# Emergent maturation from stochastic optimization in vocal development

Clément Moulin-Frier, Jules Brochard, Freek Stulp, Pierre-Yves Oudeyer

# Emergent maturation from stochastic optimization in vocal development

(Clément Moulin-Frier, Jules Brochard), Freek Stulp and Pierre-Yves Oudeyer

November 10, 2014

## 1 Introduction

In the course of infant vocal development the vocal tract articulators are not used equally. A recruitment structure is displayed where the jaw seems to play a special role, in particular regarding an important developmental change around the age of 7 months called canonical babbling. Before this period, human infants produce non-speech vocalizations, in the sense that these latter do not display the syllabic structure specific to human speech. This particular structure appears with canonical babbling, often considered as the onset of speech learning (although controversed), where the infant suddenly and robustly produce jaw cycles coupled with phonation. This results in the first proto-syllables, as we often hear from young infants: "babababa".

A number of hypotheses has been proposed to explain the particular role of the jaw in speech development. The Frame/Content theory [10, 9] suggests that jaw predominance is due to the role of feeding movement in speech evolution by providing a powerful sound modulation ability when coupled with phonation. Although controversial, this hypothesis is supported by infant data displaying statiscal vowel-consonant associations in line with the theory predictions. A relatively similar hypothesis, although arguing against the Frame/Content theory, has been proposed where the considered phylogenetic precusor would be stereotyped communicative orofacial actions like lip-smacking in macaque monkeys [5, 4]. Other works consider that such a rythmic behavior is not specific to speech because arm babbling also appears around 6 months [8]. This suggests that these rhythmic patterns could be due to a general brain dynamics reorganization during the first year of life.

Another line of works considers that rhythmic jaw movement can be the result of a sensorimotor and social learning processes. Warlaumont [20, 22, 21] proposes computational models of syllabic structure emergence based on social or intrinsic reinforcement. The model starts with random vocalizations produced using an articulatory synthesizer, i.e a computer model of the human vocal tract able to synthesize sound from articulatory movements. These random vocalizations mainly result in non-speech sounds. In the first case (social), a human subject then listens to these vocalizations and is asked to reinforce them or not according to his own judgment of speech-likeness. In the second case (intrinsic), vocalizations are reinforced using an objective salience measure (more salient sounds are reinforced more). When reinforced, a learning rule drives the system towards producing the corresponding articulation more often.

In both cases (social and intrinsic), the model converge towards the production of syllabic vocalizations.

Still based on sensorimotor and social learning, a subset of the authors of the present paper proposes a model of speech acquisition based on curiosity-driven self-exploration and imitation [12]. Here the agent drives its own vocalizations according to a learning progress maximization principle: it self-generates auditory goals according to the progress it observes in learning how to reach them. Such a mechanism has been shown to be highly efficient in learning inverse models in high-dimensional and redundant robotics setups [15, 2, 13, 6] and concretely formalize concepts of intrinsic motivation described in the psychology literature into algorithmic architectures that can be experimented in computers and robots [16, 3, 14, 1]. A side effect of this exploration strategy is to self-organize developmental pathways where the agent autonomously focuses on tasks of increasing complexity. Applied to vocal learning, we observe that the model first produces articulations mainly resulting in no sounds and then more and more complex productions from unarticulated to proto-syllabic vocalizations.

In the present paper we propose the alternative hypothesis that the predominance of the jaw could be due to even simpler learning mechanisms, without invoking curiosity-driven intrinsic motivations or social reinforcement. Our hypothesis neither rely on speculative phylogenetic precursors as in the Frame/Content theory.

## 1.1 Structuring mechanisms in sensorimotor learning

In the field of developmental robotics, we are interested in the form and formation of learning structures. Complex sensorimotor behavior acquisition by developmental agents, whatever their biological or mechanical/computational nature, implies dynamical interactions in a complex embodiment with the environment. In such a context, pure random exploration does not provide adequate data to allow an efficient learning. This is due to the high dimensionality of the involved sensorimotor spaces, the non-linearity and redundancy of the sensorimotor mappings and the significant cost in time and resources of performing informative sensorimotor interaction with the environment. These constraints force the agent to develop efficient exploration strategies, resulting in the formation of complex learning structures.

Sensorimotor maturations are examples of such structures. Let us emphasizes two examples observed in infant development. The first one concerns arm control development, where an infant recruits arm joints obeying the so-called proximodistal law: from the shoulder to the hand. Although this law has biological reasons due to myelination, progressively impacting peripheral neural structures in a proximodistal manner, it has been shown that such a law can also be explained from a developmental learning point of view. By allowing the movement of the entire arm, proximal joints produce a wider range of effect (e.g. a wider range of reached hand positions) when compared to distal joints. Therefore, an organism exploring its own sensorimotor abilities has interest in recruiting proximal joints first because such a strategy allows him to rapidly makes a reasonable approximation of the range of possible effects. It is actually not surprising that both biology, through the myelination process, and cognition, through exploration strategies, found a similar solution (sensorimotor

maturation) to the same optimization problem: how to efficiently learn complex sensorimotor behaviors.

The second example comes from speech science. During vocal development, the involved articulators (the jaw, the tongue, the lips ...) are not recruited equally. In particular, at seven months canonical babbling appears in a very robust way, where the infant experiments how jaw cycles affect vocal productions. It has been proposed that this predominance of the jaw originates through evolution from mastication and ingestion behaviors, bootstrapping the emergence of communicative orofacial gestures, such as lip-smacking, in non-human primates.

The aim of this paper is to propose and computationally support an original hypothesis regarding the predominance of jaw movements in infant speech development. We capitalize on previous research on emergent maturations from a stochastic optimization process on the arm domain [19]. This work has shown that a quite simple optimization process, allowing a progressive learning of reaching arm movements by the minimization of a cost function, can display an emergent sensorimotor maturation: without precoding it, such an algorithm naturally recruits in priority proximal joints due to their wider range of effects. Here, we adapt this work to the vocal domain, by the use of a vocal tract model, and show that a similar effect can occurs on particular conditions, recruiting the jaw in priority due to the wider range of auditory effects implied by moving this articulator.

# 2 Method

This section describes the vocal tract model and the stochastic optimization algorithm, as well as the way we connect them into the simulation loop that we will use in our experiments.

## 2.1 Vocal tract: the Maeda model

### 2.1.1 Articulatory synthesis

Our computational model involves the articulatory synthesizer of the DIVA model described in [7][1] based on Maeda's model [11]. Without going into technical details, the model corresponds to a computational approximation of the general speech production principles illustrated in Figure 1.

The model receives 13 articulatory parameters as input. The first 10 are from a principal component analysis (PCA) performed on sagittal contours of images of the vocal tract of a human speaker, allowing to reconstruct the sagittal contour of the vocal tract from a 10-dimensional vector. In this study, we will only use the 7 first parameters (the effect of the others on the vocal tract shape is negligible), fixing the 3 last in the neutral position (value 0 in the software). The effect of these 7 articulatory parameters from the PCA on the vocal tract shape is displayed **Figure 2**. Through an area function, associating sections of the vocal tract with their respective area for a given motor configuration, the model can compute the 3 first formants (see Figure 1) of the resulted signal if

[1] available online at `http://www.bu.edu/speechlab/software/diva-source-code`. DIVA is a complete neurocomputational model of speech acquisition, in which we only use the synthesizer computing the articulatory-to-auditory function.

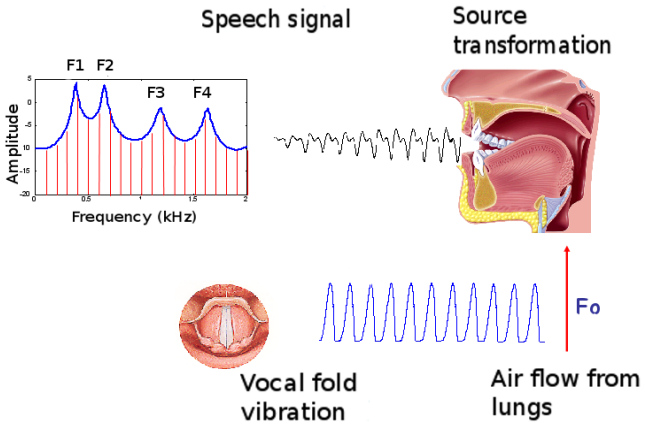



Figure 1: An articulatory synthesizer is a model of human vocal production. In this latter, the lung air flow triggers vocal fold vibration, providing a source signal with fundamental frequency $F0$. According to the vocal tract shape, acting as a resonator, the harmonics of the source fundamental frequency are selectively amplified or faded, resulting in a sound wave originating at the lips. Represented in a frequency-amplitude space, this sound wave typically displays a number of amplitude local maxima, called the formants $Fi$ and ordered from the lowest to the highest frequency. The formants are known to be important auditory features in human speech perception.

phonation occurs. Phonation is controlled through 3 others parameters, that we set at a value assuring normal phonation. It is then able to compute the formants of the signal (among other auditory and somato-sensory features) through the area function.

### 2.1.2 Motor trajectory generation

A vocalization corresponds to a trajectory of the 7 articulators displayed Figure 2 lasting $500ms$. To do so, we encode the acceleration profile of each articulator as a linear combination of predefined basis functions. Starting from the neutral position of the vocal tract, these acceleration are integrated twice over the $500ms$, thus resulting in position trajectories for each articulators, as illustrated on Figure 3.

The acceleration $\ddot{q}_{m,t}$ of the $m$-th articulator at time $t$ is determined as a linear combination of basis functions (Equation 1, 2 and 3), where the point $\theta_m$ represents the combination's weights:

$$\ddot{q}_{m,t} = g_t^\top \theta_m \qquad \text{Acc. of joint } m \qquad (1)$$

$$g_t = (\dots, [g_t]_b, \dots) = (\dots, \frac{\Psi_b(t)}{\sum_{b=1}^{B} \Psi_b(t)}, \dots) \qquad \text{Basis functions} \qquad (2)$$

$$\Psi_b(t) = \exp\left(-(t - c_b)^2 / w^2\right) \qquad \text{Kernel} \qquad (3)$$

The centers $c_{b=1\dots B}$ of the kernels $\Psi$ are spaced equidistantly in the $500ms$ duration of the movement, and all have a width of $w = 50ms$. In our experiments

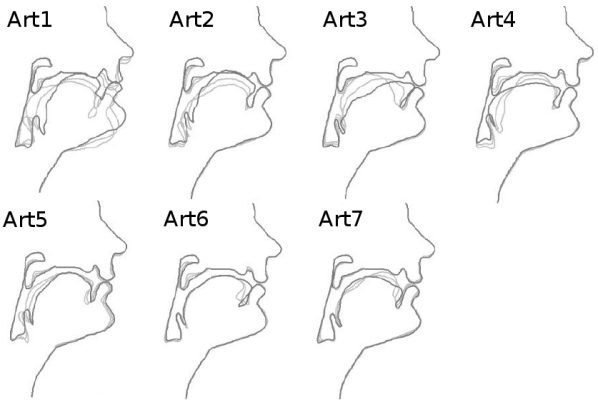



Figure 2: We use 7 articulatory dimensions to control the vocal tract shape (figure adapted from the DIVA source code documentation). Each subplot shows a sagittal contour of the vocal tract, where we can identify the nose and the lips on the right side. Bold contours correspond to a positive value of the articulatory parameter, the two thin contours are for a null (neutral position) and negative values. These dimensions globally correspond to the dimensions of movements of the human vocal tract articulators. For example, $Art_1$ mainly controls the jaw height, whereas $Art_3$ rather controls the tongue front-back position.

we will use $B = 4$ basis function for each articulator. Using 7 articulators as stated above, a full motor command is therefore a 28-dimensional vector.

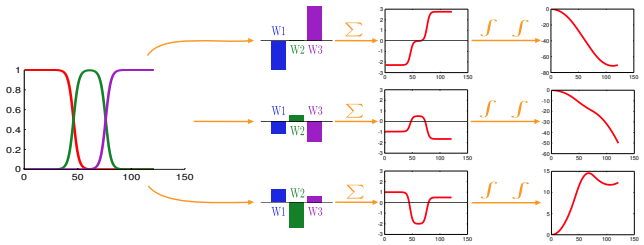



Figure 3: Examples of basis functions' uses. Left: 3 basis functions. Center: 3 triplets of weights and the weighted sum. Right: the 3 trajectories induced by the weights.

### 2.1.3 Auditory perception

The DIVA synthesizer provides formant trajectories as output, expressed in Hertz. We convert the values in Hertz into a perceptual scale, typically linear at low frequencies and logarithmic at high frequencies. We use the Bark scale defined according to the formula proposed by [17] and used in many studies on auditory perception, defined as follow:

$$Barks = 7 \times sinh^{-1}\left(\frac{Hertz}{650}\right) \tag{4}$$

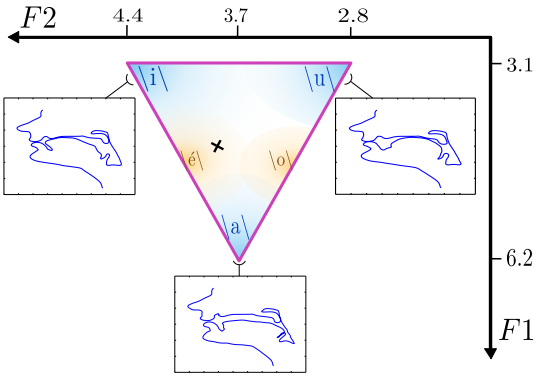



Figure 4: Variations of $F1$-$F2$ (in our Barks scale): they take place in what is called the vocalic triangle, where we represent the 5 canonical vowels, and a possible tract shape for 3 of them. The cross represent the neutral vocalization of the synthesizer.

Then, giving the same importance to all formants do not take into account spectral masking phenomena, according to which low-frequency components decrease the perceptual role of higher-frequency components. This led to the proposal [18] that $F1$ should have typically three times the weight of an "effective second formant" $F'2$ grouping of the roles of $F2$ and $F3$. Following this psycho-acoustic proposition, we use $F1$ and $F2$ as auditory features and in the distance computation in this $(F1, F2)$ plane we will weight the first dimension with a factor 3.

The perception that the vocal agent has of its own vocalization is the value in Barks of $F1$ and $F2$ at the end of the vocalization. Figure 4 displays possible values and associated vocal tract shapes.

## 2.2 Stochastic optmization: the $PI^{BB}$ algorithm

CMF: change by PI2CMAES = PI2 + CMAES?

$PI^{BB}$ is a stochastic optimization algorithm allowing to minimize an unknown cost function by exploring the search space in an evolutionary fashion. In a robotics framework, the search space typically corresponds to the motor space of the robot (e.g. the space of arm joint commands) and the cost function aims at defining a goal in a task space. For example, if one wants for the robot to reach a given hand position, the cost of a particular arm configuration can be defined as the distance between the corresponding position of the hand and the goal position, as it was in a previous paper using $PI^{BB}$ [19]. From a starting position $\theta$ in the search space, the algorithm will then generate small variations, evaluate them which respect to the cost function, update $\theta$ with respect to this evaluation, and repeat the process until convergence. It is therefore a kind of evolutionary optimization in the sense that it is based on an iterative mutation/selection process.

The first step of $PI^{BB}$ is to create some small variations around the current position $\theta$ in the search space. This is done by sampling $K$ new points for each articulator $m$ according to multivariate Gaussian probability distributions,

with means $\theta_m$ and covariance matrices $\Sigma_m$, that we note $\mathcal{N}(\theta_m, \Sigma_m)$. This results in $K$ new motor commands (28-dimensional each, see section 2.1.2) to be evaluated with respect to the cost function. Finally, a new Gaussian distribution is estimated for each articulator from these samples, by computing a weighted mean and covariance matrix, where the weight of each sample is higher when its cost is lower. This new distribution will be closer (in probabilistic term) to the goal than the previous one due to the weighting of the samples according to their respective costs. $K$ new samples are then drawn from this new distribution and the process is repeated until the cost of a sample is under a given threshold or a maximum number of iterations has been reached. The whole process is illustrated Figure 5. The main loop executes the following three steps (the name of each step is reported on the algorithm Figure 5).

**Exploration.** Sample $K$ parameter vectors $\theta_k$ from $\mathcal{N}(\theta, \Sigma)$, and determine the cost $J_k$ of each sample. In the visualization of our illustratory example task $K = 15$, and the cost $J(\theta)$ is the distance to the origin $\|\theta\|$, which lies approximately between 8 and 19 in this example.

**Evaluation.** Determine the weight $P_k$ of each sample, given its cost. Essentially, low-cost samples have higher weights, and vice versa. The normalized exponentiation function that maps costs to weights is visualized in the top-right graph. Larger green circles correspond to higher weights.

**Update.** Update the parameters $\langle \theta, \Sigma \rangle$ with weighted averaging. In the visualization, the updated parameters are depicted in red. Because low-cost samples (e.g. a cost of 8-10) have higher weights, they contribute more to the update, and $\theta$ therefore moves in the direction of the optimum $\theta_g = [0, 0]$.

In the next section, we will see how we can use this optimization algorithm in order to reach auditory goals with the vocal tract model we defined in the previous section.

## 2.3 Simulation loop

Given the vocal tract model and the stochastic optimization algorithm we just described, Figure 6 illustrates the simulation loop. Each sample of $PI^{\text{BB}}$ is composed of 7 points, one for each articulatory parameters. Firstly, through basis functions (Figure 3) and integration, each point is transformed into a motor trajectory . These 7 motor trajectories represent a complete articulatory movement. From this latter, the vocal tract model then computes the corresponding auditory trajectories in the formant space. The cost of that vocalization is computed from this articulatory and auditory trajectories, as we will explain more precisely when defining our experiments in the next section. Finally, a new Gaussian distribution is computed from these samples and their respective costs and the process repeats until convergence.

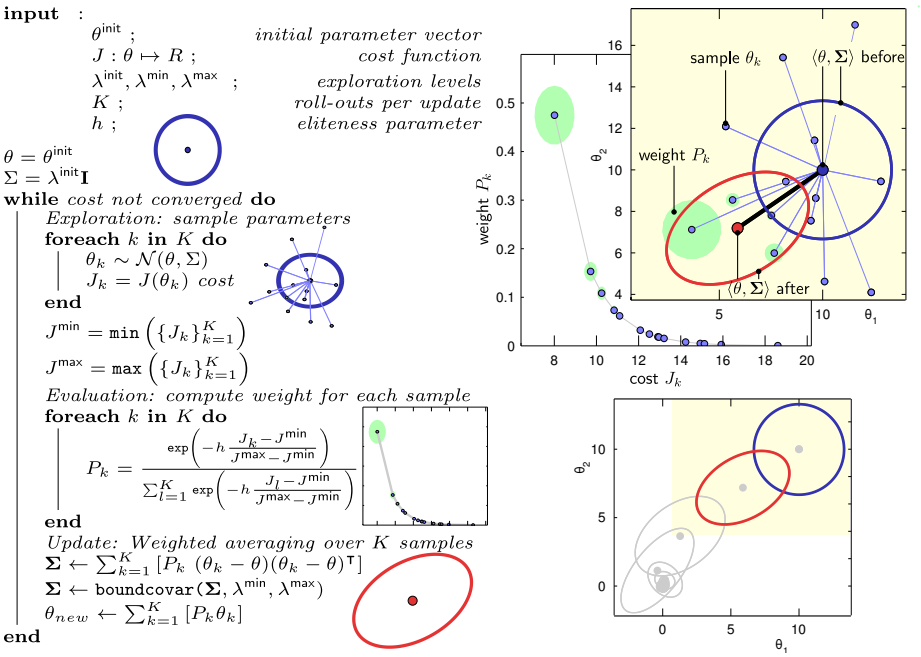



Figure 5: $PI^{\text{BB}}$ is a stochastic optimization algorithm allowing to minimize an unknown cost function by exploring the search space in an evolutionary fashion (figure extracted from [19]). For illustratory purposes, the search space is here 2-dimensional and the cost of a sample $\theta$ is simply the distance to the origin: $J(\theta) = \|\theta\|$. Left: $PI^{\text{BB}}$ pseudo-code. Top-right: Visualization of *one* parameter update with $PI^{\text{BB}}$ . Bottom-right: Evolution of the parameters over *several* updates, illustrating how the distribution converges towards the minimum $\theta_g = [0, 0]$. The algorithm is initialized by setting the mean and covariance parameters $\langle\theta, \Sigma\rangle$ to $\theta^{init}$ and $\lambda^{init}\mathbf{I}$ respectively, visualized as a dark blue ellipse. These parameters are updated at each iteration of the main loop. The red ellipse illustrates the first update.

# 3 Experiments

## 3.1 Protocols

### 3.1.1 Cost function

A crucial component of the $PI^{\text{BB}}$ algorithm is the cost function which evaluates the samples. Following Stulp's work [19], we design this function as the sum of three components. Basically the cost function favorizes auditory effects close to a given goal and penalizes articulatory positions far from the resting position and invovling a high energy cost. Thus, the first component is the distance to the goal in the formant-space: $\|(s_g) - (s)\|^2$, where $s_g = \begin{pmatrix} F1_g \\ F2_g \end{pmatrix}$ is the goal in the sensory formant space and $s = \begin{pmatrix} F1_{t_N} \\ F2_{t_N} \end{pmatrix}$ is formant values actually reached at the end of the vocalisation. Then we add a term to discard the pathological positions which are far from the resting position: $max_m\left(|P_{m,t_N}|\right)$,

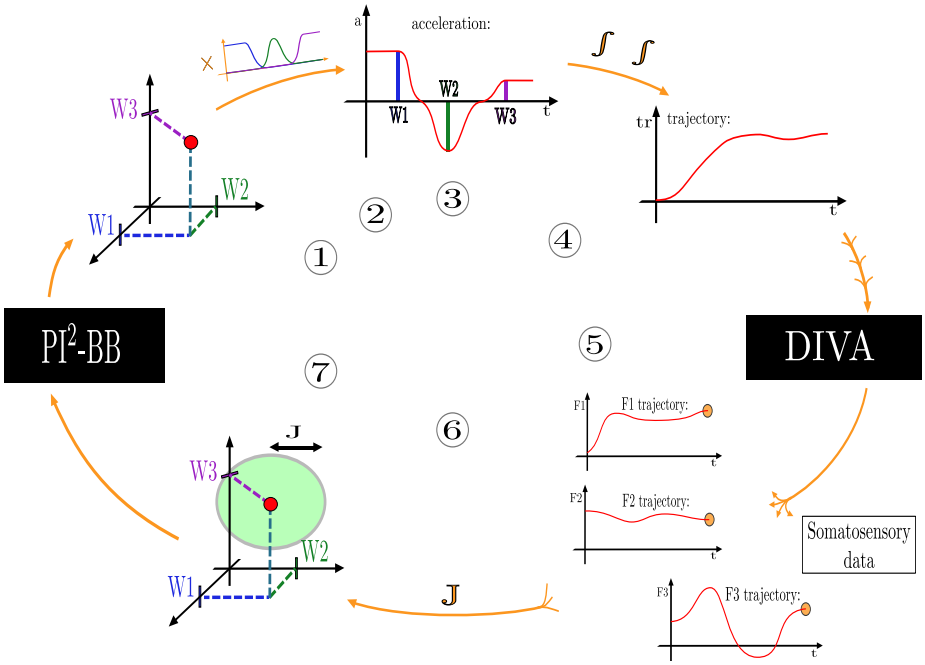



Figure 6: To reach an auditory goal in the formant space, $PI^{BB}$ is connected the DIVA synthesizer. (1) to (4) are conducted for each motor parameter $Pi$ with $i = 1..7$. (1) A point is drawn in a $B$-dimensional space, where B is the number of basis functions (here B=3 for the sake of illustration). (2) This point defines the weights of the $B$ basis functions. (3) The weighted sum of the 3 basis functions defines $Pi$ acceleration. (4) A double integration leads to the $Pi$'s position trajectory. (5) $P1$ to $P7$ trajectories are provided to the DIVA's synthesizer. (6) These trajectories and the corresponding auditory effect are used to compute the cost. (7) This cost is used to weight the sample of the 7 $B$-dimensional points (i.e the same cost for each one).

where $P_{m,t_N}$ is the position of the $m^{th}$ articulator at the end of the vocalisation. Finally, for plausibility and homogeneity of the solutions, we add an energy term: $\sum_{m=1}^{7}\sum_{t=1}^{T}\frac{a_{m,t}^2}{2}$, where $a_{m,t}$ is the acceleration of the $m^{th}$ articulator at the time step $t$ of the vocalisation.

Weighting each part in an empirical manner according to their impact on the task, we get the cost function:

$$J = 10^4 \left\| \begin{pmatrix} F1_g \\ F2_g \end{pmatrix} - \begin{pmatrix} F1_{t_N} \\ F2_{t_N} \end{pmatrix} \right\|^2 + max_m\left(|P_{m,t_N}|\right) + 10^{-1}\sum_{m=1}^{7}\sum_{t=1}^{T}\frac{a_{m,t}^2}{2} \quad (5)$$

.

A low-cost vocalisation is therefore a configuration which approaches the goal with a simple configuration and minimize its energy.

### 3.1.2 Exploration magnitudes

An interesting feature of $PI^{\text{BB}}$ is that the covariance matrix eigenvalues actually reflects the exploration magnitude on each articulator. Higher eigenvalues means a larger gaussian distribution to sample, resulting in a larger exploration on the corresponing articulator. The exploration magnitude of an articulator $P_m$ at a given update is defined as the maximum eigenvalue $\lambda_m$ of its associated covariance matrix $\Sigma_m$ at that update. Exploration magnitude will be tracked on each articulator during the optimization process in order to analyze a possible emerging maturation (i.e. the sequence of freezing and freeing of degrees of freedom over time).

Practically, we extract the eigenvalues of the gaussian, smooth them over a few updates, and normalize them to sum to one. This provides the relative exploration values for each articulator in order to detect which one is mainly explored by the algorithm and when. Finally we can sort these magnitudes according to their maximum of relative exploration, and get a recruitment order over the whole learning experiment.

Figure 7 displays the exploration magnitude tracking when the system attempts to reach the vowel /a/. We observe two major peaks of exploration. First, around the $12^{th}$ update, exploration is almost exclusively focused on the articulator P1. Second, around the $26^{th}$ update, exploration is focused on articulator P5. Negligating the other exploration peaks due to their small magnitudes (we set a treshold at 0.3), the recruitement order here is therefore P1 then P5.

## 3.2 Vocal emergent maturation

Extracting exploration magnitudes during the reaching attempts toward different goals in the vocalic triangle allows to analyze in detail possible maturation phenomena in our model. Figure 8 displays the result in two other particular simulation attempting to reach the three vowels /i/ and /u/ (a simulation for the vowel being already displayed on Figure 7, bottom). Regarding the vowel /i/, we observe 4 phases of recruitment. The 3 first ones ($P1$,$P3$ and $P2$ maxima) appear together with a relatively high total exploration magnitude. The last phase ($P6$) is more an adjustment phase. Regarding the vowel /u/, we also observe 4 phases of recruitment. The first and last ones ($P1$ and $P2$ maxima) appear together with a low total exploration magnitude and can therefore be considered as non relevant. The second and third ones ($P1$ and $P7$ maxima) appear together with a high total exploration magnitude.

Together with the exploration magnitude for the vowel /a/ (Figure 7, bottom), these results show that the most recruited parameters are those controlling the distinctive phonetic features of the reached vowel. Producing a /a/ necessitate a vocal tract constriction at the level of the throat, what is typically obtained by opening the jaw and placing the tongue in a back position, hence the recruitment of $P1$ and $P5$. Producing a /i/ necessitates a constriction of behind the teeth, what is typically obtained with a rather closed jaw and the tongue in the front position, hence the recruiting of $P1$ and $P2$. Finally, producing a /u/ necessitates construction back in the palate, what is typically obtained by opening the jaw and placing the tongue in a high-back position, hence the

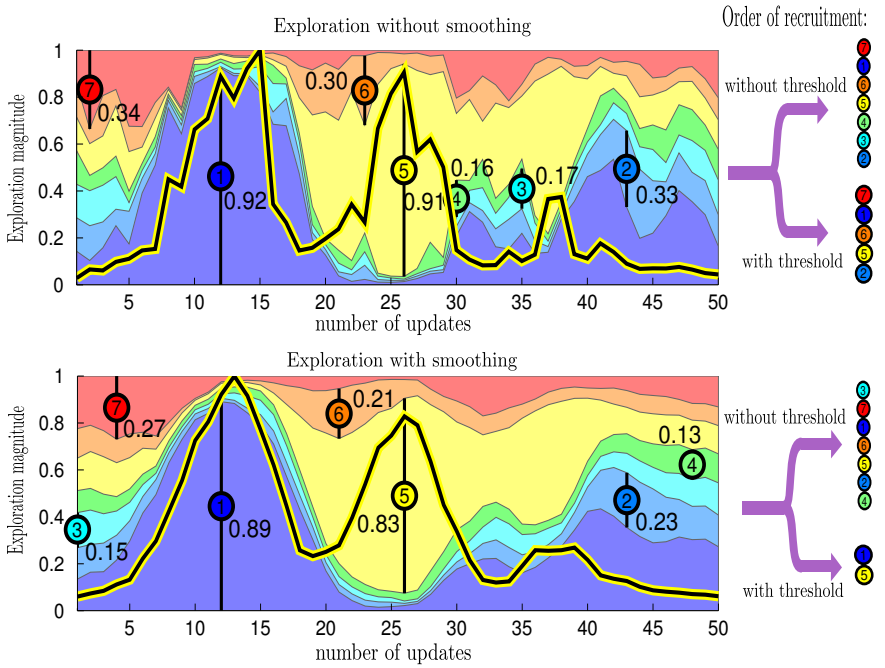



Figure 7: Exploration magnitudes when the system attempts to reach the vowel \a\. Each parameter exploration is associated to a color, its maximum being depicted by a colored circle. The total exploration magnitude, $\sum_{m=1}^{7} \lambda_m$, is represented by the yellow curve, scaled to sum at 1. Smoothing magnitudes over a time window and filtering their maximum values with a treshold allow us to extract the relevant aspects of the learning process. We set the treshold at 0.3, meaning that exploration magnitudes below this values will not be considered in the recruitment order.

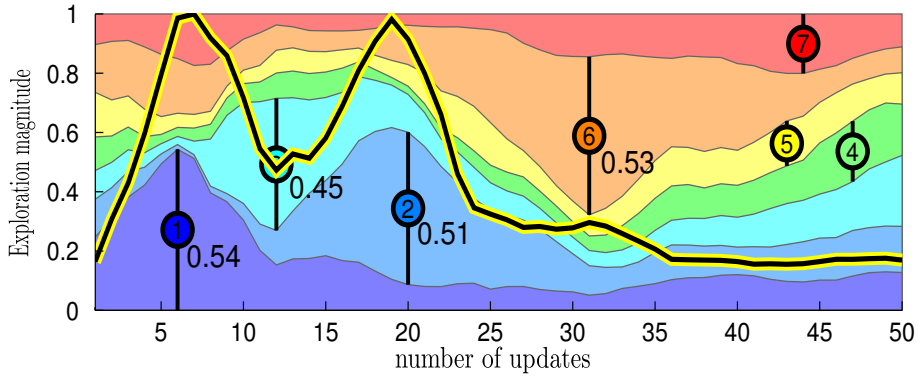
recruiting of $P1$ and $P7$ (CMF: not coherent, to check).

It therefore seems that attempting to reach different vowel goals results in different recruitment orders. Moreover the recruited articulators for a given vowel are mainly those involved in the production of the associated phonetic features.

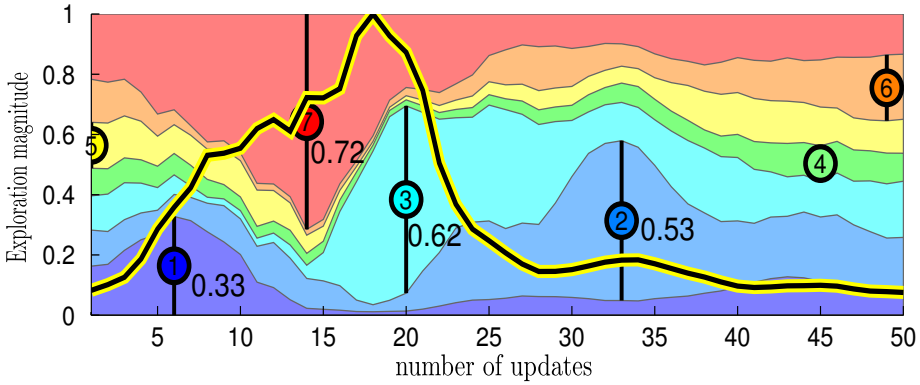
To assess the significance of these results more thoroughly, Figure 9 displays the recruitment order frequencies over a set of vowel goals drawn randomly in the vocal triangle. We observe a strong tendency to recruit $P1$ first: it is the case in almost 50% of the simulations. Our interpretation is that this is due to the following reasons:

- $P1$ mainly corresponds to a jaw movement (Figure 2).
- The jaw position has a strong influence on the $F1$ frequency.
- $F1$ has an important influence on auditory perception [17], as we modeled it using a scaling factor.

Therefore, the most important articulator to reach a variety of vowel goals is the jaw, because it allows to cover the whole range of $F1$ values, which has a




(a) Exploration magnitude analysis on a particular simulation attempting to reach the vowel \i\. Same convention as Figure 7, bottom.




(b) The goal is here the vowel \u\.

Figure 8: Recruitment representation over 3 goals. Only the values which are above the threshold of 0.3 are represented.

strong influence on the cost function because of its particular weighting (three times more than $F2$ based on psychoacoustic considerations [17]).

These results provide an original interpretation to explain the predominant role of the jaw in human speech evolution and acquisition . It has been proposed that this particular role could be due to evolutionary precursor behaviors involving jaw cycling such as mastication and ingestion as well as non-human primates communicative gestures such as lipsmacks and tonguesmacks (see the introduction). Here we show that pure learning mechanism could also be involved due to the particular role of the jaw and the first formant in vowel production.

The next section provides a sensitivity analysis emphasizing how the stochastic optimization process allows a dynamical “freezing” and “freeing” of the different articulators according to their impact on the cost function minimization.

### 3.3 Sensitivity overview

Figures 7 and 8 display different recruitment orders according to the vowel the optimization process is attempting to reach. To understand this process in more

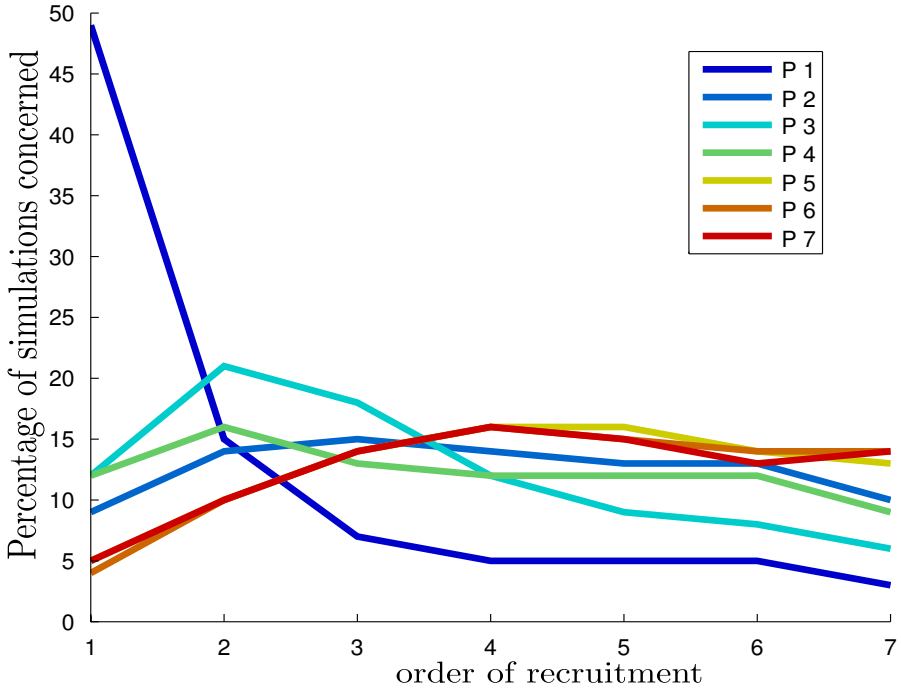



Figure 9: Frequencies of recruitment of each parameter over the whole vocalic triangle, and in each position.

detail, we rely the dynamical "freezing" and "freeing" of articulators to their relative influence on the cost function. For the sake of simplicity, we restrict our analysis to the role of $P1$ (mainly controlling the jaw position) and $P3$ (mainly controlling the tongue front/back position) on the cost funtion to reach the vowel /u/. We chose these articulators because $P1$ mainly influences $F1$ values whereas $P3$ mainly influences $F2$ ones, allowing to better distinguish their respective roles to reach the goal. Then we compute the partial derivatives of the cost function with respect to $P1$ and $P3$ at each update of the optimization algorithm. The optimization is run only on those articulatory parameters, the other ones being fixed to a position allowing the reaching of the vowel /u/. The partial derivatives are computed in an empirical manner, by looking at the cost function variations induced by small variations of the articulatory parameters.

Thus we are able to compute the respective influence of $P1$ and $P3$ on the cost function at each update and consequently to assess if $PI^{\text{BB}}$ is mainly recruiting the parameter displaying the greater influence. If it is confirmed, this will support the hypothesis that $P1$ is recruited first because of its particular influence on $F1$.

Figures 10a to 10f show this respective influence at regular time steps during a particular simulation. The auditory goal is the vowel /u/. We observe that the starting neutral configuration is in a thin region where $P1$ has much more influence than $P2$ on the cost function, i.e. where the ratio has a high value. Figure 10g shows that exploration mainly appears on $P1$ until the $9^{th}$ update. This is therefore coherent with our hypothesis that larger exploration

corresponds to higher cost function influence: $P1$ indeed both displays a higher influence on the cost function and to a larger exploration magnitude from the $1^{st}$ to the $9^{th}$ update. From this latter, the vocal tract is in a in configuration displaying a much greater influence of $P3$ rather than $P1$ on the cost function (in the blue ridge corresponding to a low ratio). Supporting our hypothesis again, we observe on Figure 10g a predominent exploration on $P3$ from the $9^{th}$ update to the end of the simulation.

This sensitivity analysis provides evidence that the stochastic optimization process will favor the recruitment of articulators having greater influence on the cost function minimization at a given update. In other words, exploration takes place on articulators which are the most useful to move closer to the auditory goal according to the current configuration of the vocal tract.

# 4 Discussion

This paper proposes an original hypothesis regarding the predominant role of the jaw in infant vocal development. Whereas this particular role has been sometime accounted as a derivation of prelinguistic behaviors in the course of human evolution, for example non-human primate orofacial communicative gestures, here we suggest that it could be a result of exploration strategies allowing the production of various auditory effects. This work took inspiration of a previous model showing that such exploration strategies implied by a stochastic optimization process provides a cognitive reason for the proximo-distal law of arm development.

For this aim, we developed a computational model using an articulatory synthesizer, movement generation and auditory perception processes, coupled with the stochastic optimization algorithm $PI^{BB}$ . We ran numbers of simulations where the system iteratively optimizes the reaching of various auditory vowel goals and performed various analyses on the underlying results: extraction of exploration magnitude and sensitivity analysis during the exploration process, as well as the order of articulator recruitment. These results show that the order of recruitment was dependent of the auditory goal to be reached and that, on average on various auditory goals, the jaw is predominantly and firstly recruited by the optimization process. Moreover, a sensitivity analysis suggested that the order of recruitment was determined by the relative influence of each articulator on the cost function at a given time step.

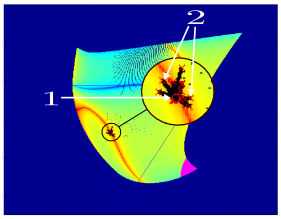
Ockam

- Citation Oller: canonical babbling is so robust that a number a forces probably act on it(our suggestion is just one possible force).

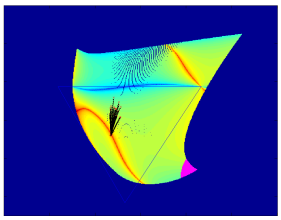
- ontogeny -¿ phylogeny

# References

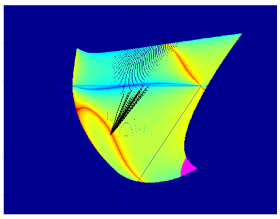
[1] G. Baldassarre. What are intrinsic motivations? a biological perspective. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–8, Aug 2011.



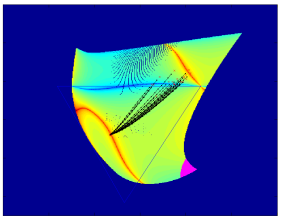

(a) Update 1: (1) the initial postion (2) the $K$ concurrent samples.



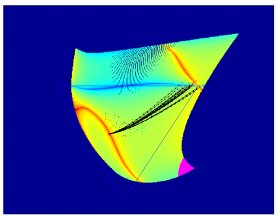
(b) Update 5



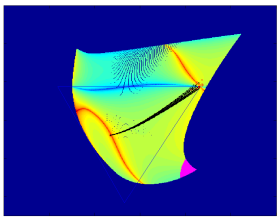
(c) Update 9



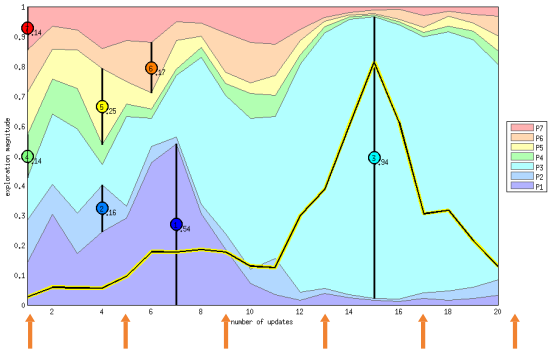
(d) Update 13



(e) Update 17



(f) Update 21




(g) Relative exploration values, where the exploration of $P3$ corresponds to the update at which the trajectories' end arrive near the blue ridge. The snapshot are signalized by the orange arrows.

Figure 10: Snapshots from a $PI^{BB}$ exploration, where the aim is the vowel /u/. The black lines represent the formants's trajectory encoded by the $K$ samples. At the beginning (update 1-7), the trajectory expands in the formant space, but do not follow the red ridge which would increase the cost. Note that the heat map displayed is computed for a particular configuration which corresponds to the aim, but not to the initial position.
Then the blue ridge is finded and $P3$ is recruited. This last until the aim is reached, even if there is a small detection of the red ridge at the update 13, but not strong enough to fork the exploration.

[2] A. Baranes and P.-Y. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 2012.

[3] A. Barto, S. Singh, and N. Chenatez. Intrinsically motivated learning of hierarchical collections of skills. In *Proc. 3rd Int. Conf. Dvp. Learn.*, pages 112–119, San Diego, CA, 2004.

[4] A. A. Ghazanfar and D. Y. Takahashi. The evolution of speech: vision, rhythm, cooperation. *Trends in cognitive sciences*, 18(10):543–553, 2014.

[5] A. A. Ghazanfar, D. Y. Takahashi, N. Mathur, and W. Fitch. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. *Current Biology*, 22(13):1176–1182, 2012.

[6] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[7] F. H. Guenther, S. S. Ghosh, and J. A. Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3):280–301, 2006.

[8] J. M. Iverson, A. J. Hall, L. Nickel, and R. H. Wozniak. The relationship between reduplicated babble onset and laterality biases in infant rhythmic arm movements. *Brain and language*, 101(3):198–207, 2007.

[9] P. MacNeilage and B. Davis. Motor mechanisms in speech ontogeny: Phylogenetic, neurobiological and linguistic implications. *Current Opinion in Neurobiology*, 11:696–700, 2001.

[10] P. F. MacNeilage. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21:499–511, 1998.

[11] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *Speech production and speech modelling*, 55:131–149, 1989.

[12] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer. Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*, 4(1006), 2013.

[13] P.-Y. Oudeyer, A. Baranes, F. Kaplan, and O. Ly. *Intrinsically Motivated Cumulative Learning in Natural and Artificial Systems*, chapter Developmental constraints on intrinsically motivated skill learning: towards addressing high-dimensions and unboundedness in the real world. Springer, to appear.

[14] P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1, 2007.

[15] P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

[16] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In J. A. Meyer and S. W. Wilson, editors, *Proc. SAB'91*, pages 222–227, 1991.

[17] M. Schroeder, B. Atal, and J. Hall. *Frontiers of Speech Communication Research*, chapter Objective measure of certain speech signal degradations based on masking properties of human auditory perception, pages 217–229. London Academic Press, 1979.

[18] J. Schwartz, L. Boë, N. Vallée, and C. Abry. The Dispersion-Focalization theory of vowel systems. *Journal of Phonetics*, 25(3):255–286, 1997.

[19] F. Stulp and P.-Y. Oudeyer. Adaptive exploration through covariance matrix adaptation enables developmental motor learning. *Paladyn. Journal of Behavioral Robotics*, 3(3):128–135, September 2012.

[20] A. Warlaumont. A spiking neural network model of canonical babbling development. In *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pages 1–6, 2012.

[21] A. Warlaumont. Salience-based reinforcement of a spiking neural network leads to increased syllable production. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE International Conference on*, pages 1–6, 2013.

[22] A. S. Warlaumont. Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38:64–95, 2013.