PCA* or *joint PCA* (this latter terminology will be used in the following to refer to their work). For each articulator, the pooled articulations are analyzed by the guided PCA as described in Table II. This generates a single matrix of articulatory predictors P, as well as a single matrix of

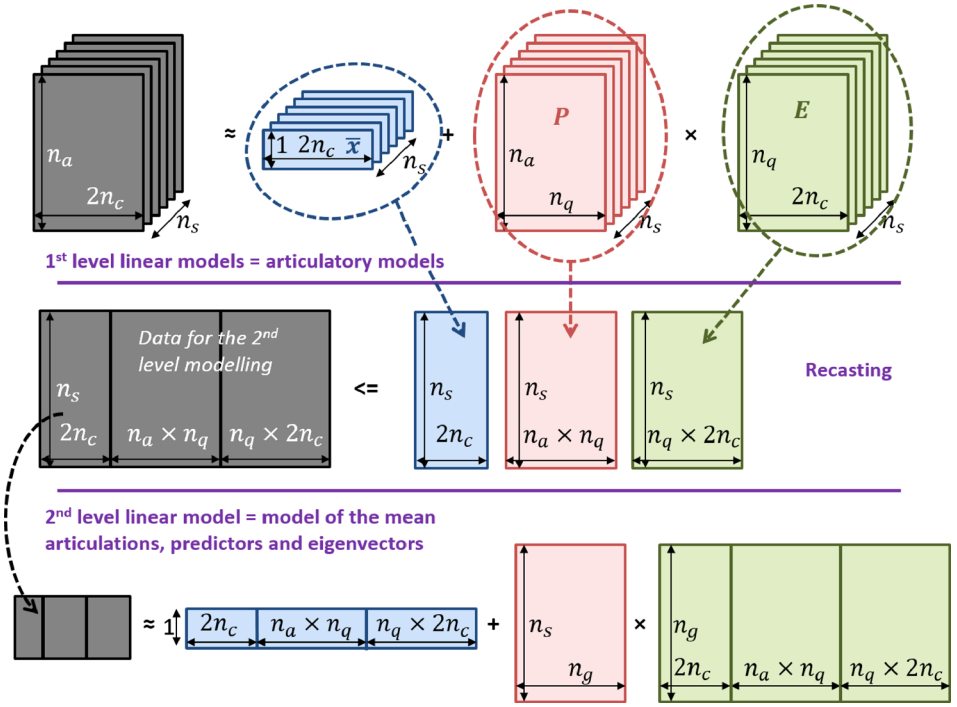

FIG. 5. (Color online) Schematic representation of the data analysis procedure for the ***general model***. Refer to Fig. 1 for the color conventions. See the text for the definitions of the various indices. As in Fig. 1, the replication of the (blue) line vectors in n_a (for the top blue line vectors) and n_s (for the bottom blue line vector) rows was omitted to enhance the comprehension.

pooled eigenvectors and a single vector of average pooled articulations. Next, the matrix of pooled eigenvectors and the vector of average pooled articulations are split and redistributed to deliver individual sets of E and $\bar{x}$ parameters for the n_s speakers, while the articulatory predictors P remain unique and common to all the speakers for each articulation. These parameters form the first level of the multi-speaker model.

In the second step, the E and $\bar{x}$ parameters for the n_s speakers are concatenated and stacked into a single matrix. The articulatory predictors P, being speaker-independent, do not need to be modelled in the second-level. As for the general model, relative weightings are first applied to the E and $\bar{x}$ parameters: E is arbitrarily chosen as reference; an empirical weighting of 1.05 obtained automatically is applied to $\bar{x}$. A PCA is subsequently applied to these data, and the n_g first speaker components are retained to establish a model, namely, the second-level model, able to estimate the speaker-specific E and $\bar{x}$ parameters of the individual articulatory models. These components will be referred to in the

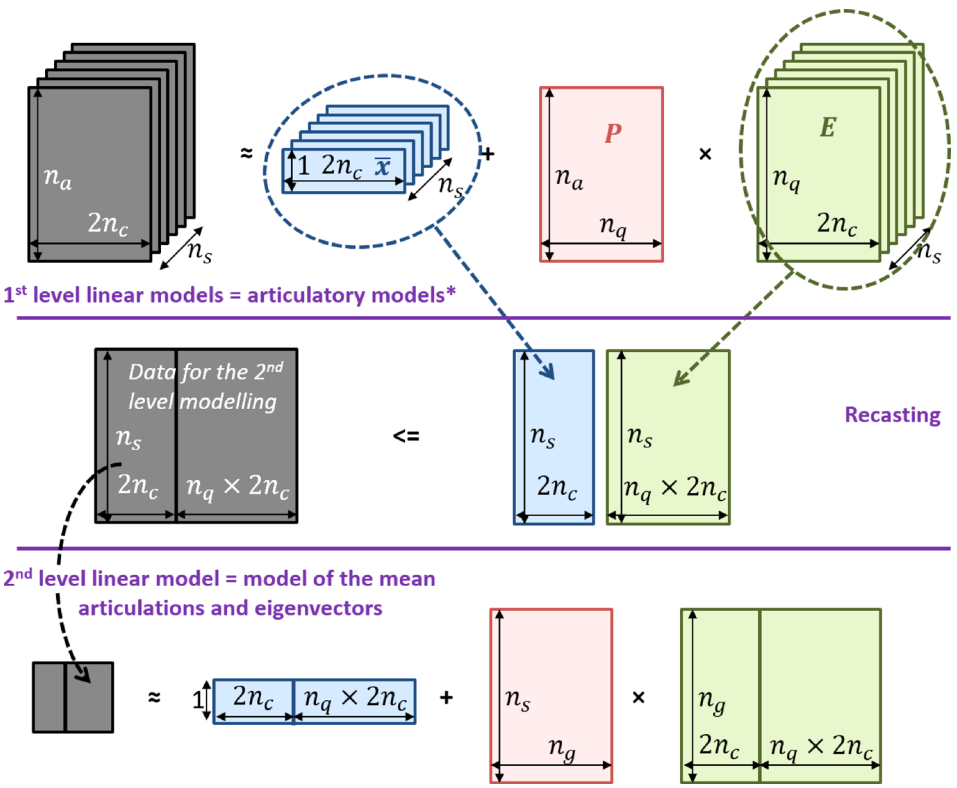

FIG. 6. (Color online) Schematic representation of the data analysis procedure for the ***universal predictor model***. Refer to Fig. 1 for the color conventions. See the text for the definitions of the various indices. The copy of the predictor matrix P (in red) for all the speakers is not represented to simplify the schema and emphasize the speaker-independent property. As in Fig. 1, the replication of the (blue) line vectors in n_a (for the top blue line vectors) and n_s (for the bottom blue line vector) rows was omitted to enhance the comprehension (color online). *Note also that the first step generating the articulatory models from pooled articulations (see the text for details) is not displayed in the figure.

following as SPup, standing for *SPeaker components of the universal predictor model*. In this model, the matrix of articulatory predictors P bears "universal predictors," i.e., the articulation of a given phoneme is controlled by the same set of predictors for all the speakers. Note that all the inter-speaker variability is thus transferred to the E and $\bar{x}$ parameters.

3. Universal eigenvector model

This third approach assumes that the differences between speakers are mainly related to the range and combination of their use of the same articulatory components rather than to the use of different components. In other words, the same set of speaker-independent eigenvectors E is controlled by speaker-specific sets of articulatory predictors P for each speaker. The procedure to achieve this model is schematized in Fig. 7.

In the first stage of the procedure, the speaker-specific mean articulations $\bar{x}$ are calculated. Then, the set of eigenvectors E of the model of the mean speaker described in Sec. III are imposed as the universal eigenvectors. Next, the matrix P of articulatory predictors corresponding to these eigenvectors are determined for each speaker and each articulation: this is basically carried out by inversing Eq. (1), knowing the speaker's data as well as the matrices E and $\bar{x}$. Specifically, for each articulator and each speaker, the mean articulation is subtracted from the data, and the predictors P are determined by pseudo-inversing iteratively the matrix E in accordance with the scenario described in Table II. For each iteration, once the predictors are determined for one component, the contribution of this component is subtracted from the residual data before estimating the predictors of the next component. This method finally yields a set of P and $\bar{x}$ parameters for each speaker. These parameters, together with the universal matrix E, form the first level of the multi-speaker model.

In the second stage of the procedure, the P and $\bar{x}$ parameters for the n_s speakers are concatenated into a single matrix. The eigenvectors E, being speaker-independent, do not need to be modelled in the second-level. As for the general and the universal predictor models, weightings need to be applied to the P and $\bar{x}$ parameters: P is arbitrarily chosen as reference, while the weighting of $\bar{x}$ is optimized, leading to a value of 3.6. A PCA is subsequently applied to these data, and the n_g first speaker components are retained to establish a model, namely, the second-level model, able to represent the speaker-specific P and $\bar{x}$ parameters of the individual articulatory models. These components will be referred to in the following as SPue, standing for *SPeaker components of the universal eigenvector model*. In this model, the matrix of eigenvectors E bears "universal eigenvectors," i.e., all the speakers are modelled with the same set of eigenvectors. Note that all the inter-speaker variability is thus transferred to the P and $\bar{x}$ matrices.

4. Mean articulation model

This model has been designed in order to explore the specific role of the mean articulation in multi-speaker modelling. In this approach, the articulatory predictors and eigenvectors are universal, whereas the mean articulations only are speaker-specific. The procedure to achieve this model is schematized in Fig. 8. The speaker-independent matrices of articulatory predictors P and eigenvectors E are those of the model of the mean speaker described in Sec. III, and are imposed to the n_s speakers. The n_s mean articulations $\bar{x}$ for the n_s speakers are then concatenated into a single matrix

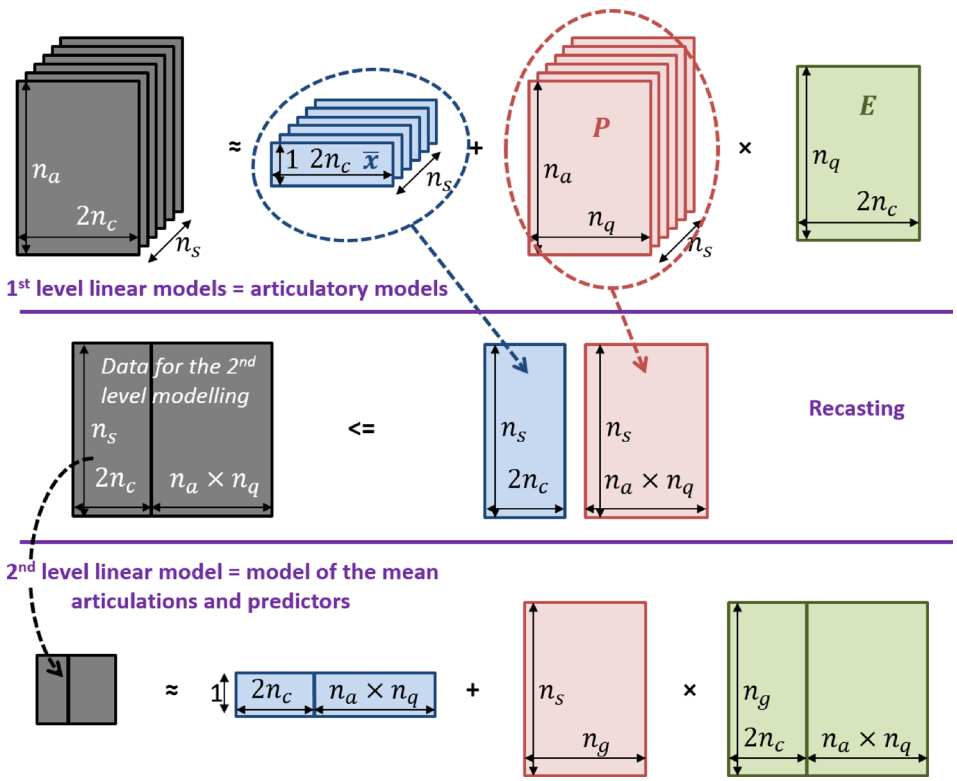

FIG. 7. (Color online) Schematic representation of the data analysis procedure for the ***universal eigenvector model***. Refer to Fig. 1 for the color conventions. See the text for the definitions of the various indices. The copy of the eigenvector matrix (in green) for all the speakers is not represented to simplify the schema and emphasize the speaker-independent property. As in Fig. 1, the replication of the (blue) line vectors in n_a (for the top blue line vectors) and n_s (for the bottom blue line vector) rows was omitted to enhance the comprehension.

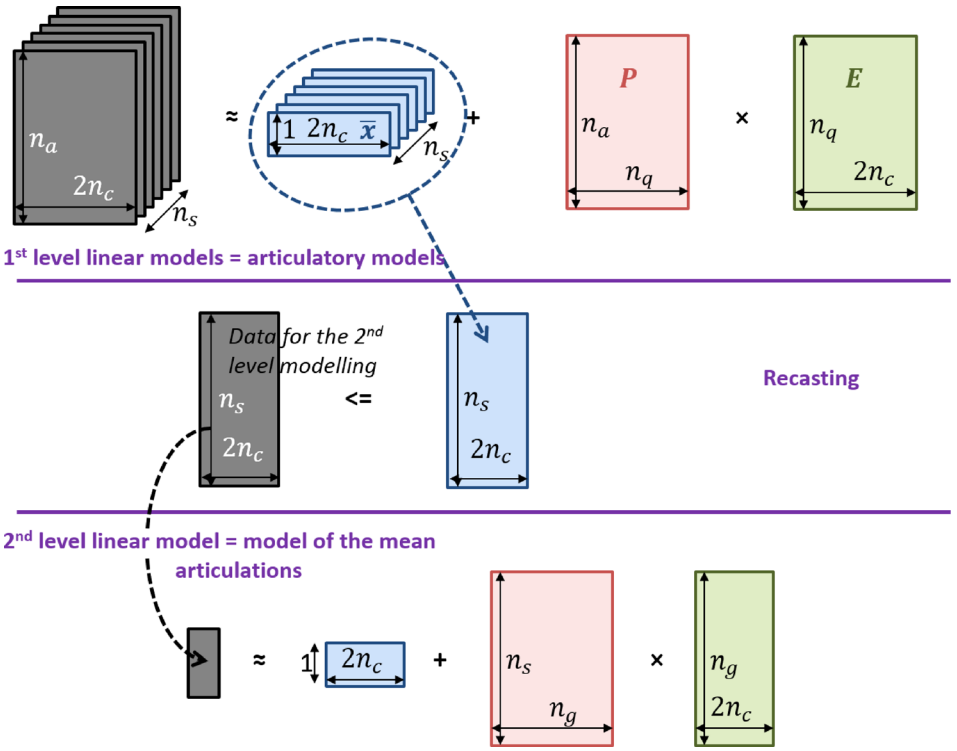

FIG. 8. (Color online) Schematic representation of the data analysis procedure for the ***mean articulation model***. Refer to Fig. 1 for the color conventions. See the text for the definitions of the various indices. The copy of the predictor (in red) and eigenvector (in green) matrices for all the speakers is not represented to simplify the schema and emphasize the speaker-independent property. As in Fig. 1, the replication of the (blue) line vectors in n_a (for the top blue line vectors) and n_s (for the bottom blue vector) rows was omitted to enhance the comprehension.

and subsequently submitted to PCA. The n_g first components are retained to establish a model, namely, the second-level model, able to represent the speaker-specific $\bar{\boldsymbol{x}}$ parameters of the individual articulatory models. These components, referred to in the following as SPm, standing for *SPeaker components of the mean articulation model*, carry only morphology information, unlike the previous multi-speaker models. Despite their close terminology, note the difference between this model, a multi-speaker model named *mean articulation model*, and the model presented in Sec. III, an articulatory model of a single speaker (the mean speaker) named *articulatory model of the mean speaker*.

B. Evaluation

1. Principles

The four multi-speaker models have been evaluated in terms of the reconstruction error through a Leave-One-Out Cross-Validation procedure (LOOCV, cf. Arlot and Celisse, 2010) applied to the whole set of speakers to assess their generalization capability. Each speaker is characterized by a set of n_g parameters that controls his/her P, E, and/or $\bar{\boldsymbol{x}}$ individual articulatory model parameters, depending on the type of MoM. Testing the generalization capability of these models consists thus in (1) building a multi-speaker model from a set of $n_s - 1 = 10$ speakers, excluding one of them in turn in a leave-one-out manner, (2) determining the n_g parameters of the excluded speaker by inversing the second level model knowing entirely the first level articulatory model of the speaker (cf. Table IV for details), (3) reconstructing the discarded speaker's articulations by means of the multi-speaker model controlled by these n_g parameters, and (4) computing the error between these reconstructions and the original data. In order to be able to compare the performances with the literature, the models for all the speakers have also been evaluated in a direct way without cross-validation.

TABLE IV. Method for the determination of the n_g parameters of the excluded speaker in a leave-one-out procedure for each multi-speaker model built on the rest of the speakers. The speaker data consist of $(\boldsymbol{x}_i)$, $i = 1 \ldots n_a$ articulations. The first-level universal parameters of each multi-speaker model, when existing, are denoted with the postfix *_universal* and the second-level parameters with the postfix "$^{(2)}$". For each model, $P^{(2)}$ of the discarded speaker is composed of n_g elements.

Steps	General model	Universal predictor model	Universal eigenvector model	Mean articulation model
Calculation of $\bar{\boldsymbol{x}}$	$\bar{\boldsymbol{x}}$ = average of $(\boldsymbol{x}_i)$, $i = 1 \ldots n_a$			
Calculation of P and E	P, E = predictors and eigenvectors of the speaker articulatory model obtained from its data $(\boldsymbol{x}_i)_{i=1\ldots na}$ (Table II)	E = Linear Regression of $(\boldsymbol{x}_i - \bar{\boldsymbol{x}})_{i=1\ldots na}$ on *P_universal* in a stepwise manner following Table II	P = result of the pseudo-inversion of equation $[(\boldsymbol{x}_i)_{i=1\ldots na} = \bar{\boldsymbol{x}} + P\,E\_universal]$ performed in a stepwise manner following Table II	
Calculation of the intermediate line vector $\boldsymbol{v}$	$\boldsymbol{v}$ = Recast of $\bar{\boldsymbol{x}}$, P and E in vector of size $2n_c + n_a n_q + n_q 2n_c$ (Fig. 5)	$\boldsymbol{v}$ = Recast of $\bar{\boldsymbol{x}}$ and E in vector of size $2n_c + n_q 2n_c$ (Fig. 6)	$\boldsymbol{v}$ = Recast of $\bar{\boldsymbol{x}}$ and P in vector of size $2n_c + n_a n_q$ (Fig. 7)	$\boldsymbol{v}$ = Recast of $\bar{\boldsymbol{x}}$ in vector of size $2n_c$ (Fig. 8)
Calculation of $P^{(2)}$	For each model, $P^{(2)}$ = result of the pseudo-inversion of equation $[\boldsymbol{v} = \bar{\boldsymbol{x}}^{(2)} + P^{(2)}\,E^{(2)}]$			

The overall errors on the $n_s \times n_a$ reconstructed articulations are expressed in terms of the rms error and percentage of variance explanation. The rms reconstruction error is calculated as

$$rms = \sqrt{\frac{1}{n_s 2n_c n_a} \sum_{l=1}^{n_s} \sum_{j=1}^{2n_c} \sum_{i=1}^{n_a} \left(x_{ijl}^2 - \hat{x}_{ijl}^2 \right)}, \quad (2)$$

where x_{ijl} and $\hat{x}_{ijl}$ represent, respectively, the measured and estimated coordinates of the contour point j of the articulation i of the speaker l of the data and its estimation. The percentage of variance explanation, usually provided as a byproduct of the modelling process for PCA or PARAFAC methods, has been calculated here as the complement to the error variance (i.e., the variance of the residuals), equivalent to the percentage of variance explanation for PCA, and referred to as such in the following for consistency and simplicity reasons.

For each model, the errors have been determined for a number n_g of speaker components SPg, SPup, SPue, and SPm varying from 0 to 10. In all cases, the first two components contribute from 40% to 45% of the variance explanation, and to about 0.25 cm of the rms error decrease relative to a model without any component. On the contrary, the contribution of the next components appears more limited, for instance about 4% of variance explanation and 0.04 cm of the rms error reduction for the third component. Therefore, a rather safe value of $n_g = 2$ was retained for the rest of the study. The overall errors for $n_g = 2$ are displayed in Table V for the LOOCV evaluation and in Table VI for the direct evaluation.

2. Results

For the present MoMs, the performances calculated directly from the individual articulatory models represent the best achievable performance, namely, 98% of overall variance explanation and 0.1 cm of overall rms reconstruction error (Table VI). On the other hand, the worst performance is achieved by simply using the articulatory model of the averaged $n_s - 1$ speakers for the discarded speaker; the performance in that case was found to be 0.66 cm of rms error and no variance explanation. These performances constitute, respectively, the *upper* and *lower bounds* of the multi-speaker models' performance.

The multi-speaker models controlled by two speaker predictors explain about 68% of the overall data variance and lead to reconstruction errors of about 0.37 cm (Table VI), which corresponds to an overall variance explanation about 30% lower than the upper bound and to an overall rms error about 0.27 cm higher.

These performances naturally depend on the empirical weightings used for the various parameters for the second-level PCA. These weightings ensure a balance between the relative importance of each of the parameters P, E and $\bar{x}$ in the resulting speaker components; they have been optimized to minimize the direct overall rms reconstruction error. A detailed analysis of the performance of the second-level PCA by means of direct evaluation for each of the four models showed a very good variance explanation of around 89% for the parameter $\bar{x}$, but of only 15 to 25% for the P and E parameters. The four modelling approaches deliver rather similar models, namely, models explaining the mean articulation and marginally the other parameters, which explains why they present similar, if not identical, results (Tables V and VI). The fact that the mean articulation has more importance than the E or P parameters in the MoMs has emerged through the automatic optimization of the empirical weightings used to achieve minimal reconstruction errors. This result emphasizes the primary importance of the mean articulation in the modeling of each speaker's articulations. In other words, it suggests that the largest source of difference between the speaker articulations lies in the difference of their average articulators' shapes rather than in their articulatory components. The potential correlation of the predictors and eigenvectors themselves over the set of speakers, analyzed separately from the mean articulation, will be explored in more detail in Sec. V. Additionally, the performances of the four models tend to diverge with the increase of the number of components n_g, demonstrating that the four models take into account the finer role of the articulatory strategies in different ways.

These global errors mask however differences between articulators, with a variance explanation varying from 1% (jaw) to 85% (pharynx) and an rms error from 0.20 cm (upper lip) to 0.48 cm (posterior supraglottis). Note that, as

TABLE V. Percentage of variance explanation (column "%") and cumulated rms reconstruction error (column "cm") estimated by LOOCV for each articulator obtained with the four multi-speaker models with $n_g = 2$.

	General model		Universal predictor model		Universal eigenvector model		Mean articulation model	
	%	cm	%	cm	%	cm	%	cm
Jaw	1	0.36	2	0.35	1	0.36	3	0.35
Tongue	56	0.46	59	0.44	57	0.45	61	0.43
Upper lip	23	0.21	29	0.20	25	0.21	32	0.20
Lower lip	6	0.33	13	0.32	8	0.33	16	0.31
Velum	60	0.33	61	0.33	60	0.33	61	0.33
Pharynx	84	0.30	84	0.30	84	0.30	85	0.29
Epiglottis	70	0.44	71	0.43	68	0.45	73	0.42
Posterior supraglottis	72	0.47	73	0.46	71	0.48	74	0.46
Overall vocal tract	**66**	**0.38**	**68**	**0.37**	**66**	**0.38**	**69**	**0.36**

TABLE VI. Percentage of variance explanation (column "%") and cumulated rms reconstruction error (column "cm") estimated directly without LOOCV for each articulator for the speaker-specific individual models and the four multi-speaker models with $n_g = 2$.

	Individual models		General model		Universal predictor model		Universal eigenvector model		Mean articulation model	
	%	cm	%	cm	%	cm	%	cm	%	cm
Jaw	90	0.11	41	0.27	39	0.28	42	0.27	38	0.28
Tongue	97	0.11	75	0.34	75	0.34	76	0.34	74	0.35
Upper lip	90	0.07	50	0.17	51	0.17	53	0.16	49	0.17
Lower lip	89	0.12	44	0.26	44	0.26	48	0.25	41	0.26
Velum	99	0.06	79	0.24	79	0.24	79	0.24	79	0.24
Pharynx	99	0.08	91	0.23	91	0.23	91	0.22	90	0.23
Epiglottis	96	0.16	82	0.34	81	0.34	81	0.35	81	0.34
Posterior supraglottis	97	0.14	85	0.35	84	0.36	84	0.35	84	0.36
Overall vocal tract	**98**	**0.10**	**81**	**0.28**	**81**	**0.28**	**81**	**0.28**	**80**	**0.29**

the various articulators differ considerably in terms of articulatory data variance, the highest rms errors (e.g., around 0.45 cm for the tongue, epiglottis and posterior supraglottis) are not necessarily associated with the lowest variance explanation rates (e.g., less than 10% for the jaw and lower lip). In general, the articulators forming the front region of the vocal tract, i.e., the jaw and lips, show the lowest rates of variance explanation, below 30%, the articulators of the middle of the vocal tract, i.e., the tongue and velum, show intermediate results, around 55%–60% of variance explanation, while the rest of the articulators in the back region of the vocal tract ranges from about 70% to 85%. The lower percentage of variance explanation for the front articulators in comparison with the back articulators can be partly ascribed to the alignment procedure that imposes the lowest edge of the upper incisors to be the same for all the speakers, leading to a lower variability in this region. This point will be revisited in the Discussion. In terms of the rms error, the articulators can be grouped into three categories: the tongue, epiglottis, and posterior supraglottis, with an error higher than 0.4 cm, the jaw, lower lip, velum and pharynx, with an error around 0.3–0.35 cm, and finally the upper lip with an error of 0.2 cm.

These results suggest that the articulations of one speaker cannot be completely reconstructed by the articulatory components derived from ten other speakers. In other words, it suggests that ten speakers are not enough to capture the complete variability of the morphology and strategy of a given speaker. Nevertheless, these errors remain still significantly lower than the overall rms error of 0.66 cm of the upper bound, which means that a part of the morphology and strategy variability of one speaker is actually borne by the other ten speakers and is exploited through the chosen modelling approach.

In addition, it can be observed that the mean articulation model presents results similar to those of the other three models that aim to estimate the specific articulatory predictors and/or eigenvectors for the discarded speaker in addition to the mean articulation. It could be expected that estimating the strategy parameters, i.e., the articulatory predictors and eigenvectors, for the discarded speaker would have led to better reconstructions. It is actually not the case, which suggests that, in the approach chosen in this study, the strategy variability of one speaker cannot easily be explained by the strategy components extracted from ten other speakers. Oppositely, the morphology variability of one speaker can, at least partly, be explained by the morphology components extracted from ten other speakers. Further studies with a larger cohort of speakers are needed to clarify these points.

3. Comparison with previous studies

As mentioned earlier, direct evaluation without leave-one-out has been performed for comparisons with the literature: the multi-speaker models explain about 80% of the overall data variance and lead to a reconstruction error of about 0.28 cm (cf. Table VI for details). As detailed in this section, this study presents at best results similar to those in previous studies and at worst lower ones, which could be ascribed to a few main facts: (1) the larger size of the dataset in terms of speakers, articulations and speech articulators, (2) the chosen approach based on the guided PCA, and (3) the inclusion of individual means in the models.

Previous multi-speaker modelling studies mostly focused on the tongue (Harshman *et al.*, 1977; Lindau, 1986; Jackson, 1988; Nix *et al.*, 1996; Hoole, 1998, 1999; Geng and Mooshammer, 2000; Hoole *et al.*, 2000; Zheng *et al.*, 2003; Hu, 2006), occasionally including the jaw and/or the lips (Johnson *et al.*, 1993), or treating them separately (Linker, 1982; Ananthakrishnan *et al.*, 2010; Valdés Vargas *et al.*, 2012; Valdés Vargas, 2013); Valdés Vargas (2013) also analyzed the velum. Although the objectives and methods differ substantially, the present work is a follow-up of Valdés Vargas (2013), and of Valdés Vargas *et al.* (2012) and Ananthakrishnan *et al.* (2010). From these three studies, only Valdés Vargas (2013) will therefore be considered for comparison and used as reference. Note that Valdés Vargas (2013) provides a LOOCV evaluation of the generalizability to articulations, whereas the present work focusses on the generalizability to speakers, assuming that the 62 phonemes of the corpus constitute a representative sampling of the French articulatory repertoire.

For the tongue, beyond these latter three studies, multi-speaker models based on contours in single- or cross-language studies have been developed by Harshman *et al.* (1977), Jackson (1988), Nix *et al.* (1996), Hoole *et al.* (2000), and Zheng *et al.* (2003) by means of PARAFAC. Two to three

factors led to variance explanations from 87% to 93% (but noticeably only 76% for Zheng *et al.*, 2003) and rms reconstruction errors, when provided, from 0.16 to 0.22 cm. Results for much sparser sampling of the tongue (for example by EMA coordinates), still based on PARAFAC decomposition, led to similar performances: from 80% to 96% of variance explanation and rms reconstruction errors from 0.11 to 0.2 cm with two or three factors (Hoole, 1999; Geng and Mooshammer, 2000; Hu, 2006). Valdés Vargas (2013) compared different linear three-way decomposition methods, namely, PARAFAC, joint PCA and Tucker. For each method, he provided the variance explanation and rms reconstruction error for a varying number of components. In the present study, the tongue articulation of any speaker can be controlled by a combination of five guided PCA articulatory components, equivalent in terms of performance to about four raw PCA components, and two speaker components, leading altogether to an equivalent of six raw PCA components. For six components, Valdés Vargas (2013) obtains from 70% to 79% of variance explanation and from 0.23 to 0.27 cm rms error. Except for the variance explanation of Zheng *et al.* (2003) and Valdés Vargas (2013), comparable to the results presented in this article, the performance of the present models appears lower than that in other studies, with a variance explanation lower by 5%–20% and an rms error larger by 0.07–0.23 cm. The closest results to the present study are observed for Zheng *et al.* (2003) and for Valdés Vargas (2013), the most comparable study in terms of data and design.

For the lips, Linker (1982) conducted an extensive cross-language analysis of articulatory measures, based on PARAFAC. For a number of factors varying between 1 and 3 depending on the language, he found a variance explanation from 87% to 96%. The models from Valdés Vargas (2013) reached from 69% to 77% of variance explanation for lip contours with PARAFAC, joint PCA, or Tucker decomposition with four components; these four components can be deemed equivalent to the three guided PCA components of the present study (equivalent to two raw PCA components) combined with two speaker components. The corresponding rms error was from 0.08 to 0.09 cm for the upper lip and from 0.14 to 0.15 cm for the lower lip. The performance of the present study appears significantly lower, with a variance explanation ranging from 41% to 53% and an rms error from 0.16 to 0.26 cm.

For the velum, Valdés Vargas (2013) obtained from 79% to 83% of variance explanation and from 0.12 to 0.13 cm of rms error with four components. The present study reaches a comparable rate of variance explanation of 79%, but with an rms error of 0.24 cm.

This study presents at best results similar to those in previous studies and at worst lower ones. This can be ascribed to a number of reasons, given the numerous differences in the data and methods. The data comprise 62 articulations, vowels and consonants, for 11 speakers, i.e., 682 articulations altogether, which is more than in most previous studies, despite the single language approach. Moreover, the data cover all the contours of the vocal tract articulators with a very detailed geometrical representation, in particular in comparison with the geometrically sparse EMA articulatory data. The number of components is also hardly comparable due to the two-level decomposition *vs* one level decomposition of previous studies. Additionally, the chosen approach is based on the guided PCA, leading to articulatory components easily interpretable, which might not be the case of those based on direct PCA decomposition. Finally, the lower performance for the front articulators can be partly ascribed, as noted for the LOOCV analysis, to the alignment process, as further discussed in Sec. VI.

The major difference lies, however, in the methodological approach that aims, in the present study, at characterizing entirely a speaker, i.e., including also his/her mean articulation with a limited number of parameters. To our knowledge, no other study has considered and explicitly modelled the mean articulations of the speakers in multi-speaker modelling. In the present approach, all the speaker morphology and strategy specificity, including the mean articulation, is captured by only two parameters. This represents a significantly smaller speaker dimensionality in comparison with previous studies, where speaker-specific means are not modelled and need, therefore, to be explicitly known to reconstruct the articulations. For instance, considering 14 predictor values corresponding to all articulators of one specific phoneme, the present approach would require in addition (using the universal predictor model) two speaker-specific values to reconstruct the speaker-specific articulation, whereas a PARAFAC approach would require the complete speaker mean articulation in addition to 14 speaker-specific weight parameters.

To summarize, with two speaker components, all general, universal predictor, universal eigenvector, and mean articulation models, present similar performances, as estimated with the direct evaluation, i.e., about 80% of variance explanation and about 0.28 cm of rms error. In a LOOCV procedure, they reach a variance explanation around 67% and an rms error around 0.37 cm, demonstrating interesting capabilities of generalization over the set of speakers.

V. MORPHOLOGY AND ARTICULATORY STRATEGY CHARACTERIZATION

Section IV described the MoMs that mutualize the information common to the individual articulatory models. In these models, the components represent indiscriminately inter-speaker variability for both morphology and strategy. The present section describes experiments aiming at disentangling morphology- and strategy-related variability. The objective is not to derive a minimal number of components for the second level model as in Sec. IV, but to analyze separately the morphology and strategy components, also by means of second level modelling.

Recall that in a linear articulatory model, the components represent the variations of the articulator shapes around their mean values. In this approach, one may assume that the mean articulation $\bar{x}$ represents the morphology of the speaker, while the articulatory predictors P and eigenvectors E represent the phoneme-specific strategy. Morphology and articulatory strategies are, however, necessarily related, as speakers must adopt strategies complying with their own

morphology to reach the required phonemic target articulations. Similarly, the mean articulation may, to a certain extent, reflect the strategy of the speaker. But, as the corpus is well balanced by design and large enough, it may be assumed that the mean articulation is free from the possible idiosyncratic articulatory strategies implemented by the speaker to achieve specific articulations, and thus truly reflects his/her morphology.

The present exploration is based on the universal predictor modelling architecture presented in Sec. IV A 2. In this approach, the first stage derives a unique set of articulatory predictors P common to all the speakers and n_s sets of speaker-specific E and $\bar{x}$ parameters. As pointed out earlier, all the variability between speakers is transferred by this method to the E and $\bar{x}$ parameters that therefore represent, respectively, the articulatory strategies and the morphologies of the speakers. In the following, these two sets of parameters are further analyzed separately using second-level modelling.

A. Morphology characterization

This section presents an analysis of the morphology variability of all the articulators' contours, including the hard palate, and shows that the variability of the mean contours of the speakers can faithfully be represented with two components. For that purpose, a PCA has been applied to the n_s sets of speakers' mean articulations $\bar{x}$ supplemented with the n_s hard palate contours. Note that this step is in practice similar to the second level modelling of the mean articulation multi-speaker model (cf. Sec. IV A 4). For the same reasons as for the MoMs described in Sec. IV, the first two components only have been retained: MP1 and MP2. These two components explain, respectively, 64% and 24% of the variance, 88% altogether, and lead to a cumulated rms reconstruction error of the mean articulation of 0.33 and 0.20 cm. The associated nomograms are displayed in Fig. 9.

The first component MP1 clearly controls the lengthening/shortening of both horizontal and vertical vocal tract dimensions, ranging from a short vocal tract with a high position of the larynx to a long vocal tract with a low position of the larynx. Figure 9 shows that MP1 controls also a variation of the depth of the hard palate, as also reported by Lammert *et al.* (2013). Note that the relation between these two speaker characteristics, vocal tract length and hard palate depth, does not seem very realistic and might be an artefactual correlation due to the limited number of speakers in the dataset. The second component MP2 represents a horizontal scaling concomitant with a rotation around the lower edge of the upper incisors much related to the head orientation (cf. Serrurier and Badin, 2008). These two components, and especially MP1, strongly reflect the well documented male-female differences (Goldstein, 1980; Fitch and Giedd, 1999; Vorperian *et al.*, 2009; Barbier *et al.*, 2015; Story *et al.*, 2018); Fig. 10 shows indeed that the male and female speakers can easily be linearly separated in the MP1-MP2 space. The next two components, not detailed here, represent less than 5% of variance explanation each and relate to residual deformations of the tongue, hard palate, and larynx region, without clear interpretation. The overall rms reconstruction error with the first two components is 0.2 cm.

Finally, as the speakers' articulatory contours have been aligned on the upper teeth and the ANS-PNS lines (cf. Sec. II D), the inter-speaker variability is lower in this region, which is reflected in the nomograms of Fig. 9. Another choice of alignment procedure could lead to different components as explained later in Sec. VI.

B. Articulatory strategy characterization

In this section, we show that the variation of the strategy is related to a certain extent to the variation of the morphology and how the strategy seems to comply with the morphology. To characterize the strategy, a PCA has been applied to the n_s sets of the eigenvectors E taken from the universal predictor model, as explained at the beginning of this section. The resulting components might be considered as articulatory strategy components associated with the variation of the matrix of eigenvectors E of individual articulatory models. As for the MoMs described in Sec. IV, the first two strategy components S1 and S2 only have been retained; they explain, respectively, 26% and 21% of the variance of the eigenvectors, altogether 47%. As could be expected, this is better than the percentage of variance of a mere 22%–25% for the eigenvectors obtained by two components in the MoMs, where the analysis is carried out simultaneously on the eigenvectors and the mean articulation, and occasionally the articulatory predictors.

Although the focus of this section is the sole strategy, the correlation of the strategy predictors with the morphology predictors obtained in Sec. V A has been analyzed to assess possible links between strategy and morphology. A

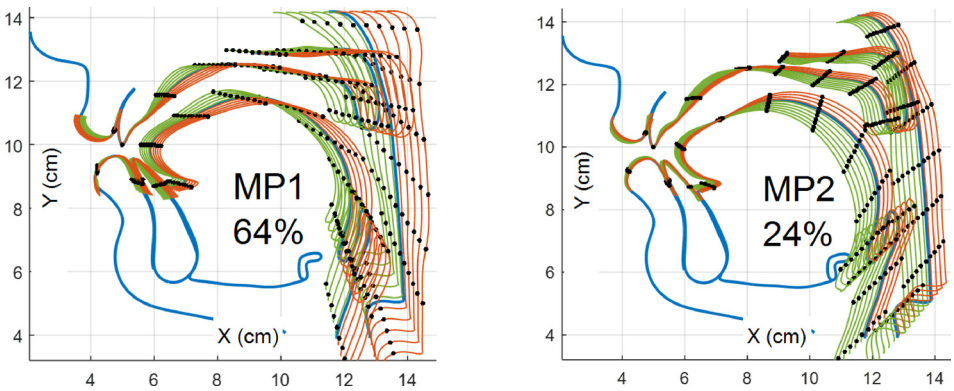

FIG. 9. (Color online) Nomograms of the contours of the mean articulation for the two morphology components for predictor values varying at regular steps between the minimal and maximal values found in the data. Contours with negative (respectively, positive) predictor values are plotted in green (respectively, orange). One every 30 point is plotted as black dot to emphasize deformation directions. The full contour of the average articulation is displayed (in blue) for better comprehension.

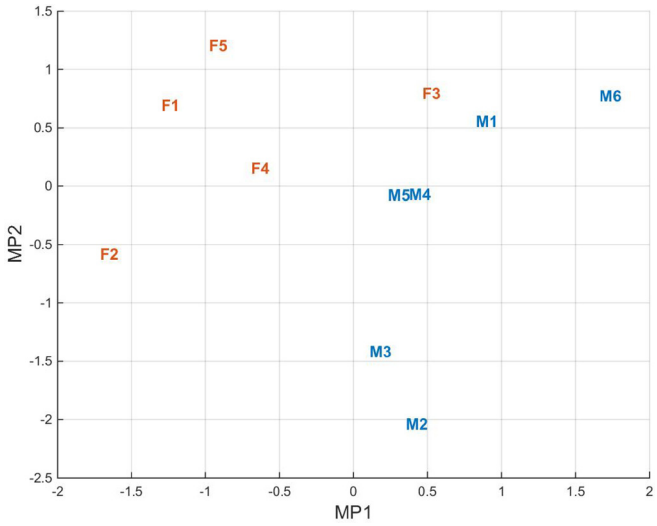

FIG. 10. (Color online) MP1–MP2 plane for the n_s speakers listed in Table I.

slight correlation (0.62) was found between the first morphology predictor MP1 and the first strategy predictor S1, suggesting that the variations of strategy captured by S1 are related to some extent to the morphology variations captured by MP1. The correlation between morphology and strategy justifies *a posteriori* the approach chosen for the multi-speaker models, where these correlations are exploited through the joint PCA of the mean articulations and the eigenvectors. However, no other significant correlation (above 0.5) was observed between the morphology and strategy predictors.

The strategy components highlight the principal inter-speaker strategy variations. Figure 11 illustrates the articulatory nomograms for the components TB, TD, and TT obtained for the two most extreme values of S1. On this figure, the contribution of S1 to the mean articulation has also been taken into consideration: a model of mean articulation obtained by linear regression of the n_s mean articulations (supplemented with the hard palate contours) on S1 has been built and is used to reconstruct the mean articulation corresponding to the value of S1 used. For space reasons, the variations of the 14 articulatory components of all the articulators as a function of the two strategy components cannot be presented; the focus is made on the three tongue components TB, TD, and TT.

The first strategy component S1 explains 26% of the variance of the eigenvector matrix E, i.e., of the inter-speaker strategy variance. The influence of the strategy on the TB component is mainly associated with a change in the orientation of the lines of deformation in the back region of the tongue, from rather oblique to horizontal, together with a change of amplitude range, especially in the front region of the tongue. For the TD component, it is mainly associated with a change of the direction of the dorsum movement, from a rather oblique to a more vertical direction, leading to a tongue-palate contact place slightly more backward or

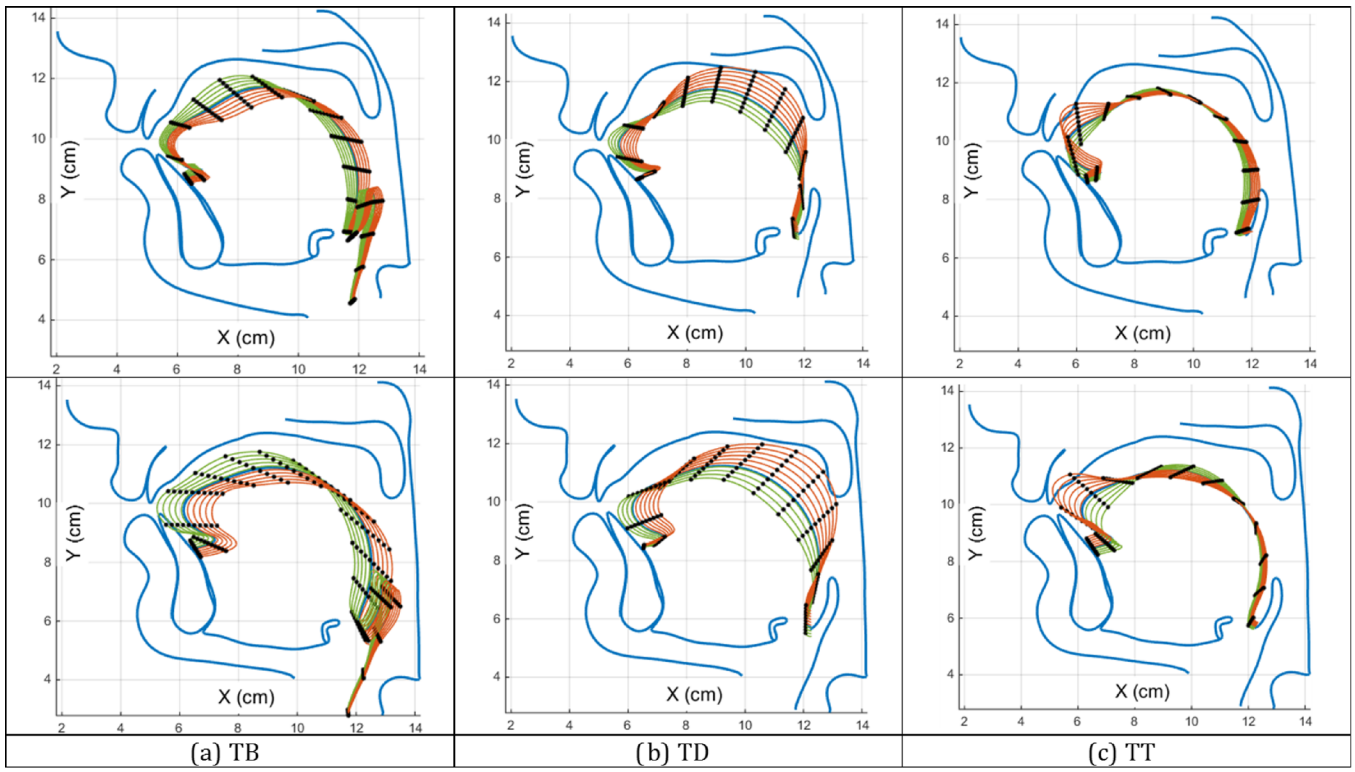

FIG. 11. (Color online) Nomograms of the contours for the tongue articulatory components TB, TD, and TT using eigenvectors reconstructed by means of the sole strategy component S1 using minimum (top) and maximum (bottom) predictor values observed in the data. The articulatory predictor values vary at regular steps between the minimal and maximal values found in the data. Contours with negative (respectively, positive) predictor values are plotted in green (respectively, orange); one every 20 points is plotted as a black dot to emphasize deformation directions. The full contour of the average articulation is displayed (in blue) for better comprehension (note that it depends on S1 and is thus different in the top and bottom rows). Note that the component TB controls the epiglottis in addition to the tongue.

frontward. This comes together with a slight change in the rotation of the tongue, the tip going more upward when the dorsum lowers in the case where the tongue-palate contact is more frontward, and vice versa. The more oblique deformation of the tongue dorsum is also associated with a large amplitude range. The influence of the strategy on the TT component is clearly mainly associated with the tongue tip deformation varying from a rather horizontal movement with a large amplitude range to a rather vertical movement with a lower amplitude range. The other components not presented here exhibit analogous ranges or types of variations as TB, TD and TT for extreme values of S1.

Interestingly, the extent to which the strategy seems to comply with the morphology can be observed in Fig. 11: for an S1 predictor value leading to a long vocal tract (bottom row), the various articulatory components tend also to have a more important frontward-backward movement than the articulatory components obtained for an S1 value leading to a short vocal tract (top row). This means that the speakers with a longer vocal tract present a larger tongue span from low or back positions like in /ɔ/ to high or front positions like in /t^{i}/. This is illustrated on the top row of Fig. 12. Another analogous trend is observed for the velar constriction/contact partly controlled by the TD parameter: its location is more backward for speakers with a longer vocal tract than for those with a shorter vocal tract. This actually maintains the constriction in the same relative position with respect to the whole tract, compensating for the larger vocal tract length variation in the vertical direction than in the horizontal direction reflecting male/female differences. The TB component tends also to adapt so that the tongue blade mirrors the shape of the palate, either domed (top row of Fig. 11) or flat (bottom row). This is for instance the case for the high front vowels like /ɛ/ as illustrated on the top of Fig. 12. For the TT component, the vertical or oblique movement of the tongue tip tends to maintain the same contact position on the palate regardless of the palate shape.

A detailed comparison of the articulations obtained using the strategy predictor S1 to control both the eigenvectors and the mean articulations with the articulations obtained using S1 to control the mean articulation only has revealed minor differences: the rms distance on the corpus is of maximally 0.1 cm for S1 varying between the minimal and maximal values found in the data. This limited influence of the strategy is illustrated on the bottom row of Fig. 12 that displays such pairs of articulations obtained for the absolute maximum value of S1 found in the data for three phonemes. This is in general agreement with the results obtained in Sec. IV, where it was found that inter-speaker variability is primarily carried by the mean articulations. This limited influence of the strategy observed in Fig. 12 seems in contradiction with the nomogram differences that can be observed

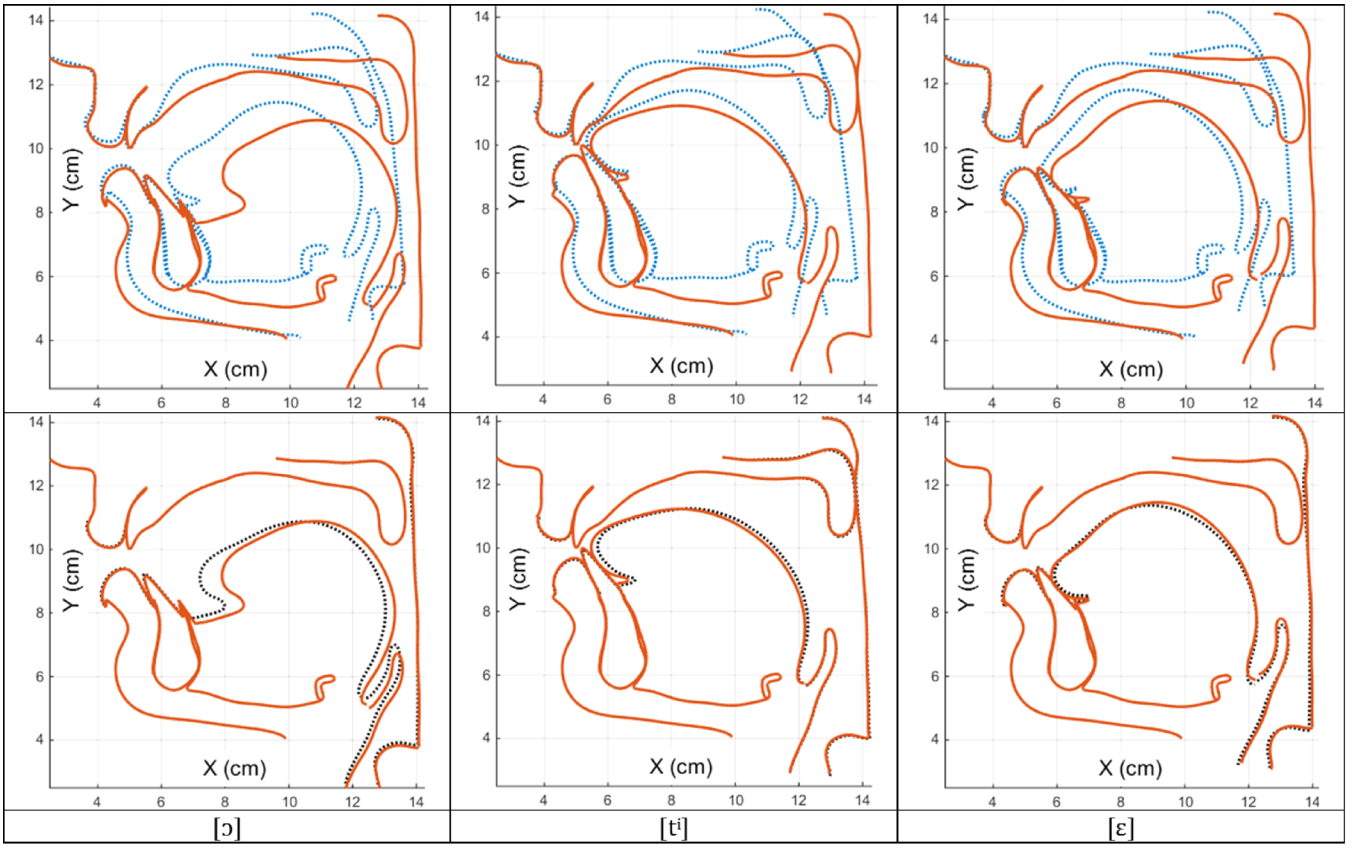

FIG. 12. (Color online) Top: superposition of the articulations obtained using the universal predictors and both the mean articulation and the eigenvectors controlled by the strategy component S1 using minimum (dashed blue) and maximum (solid orange) predictor values observed in the data for the three articulations [ɔ] (left), [t] in context [i] (middle), and [ɛ] (right). Bottom: superposition of the same solid orange contours as for the top row with the same articulations obtained by controlling only the mean articulation with the maximum value of S1 and keeping the mean eigenvectors (dashed black), so as to highlight the influence of the strategy component of the model.

between the top and bottom rows of Fig. 11. This discrepancy suggests that the strategy components would mainly model the differences between the articulatory components (Fig. 11), but that the differences would compensate for each other in the final contours of articulations as seen in Fig. 12.

The second strategy component S2 explains 21% of the inter-speaker strategy variance. As for S1, the nomograms of the various articulatory components as a function of S2 exhibit slight variations, but their representation is discarded for space reasons. The variations of the directions of deformation of the articulatory components appear, however, complementary to those induced by S1, though less easily interpretable.

In conclusion, with 47% of variance explained by the two strategy components S1 and S2, a large part of the individual strategies remains still unexplained, pointing to the importance of the individual behavior. Interestingly, the link between morphology and strategy has been further emphasized by (1) measuring the correlation between the first morphology predictor MP1 and the first strategy predictors S1—suggesting that the variations of strategy captured by S1 are related to some extent to the morphology variations captured by MP1 and related to the male-female differences—and (2) observing the adaptation of the range and directions of deformations of the articulatory components to the vocal tract morphology. The link between articulatory strategy and morphology has already been pointed out in the literature (cf. Sec. I), including by Serrurier *et al.* (2017) with the same data. The fact that no significant correlation (above 0.5) between MP2 and the strategy predictors and between S2 and the morphology predictors suggests that either speakers do not always adapt their strategies to their morphological characteristics or that these relations are not captured by the present linear modelling approach. This observation does not exclude in general the fact that the strategy may result to some extent from morphological constraints beyond the relations already brought out in this article.

As far as we know, this study constitutes the only attempt in the literature to model the strategy variations of a set of speakers. In summary, the first strategy component, accounting for about a quarter of the variance, is related to a certain extent to the first morphology component, which is related to the vocal tract dimension and palate shape. A variation of range and directions of deformation of the articulatory components related to these characteristics is logically observed in this component. The second strategy component accounts for about a fifth of the variance but seems related to the morphology to a much lesser extent.

VI. DISCUSSION AND CONCLUSION

The present article has described an approach for multi-speaker articulatory modelling based on models of models. This approach involves a two-level modelling procedure, where the parameters of the speaker-specific articulatory model are themselves controlled by another linear model. The first level, the speaker-specific articulatory model, deals with the intra-speaker variability, whereas the second level deals with the inter-speaker variability. An articulatory model can further be considered as the association of a morphology constituent[4] (the mean articulation) with a strategy constituent (the matrix of eigenvectors) controlled by a set of articulatory predictors. In Sec. IV, a full advantage of potential correlations between morphology and strategy over the set of speakers was taken in order to obtain the MoMs controlled by a minimum number of parameters. For this purpose, morphology and strategy parameters of individual articulatory models were jointly analyzed to obtain joint morphology and strategy components (the speaker components). Four approaches were considered in the study, leading to the speaker components SPg, SPup, SPue, and SPm. In Sec. V this approach was used in a complementary way to characterize independently the inter-speaker morphology and the strategy variability and to represent their main modes of variation. For this purpose, the morphology and the strategy constituents of the first level models were analyzed separately in order to obtain distinct morphology components MP and strategy components S.

This study relies on a dataset of 62 articulations collected from 11 French speakers. The extent of this dataset constitutes both an asset and a limitation to the study. On the one hand, this set constitutes to our knowledge the largest collection of contours of the whole vocal tract used for multi-speaker modelling. Previous studies usually included fewer speakers, fewer articulations, and/or fewer points on the vocal tract, typically recorded through EMA. Developments in the last decade made it possible to record real-time MRI of the vocal tract for speech production studies, leading to datasets with many more observations (e.g., Teixeira *et al.*, 2012; Narayanan *et al.*, 2014): although promising for the future, this modality still provides images of lower quality than the static ones used in current study and requires automatic image processing for contour extraction, still less accurate than the current manual processing (Labrunie *et al.*, 2018). The dataset of the present study constitutes therefore a realistic compromise to build multi-speaker models based on accurate contours. In addition, it has been proved that carefully designed corpora lead to similar articulatory models than the larger corpora based on real-time measurements (Beautemps *et al.*, 2001): the corpus designed to represent the wide range of French articulations appears therefore appropriate for the study. On the other hand, the present set of speakers might appear too small for the study of inter-speaker variability: although balanced between females and males, it cannot be ensured to be truly representative of all possible French speakers. Moreover, 11 observations represent a limited set for statistical analyses. The selection of components was therefore restricted to the first two speaker components, clearly interpretable and generalizable over the set of speakers. The next components, although possibly coding meaningful information as detailed later, were discarded in order to avoid overfitting. These limitations call for larger datasets of speakers in the future to analyze inter-speaker variability in depth.

Section III described an articulatory model of the mean speaker averaged over the set of 11 speakers, aiming at reducing the influence of individual morphologies and articulatory strategies while retaining their common background.

An overall variance explanation of 96% and an rms reconstruction error of 0.05 cm were obtained with 14 components.

Based on the articulatory modelling architecture described in Sec. III, four multi-speaker models have been explored. Depending on the parameters considered in the second level models, the articulatory predictors and/or eigenvectors could in addition be universal, i.e., speaker-independent, and therefore fixed regardless of the second level model. Recall that the second level models intend to take advantage of the correlations between the morphology and strategy constituents of the articulatory model parameters of the speaker-specific first level models. This resulted in similar performances established by LOOCV for all multi-speaker models, namely, from 0.36 to 0.38 cm of overall rms error and from 66% to 69% of variance explanation. A deeper analysis revealed that the two n_g speaker components retained for each of the four models, i.e., the components SPg, SPup, SPue, and SPm, essentially explain the mean articulation, i.e., the morphology of the speakers, rather than the articulatory eigenvectors and predictors, i.e., the speakers' articulatory strategy. The relative weighting between the morphology and strategy constituents in second level models optimized to achieve better overall results tends thus to favor the modelling of the morphology compared with the strategy. This means either that the dominant source of inter-speaker variability is related to morphology, or that the inter-speaker variability related to morphology is more suited to PCA modelling—hence better taken into consideration—than the one related to strategy. In any case, the predominance of the morphology over the strategy in the second level models can explain why the four multi-speaker models present similar performance: they all take into account in the same way the morphology constituent of the articulatory model parameters, i.e., the mean articulation, but differ in the way they take into account the strategy constituent, i.e., the articulatory eigenvectors and articulatory predictors. Further informal analyses performed on the tongue showed that the differences among the reconstruction errors obtained for the four models tend indeed to be more pronounced as the number of speaker components increases. This shows that the next components are more closely related to the strategy than to the morphology.

As emphasized earlier, the global errors mask strong disparities between front and back articulators. This can be partly ascribed to the alignment procedure. Indeed, since all the midsagittal contours have been aligned on the lowest edge of the upper incisors, this point has zero variability, and the points in its vicinity present lower variability than those farther away. This implies that the front articulators present lower variability than the back articulators (a ratio from 1 to 14 has been observed between the points of the upper lip and those of the posterior supraglottis). The first speaker components, much related to morphology components, tend to explain primarily this larger source of variance of the back articulators. This can be observed for the morphology component MP1 in Fig. 9, where the articulatory contours in the back of the vocal tract exhibit a much larger range of variation than those in the front. Once this effect has been compensated for by the first components, one can expect that the next components do not depend on the alignment procedure and appear more balanced between the articulators, which was indeed verified. However, these were finally not retained in the models for two reasons: (1) due to the low number of speakers, the risk of overfitting was important and brought us to keep only the components clearly generalizable over the speakers; (2) the gain in terms of the rms error and variance explanation appeared too marginal to justify the addition of extra components. It has been verified that another alignment procedure leads to different principal components. Nonetheless, the alignment proposed in this study on the palate, and by extension on the cranium, is in line with numerous articulatory studies and takes advantage of the only rigid structure that supports the vocal tract.

As noted earlier, the performances of the models proposed in the present study are not higher than those reported in the literature, which can be ascribed to a multitude of factors: the larger size of the corpus, the larger number of speakers, the more exhaustive description of the vocal tract contours, the two-level decomposition approach making the number of components hardly comparable, or also the guided PCA approach that intends to make articulatory components more easily interpretable. In addition, most studies provide an evaluation in terms of percentage of variance explanation, which could be influenced by the alignment procedure as explained above. For all these reasons, the results presented in this study are difficult to compare with those of previous studies. But most importantly, the primary aim of the current study, i.e., to propose a multi-speaker model able to fully characterize any speaker with a minimum of control parameters, is also eminently different from previous studies. To achieve these objectives, the study was inspired from previous studies highlighting the relationship between morphology and strategy (cf., e.g., Honda *et al.*, 1996; Fuchs *et al.*, 2008; Brunner *et al.*, 2009; Yunusova *et al.*, 2012; Rudy and Yunusova, 2013; Weirich and Fuchs, 2013; Weirich *et al.*, 2013) and involved a two-level modelling approach that takes full advantage of this relationship. Hence, unlike in previous studies, inter-speaker variability related to morphology was also taken into account.

The two-level models are composed of a chain of two linear models: one dealing with the intra-speaker variability, i.e., the articulatory model of individual speakers, and the other with the inter-speaker variability. Linear models have already proven to be efficient for modelling the articulatory variability of individual speakers. This study proposed to extend this concept to the modelling of the morphology and strategy variability of a set of speakers. The performance of the multi-speaker models appears to be intermediate between the lower and upper bounds, suggesting that these models are able to represent a part but not all of the inter-speaker morphology and strategy variability. The detailed analysis mentioned earlier revealed that the MoMs mainly explained the morphology variability rather than the strategy variability. When applied to data combining both morphology and strategy elements, second level linear models appear thus mainly efficient to model the inter-speaker morphology variability. Three possible reasons can be proposed to explain

these results: (1) there is little correlation among the strategies of the speakers, (2) the potential correlation between the strategies of the speakers cannot be represented by means of linear PCA modelling, and (3) the variance associated with the potential correlation between the strategies of the speakers remains lower than the variance associated with the correlation between the morphologies of the speakers, and requires therefore more than two components to be taken into account by the models. Further investigations involving alternative modelling approaches, including nonlinear ones, and a larger set of speakers, are required to clarify this issue.

The principles exposed for the MoMs have also been exploited to independently characterize the morphology and the strategy variability: 88% of the variance of the mean articulations could be explained by only two components, MP1 and MP2. These two components reflect the male-female differences. This suggests that the principal source of inter-speaker variability regarding the morphology is related to the size of the vocal tract, itself related to the sex.

The analysis of inter-speaker strategy variability by means of two-level modelling has shown that 47% of the articulatory eigenvector variance could be explained by two strategy components. This is significantly more than the 22%–25% explained for the eigenvector variance by two speaker components when the analysis is combined with the morphology parameters. This demonstrates that the inter-speaker strategy variability can, to a certain extent, be taken into account by linear modelling. The lower variance explanation rate observed when the analysis is combined with the morphology parameters implies either that a part of the inter-speaker strategy variability is related to idiosyncratic features and not to speakers' morphology, or that the components corresponding to this strategy variability were not retained in the two-component models. The correlation found between the first component morphology predictors and the first component strategy predictors points out the relationship between morphology and strategy and justifies the combined approach chosen in Sec. IV. Regarding the strategy components alone, their influence on the variations of the directions and ranges of deformation related to the articulatory components have been demonstrated to serve to adjust the articulatory components to comply with morphology constraints. This is particularly the case of TB and TT for the component S1, which adjust themselves according to the palate shape.

Despite all the limitations mentioned above, the multi-speaker models proposed in this study can fairly well characterize the full vocal tract contours of a speaker with two parameters. To our knowledge, such characterization has not been attempted in the literature. Beyond the multi-speaker model itself, the main contribution of the present study is the finding that inter-speaker variability is more related to the morphology, and in particular to the sex and size of the speakers, than to the idiosyncratic articulatory strategies. In addition, the present approach reveals the extent to which the articulatory components adapt themselves to comply with the morphology constraints.

The present approach opens the way to a range of applications where a generic articulatory model must be adapted to a specific user based on limited data available for this user. It is in particular the case in the domain of speech rehabilitation and language pronunciation training, where visual articulatory feedback proves to be useful. When visual articulatory feedback is used for speech rehabilitation (Roxburgh *et al.*, 2015; Fabre *et al.*, 2017), visual information regarding the articulation produced by the speaker is displayed in real-time. If this display is to be performed by means of an articulatory model, information regarding both the speaker and the articulation is necessary. As little data are available for the speaker in clinical environment, the use of multi-speaker models appears appropriate. Based on the results of this study, a proof-of-concept has been tested where a speaker-specific articulatory model was estimated from limited data from the speaker, using the MoM. Using the articulation produced by the speaker, this model has been inversed to obtain articulatory predictors, that were in turn used to control a generic articulatory model; this model could be for instance the articulatory model of the mean speaker, or a specific articulatory model. Such an approach led to an overall articulators' rms reconstruction error of 0.25 cm. While errors are still large, in particular for the tongue, this approach constitutes a promising benchmark and motivates further the development of multi-speaker models.

ACKNOWLEDGMENTS

We thank all our kind and patient speakers. We also sincerely thank J.-A. Valdés Vargas and G. Ananthakrishnan for performing the majority of the initial tracings. Finally, we would like to express our sincere thanks to Gérard Bailly for suggesting the idea of a model of models during an informal discussion. This work has been partially funded by the French ANR (Grant No. ANR-08-EMER-001-02 "ARTIS"). 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 Santé"—ANR-11-INBS-0006. The authors are also very grateful to the journal editor Zhaoyan Zhang and to two anonymous reviewers for their very insightful comments and editorial help, and to Yuchen Lin for her acute help regarding the English edition.

[1]An example of strategy-related correlation is the correlation between the larynx height and the lip protrusion that is used to maximize the acoustic distance between /i/ and /u/ (cf. Hoole and Kroos, (1998); Beautemps *et al.*, 2001)

[2]According to Toutios and Narayanan (2015) and Labrunie *et al.* (2018), the posterior supraglottis carries information regarding the height of the larynx and the voicing state. Although not a true articulator as emphasized by Labrunie *et al.* (2018), it contains however meaningful information.

[3]When needed, the signs of the eigenvectors and predictors of some components and speakers have been swapped to ensure this property. Indeed, the PCA used in Sec. III leads to arbitrary signs for eigenvectors and predictors that can happen to be swapped for any component.

[4]In this section, the term *constituent* refers to the general set of articulatory features that are related to the specific morphology or the specific strategy of a speaker (e.g., overall vocal tract length, male/female differences, articulatory components, idiosyncratic behavior).

Abry, C., and Boë, L.-J. (**1986**). ""Laws" for lips," Speech Commun. **5**, 97–104.

Ananthakrishnan, G., Badin, P., Valdés Vargas, J. A., and Engwall, O. (**2010**). "Predicting unseen articulations from multi-speaker articulatory models," in *Proceedings of Interspeech 2010*, September 26–30, Makuhari, Japan.

Arlot, S., and Celisse, A. (**2010**). "A survey of cross-validation procedures for model selection," Stat. Surv. **4**, 40–79.
Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (**2002**). "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," J. Phon. **30**, 533–553.
Badin, P., and Serrurier, A. (**2006**). "Three-dimensional linear modeling of tongue: Articulatory data and models," in *Proceedings of the 7th International Seminar on Speech Production ISSP*, December 13–15, Ubatuba, Brazil, pp. 395–402.
Badin, P., Valdés Vargas, J. A., Koncki, A., Lamalle, L., and Savariaux, C. (**2013**). "Development and implementation of fiducial markers for vocal tract MRI imaging and speech articulatory modelling," in *Proceedings of Interspeech* 2013, August 25–29, Lyon, France, pp. 1321–1325.
Barbier, G., Boë, L.-J., Captier, G., and Laboissière, R. (**2015**). "Human vocal tract growth: A longitudinal study of the development of various anatomical structures," in *Proceedings of Interspeech 2015*, September 6–10, Dresden, Germany, pp. 364–368.
Beautemps, D., Badin, P., and Bailly, G. (**2001**). "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," J. Acoust. Soc. Am. **109**, 2165–2180.
Brunner, J., Fuchs, S., and Perrier, P. (**2005**). "The influence of the palate shape on articulatory token-to-token variability," ZAS Papers Ling. **42**, 43–67.
Brunner, J., Fuchs, S., and Perrier, P. (**2009**). "On the relationship between palate shape and articulatory behavior," J. Acoust. Soc. Am. **125**, 3936–3949.
Engwall, O. (**2000**). "A 3D tongue model based on MRI data," in *Proceedings of the 6th International Conference on Spoken Language Processing*, October 16–20, Beijing, China, pp. 901–904.
Fabre, D., Hueber, T., Girin, L., Alameda-Pineda, X., and Badin, P. (**2017**). "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," Speech Commun. **93**, 63–75.
Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.
Fuchs, S., Winkler, R., and Perrier, P. (**2008**). "Do speakers' vocal tract geometries shape their articulatory vowel space?," in *Proceedings of the 8th International Seminar on Speech Production ISSP*, December 8–12, Strasbourg, France, pp. 333–336.
Geng, C., and Mooshammer, C. (**2000**). "Modeling the German stress distinction," in *Proceedings of the 5th International Seminar on Speech Production ISSP*, December 8–12, Kloster Seeon, Germany, pp. 161–164.
Geng, C., and Mooshammer, C. (**2009**). "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," J. Acoust. Soc. Am. **125**, 3278–3288.
Goldstein, U. G. (**1980**). "An articulatory model for the vocal tracts of growing children," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Harshman, R., Ladefoged, P., and Goldstein, L. (**1977**). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. **62**, 693–707.
Hashi, M., Westbury, J. R., and Honda, K. (**1998**). "Vowel posture normalization," J. Acoust. Soc. Am. **104**, 2426–2437.
Honda, K., Maeda, S., Hashi, M., Dembowski, J., and Westbury, J. R. (**1996**). "Human palate and related structures: Their articulatory consequences," in *Proceedings of the 4th International Conference on Spoken Language Processing*, October 3–6, Philadelphia, PA.
Hoole, P. (**1998**). "Modelling tongue configuration in German vowel production," in *Proceedings of 5th International Conference on Spoken Language Processing*, November 30–December 4, Sydney, Australia, pp. 1863–1866.
Hoole, P. (**1999**). "On the lingual organization of the German vowel system," J. Acoust. Soc. Am. **106**, 1020–1032.
Hoole, P., and Kroos, C. (**1998**). "Control of larynx height in vowel production," in *Proceedings of 5th International Conference on Spoken Language Processing*, November 30–December 4, Sydney, Australia, pp. 531–534.
Hoole, P., Wismüller, A., Leinsinger, G., Kroos, C., Geumann, A., and Inoue, M. (**2000**). "Analysis of tongue configuration in multi-speaker, multi-volume MRI data," in *Proceedings of the 5th International Seminar on Speech Production ISSP*, December 8–12, Kloster Seeon, Germany, pp. 157–160.
Hu, F. (**2006**). "On the lingual articulation in vowel production: Case study from Ningbo Chinese," in *Proceedings of the 7th International Seminar on Speech Production ISSP*, December 13–15, Ubatuba, Brazil.
Jackson, M. T. T. (**1988**). "Analysis of tongue positions: Language-specific and cross-linguistic models," J. Acoust. Soc. Am. **84**, 124–143.
Johnson, K. (**1991**). "Dynamic aspects of English vowels in /bVb/ sequences," UCLA Working Papers Phon. **80**, 99–120.
Johnson, K., Ladefoged, P., and Lindau, M. (**1993**). "Individual differences in vowel production," J. Acoust. Soc. Am. **94**, 701–714.
Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., and Boë, L.-J. (**2018**). "Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning," Speech Commun. **99**, 27–46.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information Conveyed by Vowels," J. Acoust. Soc. Am. **29**, 98–104.
Lammert, A., Proctor, M., and Narayanan, S. (**2013**). "Interspeaker Variability in Hard Palate Morphology and Vowel Production," J. Speech Lang. Hear. Res. **56**, 1924–1933.
Lindau, M. (**1986**). "Vowel features in Akan and English," J. Acoust. Soc. Am. **80**, S62–S62.
Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Springer, Dordrecht, the Netherlands), pp. 403–439.
Lindblom, B. E. F., and Sundberg, J. E. F. (**1971**). "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am. **50**, 1166–1179.
Linker, W. (**1982**). "Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study," UCLA Working Papers Phon. **56**, 1–134.
Maeda, S. (**1979**). "Un modèle articulatoire de la langue avec des composantes linéaires" ("An articulatory model of the tongue with linear components"), in *Actes des 10èmes Journées d'Etude sur la Parole*, Grenoble, France, 152–162.
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* (Kluwer Academic, Dordrecht, the Netherlands), pp. 131–149.
Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2007**). "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," J. Phon. **35**, 20–39.
Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y. C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., and Proctor, M. (**2014**). "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," J. Acoust. Soc. Am. **136**, 1307–1311.
Nix, D. A., Papcun, G., Hogden, J., and Zlokarnik, I. (**1996**). "Two cross-linguistic factors underlying tongue shapes for vowels," J. Acoust. Soc. Am. **99**, 3707–3717.
Passavant, G. (**1869**). "Ueber die Verschliessung des Schlundes beim Sprechen" ("About the closure of the throat while speaking"), Virchows Archiv. **46**, 1–31.
Roxburgh, Z., Scobbie, J. M., and Cleland, J. (**2015**). "Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback," in *Proceedings of the 18th International Congress of Phonetic Sciences ICPhS*, August 10–14, Glasgow, UK, 0858.
Rudy, K., and Yunusova, Y. (**2013**). "The effect of anatomic factors on tongue position variability during consonants," J. Speech Lang. Hear. Res. **56**, 137–149.
Serrurier, A., and Badin, P. (**2008**). "A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data," J. Acoust. Soc. Am. **123**, 2335–2355.
Serrurier, A., Badin, P., Boë, L.-J., Lamalle, L., and Neuschaefer-Rube, C. (**2017**). "Inter-speaker variability: Speaker normalisation and quantitative estimation of articulatory invariants in speech production for French," in *Proceedings of Interspeech 2017*, August 20–24, Stockholm, Sweden, pp. 2272–2276.
Sorensen, T., Toutios, A., Goldstein, L., and Narayanan, S. (**2016**). "Characterizing vocal tract dynamics across speakers using real-time MRI," in *Proceedings of Interspeech 2016*, September 8–12, San Francisco, CA, 465–469.
Story, B. H. (**2005**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.
Story, B. H. (**2007**). "Time dependence of vocal tract modes during production of vowels and vowel sequences," J. Acoust. Soc. Am. **121**, 3770–3789.
Story, B. H., Vorperian, H. K., Bunton, K., and Durtschi, R. B. (**2018**). "An age-dependent vocal tract model for males and females based on anatomic measurements," J. Acoust. Soc. Am. **143**, 3079–3102.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., and Shosted, R. (**2012**). "Real-time MRI for Portuguese database, Methods and applications," in *Proceedings of PROPOR 2012*, April 17–20, Coimbra, Portugal, pp. 306–317.

Toutios, A., and Narayanan, S. (**2015**). "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *Proceedings of the 18th International Congress of Phonetic Sciences ICPhS*, August 10–14, Glasgow, UK.

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

Valdés Vargas, J. A. (**2013**). "Adaptation of orofacial clones to the morphology and control strategies of target speakers for speech articulation," Ph.D. thesis, Université Grenoble Alpes, Grenoble, France.

Valdés Vargas, J. A., Badin, P., and Lamalle, L. (**2012**). "Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods," in *Proceedings of Interspeech 2012*, September 9–13, Portland, OR.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. J., and Gentry, L. R. (**2009**). "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**, 1666–1678.

Weirich, M., and Fuchs, S. (**2013**). "Palatal morphology can influence speaker-specific realizations of phonemic contrasts," J. Speech Lang. Hear. Res. **56**, S1894–S1908.

Weirich, M., Lancia, L., and Brunner, J. (**2013**). "Inter-speaker articulatory variability during vowel-consonant-vowel sequences in twins and unrelated speakers," J. Acoust. Soc. Am. **134**, 3766–3780.

Yunusova, Y., Rosenthal, J. S., Rudy, K., Baljko, M., and Daskalogiannakis, J. (**2012**). "Positional targets for lingual consonants defined using electromagnetic articulography," J. Acoust. Soc. Am. **132**, 1027–1038.

Zheng, Y., Hasegawa-Johnson, M., and Pizza, S. (**2003**). "Analysis of the three-dimensional tongue shape using a three-index factor analysis model," J. Acoust. Soc. Am. **113**, 478–486.