The effects of larynx height on vowel production are mitigated by the active control of articulators

Rick Janssen, Scott R. Moisik, Dan Dediu

The effects of larynx height on vowel production are mitigated by the active control of articulators

Rick Janssen[a], Scott R. Moisik[b,a], Dan Dediu[c,d,a,*]

[a] *Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
[b] *Linguistics and Multilingual Studies, Nanyang Technological University, Singapore*
[c] *Laboratoire Dynamique Du Langage, Université Lumière Lyon 2, Lyon, France*
[d] *Collegium de Lyon, Institut d'Études Avancées, Lyon, France*

Abstract

The influence of larynx position on vowel articulation is an important topic in understanding speech production, the present-day distribution of linguistic diversity and the evolution of speech and language in our lineage. We introduce here a realistic computer model of the vocal tract, constructed from actual human MRI data, which can learn, using machine learning techniques, to control the articulators in such a way as to produce speech sounds matching as closely as possible to a given set of target vowels. We systematically control the vertical position of the larynx and we quantify the differences between the target and produced vowels for each such position across multiple replications. We report that, indeed, larynx height does affect the accuracy of reproducing the target vowels and the distinctness of the produced vowel system, that there is a "sweet spot" of larynx positions that are optimal for vowel production, but that nevertheless, even extreme larynx positions do not result in a collapsed or heavily distorted vowel space that would make speech unintelligible. Together with other lines of evidence, our results support the view that the vowel space of human languages is influenced by our larynx position, but that other positions of the larynx may also be fully compatible with speech.

Keywords: larynx height, vowel articulation, vocal tract model

*Corresponding author
Email address: dan.dediu@univ-lyon2.fr (Dan Dediu)

2010 MSC: 00-01, 99-00

1. Introduction

The origin and evolution of language and speech are a heavily debated topic, a major division being between models proposing recent and sudden origin, restricted to modern humans only (Berwick & Chomsky, 2017; Hauser et al., 2014; Klein, 2009), versus deep origin, gradual evolution, and a wider distribution (also including archaic humans, such as the Neanderthals; Dediu & Levinson, 2013, 2018; Lieberman, 2016; Johansson, 2015). In particular, the speech capacities of archaic humans have been linked to the position of the *larynx* (itself linked to the position of the *hyoid bone*), and the corresponding *ratio* between the horizontal and the vertical parts of the vocal tract (Lieberman, 2016).

While it is currently unclear what this ratio might have been in Neanderthals and when its "modern" value evolved (Lieberman, 2016; Gokhman et al., 2017; Dediu & Levinson, 2013), a more tractable question concerns its effects on speech and language (Lieberman, 2016; Boë et al., 2002; de Boer & Fitch, 2010). More precisely, the seminal claim by Lieberman & Crelin (1971) that a high larynx (a position suggested by some for Neanderthals) reduces the vowels space, making impossible the production of the widely-used [a], [i], [u] and [ɔ], has generated a lively debate centered on the use of computer models of the vocal tract to make such inferences (de Boer & Fitch, 2010; Boë et al., 2007; Lieberman, 2007).

For example, starting from the suggestion (Honda & Tiede, 1998) that larynx height may be deduced from the shape of the oral cavity, Boë (1999) used the "variable linear articulatory model" (VLAM) (Maeda, 1990) coupled with factor analysis and a growth model to argue against (Lieberman & Crelin, 1971). Building on this and work by Ménard & Boë (2000), Boë et al. (2002) concluded that "the maximal vowel space of a given vocal tract does not depend on the larynx height index: gestures of the tongue body (and lips and jaw) allow compensation for differences in the ratio between the dimensions of the oral cavity and pharynx" (p. 481). Boë et al. (2007) reiterated that VLAM shows

a high larynx not leading to a less distinctive vowel space. However, de Boer & Fitch (2010) attributed circular reasoning to Boë et al. (2002), as the growth scaling in Boë (1999), Boë et al. (2002) and Boë et al. (2007) was applied after the articulatory factors have been extracted in the VLAM, meaning that any inferred anatomies (Neanderthals, infants) have the same degrees of articulatory freedom as modern female adults, but just with a different scaling (for example, this does not hold in the observational data from pre-babbling vocalizations of infants, which are (epilaryngeally) constricted, clearly with less degrees of articulatory freedom; Esling et al., 2015). Furthermore, such global scaling preserves the layout of the different components of the model including the angle and ratio between the pharynx and the oral cavity, but a change in this layout is precisely what has been hypothesized to set modern humans apart. Finally, de Boer & Fitch (2010) argued that the use of factor analysis in VLAM linearly extrapolates from observed to unobserved cases, likely overestimating the ability of the articulators to compensate for any effects of anatomy, and developed, in response, a model better adhering to the anatomical constraints of the vocal tract, showing that a larynx height similar to a human female would be ideal for maximally distinctive vowel inventory (Lieberman, 2012).

Here we introduce a novel computer model that has several advantages over its predecessors. First, it is based on a widely-used realistic 3D geometric model of the vocal tract (VocalTractLab 2.1) built on modern phonetic theory and calibrated with data (MRI and otherwise) from actual humans (Birkholz, 2005; Birkholz & Kröger, 2006; Birkholz, 2013a). Second, this model allows the programmatic control of multiple meaningful articulatory parameters (such as the position of the tongue tip or the degree of lip rounding), and produces the corresponding acoustic output. Third, with the author's permission, we modified this model to allow (among others) the specification of hyoid position. Fourth, we implemented a complete agent that can control this vocal tract model using a generic machine learning algorithm, and which is capable of learning to produce a set of auditorily presented target vowels (here, [ə], [ɑ], [a], [æ], [e], [i], [o] and [u]) by controlling the free articulators of the model. This allows us to sys-

tematically study the impact of larynx height on vowel production, to find the optimal height for the production of widely-used vowels, and the compensatory strategies that can mitigate the impact of extreme larynx positions.

While still far from perfect, we think that our model represents an important advance, allowing more refined answers to questions surrounding the impact of larynx height on vowel production, and providing a platform for further improvement and application to other aspects of inter-group and inter-individual variation in speech, both pathological and normal (Dediu et al., 2017). Given that the work reported here is in many ways novel, one of our main aims was to start from as "generic" and "theory-free" assumptions as possible and to write our code as easily replaceable and upgradeable modules.

2. Data and Methods

The fundamental idea is to study how learning a set of vowels is affected by controlled changes in a particular aspect of vocal tract anatomy, here, *larynx height*. Such experimental manipulations are extremely difficult to conduct with human participants, but computer simulations using realistic models of the human vocal tract may offer approximations that, while imperfect, may still be good enough for answering specific questions in an objective, repeatable and quantitative manner. For more details on the model, please see Janssen (2018).

2.1. The Vocal Tract model

We implemented a realistic 3D model of the vocal tract based on a modified version of Peter Birkholz's `VocalTractLab` version 2.1 (Birkholz, 2005, 2013b,a). VocalTractLab 2.1 is a 3D geometric model of the vocal tract where a number of articulatory parameters (such as tongue tip position or lip rounding, among others) can be manipulated, and which, for a given set of parameter values, estimates the vocal tract's area function and produces the corresponding acoustic output, resulting in naturally-sounding and intelligible speech (Birkholz, 2005,

2013b,a). We obtained VocalTractLab 2.1's source code[1] and the permission to modify it from its author, Peter Birkholz (license agreement dated 21st of March, 2014). For the work reported here, we added the functionality to adjust the larynx height and we implemented tighter constraints on the hyoid's range of motion in relation to larynx height.

In total, our model has one parameter that varies between conditions but is fixed within (LEN, controlling the vertical position of the glottis, is the length of the vertical part of the supralaryngeal vocal tract, and is used to compute the vertical position of the larynx as described below; we will refer to LEN in the following as "larynx height"), 11 parameters that are under the models' direct control (HX, HY, JA, LP, LD, TCX, TCY, TTX, TTY, TBX, TBY), 7 that have fixed values (VS, VO, W, $TS1$, $TS2$, $TS3$, $TS4$), and 2 that are automatically computed (TRX, TRY); see Table 1 and Figure 1.

The ratio between the lengths of the vertical and the horizontal parts of the vocal tract is fully determined, in our model, by the vertical position of the glottis relative to the hyoid (given that the horizontal part of the model is fixed to the VocalTractLab 2.1 default). This vertical position of the glottis represents the lower point from where we compute the length of the vertical part of the vocal tract (the upper part is also fixed to VocalTractLab 2.1 defaults) and which we denote as SVT_v *base length* or LEN (see Figure 2; SVT is an abbreviation of the Supralarygngeal Vocal Tract, and the $_v$ and $_h$ subscripts denote its vertical and horizontal parts, respectively). In practice, we implemented changes in LEN by adjusting glottis height relative to the hyoid and scaling the epilaryngeal tube and laryngopharynx[2] (Figure 3). HY is an actual

[1] `C++` using `wxWidgets` (`https://www.wxwidgets.org/`), which we re-engineered as a dynamically loaded library (DLL) compiled with `Microsoft Visual C++`® 11 x64 on Microsoft Windows® 7 64 bits.

[2] Please note that while quite realistic, VocalTractLab 2.1 is not equivalent to the actual human anatomy and our manipulations are constrained by its limits. Thus, while the epilarngeal tube is technically not present in the model, we are changing it indirectly by modifying other parameters: the glottis can be moved up and down, scaling the entire larynx and pharynx

Table 1: The parameters of the vocal tract model. LEN (a parameter determining the vertical position of the larynx – larynx height) is the initial value given for a given condition (but it varies between conditions). The next 11 parameters are under the agent's direct control (for these, specifying a default value is not informative as they are changed during learning). The next 7 (there are 4 TS parameters, $TS1$ – $TS4$) are fixed (closed velum and no tongue side elevation; wall compliance currently has no effect and is fixed to the default value of 0.0). The last 2 parameters (controlling tongue root) are automatically computed from tongue body (TCX, TCY) and hyoid (HX, HY) parameters at run time by VocalTractLab 2.1 (thus, specifying a default value is also uninformative). The range of HY depends on LEN (see Table 2). For more details on the parameters please consult Birkholz (2013a), especially Table 2 and Figure 7 therein.

Parameter	Name	Range	Unit	Default
Glottis vertical position	LEN	[-9.45, -6.45]	cm	-7.95
Hyoid x	HX	[0.0, 1.0]	relative	–
Hyoid y	HY	depends on LEN	cm	–
Jaw angle	JA	[-7.0, 0.0]	degrees	–
Lip protrusion	LP	[-1.0, 1.0]	relative	–
Lip distance	LD	[-2.0, 4.0]	cm	–
Tongue body x	TCX	[-3.0, 4.0]	cm	–
Tongue body y	TCY	[-3.0, 1.0]	cm	–
Tongue tip x	TTX	[1.5, 5.5]	cm	–
Tongue tip y	TTY	[-3.0, 2.5]	cm	–
Tongue blade x	TBX	[-3.0, 4.0]	cm	–
Tongue blade y	TBY	[-3.0, 5.0]	cm	–
Velum shape	VS	fixed	relative	0.5
Velic opening	VO	fixed	relative	-0.1
Wall compliance	WC	fixed	–	0.0
Tongue side elevation	$TS1$ – $TS4$	fixed	cm	0.0
Tongue root x	TRX	auto	cm	–
Tongue root y	TRY	auto	cm	–

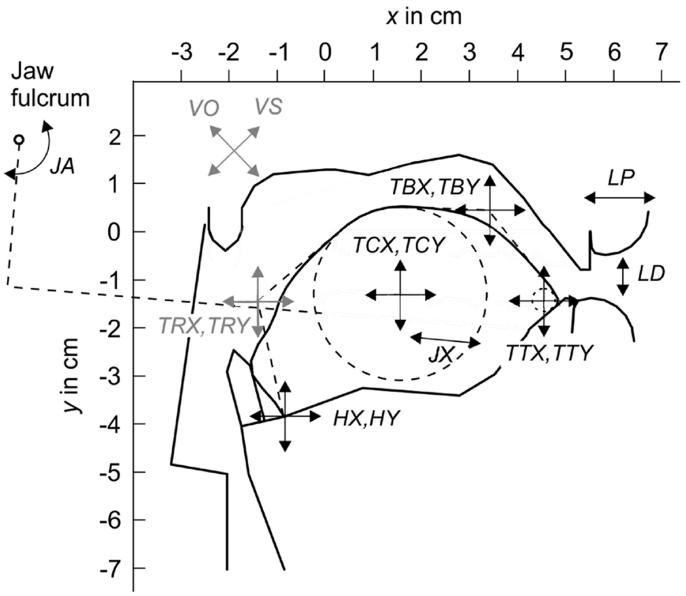

Figure 1: The geometric transformations of the vocal tract model due to articulatory parameter adjustments. Some parameters (shown in gray) are fixed (velum shape, VS, and opening, VO) or automatically adjusted by VocalTractLab 2.1's internal logic (tongue root position, TRX and TRY). The tongue side elevation parameters ($TS1$–$TS4$) are not shown. The figure is modified with permission from Birkholz (2013a).

VocalTractLab 2.1 articulatory parameter that can be dynamically adjusted for a given LEN value, and which moves the entire larynx vertically.

For a given LEN, we compute the vertical position of the glottis, g, as

$$g = HY + g_{def} - (LEN_{max} - LEN) + \frac{LEN_{max} - LEN_{min}}{2} \qquad (1)$$

where g_{def} is the default VocalTractLab 2.1 glottis vertical position (fixed at $g_{def} = -3.2$), and HY the vertical position of the hyoid (see below), with all the elements between the glottis and the bottom of the hyoid being linearly interpolated. We vary LEN within ± 1.5 cm of the default LEN value of -7.95 and, due to computational constraints, we considered seven conditions: the VocalTractLab 2.1 default value of $LEN = -7.95$ plus the six discrete equidistant values $\{-9.45, -8.85, -8.25, -7.65, -7.05$ and $-6.45\}$ between what are the lowest and highest positions currently possible with VocalTractLab 2.1.

Because human speakers can dynamically shorten and elongate the vertical portion of their supralaryngeal vocal tract SVT_v by moving the hyoid up and down, we modified VocalTractLab 2.1's default hyoid range of vertical movement by constraining HY depending on LEN and restricting the way the hyoid can change SVT_v length to more accurately reflect the anatomical and physiological reality in humans and other primates. More precisely, with a short SVT_v, the hyoid is not only positioned more cranially, but also has a shorter range of motion than with a longer SVT_v (Nishimura et al., 2006). For the two extreme configurations with a very *short* and a very *long* SVT_v, we constrained the range of vertical hyoid movement HY as follows. For the short extreme, HY varies within ± 0.5 cm centered around 3.75 cm below the uvula (thus, between $short_{min} = -6.0$ cm and $short_{max} = -5.0$ cm), while for the long extreme, HY varies within ± 1.0 cm centered around 5.5 cm below the uvula (thus, between $long_{min} = -4.0$ cm and $long_{max} = -3.5$ cm) respectively. For a given configuration, we linearly interpolate the lower and upper bounds of HY

together, and the hyoid is able to move within a certain range within the larynx (depending on glottis height), changing the ratio of the parts of the tube below and above the hyoid.

Table 2: The HY extreme values while articulating the mid central vowel [ə] for each of the considered LEN conditions. LEN is a proxy for larynx height that we control directly and that varies between conditions but is fixed within, while HY is a parameter of the VocalTract-Lab 2.1 that defines the actual vertical position of the hyoid and can be dynamically adjusted by the model during learning within constraints imposed by LEN.

LEN	Min HY	Max HY
−9.45 (lowest)	−6.00	−5.00
−8.85	−5.60	−4.70
−8.25	−5.20	−4.40
−7.95 (default)	−5.00	−4.25
−7.65	−4.80	−4.10
−7.05	−4.40	−3.80
−6.45 (highest)	−4.00	−3.50

as

$$HY_{min} = (long_{min} - short_{min})\frac{LEN - max(LEN)}{max(LEN) - min(LEN)} + long_{min} \quad (2)$$

and

$$HY_{max} = (long_{max} - short_{max})\frac{LEN - max(LEN)}{max(LEN) - min(LEN)} + long_{max} \quad (3)$$

resulting, for the seven LEN values considered here, in the ranges in Table 2.

There are two main conventions in the literature for defining the ratio (R) between the vertical (SVT_v) and the horizontal (SVT_h) parts of the vocal tract. In one, R_{vh} is defined as SVT_v/SVT_h, with $R_{vh} = 1.0$ representing a modern human vocal tract, $R_{vh} \gg 1.0$ a very low larynx, and $R_{vh} \ll 1.0$ a very high larynx (as in modern human babies, non-human primates and some reconstructions of archaic humans such as the Neanderthals; Lieberman & Crelin, 1971). However, in the alternative convention, which we will be using here, R_{ht} is defined as $SVT_h/(SVT_v + SVT_h)$ (i.e., the ratio between the horizontal and total vocal tract length), with $R_{ht} = 0.5$ representing a modern human vocal tract, $R_{ht} \ll 0.5$ a very low larynx, and $R_{ht} \gg 0.5$ a very high larynx;

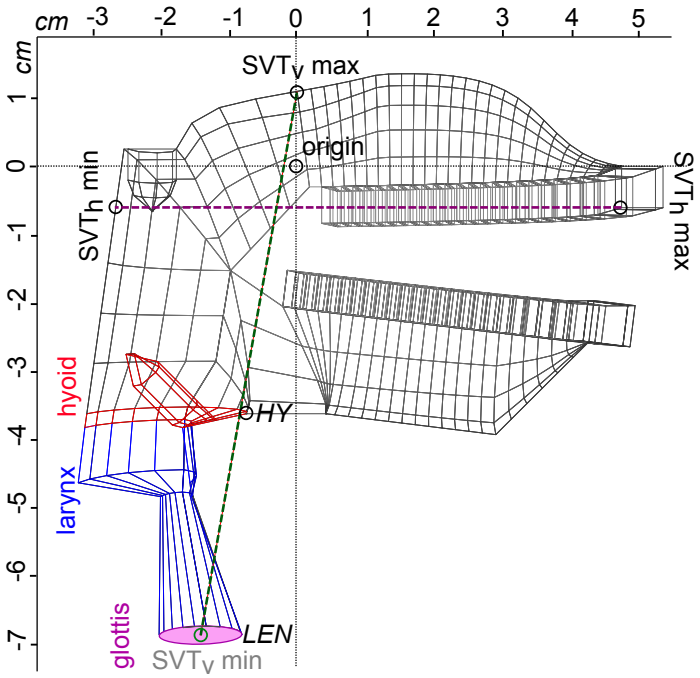

Figure 2: Wireframe lateral view of the 3D supralaryngeal vocal tract (SVT) model (thin solid lines) in the default VocalTractLab 2.1 configuration, showing the *larynx* (blue structure), the *hyoid* (and upper body of the *epiglottis*, red structure), the *glottis* (solid magenta ellipse), and the *vertical* (SVT_v) and *horizontal* (SVT_h) parts of the supralaryngeal vocal tract. The length of the horizontal part, SVT_h, (horizontal dotted magenta line) is measured linearly between the lingual incisial edge of the upper central incisors (SVT_h max, fixed at coordinates (4.7, -0.6)) and the intersection between the posterior pharyngeal wall and the horizontal line emerging from SVT_h max (SVT_h min, fixed at coordinates (-2.6, -0.6)). The length of the vertical part, SVT_v, (the oblique dotted dark green line) is measured linearly between the posterior nasal spine (SVT_v max, fixed at coordinates (0, 1.09)) and the transverse centroid of glottis (SVT_v min; gray label) which varies between conditions. The origin (0, 0) is at the intersection of the vertical through SVT_v max and the horizontal through SVT_h max (dotted black lines), and the scale on both axes is in centimeters (cm). The larynx can be moved up and down by adjusting the vertical position of the hyoid (the HY parameter); the length of the larynx itself can also be adjusted. The larynx height parameter LEN is the vertical position of the glottis relative to SVT_v max, $LEN = 1.09 - SVT_v$ min; in this image, LEN is at its default value of -7.95 cm. Please note that the tongue and lips are not shown.

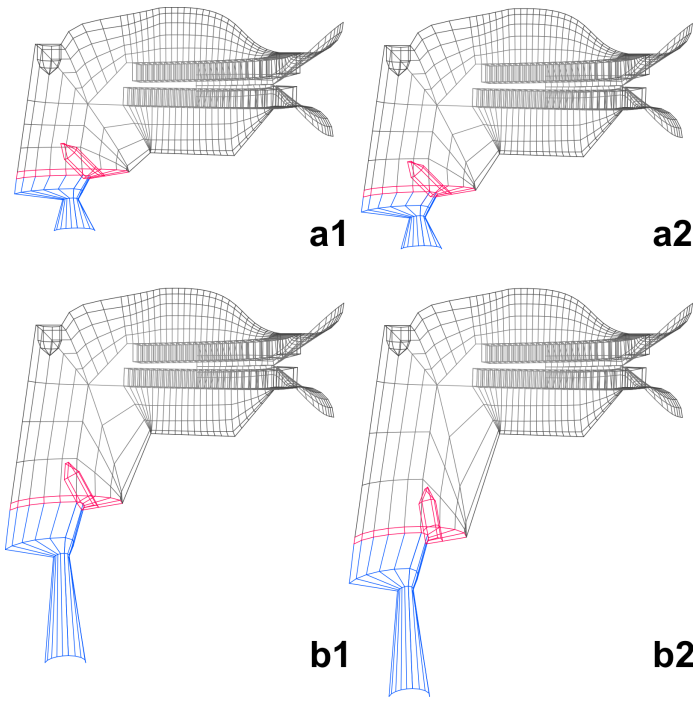

Figure 3: Two extreme larynx height (LEN) conditions illustrating the scaling of the vertical part of the vocal tract and the range of the hyoid vertical position (HY). The larynx (and laryngopharynx, in blue) can be moved vertically by adjusting the hyoid (and the upper body of the epiglottis, in red); its length can also be adjusted. Please compare with Figure 2. Top: $LEN = -6.45$, with $HY = -3.5$ (panel a1, left) and $HY = -4.0$ (panel a2, right). Bottom: $LEN = -9.45$, with $HY = -5.0$ (panel b1, left) and $HY = -6.0$ (panel b2, right). We show the VocalTractLab 2.1 configuration for producing [ə].

the value corresponding to the "standard" VocalTractLab 2.1 $LEN = -7.95$ is $R_{hl} \approx 0.44$.

2.2. Learning to articulate target vowels

For any of the seven conditions ($LEN \in \{$ -9.45, -8.85, -8.25, -7.95, -7.65, -7.05, -6.45 $\}$), there are 11 parameters (HX, HY, JA, LP, LD, TCX, TCY, TTX, TTY, TBX and TBY) that can be directly controlled by setting them to 11 real number values, prompting our modified vocal tract model to produce the corresponding vowel sound, or to fail to produce any sound at all if the configuration is impossible or results in a completely closed vocal tract. Our goal here is to implement a *learning mechanism* that, given a *target* vowel sound, is able to discover, without human intervention, a set of 11 free parameter values that allow the model to produce the same (or a very similar) vowel sound. Essentially, this models the real-world problems encountered by a child acquiring their native language(s) in the sense that, in principle we do not know *a priori* what the actual values of the 11 parameters that produced the target sound are, so that supervised learning methods cannot be applied, having to use instead *reinforcement* techniques.

Formally, we will represent a target vowel sound by its first n Bark-transformed formant frequencies $\vec{b} = \langle b_1, b_2, \ldots, b_n \rangle \in \mathbb{R}^n$. While we will focus here on the first three formants (i.e., $n = 3$), F_1, F_2 and F_3, we also considered (detailed in the Supplementary Materials in **Appendix**) $n = 5$ (i.e., also including F_4 and F_5) as they may be relevant for phenomena involving the larynx (Sundberg & Nordström, 1976); for example, in singing voice, very high spectrum peaks are often found around 3 kHz, F_4 and F_5 are largely dependent on the area function of the larynx (Sundberg, 1995), and Zhou et al. (2008) show that F_4 and F_5 differ between the canonical variants of North American English /r/.

Given the target formants $\vec{b}$, we must find the best set of $m = 11$ articulatory parameter values $\vec{p} = \langle p_1, p_2, \ldots, p_m \rangle \in \mathbb{R}^m$ (i.e, the *solution*) that allow the vocal tract model to produce the closest acoustic *reproduction* $\vec{b'} = \langle b'_1, b'_2, \ldots, b'_n \rangle \in \mathbb{R}^n$ of $\vec{b}$. Here we use the *Euclidean distance* between $\vec{b}$ and

$\vec{b'}$, $d(\vec{b}, \vec{b'}) = \sqrt{\sum_{i=1}^{n} (b_i - b'_i)^2}$, as the measure of closeness, with smaller values representing better approximations and $d(\vec{b}, \vec{b'}) = 0$ if and only if $\vec{b}$ and $\vec{b'}$ are identical. Because Euclidean distance is very general, simple and does not impose particular assumptions on the structure of the formant space, we have decided to use it in these initial simulations, but other approaches that capture more domain-specific information[3] should be investigated in future simulations (Bladon & Lindblom, 1981; Schwartz et al., 1997). However, by applying the Bark transform to the formant values, we arguably do take into account properties of human perception; moreover, as our results show, such an "unbiased" choice does produce very good outcomes of the learning process.

With these, we define an *agent* as a self-contained computational entity endowed with a "perceptual system" that extracts the first n Bark-transformed formants $\vec{b}$ from a "heard" vowel sound, a "cognitive system" that learns to immitate such sounds by mapping $\vec{b}$ to a set of m articulatory parameter values $\vec{p}$, and a "production system" that maps $\vec{p}$ to an actual sound (or nothing, due, for example, to the complete obstruction of the vocal tract). Here, the "production system" is represented by the vocal tract model discussed above, the "perceptual system" is based on VocalTractLab 2.1's extraction of formants from the produced acoustic output, and the "cognitive system" is implemented as a neutral network trained by a genetic algorithm.

2.2.1. The neural network

We use a fully connected feed-forward neural network consisting of three layers: one input layer, one hidden layer, and one output layer. The input layer has $n + 1$ input neurons, with the first n receiving as input the first n Bark-transformed formants of the "perceived" sound b_i scaled as $10\frac{b_i - b_i^{min}}{b_i^{max} - b_i^{min}} - 5$ (where (b_i^{min}, b_i^{max}) are, for $i \in 1..5$, {(2, 7), (4, 15), (14, 16), (15.5, 17.5), (16.5,

[3] Such as the *cepstral distance* (Tohkura, 1987), the *dispersion-focalization distance* (Schwartz et al., 1997) or the $F_1 \times F_2$ *weighted Euclidean distance* in De Boer (2000).

19)} bark[4]); the $n+1$ neuron is a bias neuron allowing the network to cope with saturated gradient input. The hidden layer consists of $h+1$ neurons, with $h = Round(\frac{n+m}{2})$, where neuron $h+1$ is a bias neuron, and each of the first h neurons receiving activation from all the $n+1$ input neurons. The output layer consists of m neurons, each receiving activation from all $h+1$ neurons in the hidden layer and controlling, in turn, one of the free articulatory parameters of the vocal tract model by feeding their output value x (normalized to the parameter's range $[p_{min}, p_{max}]$ using $(p_{max} - p_{min})x + p_{min}$). The activation of the hidden and output neurons is computed using the commonly used sigmoid function $\sigma(x) = \frac{1}{(1+e^{-x})}$ applied to the sum of all the inputs to a given neuron weighted by the strength of the connections (the "synapses") through which these inputs flow. Formally, for a neuron with k inputs u_i and synaptic weights w_i, $1 \leq i \leq k$, the activation is $a = \sigma(\sum_{i=1}^{k} w_i u_i)$.

2.2.2. Training the neural network

The neural network's architecture, activation function and synaptic weights can be conceptualized as implementing a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ mapping the n Bark-transformed first formants of the "perceived" sound $\vec{b} = \langle b_1, b_2, \ldots, b_n \rangle$ to the m free parameters of vocal tract $\vec{p} = \langle p_1, p_2, \ldots, p_m \rangle$. We are interested in finding such a function f that, given a sound described by $\vec{b}$, produces a vector of parameters values $\vec{p} = f(\vec{b})$ such that the sound produced by the vocal tract model using these parameter values is as close as possible to the original sound. If we denote the n first Bark-transformed formants of this produced sound as $\vec{b'} = \langle b'_1, b'_2, \ldots, b'_n \rangle$, then the distance $d(\vec{b}, \vec{b'})$ is the minimum across all possible produced sounds $\vec{b'}$.

However, given that the mapping from the m free parameter values $\vec{p}$ and the first n Bark-transformed formants of the resulting sound $\vec{b'}$ is extremely complex

[4]This scaling ensures that the range of each of the n formants considered maps produces an activation between approximately 0 and 1; for details please see Sections 4.3.4.2 and 4.5.3 and in Janssen, 2018.

and completely opaque to the learning algorithm (just as the inner workings of the body are opaque to the child's brain), the agent must somehow discover a function $\bar{f}$ that approximates as well as possible the best function f which truly minimizes the distance $d(\vec{b}, \vec{b'})$ (just as the child's brain must find ways of controlling its body, including the vocal tract). More precisely, as we have an acoustic target but we must discover an "intermediate" articulatory solution that maps in very complex way to acoustics, we cannot perform supervised learning (such as standard back-propagation) and must rely instead on some form of reinforcement learning. While there are very interesting approaches to this problem, including recent developments in *curiosity-driven learning* (Moulin-Frier et al., 2014; Oudeyer & Smith, 2016), we decided to implement here an as-domain-general and standard method as possible, at the cost of reduced biological plausibility and computational inefficiency. Previous work (Prom-on et al., 2014) used VocalTractLab 2.1 and stochastic gradient descent to learn Thai vowels, and we are planning comparisons between various learning algorithms (including our current implementation described here, gradient descent and curiosity-driven search) in terms of their computational costs and accuracy.

The approach we opted for here was to use an off-the-shelf genetic algorithm[5] where the "genome" encodes the neural network's synaptic weights, one weight per floating-point "gene". The population size is fixed to 100 genomes.

The first generation is initialized by randomly generating each of the 100

[5] *Genetic algorithms* (see, for example, Eiben & Smith, 2003 or `https://en.wikipedia.org/wiki/Genetic_algorithm`) are inspired from biological evolution in the sense that they evaluate whole "populations" (sets) of solutions, and these solutions are allowed to "reproduce" in relation to their "fitness" (i.e., how "good" they relatively are at solving the problem at hand). A solution is represented by the particular values of the "genes" in a "genome", which really are the values of a set of parameters used to estimate the value of a function (the "fitness" function). New "genomes" are produced by "mutation" (e.g., randomly changing the value of a "gene" in a "genome") and "cross-over" (producing a combination of two parental "genomes"). Just as in biological evolution, after several generations the algorithm explores the (usually very complex and multi-dimensional) "fitness space" and finds one or more (usually local, but sometimes global) optima.

"genomes": more precisely, we draw random values for all the "genes" in a "genome", then use the corresponding neural network to control the vocal tract model in order to produce acoustic output and, if there is no acoustic output we restart by drawing a new set of random "genes", until all "genomes" in the population are able to produce some sort of sound.

The fitness function (i.e., the function that is optimized by the genetic algorithm) of a given "genome" is the distance $d(\vec{b}, \vec{b'})$ between the target sound $\vec{b}$ and the sound produced using the "genomes"'s "genes" as neural network weights, $\vec{b'}$, $d(\vec{b}, \vec{b'})$; the fitness of a "genome" that does not result in a sound is set to $+\infty$ (the worst possible fitness given that we perform fitness *minimization*). In each generation, the potential parents are selected using stochastic universal sampling[6] (Baker, 1987) with elitism[7] (Eiben & Smith, 2003), and these selected parents produce the next generation of 100 offspring genomes.

Mutation (i.e., the creation of new "genomes" with new values of the parameters) is handled using evolution strategies (Beyer & Schwefel, 2002) which first evolve a set of strategy parameters σ that control the step size of the mutation operator as applied to continuous "genes" (i.e., the standard deviation of a Gaussian distribution used to generate mutated values). In this approach, the rate with which the "genes" mutate can itself evolve as well, increasing the ability to escape local optima.

We ran the genetic algorithm for a maximum of 500 "generations", but with the possibility of an early stopping if the fitness of the best "genome" in the

[6] Stochastic Universal Sampling (or SUS) is an alternative to Fitness Proportionate Selection (FPS or "roulette wheel selection") that ensures no bias and less spread, allowing genomes with worse fitness a fairer chance to reproduce. It works by mapping genomes to a line such that each genome's segment is equal to its fitness (as in FPS), but then places on this line n (= next generation's population size) equally spaced points starting at a random location (the genomes where the points fall are selected for reproduction).

[7] A technique ensuring that the best genomes in a generation are not lost to random fluctuations in the selection process by copying a predefined number of the fittest genomes into the next generation.

population seems to stabilize, apparently reaching an optimum value (this early stopping was implemented due to the relatively high computational costs of the genetic algorithm). We decided to stop when the improvement $\Delta t_i = t_{i+w} - t_i$ becomes 0.0, where t_i is the best fitness value of in generation i and w is the sliding window size (set here to $w = 100$ generations).

2.3. Putting everything together: the agent learns to speak

Thus, we implemented a computational agent which, when exposed to a given target sound extracts its first n Bark-transformed formants $\vec{b}$, maps them to the m free parameters $\vec{p}$ using a neural network, and feeds these m values to the vocal tract model to produce an acoustic output characterized again by n Bark-transformed formants $\vec{b'}$. The agent autonomously learns to control its vocal tract and to produce an output sound $\vec{b'}$ (that matches as well as possible the target sound $\vec{b}$) using a generic learning mechanism, in this implementation a genetic algorithm with evolution strategies (due to our modular design, this is easily replaceable by other optimization techniques). Figure 4 gives a visual overview of the agent.

However, just like human children, our agent should be able to learn a whole set of target sounds, $T = \{\vec{b}_1, \vec{b}_1, \ldots, \vec{b}_k\}$, with $k > 1$. The obvious way to implement this would be to feed the whole set of targets T into the same neural network, searching for a single mapping function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that optimizes simultaneously the production of all k targets, i.e., that minimizes simultaneously all distances $d(\vec{b}_i, \vec{b'}_i)$, $i \in \{1, 2, \ldots, k\}$. Unfortunately, this simultaneous training results in interference between the different targets, suboptimal performance and convergence issues, due to the lack of a mechanism for keeping the internal representations of the k targets (and their productions) separate within the neural network.

Therefore, we decided here to implement a separate neural network for each target sound $\vec{b}_i$, individually trained to find the best mapping $f_i : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that optimizes the production of this particular target, i.e., that minimizes the distance $d(\vec{b}_i, \vec{b'}_i)$. In this way, we allow each target sound to be learned indepen-

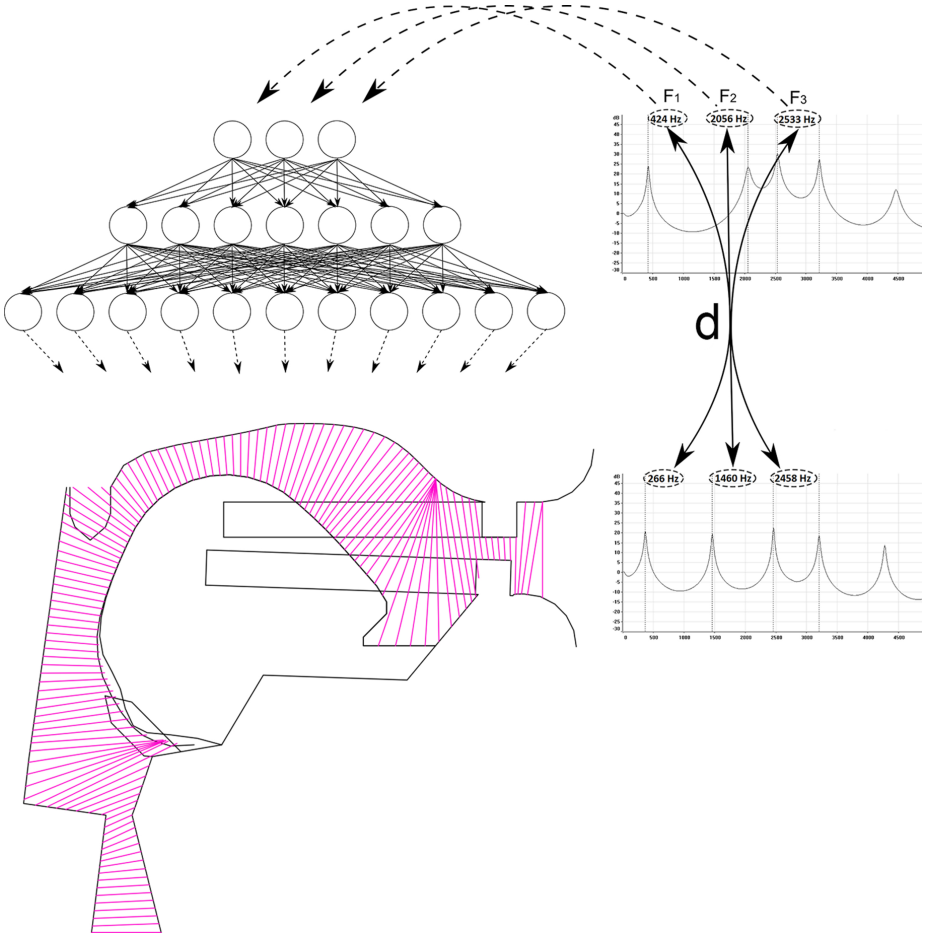

Figure 4: Overview of the agent model. A target vowel (top right) represented here only by its first three formants $F_1 - F_3$, is fed into a input layer of the neural network (top left) where each input neuron is mapped onto a given formant. The information is then propagated through the network's hidden layer to the output layer (as modulated by the network's synaptic weights), where each output neuron maps uniquely onto a single articulatory parameter of the vocal tract model (bottom left). The area function of the model (magenta lines perpendicular to the centerline of the vocal tract airway) is then computed depending on the actual positions of the articulators as described by the current parameter values, possibly producing, in turn, an acoustic output (bottom right). The agent then estimates the error (d) between target and produced sounds, which represents the inverse of the fitness to be maximized by the genetic algorithm (not shown). This cycle is repeated multiple times, until a plateau of small errors is reached. Please note that while, for clarity, we represented here only the three first formants, this model is generalizable to any number of formants n.

dently of the other targets, while not having to impose artificial mechanisms for keeping them distinct; metaphorically, we provide a way for the agent's brain to "label" these internal representations as distinct (e.g., as different phonemes), without explicitly modeling the acquisition of this labeling mechanism. Given our specific research question here, focusing on the influence of variation in vocal tract anatomy on the articulation of vowels while trying to minimize the effects of other levels, we would argue that these design decisions are appropriate. Nevertheless, our use of generic (even if not biologically plausible) learning techniques and our keeping of the vowels separate, can be easily changed, and the effects of these changes can be tested[8].

2.4. The target vowels

We predefined eight target vowels, namely [ə], [ɑ], [a], [æ], [e], [i], [o] and [u]. These vowels were recorded three times by SRM, each vowel being produced for approximately 2 seconds. The vowel spectra were obtained from 1 second windows positioned during the most stable portion of each vowel. Vowel formants were calculated using PRAAT (Boersma & Weenink, 2018) over comparable ∼1 second windows and we computed their averages (some manual measurement of F_3 and especially F_4 was necessary as well). The vowels were then recreated using the default configuration of VocalTractLab 2.1 by manually adjusting its free parameters. The recreated vowels thus represent reasonable approximations based (primarily) on acoustic criteria. Table 3 gives the formant values and the VocalTractLab 2.1 (in its default configuration) parameter settings for each vowel.

Please note that while usually [i], [a], and [u] are considered to be the "extreme" vowels (Maddieson & Disner, 1984), we also included [ə] as a neutral

[8] As a note, because we learn here individual vowels using separate neural networks, it could be possible to disregard entirely the neural network and use instead the Genetic Algorithm to discover directly the 11 articulatory parameters themselves. While this would greatly increase the efficiency of the search, it would probably not alter the results, but we decided to design our code as "future proof" as possible by allowing the simultaneous learning of sets of vowels.

(or control) vowel[9]. We decided to also include [ɑ] and [æ] besides [a] primarily because we aim at an increased coverage and resolution of the low vowel space, but also due to the ambiguity of the [a] notation, probably usually meant as a rather more broad interpretation of the "low" vowel (see, for example, Honda, 1996; Ladefoged & Johnson, 2010; Barry & Trouvain, 2008). We are focusing on these eight vowels because they are widely distributed cross-linguistically (Moran et al., 2014), represent different combinations of tongue height and fronting, are relatively extreme in terms of the modern human vowel space (see also Figure 5), and the lively debate concerning the evolution of speech and language makes reference to at least some of them (e.g., Lieberman & Crelin, 1971; Boë et al., 2002, 2013).

2.5. Implementation and availability

In total, we ran 7 conditions (larynx height LEN values) $\times$ 8 target vowels ([ə], [ɑ], [a], [æ], [e], [i], [o] and [u]) $\times$ 2 sets of formants (first three, $n = 3$ versus first five, $n = 5$) $\times$ 100 independent replications each = 11200 runs. One replication was run as a single thread occupying one (real or virtual) core; the whole simulation took about 1.5 months wall-clock time on a dedicated server with two Intel Xeon E5 2620 CPUs (2.0 GHz, up to 2.5 GHz Turbo Boost, 6 physical cores with HyperThreading) and 64Gb RAM running Windows 7 64 bits.

The actual implementation (a) uses a version of the `Vocal Tract Lab` 2.1 Birkholz (2013a) modified by us to allow the specification of the vertical position of the larynx, and refactored as a Windows 64 bit Dynamically Linked Library (DLL) encapsulating the relationship between a set of articulatory parameter values and the output acoustics in terms of formant values (`C++` compiled with `Microsoft Visual C++ 11` 64 bits on `Windows 7`); (b) the agent

[9]We consider [ə] as a "control" in the sense that we wanted to have a "reference point" relatively unaffected by the anatomical manipulations we study here to which we can compare the other vowels.

Table 3: The target vowels. The first five Bark-transformed formants (F_1–F_5), followed by values of the free articulatory parameters in the standard VocalTractLab 2.1 configuration that produce these formants, and the corresponding SVT_v and $R_{ht} = SVT_h/(SVT_h + SVT_v)$; please note that $SVT_h = 7.32$ throughout. The formant values used here come from formant measurements taken in PRAAT (Boersma & Weenink, 2018) (using the automated method but with each formant value checked manually using spectral slices) of the careful phonetic productions by author SRM; we checked that VocalTractLab 2.1 can reproduce these accurately.

Parameter	[ə]	[ɑ]	[a]	[æ]	[e]	[i]	[o]	[u]
F_1	5.13	6.59	6.94	6.69	4.29	2.29	3.99	2.72
F_2	9.50	8.34	10.31	12.01	13.12	14.05	6.10	5.07
F_3	14.55	15.11	14.45	14.72	14.76	15.63	15.15	15.14
F_4	16.08	15.91	15.95	16.33	16.28	16.52	15.70	15.79
F_5	18.0	18.03	18.01	18.20	18.05	17.88	17.73	16.96
HX	0.99	0.13	0.52	0.34	0.90	1.00	0.48	1.00
HY	-4.14	-4.75	-4.31	-3.55	-3.79	-4.68	-4.80	-5.70
JA	-1.93	-6.73	-6.02	-7.00	-2.89	-0.98	-3.44	-4.57
LP	0.24	0.22	-0.56	0.46	-0.55	-0.55	0.88	1.00
LD	1.66	2.48	1.12	2.54	0.50	0.28	0.30	0.14
TCX	0.24	-0.77	0.08	1.06	1.50	2.20	-0.76	-0.50
TCY	-1.33	-2.53	-2.44	-2.06	-0.88	-0.71	-1.67	-1.32
TTX	2.61	2.17	3.22	3.82	3.38	4.42	1.90	1.87
TTY	-1.70	-2.16	-3.00	-1.75	-1.11	-1.17	-1.48	-0.41
TBX	2.00	1.01	1.41	2.83	3.43	3.77	0.62	1.40
TBY	-0.48	-1.28	0.27	-0.83	0.05	0.27	0.31	0.46
$SVTv$	8.56	9.17	8.73	8.00	8.20	9.10	9.22	10.14
R_{ht}	0.46	0.44	0.46	0.46	0.47	0.45	0.44	0.42

was developed in `Java` in `Eclipse Mars` using the `Encog 3.2` library (Heaton, 2015) for the neural network and the `Watchmaker Framework` version 0.7.1 (`https://watchmaker.uncommons.org/`) for the genetic algorithm, compiled into an executable `JAR` file with the `Java Development Kit 1.7`; and (c) the various conditions and replications were controlled from a custom `Python` (version 2.7.6) script.

All our source code is freely available under an open source license (GPL) in a dedicated GitHub repository (`https://github.com/ddediu/larynx-height-vowels`), which also contains pre-compiled binaries for Microsoft `Windows` with installation instructions, the configuration files needed to reproduce the results reported here, and the `R` and `Rmarkdown` scripts used to analyze and plot them.

3. Results

The analyses and plots reported here used `R 3.4.4` (R Core Team, 2017). The full analysis (including aspects and details, including considering $n = 5$ formants, not reported here due to space constraints) can be found in the Supplementary Materials in **Appendix**. The patterns obtained considering $n = 3$ and $n = 5$ formants are roughly similar, so that we will be focusing here on the first.

We will describe first the tight relationship between the dynamically-adjusted continuous vocal tract ratio R_{ht} (defined as the ratio between the horizontal and the total supralaryngeal vocal tract lengths: $R_{ht} = SVT_h/(SVT_v + SVT_h)$) and the predefined discrete values of larynx height LEN, turning then to the acoustic properties of the produced vowels function of larynx height (i.e., how similar the individual vowels and the whole vowel system is to the predefined targets), ending with the compensation of non-optimal larynx heights by other articulators.

3.1. Vocal tract ratio versus larynx height condition

While the 7 larynx height position conditions (the values of LEN) are predefined and fixed for any given run, our model can adjust the various components

Table 4: Summary of regressing the errors per formant (i.e., $(F_{produced} - F_{target})^2$) on vowel, R_{ht} (the ratio between the vertical and the total lengths of the supralryngeal vocal tract) and its squared value R_{ht}^2, and their interactions. We show here the β's for R_{ht} and R_{ht}^2 only, and adjusted R^2; all p-values $\leq 2 \cdot 10^{-16}$. Full details are in the Supplementary Materials in **Appendix**.

Predictor	F_1	F_2	F_3
R_{ht}	-5.7	-18.6	-95.3
R_{ht}^2	6.1	19.5	99.7
adjusted R^2	16.6%	71.4%	78.2%

of the vocal tract (as discussed later), including the hyoid vertical position HY, which changes, in turn, the vertical length of the vocal tract, SVT_h, allowing thus the vocal tract ratio R_{ht} to be dynamically adjusted by the learning mechanism. Thus, while LEN is fixed, R_{ht} is actively changed by the agent (within constraints), in a manner that is well approximated by a linear relationship: the linear regression of R_{ht} on LEN, the vowels, and their interactions results in an adjusted $R^2 = 97.5\%$, $\beta_{LEN} = 0.042$, $p < 2 \cdot 10^{-16}$.

3.2. The acoustics of produced vowels

We compared the acoustics of the target and of the produced vowels as a function of LEN (or R_{ht}), either for each vowel separately or for the whole system composed of all vowels simultaneously.

The **Top panel** of Figure 5 shows the actual formant values produced for each vowel against the corresponding target (the **bottom panels** give a different representation of the same data): the best approximation of the target vowels is obtained around the "standard" LEN, with deviations from this region affecting mostly F_3 (see Table 4 for quadratic regressions; as a note, deviating from the "standard" LEN does not result in a simple collapsing of the system towards [ə]).

For each vowel, we computed the Euclidean distance between the target and the production (i.e., the "acoustic error"; the **mid panel** of Figure 5),

and we found a non-linear relationship with R_{ht} that varies between vowels (quadratic regression on R_{ht}, vowels and their interactions results in an adjusted $R^2 = 82.1\%$). Overall, the error is high for $R_{ht} \approx 0.40$ but rapidly drops towards 0.0 as R_{ht} approaches 0.44 and remains low for most vowels except [i] and [u], where it increases again with increasing R_{ht}. This suggests that reproduction is best around $0.42 \lesssim R_{ht} \lesssim 0.45$ for all vowels, the worst for $R_{ht} \lesssim 0.40$ for all vowels except [u] (where the worst seems to be at $R_{ht} \approx 0.41$), while for $R_{ht} \gtrsim 0.46$ there is either a plateau or a slight worsening.

Ordinary *Procrustes analysis* (Zelditch et al., 2012) is a widely used technique that matches two configurations of corresponding points using translation, rotation and scaling (by minimizing the squared distances, SSD, between these corresponding points), separating thus *shape* from *size*. The $\sqrt{SSD}$ (the Procrustes distance) is thus an estimate of how well the two configurations match as a whole. We considered the system of the target vowels as a configuration and the system of the corresponding productions as another, and we computed the Procrustes distance between them as a measure of how well they match *as systems*. These Procrustes distances show a "U"-shaped pattern (image shown in the Supplementary Materials in **Appendix**; the quadratic regression on R_{ht} produces an adjusted $R^2 = 81.8\%$) of high values for low R_{ht}, reaching a minimum (close to 0.0) in the neighborhood of $R_{ht} = 0.45$, increasing again for larger R_{ht} (but well below the values for low R_{ht}; the same pattern obtains for $n = 5$). This suggests that, as a whole, the produced vowel system is most similar (almost identical) to the target one for $R_{ht} \approx 0.44 - 0.46$, being far worse for lower than for higher R_{ht} values.

For each *pair* of vowels, we computed the Euclidean distance between their productions (see the Supplementary Materials in **Appendix**; thus, how different acoustically the produced vowels are from each other). These confirm that R_{ht} in the neighborhood of 0.45 results in pairwise distances close to the distances between the target vowels, that, in general, lower R_{ht} values have larger effects than higher values, and that the effects of the vowels are more complex than a simple collapse of the system towards [ə].

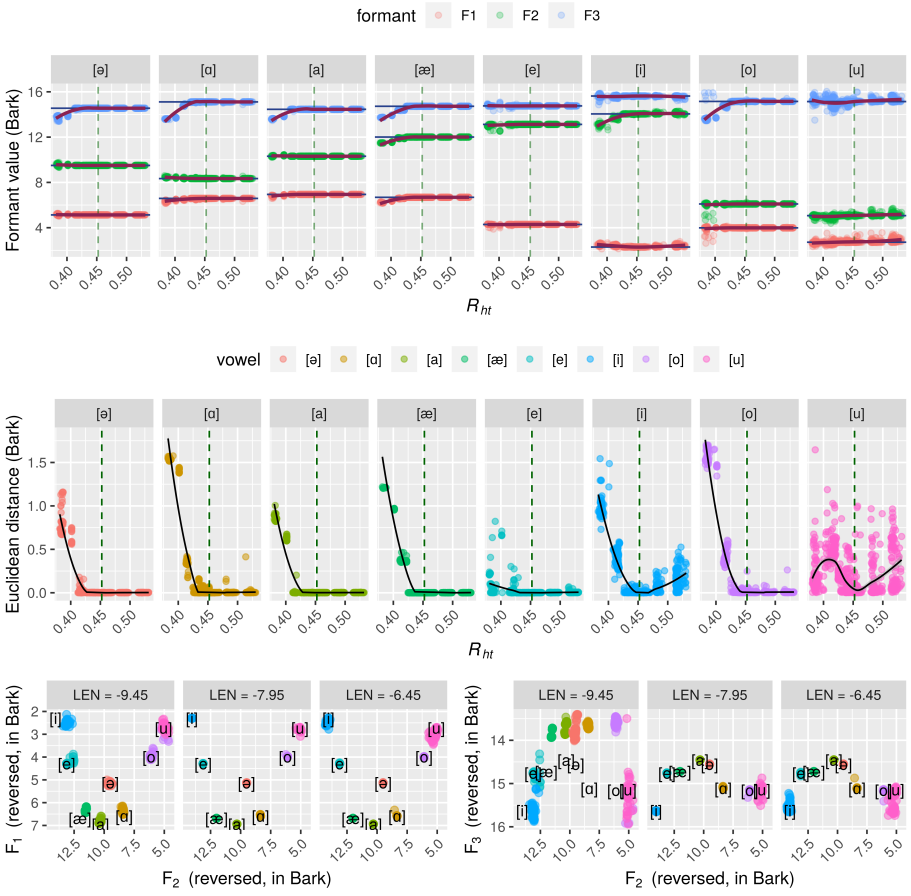

Figure 5: The acoustics of the produced vowels as a function of larynx height LEN and the vertical to total supralaryngeal vocal tract lengths ratio R_{ht}. **Top panel**: the first three formants as a function of R_{ht} (x axis) per vowel with loess regression (black curves); vertical dotted dark green lines represent $R_{ht} = 0.452$ which is the average of target R_{ht} values in the standard configuration; horizontal dark blue solid lines are the target values. **Mid panel**: the Euclidean distances between the produced and target vowels individually function of R_{ht} (x axis) with loess regression (black curves); please note that the vowel legend applies to the bottom panels as well. **Bottom panels**: the produced vowels (colored dots) in the $F_1 \times F_2$ (**left**) and $F_2 \times F_3$ (**right**) spaces for the highest ($LEN = -9.45$), standard ($LEN = -7.95$) and lowest ($LEN = -6.45$) larynx positions (it also shows the target vowels as black IPA symbols with constant positions across conditions). Please see the Supplementary Materials in **Appendix** for the Procrustes distances and for $n = 5$ formants (where the effects of higher LEN and R_{ht} are more important).

The resulting vowels, even those with the highest distance from the intended target for the most extreme values of LEN, are auditorily very similar to the intended targets[10] (the worst and best productions for each condition and target vowel can be found in the Supplementary Materials in **Appendix**).

3.3. The articulatory parameter values

The various articulators in our model respond differently to changes in LEN depending on the target vowel. It can be seen in Figure 6 that:

- for most vowels and articulatory parameters, the parameters' values change function of the ratio R_{ht} (please note that HY's almost linear dependency on R_{ht} is unsurprising given the setup of our model);
- these changes range from the very dramatic (e.g. LP for [e], JA for [e], or TCX for [i]) to the extremely small or the arguably non-existent (e.g., TCX for [ɑ], TBY for [e] and [i], or LD for [u]);
- in some cases there is very little variance between runs for the same R_{ht} value (e.g., TCX for most vowels), but in others this variance is exceedingly large (e.g., LP, TTX and TTY for most vowels);
- the shape of the dependency varies from almost linear (e.g., LP for [a] or TCY for [ɑ]) to monotonic + plateau (e.g., LP for [ə] or TBX for [e]) to more complex shapes that show inflections (e.g., LP for [e], TTY for [ə] and [e] or HX for [o] and [u]);
- while some articulators seem to behave in relatively similar ways across vowels (e.g., LP) other change dramatically (e.g., HX, TCX or TCY).

To understand how the articulators accommodate varying larynx height conditions, we performed Principal Component Analysis (PCA) on LEN, R_{ht}, the

[10] Except for a few, such as [a](in condition F_1–F_5; for LEN=-6.45) which sounds more like [ɐ]; [e](F_1–F_3; -9.45) and [e](F_1–F_5; -9.45) ∼ [ø], and [i](F_1–F_3; -9.45) and [i](F_1–F_5; -9.45) ∼ [y] (rounding); and [u](F_1–F_3; -6.45 – -7.65) and [u](F_1–F_5; -6.45 – -7.65) ∼ [o].

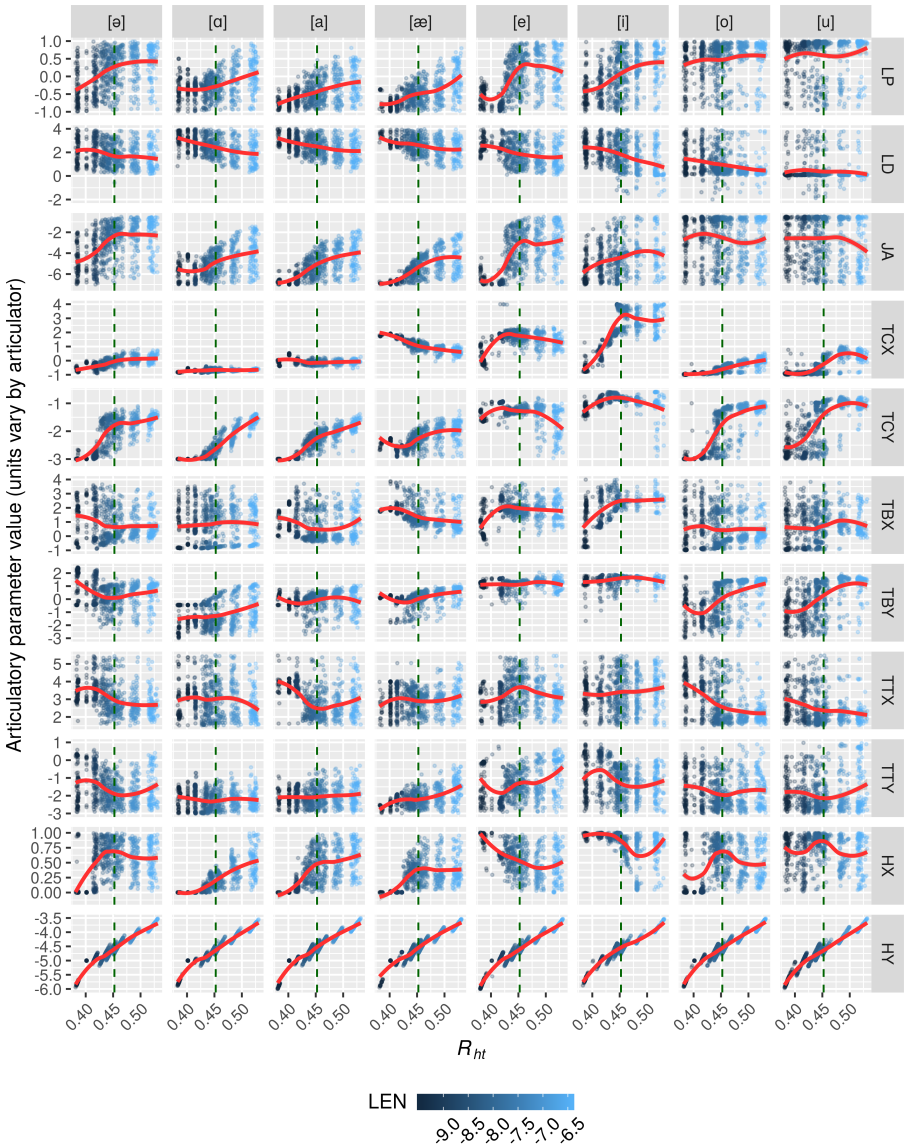

Figure 6: The 11 articulators function of larynx height LEN (dot color), vertical to total supralaryngeal vocal tract lengths ratio R_{ht} (x axis) and vowel, with loess regression (dark red curves) and standard VocalTractLab 2.1 R_{ht} (vertical dashed lines). The 8 vowels are represented by the columns, while the 11 articulators by the rows, each panel representing that distribution across runs of the values of a given articulatory parameter for a given vowel function of the ratio R_{ht}. For example, the top-left panel shows that for [ə], LP (lip protrusion) tends to be negative for very low R_{ht}, positive for larger R_{ht}, and reach a plateau for $R_{ht} >\approx 0.47$, but there is quite a spread of its values for each R_{ht}.

three formants, and the 11 free articulatory parameters (Figure 7). The first three Principal Components (PCs) explain together 61.2% of the variance and capture familiar aspects of vowel articulation such as the inverse relationship between F_1 and the vertical location of tongue constriction (TCY, TBY and less so TTY; thus high F_1 characterizing low vowels, and low F_1 high vowels) captured by PC1 (and less strongly by PC3), the positive relationship between F_2 and the anterior location of the tongue constriction (TCX, TBX and TTX; thus high F_2 characterizing front vowels, and low F_2 back vowels) captured by PC2; while obvious, finding these familiar relationships is a sanity check for our approach.

As expected, LEN, R_{ht} and HY behave in very similar ways, and may be considered for practical purposes as a single "unit" related to larynx height. This combined "larynx height" unit has positive and strong contributions (around $10 - 15\%$) to both PC1 and PC3 (thus, acoustically, on F_1 and F_3), and a negligible negative one (below 5%) to PC2 (see also Moisik, 2013, p. 300).

The jaw (JA) and lips (LP and LD) seem to also form a unit and contribute to F_2 (PC2; negative for jaw angle and lip protrusion and positive for lip distance) and less so to F_1 and F_3 (PC1).

However, this PCA does not differentiate between the target vowels except through their particular combinations of formant values, probably explaining the structure of PC1 and PC3 with respect to F_1 (the first capturing a negative effect of "larynx height" while the second a positive one).

Another limitation of such approaches is that it cannot capture the *causal asymmetry* between variables, here the fact that LEN is manipulated by the experimenter and that the other articulators react to this manipulation, "attempting" to compensate for its effects on the acoustics of the produced vowels. We will use here an approach that draws on advances in the study of causality using Directed Acyclic Graphs (DAGs) and the inference of such DAGs from observed data (Pearl, 2000; Pearl & Mackenzie, 2018). The fundamental idea is to model the relationships between measured variables as connections (or edges) between the nodes representing the variables in a directed network (or graph),

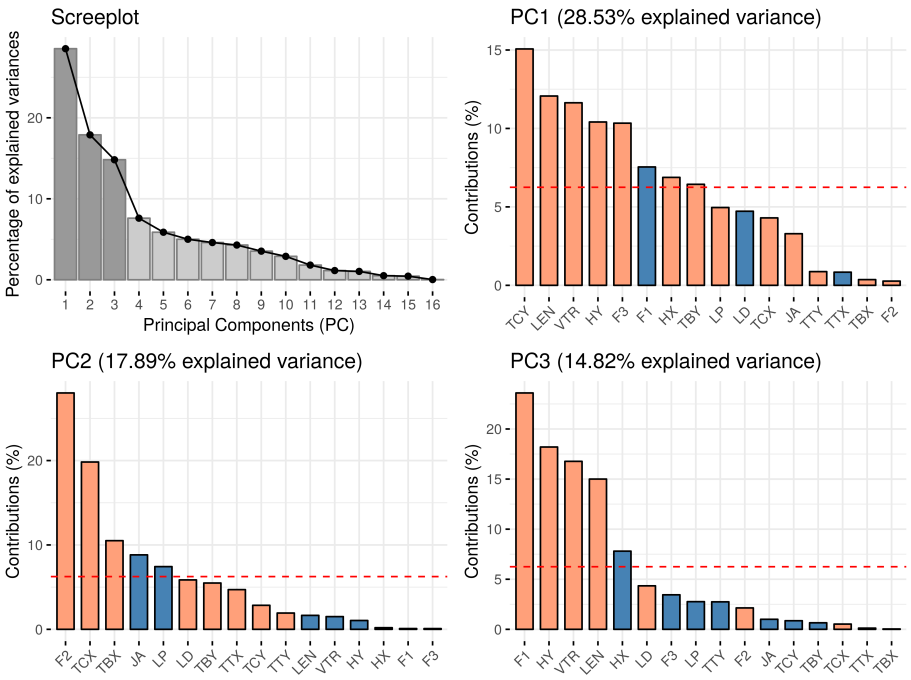

Figure 7: Combined Principal Components Analysis (PCA) of the larynx height condition (LEN), the vertical to total supralaryngeal vocal tract lengths ratio (R_{ht} denoted as VTR in the plots), the acoustics (the 3 formants, $F_1 - F_3$) and the articulator parameter values (HX, HY, JA, LP, LD, TCX, TCY, TTX, TTY, TBX and TBY) for the produced vowels across all eight vowels. Top-left planel: the scree plot showing the % explained variance by PC and highlighting the first 3 PCs. The following three panels show the contributions of each variable to the first three PCs; light red = positive contribution, light blue = negative contribution (please note that the signs themselves are arbitrary but the differences in signs are meaningful); horizontal dashed line = reference line of expected contribution if all contributions were equal; variables are sorted by absolute value of contribution.

such that a directed edge between two variables, V1 → V2, represents the dependency of V2 on V1; with certain assumptions, these can be conceptualized as causal links, with V1 causing V2 (Pearl, 2000; Pearl & Mackenzie, 2018). Importantly, such directed graphs cannot contain cycles (such as V1 → V2 → V3 → V1), making them *acyclic directed graphs* (or DAGs).

We considered as *input variables* LEN and the target formants F_{1t} – F_{3t}, as *mediators* the 11 *articulatory parameters* JA, LP, LD, TCX, TCY, TTX, TTY, TBX, TBY, HX and HY, and as *output variables* the produced formants F_1 – F_3. We defined the following constraints on acceptable DAGs (reflecting our prior beliefs about the causal processes) as forbidden edges:

a) between the input variables (i.e., LEN and the target formants do not influence each other),

b) between the output variables (i.e., the produced formants do not directly influence each other),

c) directly from the input to the output variables (i.e., the influences must pass through the articulatory parameters),

d) no back influences from the output variables to the articulatory parameters (i.e., the acoustics does not directly affect articulation) and the input variables (as these are predefined), and

e) no back influences from the articulatory parameters to the input variables (again, as these are predefined).

In summary, we modeled a uni-directional flow from the predefined input to the realized acoustic ouput strictly mediated by the articulatory parameters. While such a model is missing potentially important variables and fails to properly capture the complex feedback loops between articulators during learning, it does represent a testable hypothesis that arguably approximates reality to an acceptable degree for our purposes here (Pearl & Mackenzie, 2018).

We then searched automatically for the DAGs that meet these constraints and fit our z-scored data using two methods implemented by the `bnlearn R` package (Scutari, 2010): the constraints-based Incremental Association Markov

Table 5: The direct effects of LEN on the articulatory parameters across all vowels together and for each vowel separately, as estimated by *iamb* (Incremental Association Markov Blanket) search (empty cells are dropped edges).

V	HX	HY	JA	TCX	TCY	TBX	TBY	TTX	TTY	LP	LD
all		0.92	0.29		0.44					0.30	-0.24
[ə]		0.93	0.33		0.80	-0.23				0.45	
[ɑ]	0.79				0.87						-0.45
[a]		0.91	0.70							0.48	-0.38
[æ]		0.82	0.71	-0.68						0.61	
[e]		0.93	0.38		-0.30		0.11		0.31	0.55	-0.37
[i]	-0.47	0.94	0.38							0.49	-0.40
[o]		0.94		0.82	0.81						
[u]		0.93		0.31	0.72						

Blanket algorithm (denoted in the following as *iamb*) and the score-based Tabu Search (denoted *tabu*; we focus here on *iamb*, as it tends to produce sparser networks, but the full details are available in the Supplementary Materials in **Appendix**). As all variables are continuous and z-scored[11], the edge coefficients are the partial regression coefficients β that can be meaningfully compared across DAGs; Figure 8 shows a representative sample of such DAGs, while Tables 5 and 6 summarize the β's of the direct effects *of* LEN on the articulatory parameters and of these *on* the formants F_1 – F_3.

These results suggest that LEN, R_{ht} and HY are strongly coupled, behaving as a unit describing larynx height, as do (to a lesser extent) LP, LD and JA. The articulators (except for HY) behave differently (and non-linearly) for different vowels and larynx heights, suggesting that compensatory strategies differ between extreme larynx positions and target vowels. The condition LEN affects directly hyoid position (mostly through HY), but also the jaw (JA), the tongue body (TCX and TCY) and the lips (LP and LD), suggesting that these articulators may assume the most important roles in compensating for

[11] Here, we z-scored the variables to minimize the potential influence of their different measurement scales on the results and to allow the comparability of the edge coefficients across models.

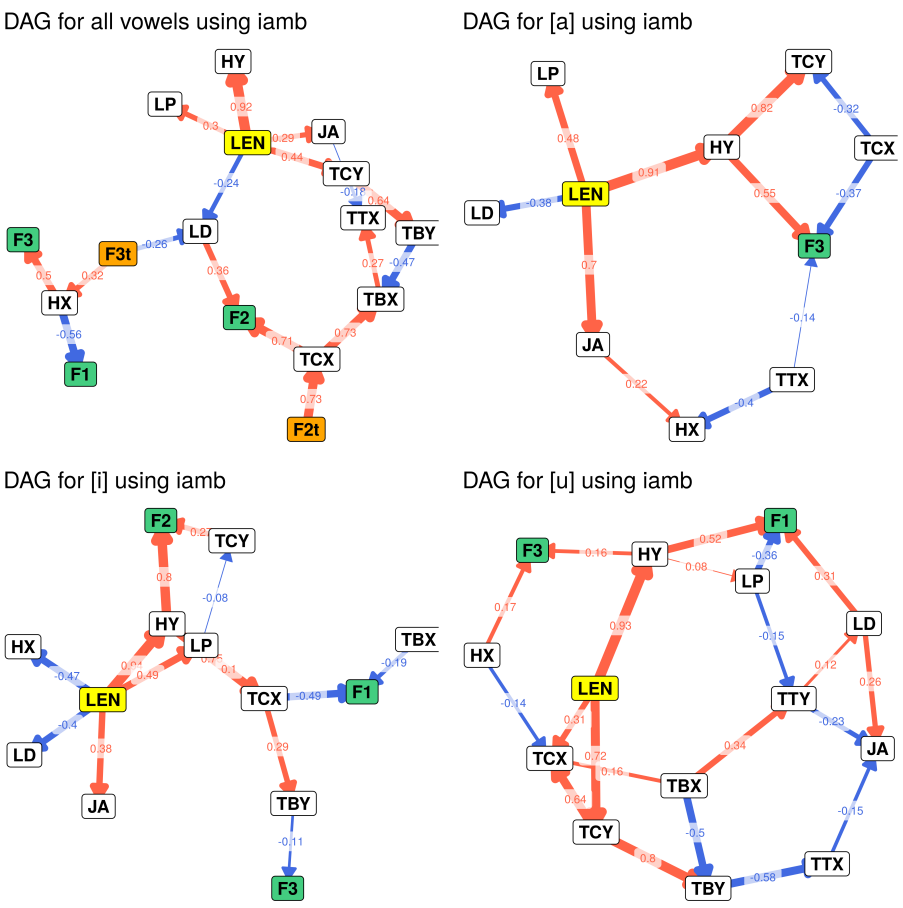

Figure 8: A few examples of Directed Acyclic Graphs (DAGs) capturing the effects of larynx height LEN on the articulators and of these on the produced formants F_1 – F_3. From top-left to bottom right: for all vowels together, for [a], [i] and [u], respectively. We show here the results when searching for DAGs using the *iamb* (Incremental Association Markov Blanket) algorithm. The target formants F_{1t} – F_{3t} are meaningful only for all vowels together (as they are constant for a given vowel). *Nodes*: color represents the variable type (yellow = LEN, varying between conditions; orange = target formants F_{1t} – F_{3t}, varying between vowels; white = articulatory parameters; green = produced formants F_1 – F_3). *Edges*: numbers are the partial regression coefficients β, width represents the absolute value of the coefficients, color represents the sign of these coefficients (blue = negative, red = positive). All DAGs can be found in the Supplementary Materials in **Appendix**.

Table 6: The direct effects of the articulatory parameters on the formants across all vowels together and for each vowel separately, as estimated by iamb search (empty cells are dropped edges; rows with only empty cells are not shown to avoid clutter).

V	HX	HY	JA	TCX	TCY	TBX	TBY	TTX	TTY	LP	LD	F
all	-0.56											F_1
[ə]									0.21			F_1
[ɑ]		0.71										F_1
[i]				-0.49		-0.19						F_1
[o]	-0.09											F_1
[u]		0.52								-0.36	0.31	F_1
all				0.71							0.36	F_2
[ə]	-0.42					0.13	0.16					F_2
[e]	-0.30											F_2
[i]		0.80			0.27							F_2
all	0.50											F_3
[ə]	0.25	0.53			0.06					0.09		F_3
[a]		0.55		-0.37				-0.14				F_3
[æ]	0.51		0.44			0.02						F_3
[i]							-0.11					F_3
[o]		0.75										F_3
[u]	0.17	0.16										F_3

larynx position. As such, concerning the articulators mentioned by Ménard & Boë (2000) and Boë et al. (2002), our results suggest some general trends: for a high larynx, the lips protrude (high LP) and close (low LD) (Badin et al., 2014) (less marked for [o] and [u]), the tongue body rises (high TCY) (except for [i] and [e]), but the tongue tip (TTX and TTY) is less actively involved than expected. In what concerns the acoustics, the produced formants F_1 – F_2 are directly affected mostly by larynx position, the tongue and the lips, but these effects vary among vowels and for different formants. Thus, these results suggest that (a) multiple articulators interact, forming "units" that may be responsible for certain aspects of the acoustics, (b) these units might be complex and (c) not unique (i.e., there might be more than one way of achieving the specific output formant frequencies), and (d) compensating for LEN might involve several such (possibly equivalent) units.

4. Discussion and conclusions

We focused here on the systematic variation of larynx height and on its effects on vowel acoustics and on the articulatory mechanisms engaged in compensating for it. Our computational agents, using a generic machine-learning mechanism that controls a realistic geometric model of the vocal tract, did learn to a very high degree of accuracy eight target vowels ([ə], [ɑ], [a], [æ], [e], [i], [o] and [u]) widely attested cross-linguistically and covering the modern human vowel space. However, this is a complex task that can be captured by local optima: while potentially a problem for finding the globally optimal solution(s), this can be seen as a realistic counterpart to actual human acquisition strategies, which can likewise settle for local optima (for example, articulatory idiosyncrasies that might even require speech therapy).

The attained accuracy is best around values of the vertical to total supralaryngeal vocal tract lengths ratio R_{ht} of approximately 0.45, degrading with deviations from this optimal region (especially strong for extremely low R_{ht}) in different manners for different vowels. Likewise, the maximal distinctiveness

between vowels is reached for similar R_{ht} values, with reduced distinctiveness for extreme ratios. The first formant F_1 is little affected by R_{ht}, while F_3 shows strong effects (especially for low R_{ht}), and while the vowel system as a whole is affected however, there is no simple collapse towards [ə].

Interestingly, the strongest effects on acoustics are for very low R_{ht} values (close to 0.40) and much less marked for very high values (close to 0.60), except for [u] (and, to a lesser degree, [i]). While this could be caused by us failing to model even more extremely high R_{ht} values (due to constraints inherent in the VocalTractLab 2.1 that remain to be addressed by future work), we think that is at best only part of the explanation (our larynx height conditions, LEN, are a sample of an equal number of positions equally displaced below and above the "standard" VocalTractLab 2.1 position). We suggest instead that higher larynx positions are more effectively compensated by other articulators, especially the tongue and the lips, than lower positions. Our findings of the active (and mostly effective) compensation for larynx height by other articulators fit previous reports such as speaking without a tongue (Gerdeman & Fujimura, 1990) or after partial resection of the larynx (Crevier-Buchman et al., 2012), with an artificial hard palate (Brunner et al., 2006; McFarland et al., 1996) or with a bite-block (Fowler & Turvey, 1980). However, also considering the fourth and the fifths formants (F_4 and F_5) uncovers effects of both very low and very high larynx positions R_{ht} on them (especially clear for [u]).

Nevertheless, while reinforcing the widespread capacity for compensation of even pathological variation in vocal tract anatomy, our results also show that compensation is not total. This adds support to our proposal that widespread inter-individual and inter-group normal variation in vocal tract anatomy can result in audible effects that may *bias* sound change, ultimately playing a role in explaining the observed linguistic diversity (Dediu et al., 2017; Moisik & Dediu, 2017; Dediu & Moisik, accepted).

The optimal ratio found here, ≈ 0.45, is lower than that reported by previous

modeling studies[12] (see Table 7), but this may be partly due to the simpler models used there. For example, the model in de Boer (2010a) lacks lips, which are argued by Badin et al. (2014) to reduce its capacity to compensate larynx height (compensation that is clearly happening in our model). However, the bulk of the differences in vocal tract ratios are probably due to the different definitions and measurements of the vertical and horizontal parts of the vocal tract: for example, Boë et al. (2002) use the arytenoid apex as the lower limit of the vertical part, which is above the vocal folds used by Nishimura et al. (2006), resulting in shorter estimated vertical lengths and an inflated vocal tract ratio. Nevertheless, despite these discrepancies in the actual values, our results agree with the previous findings that, while there seems to be a "sweet spot" of larynx heights, "suboptimal" values do not preclude vowel production. While statistically significant different from their targets (on the scale of under a bark for F_1 and F_2 and up to a few barks for the higher formants), the vowels produced with extreme vocal tract ratios are acoustically very similar to their targets (see Supplementary Materials in **Appendix**).

While our model does not directly test very high (supposedly "Neanderthal"-like) vocal tract ratios (as per Lieberman & Crelin (1971)), it supports a more nuanced view, in line with suggestions by Fitch (2000), where even extreme positions of the larynx may be actively compensated by other articulators and probably do not preclude vowel production. Currently available data shows that the Neanderthal hyoid bone was anatomically and biomechanically very similar to that of modern humans (D'Anastasio et al., 2013; Martínez et al., 2008) and that their hearing, despite some differences in the anatomy of the ear and the ear ossicles, was essentially modern (Stoessel et al., 2016; Martínez et al., 2013), much less is known about the position of the larynx in the throat

[12]This is also very close to values obtained from actual human data (e.g., Nishimura et al., 2006; Lieberman et al., 2001; Xue & Hao, 2006 and our own – not yet published – MRI data) and to the "default" VocalTractLab 2.1 R_{ht} (which is not surprising given that it is based on MRI scans).

Table 7: Various values of the human vertical to total supralaryngeal vocal tract lengths ratio R_{ht} reported in the literature (please note that different studies do not necessarily use the same definitions, measures and techniques).

Source	Sample	R_{ht}
Boë et al. (2002)	adult males	0.50–0.64
Xue & Hao (2006)	adult males	0.46–0.48
Boë et al. (2002)	adult females	0.54–0.63
Xue & Hao (2006)	adult females	0.45–0.50
Nishimura et al. (2006)	children ($\sim$ 9 years old)	$\lesssim 0.5$
Lieberman et al. (2001)	children ($\sim$ 8 and older)	$\lesssim 0.5$
de Boer (2010b)	optimal ratio	0.53
this study	optimal ratio	≈ 0.45

and the associated vocal tract ratio R_{ht}. Early reconstructions (Lieberman & Crelin, 1971) suggested a very high larynx in Neanderthals, precluding the articulation of vowels such as [a], [i] and [u] (c.f., Lieberman & Crelin, 1971, p. 177; however, note that these authors highlight that there's more to speech and language than these vowels and suggest that Neanderthals might have been linguistic humans) and despite criticisms and further work (see Dediu & Levinson, 2013 and Dediu & Levinson, 2018), there is recent data coming from the reconstruction of archaic epigenomes (Gokhman et al., 2017) suggesting that the modern human vocal tract might differ in certain respects from that of archaic humans. Even assuming that the Neanderthal larynx was higher than in modern humans, previous modeling work questions the inference that they were not capable of producing the whole human vowel space (Boë et al., 2002, 2007) (but such work has been, in turn, criticized for using anti-conservative assumptions; de Boer & Fitch, 2010; Lieberman, 2007). However, recent work (Fitch et al., 2016) suggests that even a macaque vocal tract may be capable of producing a wide range of speech sounds, being, in this sense, "speech-ready".

Our model shows that a generic learning mechanism is able to control a realistic vocal tract to reproduce with high accuracy a set of target vowels

by discovering compensatory strategies involving multiple articulators despite perturbations in the position of the larynx. In comparison with previous work, our model must control more articulatory parameters (11) than either that of de Boer (2010b) (5 parameters) or Boë et al. (2002) (7 parameters). This precludes the use of uniform or random sampling and, coupled with the *a priori* non-linear nature of the mapping between articulatory parameters and acoustics, requires better search strategies. However, this induces issues of its own, such as the low probability of finding the true global optima (but arguably not unlike how humans learn to control their motor system). It may be the case that given a sufficiently human-like vocal tract and a generic learning mechanism capable of controlling body movement, learning to reach articulatory vowel targets is feasible given sufficiently fine motor control over the articulators. Further work should test this hypothesis by introducing larger distortions to our vocal tract model, using other domain-general learning mechanisms, and by degrading the degree of fine motor control over various articulatory parameters in the model.

Acknowledgements

We wish to thank Peter Birkholz for sharing the source code of VocalTactLab 2.1, for allowing us to modify it and for answering our questions, and to three anonymous reviewers whose comments and suggestions greatly improved the paper. This work was Funded by the Netherlands Organisation for Scientific Research (NWO) VIDI grant 276-70-022 to DD. During the writing of this paper, DD was supported by an European Institutes for Advanced Study (EURIAS) Fellowship (2017-2018) and an IDEXLyon Fellowship, Université de Lyon (2018-2021).

References

Badin, P., Boë, L.-J., Sawallis, T. R., & Schwartz, J.-L. (2014). Keep the lips to free the larynx: Comments on de Boer's articulatory model (2010). *Journal of Phonetics*, *46*, 161–167. doi:10.1016/j.wocn.2014.07.002.

Baker, J. E. (1987). Reducing bias and inefficiency in the selection algorithm. In *Genetic algorithms and their applications: proceedings of the second International Conference on Genetic Algorithms* (pp. 14–21).

Barry, W. J., & Trouvain, J. (2008). Do we need a symbol for a central open vowel? *Journal of the International Phonetic Association*, *38*, 349–357. doi:10.1017/S0025100308003587.

Berwick, R. C., & Chomsky, N. (2017). Why only us: Recent questions and answers. *Journal of Neurolinguistics*, *43, Part B*, 166–177. doi:10.1016/j.jneuroling.2016.12.002.

Beyer, H.-G., & Schwefel, H.-P. (2002). Evolution strategies: A comprehensive introduction. *Natural Computing*, *1*, 3–52. doi:10.1023/A:1015059928466.

Birkholz, P. (2005). *3D-artikulatorische Sprachsynthese*. Logos.

Birkholz, P. (2013a). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, *8*, e60603. doi:10.1371/journal.pone.0060603.

Birkholz, P. (2013b). Vocaltractlab 2.1 user manual. URL: http://www.vocaltractlab.de/download-vocaltractlab/VTL2.1-Manual.pdf.

Birkholz, P., & Kröger, B. J. (2006). Vocal tract model adaptation using magnetic resonance imaging. In *7th International Seminar on Speech Production (ISSP06)* (pp. 493–500). URL: https://pdfs.semanticscholar.org/1d69/8b6e207d108c6baee014aa92147204356767.pdf.

Bladon, R. a. W., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *The Journal of the Acoustical Society of America*, *69*, 1414–1422. doi:10.1121/1.385824.

Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis. In *Proceedings of the International Congress of Phonetic Sciences*

(pp. 1–25). URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_2501.pdf.

Boë, L.-J., Badin, P., Ménard, L., Captier, G., Davis, B., MacNeilage, P., Sawallis, T. R., & Schwartz, J.-L. (2013). Anatomy and control of the developing human vocal tract: A response to Lieberman. *Journal of Phonetics*, *41*, 379–392. doi:10.1016/j.wocn.2013.04.001.

Boë, L.-J., Heim, J.-L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, *30*, 465–484. doi:10.1006/jpho.2002.0170.

Boë, L.-J., Heim, J.-L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007). *Journal of Phonetics*, *35*, 564–581. doi:10.1016/j.wocn.2007.06.006.

de Boer, B. (2010a). Investigating the acoustic effect of the descended larynx with articulatory models. *Journal of Phonetics*, *38*, 679–686. doi:10.1016/j.wocn.2010.10.003.

de Boer, B. (2010b). Modelling vocal anatomy's significant effect on speech. *Journal of Evolutionary Psychology*, *8*, 351–366. doi:10.1556/JEP.8.2010.4.1.

de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*, 36–47. doi:10.1177/1059712309350972.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer. URL: http://www.praat.org/.

Brunner, J., Hoole, P., Perrier, P., & Fuchs, S. (2006). Temporal development of compensation strategies for perturbed palate shape in German/sch/-

production. In *Proceedings of the 7th International Seminar on Speech Production* (pp. 247–254). URL: http://publikationen.ub.uni-frankfurt.de/opus4/frontdoor/index/index/docId/14086.

Crevier-Buchman, L., Pillot-Loiseau, C., Rialland, A., Narantuya, Vincent, C., & Desjacques, A. (2012). Analogy between laryngeal gesture in Mongolian Long Song and supracricoid partial laryngectomy. *Clinical Linguistics & Phonetics*, *26*, 86–99. doi:10.3109/02699206.2011.590920.

D'Anastasio, R., Wroe, S., Tuniz, C., Mancini, L., Cesana, D. T., Dreossi, D., Ravichandiran, M., Attard, M., Parr, W. C. H., Agur, A., & Capasso, L. (2013). Micro-biomechanics of the Kebara 2 hyoid and its implications for speech in Neanderthals. *PLoS ONE*, *8*, e82261. doi:10.1371/journal.pone.0082261.

De Boer, B. (2000). Self organization in vowel systems. *Journal of Phonetics*, *28*, 441–465. doi:10.1006/jpho.2000.0125.

Dediu, D., Janssen, R., & Moisik, S. R. (2017). Language is not isolated from its wider environment: vocal tract influences on the evolution of speech and language. *Language and Communication*, *54*, 9–20. doi:doi:10.1016/j.langcom.2016.10.002.

Dediu, D., & Levinson, S. C. (2013). On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in Language Sciences*, *4*, 397. doi:10.3389/fpsyg.2013.00397.

Dediu, D., & Levinson, S. C. (2018). Neanderthal language revisited: not only us. *Current Opinion in Behavioral Sciences*, *21*, 49–55. doi:10.1016/j.cobeha.2018.01.001.

Dediu, D., & Moisik, S. R. (accepted). Pushes and pulls from below: anatomical variation, articulation and sound change. *Glossa*, (pp. –).

Eiben, A. E., & Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Natural Computing Series. Springer-Verlag. doi:10.1007/978-3-662-05094-1.

Esling, J. H., Benner, A., & Moisik, S. R. (2015). Laryngeal articulatory function and speech origins. In *18th International Congress of Phonetic Sciences [ICPhS 2015] Satellite Event: The Evolution of Phonetic Capabilities: Causes, Constraints and Consequences* (pp. 2–7). ICPhS. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.866.452&rep=rep1&type=pdf.

Fitch, W. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, *4*, 258–267. doi:10.1016/S1364-6613(00)01494-7.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *Science Advances*, *2*, e1600723. doi:10.1126/sciadv.1600723.

Fowler, C. A., & Turvey, M. T. (1980). Immediate compensation in bite-block speech. *Phonetica*, *37*, 306–326. doi:10.1159/000260000.

Gerdeman, B., & Fujimura, O. (1990). Speaking without a tongue. *The Journal of the Acoustical Society of America*, *87*, S89–S89. doi:10.1121/1.2028415.

Gokhman, D., Agranat-Tamir, L., Housman, G., Garcia-Perez, R., Nissim-Rafinia, M., Mallick, S., Nieves-Colón, M., Li, H., Alpaslan-Roodenberg, S., Novak, M., Gu, H., Ferrando-Bernal, M., Gelabert, P., Lipende, I., Kondova, I., Bontrop, R., Quillen, E. E., Meissner, A., Stone, A. C., Pusey, A. E., Mjungu, D., Kandel, L., Liebergall, M., Prada, M. E., Vidal, J. M., Prüfer, K., Krause, J., Yakir, B., Pääbo, S., Pinhasi, R., Lalueza-Fox, C., Reich, D., Marques-Bonet, T., Meshorer, E., & Carmel, L. (2017). Extensive regulatory changes in genes affecting vocal and facial anatomy separate modern from archaic humans. *bioRxiv*, (p. 106955). URL: https://www.biorxiv.org/content/early/2017/10/03/106955. doi:10.1101/106955.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M., Watumull, J., Chomsky, N., & Lewontin, R. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*, 401. doi:10.3389/fpsyg.2014.00401.

Heaton, J. (2015). Encog: Library of interchangeable machine learning models for Java and C#. *Journal of Machine Learning Research*, *16*, 1243–1247. URL: http://jmlr.org/papers/v16/heaton15a.html.

Honda, K. (1996). Organization of tongue articulation for vowels. *Journal of Phonetics*, *24*, 39–52. doi:10.1006/jpho.1996.0004.

Honda, K., & Tiede, M. K. (1998). An MRI study on the relationship between oral cavity shape and larynx position. In *5th International Conference on Spoken Language Processing (ICSLP 98)* (p. 4). Sydney, Australia: ISCA Archive. URL: https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0686.pdf.

Janssen, R. (2018). *Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny*. PhD Thesis Radboud University/Max Planck Institute for Psycholinguistics Nijmegen, The Netherlands. URL: https://github.com/ddediu/let-the-agents-do-the-talking.

Janssen, R., Moisik, S. R., & Dediu, D. (2018). Modelling human hard palate shape with Bézier curves. *PLoS ONE*, *13*, e0191557. doi:10.1371/journal.pone.0191557.

Johansson, S. (2015). Language abilities in Neanderthals. *Annual Review of Linguistics*, *1*, 311–332. doi:10.1146/annurev-linguist-030514-124945.

Klein, R. G. (2009). *The Human Career: Human Biological and Cultural Origins*. (3rd ed.). University of Chicago Press.

Ladefoged, P., & Johnson, K. (2010). *A Course in Phonetics*. (6th ed.). Cengage Learning.

Lieberman, D. E., McCarthy, R. C., Hiiemae, K. M., & Palmer, J. B. (2001). Ontogeny of postnatal hyoid and larynx descent in humans. *Archives of Oral Biology*, *46*, 117–128. doi:10.1016/S0003-9969(00)00108-4.

Lieberman, P. (2007). Current views on Neanderthal speech capabilities: A reply to Boë et al. (2002). *Journal of Phonetics*, *35*, 552–563. doi:10.1016/j.wocn.2005.07.002.

Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*, *40*, 608–622. doi:10.1016/j.wocn.2012.04.001.

Lieberman, P. (2016). The evolution of language and thought. *Journal of Anthropological Sciences*, *94*, 127–146. doi:10.4436/JASS.94029.

Lieberman, P., & Crelin, E. S. (1971). On the speech of Neanderthal man. *Linguistic Inquiry*, *2*, 203–222. URL: http://www.haskins.yale.edu/Reprints/HL0104.pdf.

Maddieson, I., & Disner, S. F. (1984). *Patterns of sounds*. Cambridge University Press.

Maeda, S. (1990). Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Speech Modelling* NATO ASI Series (pp. 131–149). Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-2037-8_6.

Martínez, I., Arsuaga, J., Quam, R., Carretero, J., Gracia, A., & Rodríguez, L. (2008). Human hyoid bones from the middle Pleistocene site of the Sima de los Huesos (Sierra de Atapuerca, Spain). *Journal of Human Evolution*, *54*, 118–124. doi:10.1016/j.jhevol.2007.07.006.

Martínez, I., Rosa, M., Quam, R., Jarabo, P., Lorenzo, C., Bonmatí, A., Gómez-Olivencia, A., Gracia, A., & Arsuaga, J. (2013). Communicative capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Quaternary International*, *295*, 94–101. doi:10.1016/j.quaint.2012.07.001.

McFarland, D. H., Baum, S. R., & Chabot, C. (1996). Speech compensation to structural modifications of the oral cavity. *The Journal of the Acoustical Society of America*, *100*, 1093–1104. doi:10.1121/1.416286.

Ménard, L., & Boë, L.-J. (2000). Exploring vowel production strategies from infant to adult by means of articulatory inversion of formant data. In *International Congress of Spoken Language Processing, Beijing (Chine)* (pp. 465–468). URL: https://pdfs.semanticscholar.org/f1ed/89d8b405029a086b0f2166c2e599a57fb75e.pdf.

Moisik, S. R. (2013). *The Epilarynx in Speech*. PhD Thesis University of Victoria Victoria, British Columbia, Canada. URL: https://dspace.library.uvic.ca:8443//handle/1828/4690.

Moisik, S. R., & Dediu, D. (2017). Anatomical biasing and clicks: Evidence from biomechanical modeling. *Journal of Language Evolution*, *2*, 37–51. doi:doi:10.1093/jole/lzx004.

Moran, S., McCloy, D., & Wright, R. (2014). PHOIBLE online. URL: http://phoible.org.

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Cognitive Science*, *4*, 1006. doi:10.3389/fpsyg.2013.01006.

Nishimura, T., Mikami, A., Suzuki, J., & Matsuzawa, T. (2006). Descent of the hyoid in chimpanzees: evolution of face flattening and speech. *Journal of Human Evolution*, *51*, 244–254. doi:10.1016/j.jhevol.2006.03.005.

Oudeyer, P.-Y., & Smith, L. B. (2016). How evolution may work through curiosity-driven developmental process. *Topics in Cognitive Science*, *8*, 492–502. doi:10.1111/tops.12196.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Penguin UK.

Prom-on, S., Birkholz, P., & Xu, Y. (2014). Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*. doi:10.1186/1687-4722-2014-23.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, *25*, 255–286. doi:10.1006/jpho.1997.0043.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*, 1–22. doi:10.18637/jss.v035.i03.

Stoessel, A., David, R., Gunz, P., Schmidt, T., Spoor, F., & Hublin, J.-J. (2016). Morphology and function of Neandertal and modern human ear ossicles. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 11489–11494. doi:10.1073/pnas.1605881113.

Sundberg, J. (1995). The singer's formant revisited. *Voice*, *4*, 106–119. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.642.3861&rep=rep1&type=pdf.

Sundberg, J., & Nordström, P. E. (1976). Raised and lowered larynx - the effect on vowel formant frequencies. *Speech Transmission Laboratory - Quarterly Progress and Status Report*, *17*, 35–39. URL: https://pdfs.semanticscholar.org/0fd1/0b3243459dc0cfcdf36a17f23c64c8243ff3.pdf.

Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*, 1414–1422. doi:10.1109/TASSP.1987.1165058.

Xue, S. A., & Hao, J. G. (2006). Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice*, *20*, 391–400. doi:10.1016/j.jvoice.2005.05.001.

Zelditch, M. L., Swiderski, D. L., & Sheets, H. D. (2012). *Geometric Morphometrics for Biologists: A Primer*. Academic Press.

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" american english /r. *The Journal of the Acoustical Society of America*, *123*, 4466–4481. doi:10.1121/1.2902168.

Appendix

The **Supplementary Materials** included here are composed of the following items (see also the GitHub repository `https://github.com/ddediu/larynx-height`).

- **simulation.7z**: a 7zip-compressed archive with the following directory structure:
  - **agent**: contains the pre-compiled code and configuration files needed to run the simulations:
    - `NativeInterface.dll`: the pre-compiled (as a Dynamically-Linked Library for Microsoft Windows 7 or newer, 64 bits) version of the VocalTractLab 2.1 refactored and modified by us (the C++ cource code is available upon request conditional on the acceptance of a custom license agreement mirroring the original VocalTractLab 2.1 source code license);

    * `Agent.jar`: the Java implementation of the agent;
    * `chain.py`: the Python implementation of a simulation;
    * `cyBezier.pyd`: the compiled Python Bézier model of the hard palate (see Janssen et al., 2018);
    * `config.csv`: CSV file controlling various parameters of the simulation such as the target vowels, the anatomical configurations, the number of replications and the number of parallel threads.
    * **config**: a folder containing extra configuration files:
      · `anatomy.csv`: definitions of various anatomical configurations (referred to from `config.csv`); here, the only relevant column is "SVTV length";
      · `targets.csv`: definitions of the target sounds (referred to from `config.csv`)in terms of the articulator values needed to produce them in the "standard" VocalTractLab 2.1 configuration.
  – **data**: contains the Python script `summarize.py` and the speaker definition file `JD2.speaker` needed to summarize the simulation results for further analyses;
  – **Kits**: contains a list of installation kits needed to run the simulations (but not the kits themselves due to their size and licenses);
  – **results**: contains the summary results (here the XZ-compressed TAB-separated files `results_F1_F3.csv.xz` and `results_F1_F5.csv.xz`) as well as an `R` script `preprocess-results.R` that further prepares these results for the final statistical analysis and produces the files (included here for convenience): `results_F1_F3.rds`, `results_F1_F5.rds`, `euclid_dist_vowels_F1_F3.rds`, `euclid_dist_vowels_F1_F5.rds`, `procrustes_dist_target_elite` and `procrustes_dist_target_elite_F1_F5.rds`;

- **analysis.7z**: a 7zip-compressed archive containing the R Markdown script `Rscript.Rmd` and the resulting full analysis report as a HTML document

`Rscript.html` (during the first compilation of the Rmarkdown script various directories, cached files and images will be also created);

- **sounds.7z**: a 7zip-compressed archive containing the Praat script used to produce the actual acoustic output corresponding to a set of three or five formant values (`PraatFormants2Wav`) and the WAV files corresponding to the target formants and to the best and worst productions for each vowel in each condition.