

A mathematical model of the vowel space

Frédéric Berthommier

▶ **To cite this version:**

Frédéric Berthommier. A mathematical model of the vowel space. 2021. hal-03384303v2

HAL Id: hal-03384303
https://hal.science/hal-03384303v2

Preprint submitted on 29 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A mathematical model of the vowel space

Frédéric Berthommier
Univ. Grenoble Alpes, CNRS, Grenoble INP,
GIPSA-lab, 38000 Grenoble, France

October 29, 2021

Abstract

The articulatory-acoustic relationship is many-to-one and non linear and this is a great limitation for studying speech production. A simplification is proposed to set a bijection between the vowel space (f_1, f_2) and the parametric space of different vocal tract models. The generic area function model is based on mixtures of cosines allowing the generation of main vowels with two formulas. Then the mixture function is transformed into a coordination function able to deal with articulatory parameters. This is shown that the coordination function acts similarly with the Fant's model and with the 4-Tube DRM derived from the generic model.

1 Introduction

Establishing a causal relationship between articulatory movements and speech sounds is a major goal of the speech sciences. The main theoretical advances were made during the 1960s, devoted to understanding the actual shapes of the vocal tract (VT) in sound production. This led to the design of articulatory models as well as to the synthesis of speech sounds, but this effort was slowed down by the complexity of this relationship. Even if we limit the domain to vowels and their first two formants (f_1, f_2), this problem remains poorly understood. This is due to the many-to-one relation between the very large space of VT forms and the two-dimensional space of formants as well as to the non-linearities. Starting from real VT forms, this problem is so difficult that theoretical researches like [11, 1, 22] have failed to produce a clear conclusion. De facto, there is no reliable inversion technique applicable to any vocal tract, even with modern Bayesian or neural network approaches.

We are therefore unable to clearly understand how speech evolved from the shapes of the monkey vocal tract, or to understand how movements are generated by the brain to control the vocal tract. This lack of understanding is at the root of long-standing conflicts such as the one examined and seemingly closed by [6]. After the development of articulatory models such as Maeda's or Mermelstein's, the community split into two approaches: those who work directly with the area function to synthesize speech sounds and those who work with human-like models controlled by articulatory parameters. Both fields have thus seen many advances, but without much convergence. Until today, only the first approach has allowed to realize a bijection between the parametric space of the area function and the vowel space (f_1, f_2) [7, 17]. The aim of this paper is to revisit the roots in order to construct a link with articulatory modelling and to understand which causality and structure these models share. First, the technique adopted in section 2 avoids real VT data in order to perfectly control the area function. This allows for a generation of vowel spaces that are formed from a seed of three well chosen vectors and not empirical functions derived from data as in [17]. In section 3, the simplest model, which we call the generic model, allows the synthesis of 8 vowels with two concise formulas having 2 variables and length as the only static parameter. Since the vowel space is filled from its periphery and not generated recurrently from its center as in [7], the use of the sensitivity function [9] is avoided. Thus, we free ourselves from any consideration of the validity of this principle and the generic model can acquire the status of an independent mathematical object. Moreover, the first generative function, which we call the three-phase mixing function, is transformed in section 4 into a more comprehensible coordination function. This is the coordination of human-like articulators, which is tested in section 5 with a 4-tube model controlled by articulatory parameters. The seed of this model is created according to intuitive principles. Comparison with the DRM [13], a well-known 4-tube model here derived from the generic model, provides insight into the role played by the coordination function in reducing the many-to-one relationship.

2 The three-phase mixing function

A generative function $\boldsymbol{P}(\theta)$ is constructed with the principle of the Gibbs triangle used in chemistry for the representation of ternary mixtures of components [21]. The dimension of this graph is reduced to two because the three proportions a, b and c are reduced to a constant K fixed at 1 and the components are located at the three corners of the triangle. Since the vowel space is also considered a triangle with the extreme vowels [u,a,i] at

the corners (see [6]) and the neutral vowel [ə] at the center, the generative function of the vocal tract shapes takes three vectors corresponding to these vowels as input and defines the neutral vowel as their average. While the acoustic space is a triangle, the vocal tract shape space (VT) will be defined as a circular domain with a center, this for each scalar value of a vector. Each vector of a cycle including the three vectors $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ is given by the three-phase mixing function according to its angle θ:

$$\boldsymbol{P}(\theta) = \frac{1}{3}(\boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}) + \frac{2}{3}(\boldsymbol{i}\, cos(\theta - \frac{\pi}{3}) + \boldsymbol{j}\, cos(\theta - \pi) + \boldsymbol{k}\, cos(\theta - \frac{5\pi}{3})) \tag{1}$$

The first property of this function is to verify $\boldsymbol{P}(\frac{\pi}{3}) = \boldsymbol{i}$, $\boldsymbol{P}(\pi) = \boldsymbol{j}$ and $\boldsymbol{P}(\frac{5\pi}{3}) = \boldsymbol{k}$ and setting $\boldsymbol{\Omega} = \frac{1}{3}(\boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k})$, the Eq. (1) is rewritten:

$$\begin{aligned} \boldsymbol{P}(\theta) = \boldsymbol{\Omega} + \frac{2}{3}(&\boldsymbol{i}\,(cos\,\theta\, cos\,\frac{\pi}{3} + sin\,\theta\, sin\,\frac{\pi}{3}) - \boldsymbol{j}\, cos\,\theta \\ &+\boldsymbol{k}\,(cos\,\theta\, cos\,\frac{5\pi}{3} + sin\,\theta\, sin\,\frac{5\pi}{3})), \end{aligned} \tag{2a}$$

$$\rightarrow \boldsymbol{P}(\theta) = \boldsymbol{\Omega} + \frac{2}{3}(((\boldsymbol{i} + \boldsymbol{k})\, cos\,\frac{\pi}{3} - \boldsymbol{j})\, cos\,\theta + (\boldsymbol{i} - \boldsymbol{k})\, sin\,\frac{\pi}{3}\, sin\,\theta). \tag{2b}$$

This function is used as a generator of closed-open tube shapes, considering that $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ are the vectors defining the shapes of the tubes producing [u,a,i] placed at given angles $\theta \in \{\frac{\pi}{3}, \pi, \frac{5\pi}{3}\}$ of an equilateral triangle. The angle θ varies from 0 to 2π in order to mix these components. Having a triangle on one side and a circle on the other is a main support for this paper. The aim is to address the nonlinear and multiple characteristics of the articulatory-acoustic relationship and to obtain a bijection between a large space (f_1, f_2) of formant frequencies and a domain included in the very large VT shape space. This domain is constrained to incorporate the configurations of the 3 extreme vowels and to establish a continuum between them.

3 The generic vocal tract model

Following the distinction proposed by [18] between VT modeling approaches, the objective here is to explore the formal counterpart of the empirical studies conducted by most researchers. The vectors $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ are chosen from a theoretical point of view and not in close relation to real forms of VT. The first objective is to establish a link between the physics of closed-open tubes and the structure of the vowel space, and then to propose a unification of the two approaches. [15] established from Ehrenfest's theorem $\frac{\delta f_i}{f_i} = \frac{\delta E_i}{E_i}$ that for

small perturbations of the neutral tube, the first 2 odd cosine Fourier coefficients of the area function are related to the resonant frequency variations: $\frac{\delta f_i}{f_i} = -\frac{1}{2}a_i$ (a_1 being the amplitude of $cos(\frac{\pi x}{L})$ and a_2 of $cos(\frac{3\pi x}{L})$). This has a physical meaning independent of a specific transfer function between the VT shape space and the acoustic space, and others [12, 11] have derived the same relation from Webster's Horn equation after [19]. Let us evaluate these terms in the three-phase mixtures provided by Eq. 1. To simplify the mathematical expressions, $cos(\frac{\pi x}{L})$ is renamed to $\boldsymbol{v_1}$ and $cos(\frac{3\pi x}{L})$ to $\boldsymbol{v_2}$. Then $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ are defined as $\{\mathbf{1} + \boldsymbol{v_1} + \boldsymbol{v_2}, \mathbf{1} - 2\,\boldsymbol{v_1}, \mathbf{1} + \boldsymbol{v_1} - \boldsymbol{v_2}\}$. These are the simplest components compatible with the three-phase mixing function Eq. 1 and at the same time able to maximize and minimize the first two frequency resonances by extrapolation of the Schroeder-Ehrenfest relation. First, this function satisfies the condition $\boldsymbol{\Omega} = \frac{1}{3}(\boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}) = \mathbf{1}$. Then, these terms are introduced in Eq. 2b to identify $a_1(\theta)$ and $a_2(\theta)$:

$$\begin{aligned}\boldsymbol{P}(\theta) = \mathbf{1} + \frac{2}{3}\,((\mathbf{1} + \boldsymbol{v_1} + \boldsymbol{v_2} + \mathbf{1} + \boldsymbol{v_1} - \boldsymbol{v_2})\,cos\,\frac{\pi}{3}\; - \mathbf{1} + 2\,\boldsymbol{v_1})\,cos\,\theta \\ + \frac{2}{3}\,(\mathbf{1} + \boldsymbol{v_1} + \boldsymbol{v_2} - \mathbf{1} - \boldsymbol{v_1} + \boldsymbol{v_2})\,sin\,\frac{\pi}{3}\,sin\,\theta,\end{aligned} \tag{3a}$$

$$\rightarrow \boldsymbol{P}(\theta) = \mathbf{1} + 2\,\boldsymbol{v_1}\,cos\,\theta + \frac{4}{3}\,\boldsymbol{v_2}\,sin\,\frac{\pi}{3}\,sin\,\theta, \tag{3b}$$

$$\text{then}\, a_1(\theta) = 2\,cos\,\theta,\; a_2(\theta) = \frac{4}{3}\,sin\,\frac{\pi}{3}\,sin\,\theta\,\text{with}\,\boldsymbol{P}(\theta) = \mathbf{1} + a_1(\theta)\,\boldsymbol{v_1} + a_2(\theta)\,\boldsymbol{v_2}. \tag{3c}$$

For main vowel synthesis, there is no other parameter than the length L given in cm and the amplitude is already scaled in cm^2. A suitable soft rectifier Eq. 4b is applied to the continuous functions $P(x, \theta)$ arising from Eq. 3c to obtain positive valued area functions $A(x, \theta)$. This is the only explicit nonlinearity introduced in the model that can distort the Schroeder-Ehrenfest relation away from the neutral tube. These are sampled to apply the transfer function of a transmission line model. For the present simulations, we chose the Badin and Fant model [2] defined as lossless. The choice of transmission line model is not critical to reproduce the present results with these equations. For the sampling rate, a large value of $n \geq 100$ tubelets of length $\frac{L}{n}$ is preferable to avoid the distortions introduced at this stage. The length L is fixed at $17.5\,cm$ for comparison with other works. The resulting concise vowel equation is as follows:

$$\begin{cases} A(x, \theta) = expif(1 + a_1(\theta)\,cos(\frac{\pi x}{L}) + a_2(\theta)\,cos(\frac{3\pi x}{L})), & \text{(4a)} \\ \text{if}\,(y < 1)\,\text{then}\,expif(y) = exp(y - 1)\,\text{else}\,expif(y) = y. & \text{(4b)} \end{cases}$$

This equation associates the main vowels with θ angles corresponding to pairs of Fourier coefficients a_1 and a_2. In Fig. 1, the series of eight vowels [ɨ, u, o, ɔ, a, ɛ, e, i] is produced with $\theta \in \{0, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \pi, \frac{4\pi}{3}, \frac{3\pi}{4}, \frac{5\pi}{3}\}$ and with Fourier coefficients (partially) given by Eq. 3c $(a_1, a_2) \in \{(2,0), (1,1), (0,\alpha), (-1,1), (-2,0), (-1,-1), (0,-\alpha), (1,-1)\}$ with $\alpha = \frac{4}{3}\, sin\, \frac{\pi}{3}$. It is logical to find for [u, a, i] the coefficients given for the definition of $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$. Let us mention the analogy with an octatonic musical scale having W-H-H-W-W-H-H-W intervals which suggests that a common structure to speech and music is revealed by this representation (see [20]). Equation. 1 has a second important property for the expression of the structure of the vowel space which is the anti-symmetry with respect to $\boldsymbol{\Omega}$: we have $\boldsymbol{P}(\theta + \pi) = 2\,\boldsymbol{\Omega} - \boldsymbol{P}(\theta)$ then the coefficients of [ɛ, ɨ, ɔ] can be directly deduced from those of [u, a, i] because $a_i(\theta + \pi) = -a_i(\theta)$. The extrema of the acoustic space are well reached in spite of the nonlinearities transforming the circle into a triangular shape (Fig. 1): [ɨ, a, o, e] correspond well to $\{min(f_1), max(f_1), min(f_2), max(f_2)\}$ and this is determined by the values of a_i. To generate the complete vowel space, the extension to a domain is performed by multiplying $a_1(\theta)$ and $a_2(\theta)$ by a scalar ρ varying between 0 and 1. Thus, the central vowel [ə] corresponds to $\rho = 0$ and except for this point, a bijection is established between the surfaces $\boldsymbol{P}(\rho, \theta) \leftrightarrow (\rho\, a1, \rho\, a2)$ and this is observed between all of them and the vowel space (f_1, f_2).

4 The coordination function

Another more practical equation for n-tubes and articulatory modeling can be rewritten from the three-phase mixing function (Eq. 1). Let us insist that this is mathematically equivalent. This highlights the properties of the resulting parameter vector $\boldsymbol{P}(\theta)$ and its terms are significant: they express the coordination between the parameters along the θ cycle. In other words, for articulatory modeling, they explain how the phases and amplitudes of the different parameters of a model are locked together in a continuous domain. The area function is a side product that is analyzed section 5 to find a link between these two classes of models. As developed in section 5, it is also possible to apply it not only with the vectors $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$, but also with their three dual vectors $\{\boldsymbol{\Omega}, \boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}\}$ appearing in this function. As it is a rewriting, we give a priori the new form in which the two unknown terms are identified from those of Eq. 2b with element wise operations (we already

know the third $\boldsymbol{\Omega} = \frac{1}{3}(\boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k})$):

$$\boldsymbol{P}(\theta) = \boldsymbol{\Omega} + \boldsymbol{\Psi_1}\, cos(\boldsymbol{\Psi_2} - \theta) \tag{5a}$$

$$\rightarrow\ \boldsymbol{P}(\theta) = \boldsymbol{\Omega} + \boldsymbol{\Psi_1}\,(cos\,\boldsymbol{\Psi_2}\, cos\,\theta + sin\,\boldsymbol{\Psi_2}\, sin\,\theta) \tag{5b}$$

After identification, we obtain: $\boldsymbol{\Psi_1}\, cos\,\boldsymbol{\Psi_2} = \frac{2}{3}((\boldsymbol{i}+\boldsymbol{k})\, cos\,\frac{\pi}{3} - \boldsymbol{j})$ and $\boldsymbol{\Psi_1}\, sin\,\boldsymbol{\Psi_2} = \frac{2}{3}(\boldsymbol{i} - \boldsymbol{k})\, sin\,\frac{\pi}{3}$. The values of the vectors $\{\boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}\}$ are derived from those of $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$:

$$\begin{cases} \boldsymbol{\Psi_2} = atan(((\boldsymbol{i} - \boldsymbol{k})\, sin\,\dfrac{\pi}{3})/(\dfrac{1}{2}(\boldsymbol{i} + \boldsymbol{k}) - \boldsymbol{j})), & \text{(6a)} \\ \text{and } \boldsymbol{P}(\pi) = \boldsymbol{j} \rightarrow \boldsymbol{\Psi_1} = (\boldsymbol{\Omega} - \boldsymbol{j})\, cos^{-1}\boldsymbol{\Psi_2}. & \text{(6b)} \end{cases}$$

The antisymmetry with respect to $\boldsymbol{\Omega}$ is easily proved in this second form: we have $\boldsymbol{P}(\theta + \pi) = 2\,\boldsymbol{\Omega} - \boldsymbol{P}(\theta)$ since $cos(\boldsymbol{\Psi_2} - (\theta + \pi)) = -cos(\boldsymbol{\Psi_2} - \theta)$. In particular, the effect of each scalar parameter included in $\boldsymbol{P}$ at each θ is easy to identify as agonist or antagonist and this is useful for controlling articulatory models. The variations of each parameter along the cycle can be coded according to a rule for assigning a parameter extremum to a phase. For example, the length variation is coded as $L = \overline{L} + \Delta L\, cos(\frac{\pi}{3} - \theta)$ for having a maximum $\overline{L} + \Delta L$ for the phase $\frac{\pi}{3}$ of vowel [u]. Thus, the direct method is the conversion of the known vectors of the parameters of [u,a,i] with Eq. 6a and 6b. As before, the coordination function is extended to a domain by multiplying the amplitude $\boldsymbol{\Psi_1}$ by ρ:

$$\boldsymbol{P}(\rho, \theta) = \boldsymbol{\Omega} + \rho\, \boldsymbol{\Psi_1}\, cos(\boldsymbol{\Psi_2} - \theta) \tag{7}$$

This formula is easy to use and if applied to real VT data published by [17], it generates a mapping from 3 vectors corresponding to extreme vowels including the length component.

5 Comparison of two 4-tube models

While the generic model has 2 degrees of freedom to generate a large vowel space (f_1, f_2), the 4-tube models [16, 8] require 3 homogeneous parameters with $\{X_c, A_c, A_l\}$ (constriction position and area, lips area). It has been shown that anthropomorphic models also require at least these parameters for inversion [5]. On the other hand, with Monte Carlo simulations with 4-tube with a total length set at $17.5\,cm$, [6] noticed that the vowel space is maximized with 7 parameters and that having $N > 4$ sections is unnecessary. This maximized vowel space was defined as the articulatory potential of a closed-open tube having a given length. The generic model has somewhat smaller (f_1, f_2) (Fig. 1) and we suspect that the extra 5 of parameters

are the source of the undesirable many to one relationship, partly due to the fact that the 4-tube cut is determined by 3 of them. Note that the third formant is not incorporated into the current definition of vowel space. Therefore, to obtain its control, the generic model must be defined with $cos(\frac{5\pi x}{L})$ terms leading to an additive parameter. This would be consistent with $\{X_c, A_c, A_l\}$ but this improvement is to be developed later. We derive from the generic model the well-known 4-tube DRM having an optimal cut [14] and compare it with a circularized Fant model with 4 parameters $\{X_c, A_c, A_l, L\}$. The goal of the DRM is to propose a division into 8 regions and a rule to vary their areas by knowing the direction of variation of each of the first 3 formants. For the 2 formants only, we limit ourselves to 4 regions and the cut of the closed-open tube is done at $\{\frac{L}{6}, \frac{L}{2}, \frac{5L}{6}\}$ from glottis to lips. The parameters of the DRM are the areas of each section named $\{A_1, A_2, A_3, A_4\}$ while the corresponding values of the generic model are the P_i values before rectification (Eq. 4b). Since the cutting is done at zero crossings of the $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ components, these values are taken in the generic model at the $x \in \{0, \frac{L}{3}, \frac{2L}{3}, L\}$ points to compute P_i. Moreover, the values of $\{P_1, P_2\}$ are antisymmetric by construction from the generic model with those of $\{P_3, P_4\}$. This is an internal antisymmetry about the midpoint $\frac{L}{2}$, well described by [14] and not to be confused with that created by the coordination function, although it has common effects. Because $v_i(0) = -v_i(L)$ and $v_i(\frac{L}{3}) = -v_i(L - \frac{L}{3}) = -v_i(\frac{2L}{3})$ we obtain $P_3 = 2 - P_2$ and $P_4 = 2 - P_1$ from Eq. 3c. What is called the reduction is the systematic calculation of $\{P_3, P_4\}$ from $\{P_1, P_2\}$ so that this property explicitly reduces the number of free parameters to 2. When this reduction is not applied, the coordination function preserves the internal antisymmetry introduced by the components $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. This is because their amplitudes respect $a_i(\theta + \pi) = -a_i(\theta)$ along the cycle as seen in section 3. There is an equivalence between the explicit reduction and the coordination and this is encoded in $\boldsymbol{\Psi_1}$ and $\boldsymbol{\Psi_2}$. We find here the main property of the coordination function which consists in relying on the odd cosine Fourier components. With the generic model and the present DRM reduction, these 2 components are controlled variables so that only two degrees of freedom are needed to determine the variations of (f_1, f_2) away from the neutral tube, even if they are non linear. This is a baseline for observing the effect of the coordination function when the $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ components and many others are not controlled. Its configuration for $\{P_1, P_2\}$ after the transformation $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\} \rightarrow \{\boldsymbol{\Omega}, \boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}\}$ is detailed in Table 1. Following the same notation as for the generic model, for $\{P_3, P_4\}$, we have $\Psi_1(\frac{2L}{3}) = -\Psi_1(\frac{L}{3})$, $\Psi_2(\frac{2L}{3}) = \Psi_2(\frac{L}{3})$ and $\Psi_1(L) = -\Psi_1(0)$, $\Psi_2(L) = \Psi_2(0)$. This is easy to check that $P_3 = 2 - P_2$ and $P_4 = 2 - P_1$ as for the reduction. The version of Fant's model that was implemented by [3] is embedded

in the coordination function format. Its main difference is that a constant length section is sliding (Fig. 3a) whereas the DRM is a piston model with a fixed cut: it varies the areas of sections only. The advantage pointed out by [4] is to represent the tongue movements by variations of $\{X_c, A_c\}$ and this has been considered until now as a divergence between these two classes of models. However, the geometry of this model does not have the symmetries introduced by the cut and at the same time it reduces the antisymmetry of the area function. This does not mean that the latter is unrecoverable or that it has no role. Indeed, the vocal tract does not function as a piston model but there is no evidence that they have nothing in common as defended by [4]. To equalize their number of parameters to 4, we also vary the length L in the Fant model. We have seen in section 4 how to find $\{\mathbf{\Omega}, \mathbf{\Psi_1}, \mathbf{\Psi_2}\}$ for this parameter. The principle is the same for the 3 others: (1) choice of the offset (2) fixing the variation amplitude around the offset before soft rectification (3) assigning an extremum of each parameter to a vowel phase. For example Table 2, X_c is maximal for [i] so that the configuration acquires a large posterior Helmholtz cavity specific to this vowel [6]. Note that the language adopted to specify configurations is different in articulatory modeling because it is goal-oriented and qualitative. Based on the geometry of the model, the coordination function specifies how the parameters are cycled. Here, variations in $\{X_c, A_c\}$ correspond to an anatomically relevant rotation of the tongue in the mouth. This is also easy to recover $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ from $\{\mathbf{\Omega}, \mathbf{\Psi_1}, \mathbf{\Psi_2}\}$. At this point, the coordination function is fully compatible with both classes of models. Nevertheless, there is a reconstruction step to have a sampled $P(i, \rho, \theta)$ which is described in the captions of Table 1 and 2. Then, soft rectification is applied with Eq. 4b, changing the DCT spectrum for both models. Remarkably, a small amount of even components are introduced into the area function of the DRM. The transmission line model is set lossless for the DRM and lossy for the Fant model. The models are first simulated at $\rho = 1$ and θ varying from 0 to 2π with the coordination function (labelled $\boldsymbol{C2}$ in Fig. 2 and 3). Both vowel spaces have a similar structure with the 8 vowels placed comparatively (Fig.2,3b). The size of the vowel space for Fant's model appears to be reduced. The two 4-tube models are simulated under 2 conditions: (1) $\boldsymbol{C1}$, with their 4 parameters uncorrelated and uniformly distributed within $\mathbf{\Omega} \pm \mathbf{\Psi_1}$ (2) $\boldsymbol{C2}$ using the coordination function with $\rho = 1$ together with (ρ, θ) randomnesses. We observe Fig.2 that for the DRM, the space (f_1, f_2) is identically covered except at the corners of the uniform distribution. For Fant's model, more points of $\boldsymbol{C1}$ escape from the $\boldsymbol{C2}$ space but this is globally confined. Then $\boldsymbol{C2}$ does not decrease so much the articulatory potential of Fant's model as defined by $\boldsymbol{C1}$. We evaluate the two cosine Fourier coefficients of the area function which are here the most

informative components of the DCT spectrum:

$$\begin{cases} \tilde{a}_1(\rho,\theta) = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n} A(i,\rho,\theta)\, cos(\frac{\pi i}{n}), & \text{(8a)} \\ \tilde{a}_2(\rho,\theta) = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n} A(i,\rho,\theta)\, cos(\frac{3\pi i}{n}). & \text{(8b)} \end{cases}$$

For each of the models, the formants (f_{1n}, f_{2n}) of the neutral position $\boldsymbol{P}(0,0)$ are calculated. Let remark this is not an uniform area function for the Fant model. In Fig.4, the deviations from the neutral position in relative values $df_1 = (f_1 - f_{1n})/f_1$ and $df_2 = (f_2 - f_{2n})/f_2$ are plotted as a function of $(\tilde{a}_1, \tilde{a}_2)$. This shows that $\boldsymbol{C2}$ is represented by surfaces whereas $\boldsymbol{C1}$ is represented by disparate dots distant from these surfaces. Considering that $(\tilde{a}_1, \tilde{a}_2)$ reflects control parameters the key point is that we observe for $\boldsymbol{C1}$ the many to one relationship whereas it disappears for $\boldsymbol{C2}$: the intrinsic dimension is reduced at 2 for both models. By combining the two surfaces which cross in each graph, a bijection is established between pairs $(\tilde{a}_1, \tilde{a}_2)$ and (df_1, df_2). Note, however, that there is a non-visible deviation from bijectivity between [a] and [ɔ] for Fant's model, but that this does not affect the validity of this observation. For DRM, we have seen that the dimension can be explicitly reduced due to internal antisymmetry. Here, the only reason to have a many-to-one relationship is the decorrelation between the areas A_i. The coordination function only generates the internal antisymmetry as does the explicit reduction. Moreover, the variations of (f_1, f_2) according to $(\tilde{a}_1, \tilde{a}_2)$ from one model to another are well correlated in the $\boldsymbol{C2}$ condition. For the Fant model, the coordination function generates a geometrically consistent set of configurations that are antisymmetric because $(\tilde{a}_1, \tilde{a}_2)$ vary circularly. Despite the weaker internal antisymmetry and then weaker values of $(\tilde{a}_1, \tilde{a}_2)$, the two vowel spaces acquire a similar structure. The first explanation for the disparity of the $\boldsymbol{C1}$ points in Fig.4a, which is the signature of the many-to-one relation, is the growth of the even cosine Fourier coefficients associated with a degree of internal symmetry. [15] showed as a corollary that these coefficients have no first order effect on (f_1, f_2) and we extrapolate here by assuming that they have a small effect in the range of odd coefficients. After [12], [22] tried to disentangle them using constraints, knowing they cannot be evaluated from the formants f_i. Then, the decorrelation between areas A_i in DRM $\boldsymbol{C1}$ is a source of many-to-one relationship as shown in Fig.4a. A remaining question is the relationship between the resonant frequencies and the cosine Fourier coefficients. It turns out that for the generic model as well as for the reduced DRM, this presents a quadratic

term biasing the Schroeder-Ehrenfest relation only for df_1. The variations away from the neutral tube are approximated fairly well by $df_1 = -\frac{\tilde{a}_1}{2} - \frac{\tilde{a}_2^2}{4}$ and $df_2 = -\frac{\tilde{a}_2}{2}$ (this estimate is plotted Fig. 2). This has the main effect of transforming a circular shape into a triangular shape, providing physical support for the common shape of the vowel space. This cannot be due to soft rectification because it does not produce such a bias but it is an interference in the transfer function. This fitting is a surprising simplification to explain. This depends on the specific cut $\{\frac{L}{6}, \frac{L}{2}, \frac{5L}{6}\}$ of the reduced DRM and any perturbation of it will change $(\tilde{a}_1, \tilde{a}_2)$ and invalidate the previous estimate. In the Fant model, the X_c parameter controlling the length of the sections is involved, precluding the fit between $(\tilde{a}_1, \tilde{a}_2)$ and (f_1, f_2). We conclude that implicit nonlinearities are difficult to analyze, emphasizing the value of having the simplest models in hand as a baseline. Moreover, even coefficients linked with symmetric configurations are enhanced in the Fant model and this is combined with length variations for explaining the many-to-one relationship observed in $\boldsymbol{C1}$. All these factors are present with real VT forms and this makes the existence of an analytical solution doubtful. In this context, the coordination function has the main property of generating for the non-optimal VT forms the regular relation existing with the generic model and the reduced DRM. In other words, the coordination function is an optimal way to exploit the articulatory potential of a vocal tract to create a vowel space as well as to generate a one-to-one correspondence. The residue of $\boldsymbol{C1}$ points (Fig. 3b) with better contrast might be unrecoverable. Following this reasoning, most of the points produced with 7 parameters [6, Fig. 4 with $L = 17.5\,cm$] and especially those located in the crown outside the DRM $\boldsymbol{C2}$ could be excluded from the definition of the articulatory potential of a closed-open tube. It appears that these points are outside the human vowel space, with the exception of [i] which is outside of the DRM $\boldsymbol{C2}$. This suggests that the third parameter we have discarded has a role for this vowel and is a call to extend the present analysis.

6 Conclusion

After [7], we showed that the vowel space can be described as a pre-existing mathematical object. After [4], this has the disadvantage that it bears little resemblance to the human VT. To resolve this contradiction, a Platonic proposal is to consider that humans have discovered how to make vowels with their non-optimal vocal tract by using the coordination function. The first way of coding $\{\boldsymbol{\Omega}, \boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}\}$ from $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ as in Table 1 is abstract but the second Table 2 is on the contrary adapted to think about this evolutionary

process because it is geometrically constrained. This means that the acoustic contrast df_i is determined by the volume $[\mathbf{\Omega} \pm \mathbf{\Psi_1}]$ which is indexed by $(\tilde{a}_1, \tilde{a}_2)$. For example, a restriction of the interval of X_c related to the ability to rotate the tongue in the mouth has a roughly proportional effect on $(\tilde{a}_1, \tilde{a}_2)$ and then on the size of the space (f_1, f_2). For monkeys, this capacity is strongly reduced because they have a fixed larynx and a small pharyngeal cavity. The anatomical ranges observed by [10] are quite small and this provides a new objective criterion for judging whether they are truly "speech ready".

References

[1] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978.

[2] P. Badin and G. Fant. Notes on vocal tract computations. Stl- qpsr 2-3/1984, Royal Institute of Technology, Stockholm, Sweden, 1984.

[3] P. Badin, P. Perrier, L.-J. Boë, and C. Abry. Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *The Journal of the Acoustical Society of America*, 87(3):1290–1300, 1990.

[4] Louis-Jean Boë and Pascal Perrier. Comments on “distinctive regions and modes: A new theory of speech production” by m. mrayati, r. carré and b. guérin. *Speech Communication*, 9(3):217–230, 1990.

[5] Louis-Jean Boë, Pascal Perrier, and Gérard Bailly. The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1):27–38, 1992.

[6] Louis-Jean Boë, Thomas R. Sawallis, Joël Fagot, Pierre Badin, Guillaume Barbier, Guillaume Captier, Lucie Ménard, Jean-Louis Heim, and Jean-Luc Schwartz. Which way to the dawn of speech?: Reanalyzing half a century of debates and data in light of speech science. *Science Advances*, 5(12):eaaw3916, 2019.

[7] René Carré. Dynamic properties of an acoustic tube: Prediction of vowel systems. *Speech Communication*, 51(1):26–41, 2009.

[8] G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1960.

[9] G. Fant and S. Pauli. Spatial characteristics of vocal tract resonance modes. In *Proceeding of the Speech Communication Seminar*, 1974.

[10] W. Tecumseh Fitch, Bart de Boer, Neil Mathur, and Asif A. Ghazanfar. Monkey vocal tracts are speech-ready. *Science Advances*, 2(12):e1600723, 2016.

[11] J. M. Heinz. Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues. Stl- qpsr 1/1967, Royal Institute of Technology, Stockholm, Sweden, 1967.

[12] P. Mermelstein. Determination of the vocal-tract shape from measured formant frequencies. *The Journal of the Acoustical Society of America*, 41(5):1283–1294, 1967.

[13] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: A new theory of speech production. *Speech Commun.*, 7(3):257–286, October 1988.

[14] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: A new theory of speech production. *Speech Commun.*, 7(3):257–286, October 1988.

[15] M. R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *The Journal of the Acoustical Society of America*, 41(4B):1002–1010, 1967.

[16] Kenneth N. Stevens and Arthur S. House. Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, 27(3):484–493, 1955.

[17] Brad H. Story. A parametric model of the vocal tract area function for vowel and consonant simulation. *The Journal of the Acoustical Society of America*, 117(5):3231–3254, 2005.

[18] Brad H. Story and Ingo R. Titze. Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics*, 26(3):223–260, July 1998.

[19] G. Ungeheuer. *Elemente einer akustischen Theorie der Vokalartikulation*. Springer-Verlag, Berlin, 1962.

[20] Wikipedia contributors. List of musical scales and modes — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?`

title=List_of_musical_scales_and_modes&oldid=1007710677, 2021. [Online; accessed 20-October-2021].

[21] Wikipedia contributors. Ternary plot — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Ternary_plot&oldid=1027347669, 2021. [Online; accessed 20-October-2021].

[22] Hani Yehia and Fumitada Itakura. A method to combine acoustic and morphological constraints in the speech production inverse problem. *Speech Communication*, 18(2):151–174, 1996.

7 Figures and Tables

Parameter	x	$i(x)$	$j(x)$	$k(x)$	$\Omega(x)$	$\Psi_1(x)$	$\Psi_2(x)$
P_1	0	3	-1	1	1	$2\,cos^{-1}(atan(\frac{2}{3}\,sin\,\frac{\pi}{3}))$	$atan(\frac{2}{3}\,sin\,\frac{\pi}{3})$
P_2	$L/3$	0.5	0	2.5	1	$cos^{-1}(atan(-\frac{4}{3}\,sin\,\frac{\pi}{3}))$	$atan(-\frac{4}{3}\,sin\,\frac{\pi}{3})$

Table 1: Setup of the coordination function (Eq. 7) for the 2 first parameters of the DRM controlling the areas of the two first sections from glottis and the two others are set in the text. The model is derived from $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$ of the generic model at given points x. The reconstruction of the 4-tube $P(i, \rho, \theta)$ function is realized with $n = 120$ tubelets having length $\frac{L}{n}$ with $L = 17.5\,cm$. The proportion of n composing each of the 4 sections is $\{\frac{n}{6}, \frac{n}{3}, \frac{n}{3}, \frac{n}{6}\}$ and the value of $P(i, \rho, \theta)$ for each is $\{P_1, P_2, P_3, P_4\}$.

Parameter	$\boldsymbol{\Omega}$	$\boldsymbol{\Psi_1}$	$\boldsymbol{\Psi_2}$
X_c	$L_c/2$	$0.3\,(L_c - l)$	$5\Pi/3$
A_c	$-1.5\,A$	$2\,A$	Π
A_l	$A/2$	$-A$	$\Pi/3$
L	$\overline{L}$	ΔL	$\Pi/3$

Table 2: Specification of $\{\boldsymbol{\Omega}, \boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}\}$ of Eq. 7 for the 4 parameters of the Fant model with the rules described in the text. For parameters $\{A_c, A_l, L\}$ the underlying geometrical parameters are: $A = 1\,cm^2$, $\overline{L} = 17.5\,cm$ and $\Delta L = 1.5\,cm$. The length of sections is defined in number of tubelets so that for X_c the underlying geometrical parameters are proportions of $n = 200$: $L_c = 0.9$ and $l = 0.3$. The reconstruction of the 4-tube $P(i, \rho, \theta)$ function is realized after rounding, merging and before applying the soft rectification.

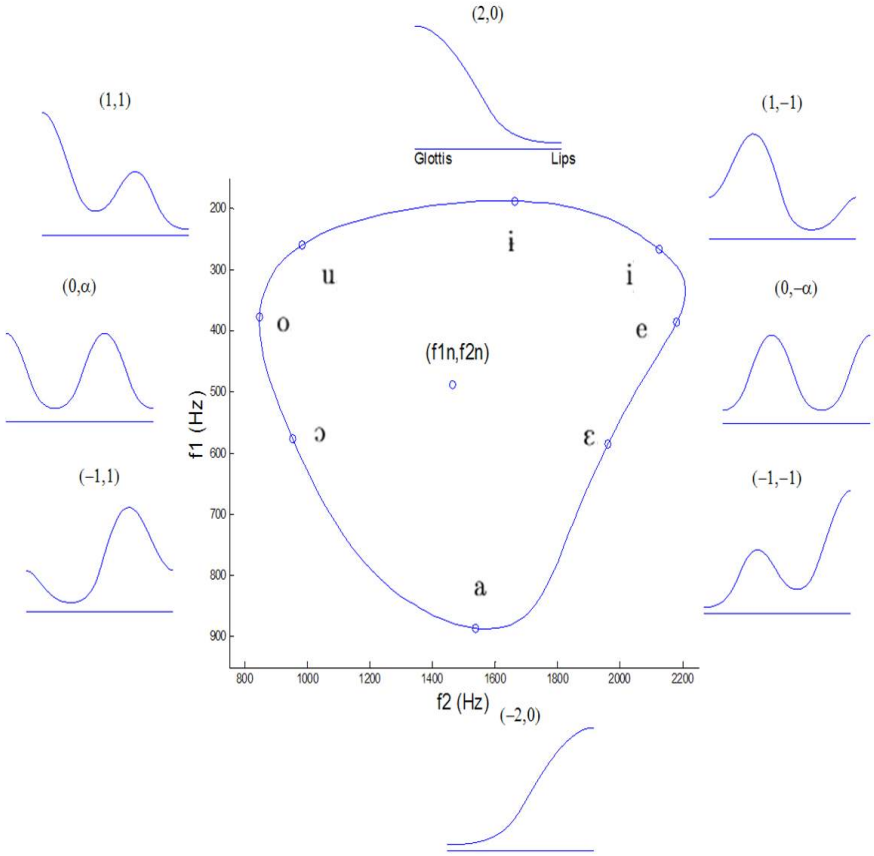


Figure 1: The vowel space of the generic vocal tract model and its 8 characteristic vowels. The area functions are plotted for each vowel with, at the top, the pairs (a_1, a_2) of cosine Fourier coefficients which are introduced in the vowel equation (Eq. 4ab). For [o,e], we have $\alpha = \frac{4}{3}\, sin\, \frac{\pi}{3}$.

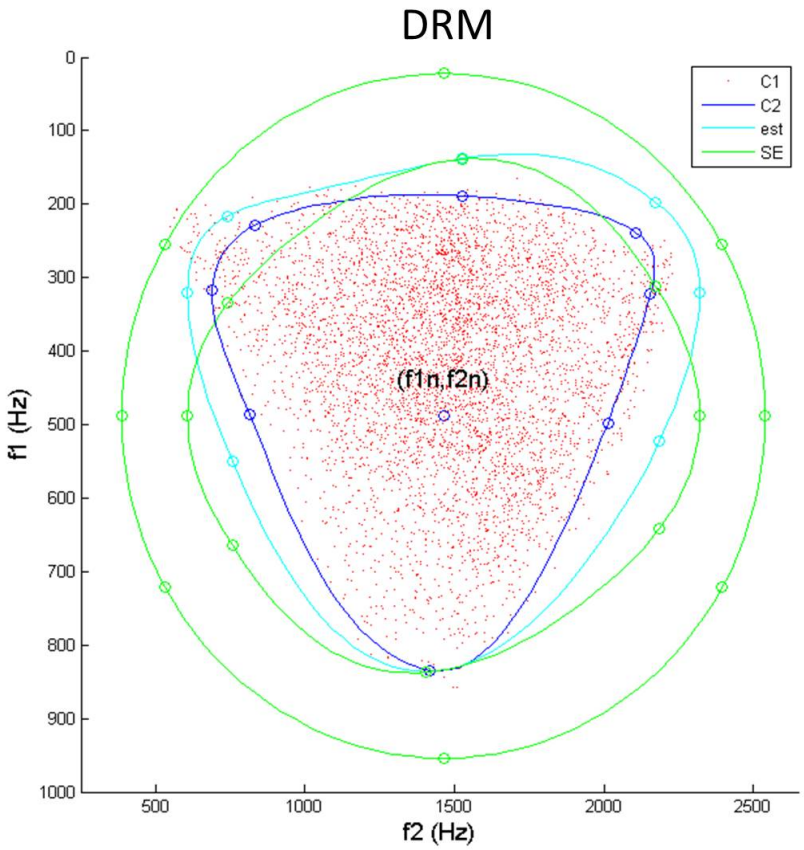


Figure 2: Simulation in condition $\boldsymbol{C1}$ together with the vowel space $\boldsymbol{C2}$ with $\rho = 1$ obtained with the coordination function. All (non figured) points of $\boldsymbol{C2}$ are inside of this. The spaces derived from the Schroeder-Ehrenfest relation and Fourier coefficients before and after rectification ("SE" from (a_1, a_2) and $(\tilde{a}_1, \tilde{a}_2)$) and the biased relation ("est" from $(\tilde{a}_1, \tilde{a}_2)$) are plotted relative to the neutral reference (f_{1n}, f_{2n}).

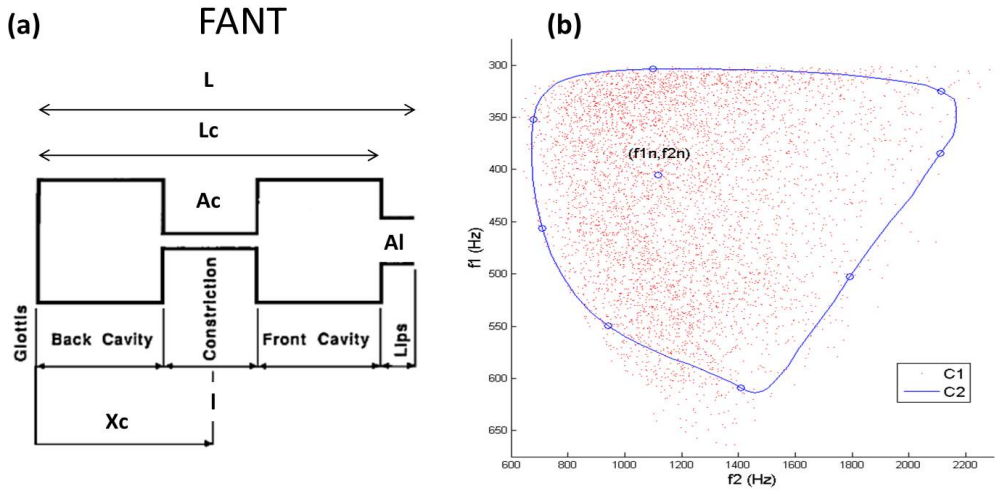


Figure 3: (a) The design of the Fant model redrawn from [3, Fig. 1] and the definition of geometrical parameters used in Table 2 (b) Simulation in condition $\boldsymbol{C1}$ together with the vowel space $\boldsymbol{C2}$ with $\rho = 1$ obtained with the coordination function. All (non figured) points of $\boldsymbol{C2}$ are inside of this.

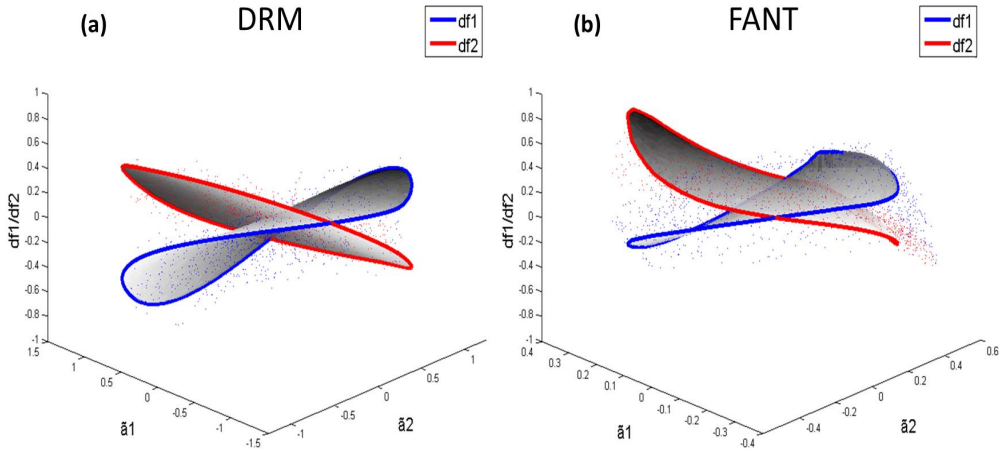


Figure 4: Simulations of df_i in condition $\boldsymbol{C1}$ and $\boldsymbol{C2}$: (a) DRM (b) Fant model. Abscissa: Fourier cosine coefficients of the area function. Ordinate: Deviations from neutral in relative values $df_i = (f_i - f_{in})/f_{in}$ for the two first formants. The circled gray surfaces represent the configurations produced by the coordination function in $\boldsymbol{C2}$ and the dots represent $\boldsymbol{C1}$ which is a simulation of the 4 parameters of each model but uncorrelated in the same domain of variation $\boldsymbol{\Omega} \pm \boldsymbol{\Psi_1}$.