Canary Vocal Sensorimotor Model with RNN Decoder and Low-dimensional GAN Generator

Silvia Pagliarini, Arthur Leblois, Xavier Hinaut

Canary Vocal Sensorimotor Model with RNN Decoder and Low-dimensional GAN Generator

Silvia Pagliarini
Inria Bordeaux Sud-Ouest, Talence, France.
LaBRI, UMR 5800, CNRS, Bordeaux INP,
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
silvia.pagliarini@inria.fr

Arthur Leblois*
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
arthur.leblois@u-bordeaux.fr

Xavier Hinaut*
Inria Bordeaux Sud-Ouest, Talence, France.
LaBRI, UMR 5800, CNRS, Bordeaux INP,
IMN, UMR 5293, CNRS,
Université de Bordeaux, France.
xavier.hinaut@inria.fr

***Abstract*—Songbirds, like humans, learn to imitate sounds produced by adult conspecifics. Similarly, a complete vocal learning model should be able to produce, perceive and imitate realistic sounds. We propose (1) to use a low-dimensional generator model obtained from training WaveGAN on a canary vocalizations, (2) to use a RNN-classifier to model sensory processing. In this scenario, can a simple Hebbian learning rule drive the learning of the inverse model linking the perceptual space and the motor space? First, we study how the motor latent space topology affects the learning process. We then investigate the influence of the learning rate and of the motor latent space dimension. We observe that a simple Hebbian rule is able to drive the learning of realistic sounds produced via a low-dimensional GAN.**

I. Introduction

Vocal learning represents the ability to produce new sounds via imitation. In humans, vocal learning allows infants to learn to produce speech through the parallel development of speech perception and production ability [1], [2]. Among complex vocal learners, songbirds represent the most studied model organisms for vocal learning. Songbirds share with humans similar vocal development [3]. Vocal learning starts with a sensory learning phase, when infants and juvenile songbirds learn to discriminate the sounds they hear from conspecific adults. Then, vocal learning resumes with a sensori-motor phase during which the infants/juvenile start to produce their own vocalizations. Human babies start producing non-speech sounds (e.g., cries) considered as precursor of speech at birth, then produce vowel-like sounds around the third and seventh month of their life [4], [5] and produce canonical babbling after seven months [1], [5]. Vocal production in juvenile songbirds is also gradual: they start with a variable babbling behavior, then slowly adapt their vocalizations to incorporate elements of the tutor song and finally produce highly complex, stereotyped motifs in adulthood [6]. While humans and songbirds share analogous brain circuits for vocal learning [7], a circuit dedicated to song learning in birds facilitates the study of the underlying neuronal mechanisms [6].

The basic structure of a vocal learning schema involves three spaces (motor, sensory, perceptual), the motor control function, the sensory response function, and the learning architecture [8], [9]. The motor space contains the motor coordinates. The sensory space contains the real sounds. The perceptual space represents the encoding of the sounds categories: such perceptual categories can be seen as perceptual goals. The motor control function allows the production of sound. The sensory response function processes the sound and encodes it in a low-dimensional space (the perceptual space). Alternatively, the sensory response function can provide a reward to the model. The learning architecture defines the learning algorithm, and the exploration strategy. Several learning frameworks have been proposed to model vocal learning in humans and birds. A recent comparative review summarizes a wide set of models, their objectives and how the various components have been defined [9].

In the songbird literature, the motor control function has been often defined using a system of ordinary differential equations that model the anatomy of the syrinx (i.e., the birds' vocal organ) [10], or the features of sound [11]. Recently, generative networks have been introduced to solve tasks such as image, music, and speech generation or classification and have been used to investigate visual pathways in the brain [12]. The advantages of using generative neural networks are to obtain resemblance of the generated data with the real data from an uniformly distributed low-dimensional motor space.

We propose a canary sensorimotor model where the motor function and the sensory response function are implemented in a novel way. On the one hand, the motor function is implemented using a low-dimensional generator model obtained from a Generative Adversarial Network (GAN) generator. As shown in [13], such a model shows the ability of producing realistic sounds (canary syllables), and represents an alternative to previously proposed vocal tract models. On the other hand, the sensory response function is defined as a Recurrent Neural Network (RNN) classifier [14] implemented with the ReservoirPy library [15]. For this study, the classifier and the GAN are pre-trained. The connections between perceptual and motor spaces – that form the inverse model – are learned through activity-dependent plasticity. We check whether or not a simple Hebbian learning rule combined with random motor exploration are sufficient to build an inverse model between perceptual and motor representations of a 16-syllables canary

* Corresponding authors that co-supervised the study.

repertoire.

Section II introduces the components of the proposed vocal learning model. Section III shows the results, including the exploration of the structure of the motor latent space, and the influence of various conditions on learning. Section IV summarizes the advantages and the limitations of the model, and discusses possible perspectives to expand this work.

All the details of the implementation are available at github.com/spagliarini/canary-vocal-sensorimotor-model

II. Methods

The proposed model contains three spaces (perceptual, motor, sensory), a motor control function, a sensory response function, and an inverse model (Figure 1). Section II-A introduces the structure of the model. Section II-B details the learning algorithm in the inverse model, Section II-C the motor control, and Section II-D the sensory system. Finally, Section II-E contains the experimental setup.

A. General architecture

A one-layer perceptron models the connections between the perceptual space and the motor space, see Figure 1. The first layer $(P_1, .., P_{n_P})$ represents the perceptual space **P**. The second layer $(M_1, .., M_{n_M})$ represents the motor space **M**. At each time step t, the perceptual units are defined as a n_P-dimensional vector P_t, where n_P represents the size of the perceptual layer. The motor units are defined as a n_M-dimensional vector M_t, where n_M represents the number of motor parameters. The synaptic weights at t of the inverse model describing the connections between the motor and the perceptual space are defined by matrix W_t. Given a motor pattern M_t, the motor control function G provides a real sound S_t (i.e., an element of the sensory space). The sensory space **S** is the domain of the sensory response function: at each time step t, the sensory response A is a function of the actual sounds produced S_t (i.e., $P_t = A(S_t)$).

B. Inverse model and Hebbian learning

The aim of the inverse model I is to learn the link between perceptual and motor space. At each time step t, a motor pattern M_t is drawn from $[-1, 1]^{n_M}$ and enables a sensory response P_t. Learning is driven by the Hebbian learning rule

$$\Delta W_t = \eta M_t P_t, \quad (1)$$

where W_t represents the synaptic weight and η the learning rate. The synaptic weights $W_{t=t_0}$ are initialized as random uniform values and vary according with Equation 1 until time $t = t_f$. The motor space is explored using random exploration.

C. Motor control

As motor control function G, the generator part of a n_M-dimensional GAN (where $n_M \in \{1, 2, 3, 6\}$) is used to produce sounds: e.g. during motor exploration or when we want to evaluate the architecture during training. The generator model has been previously obtained by training WaveGAN [17] on a dataset of canary syllables [18]: it is able to provide syllables similar to the real ones [13]. As one could expect, one syllable class can be produced using multiple motor configurations.

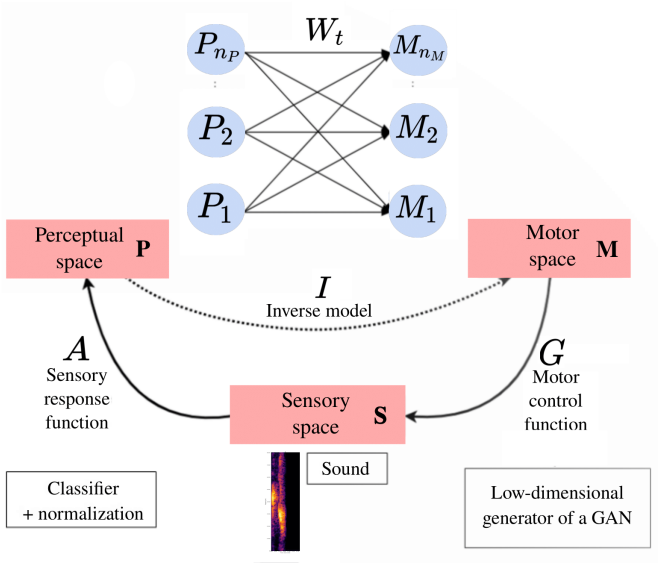


Fig. 1: **Vocal learning model schema.**. The model contains three spaces: the perceptual space, the motor space, and the sensory space. A one-layer perceptron connects the perceptual space $(P_1, .., P_{n_P})$ to the motor space $(M_1, .., M_{n_M})$. W_t represents the synaptic connections between perceptual and motor spaces at each time step t. The motor control function G is a n_M-dimensional generator of a GAN that enables sound production. At each time step t, the sensory response P_t is a function of the actual sound production (i.e., $P_t = A(S_t)$, where S_t is the actual sound produced by G at time step t. The sensory response function is composed of a RNN-classifier and a normalization layer to restrict the obtained activation in the interval $[0, 1]$.

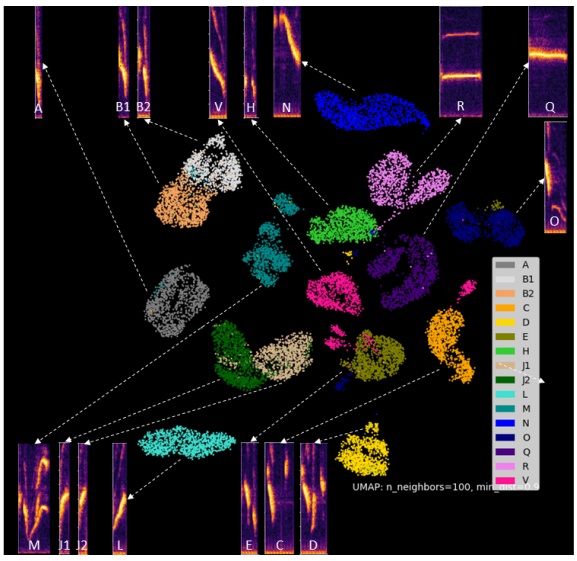


Fig. 2: **Repertoire and training dataset.** Low-dimensional representation of the training dataset obtained using Uniform Manifold Approximation and Projection (UMAP) [16]. Each point represents the spectrogram of a syllable, and each colored cluster represents a class $\mathbf{R_i}$. A template syllable of each class is highlighted with the corresponding spectrogram (an arrow connects each cluster to the corresponding template). Image from [13].

D. Sensory system

The sensory system is able to detect syllables belonging to a canary repertoire **R** composed of N different syllable classes

and an *alternative* class *X*. We define the vocabulary $\mathbf{V}$ as the set containing all classes $\mathbf{V} = \mathbf{R} \cup \{X\}$. The repertoire $\mathbf{R}$ is composed by N classes R_i for $i \in [1, N]$. The class *X* represents distorted syllables that the classifier is not able to assign to a class of the repertoire (e.g. inter-syllabic transitions) and has been built using bad WaveGAN generations (e.g. early epochs of GAN training) and white noise [13]. Note that the generator model G was trained with R (not V).

In our experiments, the dimension of the perceptual space is given by $n_P = N$, as the repertoire contains N different syllable classes. The output of the sensory response function has always dimension $N + 1$ since class X is included. The sensory response function A (see Fig 1) is made of two components. First, a *RNN classifier* takes a sound S_t (a syllable) as input and provides a distribution of classes - the probability of S_t to belong to each class of the repertoire [14]. Then, a normalization layer scales the obtained activation to values in $[0, 1]$. As the RNN-classifier produces a distribution of activations that depends on the syllables class (some are often very high and some often very low), we need to normalize the classifier outputs in order to have a more balanced distribution among syllable classes. This will reduce the bias of the perceptual part of the architecture during learning. Hereafter, we explain the normalization that we performed. First, we precomputed the 95-percentile for each class of the vocabulary as follows:

1) for all the precomputed motor patterns M_i ($16k$ in total) the sensory output S_i has been generated and processed by the classifier;
2) from the classifier outputs, we selected[1] the peak of the most active output $Y_M(S)$;
3) for each class in the repertoire, we computed the 95-percentile $p95$ from all $Y_M(S)$ obtained (from all generated motor patterns in step 1).

Then, during motor random exploration at time step t, the perceptual activation $P_t^{R_i}$ of a sound is computed, for each class $\mathbf{R_i}, i \in [1, n_P]$, as the maximum activation Y_M divided element-wise by the global 95-percentile obtained for class $\mathbf{R_i}, i \in [1, n_P]$:

$$P_t^{R_i} = A^{R_i}(S) = \begin{cases} 1 & if \quad Y_M^{R_i}(S) > p95_{R_i}, \\ \frac{Y_M^{R_i}(S)}{p95_{R_i}} & otherwise, \end{cases} \tag{2}$$

where $\mathbf{R_i}$ represents each class of the vocabulary, $Y_M^{R_i}(S_t)$ represents the maximum activation of sound S_t provided by the classifier for $\mathbf{R_i}$, and $p95_{R_i}$ represents the precomputed global 95-percentile of $\mathbf{R_i}$. Such definition for auditory activation leads to $\sim 5\%$ of the motor latent space leading to the production of a given syllable class. This final vector corresponds to the activation of the perceptual layer of the inverse model.

[1]Please note that one generated syllable has a length of 1 sec. and in average the longest syllables have a duration of 300 ms. Thus, we restricted the output activity to the first 500 ms before selecting the maximum activation.

E. Experimental setup

The training dataset used for this work has been obtained from a larger set of adult canary recordings [18]. In [13], WaveGAN was trained with the syllables classes containing enough samples leading to a $n_P = 16$ classes repertoire R. Syllable examples of each class $\mathbf{R_i}$ can be seen in Figure 2. The representation obtained using Uniform Manifold Approximation and Projection (UMAP) [16] shows the clusters that compose the training dataset.

During learning, at each time step t, a random motor vector M_t (i.e. n_M-dimensional vector taking random values in $[-1, 1]$) is given as input to the motor control function G. This generator model produces a syllable sound $S_t = G(M_t)$, and the sensory response function computes the corresponding perceptual representation P_t. For simplicity, a set of $16k$ motor vectors, the corresponding syllables waveforms, and the corresponding perceptual representations have been precomputed beforehand. The inverse model synaptic weights $W \in M^{16 \times n_M}$ are initialized as $W_{t_0} \in U[-0.001, 0.001]$.

We studied (1) the influence of the motor space dimension on the learning, (2) the influence of using different learning rates ($\eta = 0.01$ versus $\eta = 0.1$) on the learning. We stopped the learning after $3k$ time steps. We trained three model instances for each condition keeping fixed the initial synaptic weights W_{t_0} between motor and perceptual spaces, but varying the random motor exploration across instances. Indeed, as the initial weights are close to zero, they have negligible influence on the learning. Conversely, the random motor exploration affects more the initial phases of learning.

F. Evaluation

To evaluate the inverse model during learning, we need to see if each given perceptual neuron (each corresponding to one syllable class) can activate the motor layer and consequently produce the sound of the correct syllable. Thus, we evaluate (for each syllable) what would be the evoked activation of the perceptual layer obtained via the full sensorimotor loop: we evaluate every 15 time steps. To do so, we (1) activate one given perceptual unit i of the inverse model $\hat{P}^{R_i}$ (i.e. the ideal perceptual pattern of syllable i), (2) record the motor pattern $\tilde{M}_t^{R_i}$ produced through W_t, (3) use this motor pattern to generate a sound S_t through the GAN generator, (4) record the evoked perceptual activation $\tilde{P}_t^{R_i}$. If the evoked activation is at 1 (i.e. the perceptual goal is reached), then the syllable is considered to be learned by the inverse model. Figures 7 and 8 show these evoked perceptual activations for each given syllable. For each class $\mathbf{R_i}$ ($i \in [1, N]$) of the repertoire, the ideal perceptual activation is encoded as one-hot vector. During the perceptual evaluation, the motor pattern at time t is defined as $\tilde{M}_t^{R_i} = f(W_t \hat{P}^{R_i})$, where W_t is the matrix of the synaptic weights at time step t, $\hat{P}^{R_i}$ is the ideal auditory pattern of class $\mathbf{R_i}$[2], and f a piecewise

[2]We do no use goal-babbling here, but if we were to do so, $\hat{P}^{R_i}$ would be the perceptual goal driving the exploration to produce target syllable R_i.

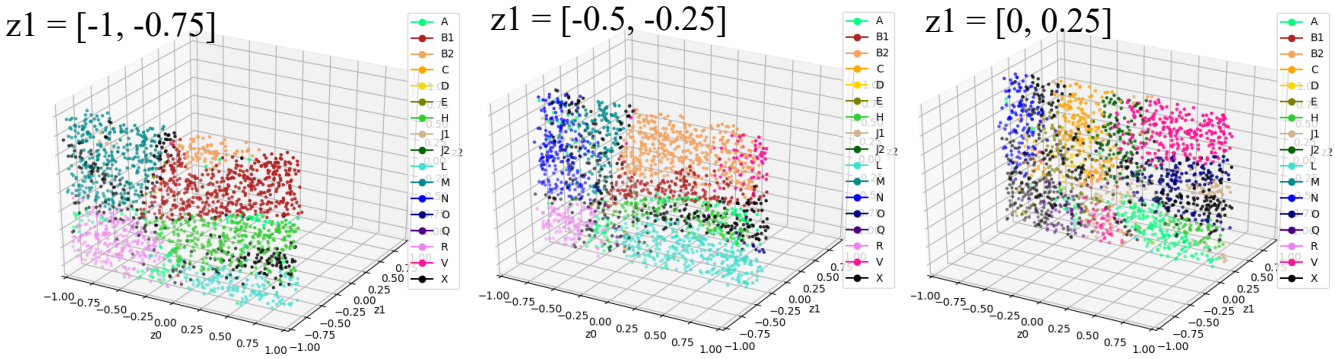


Fig. 3: **Three-dimensional motor latent space**. Each point represents a 3-dimensional motor pattern M belonging to the motor latent space. Each motor pattern takes values in $[-1, 1]^3$. Each figure represents a slice of the three dimensional cube: z_1 component has been fixed to a given interval (e.g., $z_1 \in [-0.25, 0]$ in the right panel), while components z_0 and z_2 are free. Black points represent motor patterns that lead to the production of syllables belonging to class *X*, each other color correspond to a class of the repertoire. These patterns are often making junctions between patterns which lead to the production of two different syllables of the repertoire.

	A	**B1**	**B2**	**C**	**D**	**E**	**H**	**J1**	**J2**	**L**	**M**	**N**	**O**	**Q**	**R**	**V**	**X**
$n_M = 1$	5.02	7.19	4.79	6.38	5.48	5.04	4.86	5.86	5.82	4.87	4	4.7	6.85	**1.31**	2.8	7.39	**17.64**
$n_M = 3$	5.43	6.84	6.64	5.38	5.89	5.23	4.76	5.4	5.4	5.77	5.12	3.81	4.09	**1.92**	3.64	6.03	**18.64**

TABLE I: **Motor latent space composition.** Percentage of syllables belonging to each class for motor dimensions 1 and 3.

linear function[3] that restricts the values of $\tilde{M}_t^{R_i}$ in the interval $[-1, 1]$. By construction, the input space for the generator lies in $[-1, 1]$.

III. Results

This section shows the results, including the exploration of the structure of the motor latent space, and the influence of various conditions on learning.

A. Structure of the motor latent space

We remind that at each time step t, the model explores a motor pattern M_t which enables the generations of a sound and, consequently, the corresponding sensory response. Each motor pattern is randomly chosen from a predefined set containing $16k$ motor patterns, and is a n_M-dimensional vector in $[-1, 1]^{n_M}$. In Figures 3 and 4, we observe that motor patterns are organized in clusters in the motor latent space, when coloring each motor pattern with the class of sound it is generating. We observe that transitions between two clusters (i.e. between two different classes) are often characterized by the presence of patterns that lead to the production of syllables belonging to class the alternative class X.

Interestingly, the structure of the latent space can help understanding the learning dynamics (see Figures 7 and 8). For instance, when $n_M = 3$, syllable O (blue navy dots in Figure 3) is characterized by a sparse "cluster" intermixed with the alternative classe. This may result in O being a more challenging syllable to learn for the model. This is coherent with the fact that, in biological and robotic systems, some gestures may be easier to learn than others. Additionally, this

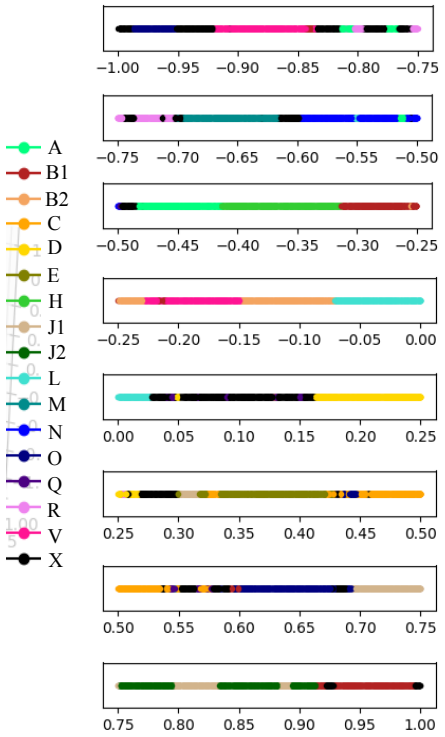


Fig. 4: **One-dimensional motor latent space**. Structure of the motor latent space when $n_M = 1$: each point represents a 1-dimensional motor pattern M. Legend is the same of Figure 3. Similarly to the 3-dimensional space (see Figure 3), class *X* is spread in space and often makes the junction between two different clusters. It shows that the GAN is interpolating in-between existing classes. The pairs of syllables B1/B2 and J1/J2 are very similar, which explains why they are intermixed in the latent space.

representation shows that the latent space does not contain a

[3]The piecewise linear function we use in this work is defined as follows: $\forall j \in [0, n_M]$, $\tilde{M}_j^{R_i} = 1$ if $\tilde{M}_j^{R_i} > 1$, $\tilde{M}_j^{R_i} = -1$ if $\tilde{M}_j^{R_i} < -1$, and $\tilde{M}_j^{R_i} = \tilde{M}_j^{R_i}$ otherwise.

neutral position: any point of the space will generate a sound that can be classified. This is why the sensory response can be greater than zero at the beginning of learning for some syllables (i.e. when the synaptic connections are weak and close to 0). In particular, for motor patterns close to the origin $\mathbf{O} \in \mathbf{R}^3$, syllables *J2* and syllables *J1* are produced for $n_M = 3$. Respectively, syllables *L* and syllables *J2* are produced for $n_M = 1$.

Finally, the motor latent space is not a balanced space: syllables are not equally represented (in percentage) in the motor latent space (Table I). Extreme values are represented in bold: for both $n_M = 1$ and $n_M = 3$, class *Q* is the least represented whereas class *X* is the most represented. In particular, the percentage of syllables belonging to class X represents $\sim 18\%$ of the total amount: this introduces learning difficulties because these syllables should not be learned. Both for $n_M = 1$ and $n_M = 3$, patterns producing class *X* syllables (i.e. black points in Figures 3 and 4) lies at the junction of repertoire syllables. Such a distribution of *X* syllables shows that the generator model is able to generalize and produce not only syllables belonging to the training dataset but also intermediate syllables interpolating from different classes [13].

B. Evolution of learning and sounds produced

For each class $\mathbf{R_i}, i \in [1, n_P]$ of the repertoire $\mathbf{R}$, the perceptual activation P^{R_i} across time is described by Equation 2 and takes values in $[0, 1]$. The learning is considered achieved when P^{R_i} stabilizes at 1. For most syllables, $P^{R_i} = 0$ at time $t = t_0$. Then, oscillatory dynamics can be observed until the activation stabilizes, as can be observed in the left panel of Figure 5 for syllable *R*. A similar example can be found in the right panel of Figure 5 for class *B1*. Although realistic syllables can be produced since the beginning (due to the absence of neutral position), the random motor exploration allows to produce the correct syllables towards the end of the training (bottom panel of Figure 5). One can notice that at $t = 0$ the same syllable is produced both for class *B1* and *R*: indeed, as mentioned in Section III-A, there is no *neutral position* for the motor pattern. Instead, syllable *J1* is produced. Towards the end of the training, the produced syllables become similar to the corresponding target in the repertoire.

In Figure 6, we see that the sounds produced, when the perceptual activation is higher than a certain threshold $ths_P = 0.99$, are stable for the majority of the classes. Empty boxes mean that the perceptual activation never crosses ths_P (for syllables *O* and syllable *B2*). In the top-left panel of Figure 6, we see that class *A* represents an exception with respect to the other classes: syllables belonging to other classes (in particular, to class *R*) influences the mean spectrogram[4]. In Figure 3, we can see that *A* (lime green points) and *R* (light pink points) can be blended with alternative class *X* in the motor latent space. Thus, it is probable that some sounds generated by the GAN are actually interpolations between *A* and *R* which are classified as *A*[5]. This is a potential explanation for the mixed shape of the mean spectrogram of class *A*.

C. Evolution of learning and learning rate

A learning rate of $\eta = 0.1$ (red lines in Figure 7) can induce faster changes in the synaptic weights (i. e., in W_t) with respect to $\eta = 0.01$ (blue lines in Figure 7). Nevertheless, both for $\eta = 0.1$ and $\eta = 0.01$, the perceptual activation increases and reaches the optimal plateau (a value of 1) for 14 syllables over 16. Moreover, one can expect that syllable *B2* can be learned if a longer simulation is performed. Some syllables, like C and H have a decay after having reached the plateau. Such decay is due to the fact that the learning is driven by a simple Hebbian learning rule which is not expected to converge (due to the absence of normalization). Thus, a decay could also be expected for all syllables if the simulation would go on, except for the syllables that are at the boundaries of the motor latent space (-1 or 1 for each coordinate). That is, if the time is long enough, we expect a decay for all the classes.

D. Evolution of learning and motor space topology

The structure of the motor latent space has an influence on the learning. Observing both (Table I) and Figure 7, one can see that even if syllable *Q* is the least represented the model is able to reach $P^Q = 1$ during learning. Alternatively, the model struggles in learning syllable *O*: although it is well represented in the motor latent space. As mentioned in Section III-A, the "cluster O" (blue navy points in Figure 3) seems to introduce a challenge for the model because it is not convex[6]. Clear convex clusters are probably not necessary for the simple learning rule we used, but convexity of the motor space helps the learning.

E. Evolution of learning and motor space dimension

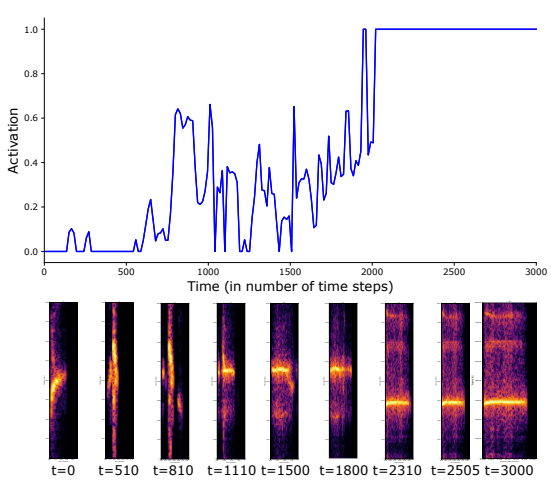
A higher motor dimension allows the model to learn a higher number of classes of the repertoire. As we saw previously, a 3-dimensional motor space (yellow lines in Figure 8) allows the learning of almost all the syllables. A similar behavior can be observed when $n_M \in 2, 6$ (respectively, blue and blak lines in Figure 8). Alternatively, a 1-dimensional motor space (red lines) prevents the learning of most classes of the repertoire: the perceptual activation never reaches one, but for syllables *J2* and syllable *L*. A higher learning rate does not result in the learning of a higher number of classes (see Figure 7). The 1-dimensional null motor space produces a syllable classified as *L* (light green blue points in the fourth panel of Figure 4).[7]

[4]Shape features of syllable *A* can be seen in Figure 2.

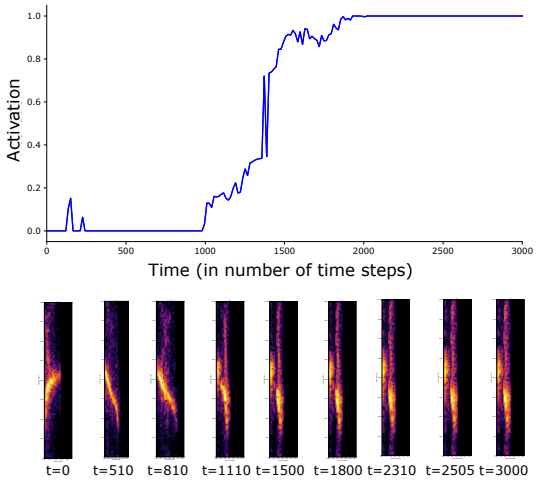
[5]Remind that the perceptual space P does not include the alternative class X.

[6]If we were to take an intermediate point between two random points of the "cluster O", this intermediate point may not lie inside cluster O.

[7]However, it is not because the produced sound is classified as *L* that the perceptual activation should be necessarily at 1 (i.e. indicating a well produced sound similar to a real canary syllable *L*): the colored dot obtained by the classification only indicates to which class this sound is the closest. Indeed, as the learning goes on, the perceptual activation decays quickly.



(a) Syllable R



(b) Syllable $B1$

Fig. 5: **Evolution of sound produced during learning for syllables *R* and *B1***. (top) Evolution of the sensory response activation P^R of unit R (left) and B1 (right) obtained during one instance of training. (bottom) Evolution of the corresponding sounds produced over time for 9 selected time steps. Parameter values: $n_P = 16$, $n_M = 3$, $\eta = 0.01$, $W_{t_0} \in U[-0.001, 0.001]$, $t_f = 3000$.

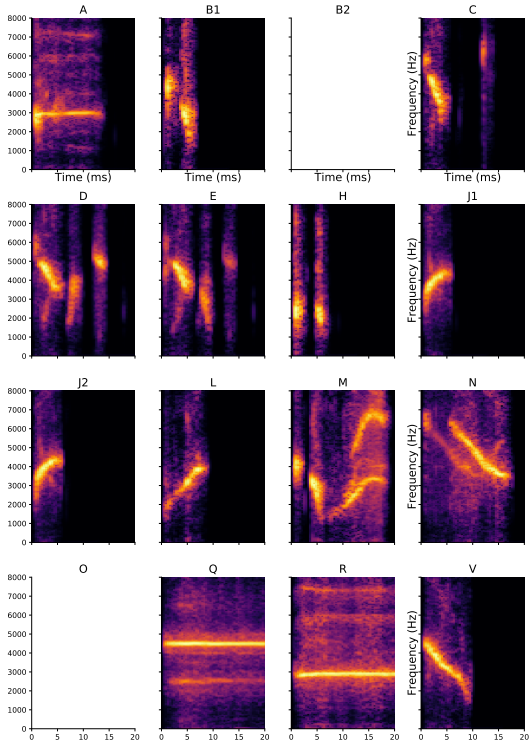


Fig. 6: **Syllables produced by the learning model**. Mean spectrogram obtained from all the sounds generated when the perceptual activation is higher than 0.99 during learning (i.e., if a sound S_t produced at a certain time is such that $P^{R_i} = A^{R_i}(S_t) \geq 0.99$). An empty box in panel means that for all the produced sounds of that class we obtain $P^{R_i} = A^{R_i}(S) < 0.99$, $\forall S$, $\forall t_0 < t < t_f$. Parameter values: $n_P = 16$, $n_M = 3$, $\eta = 0.01$, $W_{t_0} \in U[-0.001, 0.001]$, $t_f = 1500$.

IV. Discussion

We built a vocal learning model with a full action-perception loop [9]. The aim of the model is to learn a repertoire of 16 different classes of canary syllables. The motor space is a low-dimensional latent space obtained from training WaveGAN [17] on a dataset of canary syllables [13]. The motor control function is a generator model that enables the production of syllables resembling real recordings. The sensory space is the actual produced sound (and not a spectrogram or formants as it could happen in other models [9]). The sensory response function encodes the sound in a rather low-dimensional space, i.e. the perceptual space which plays the role of a goal space. We used the normalized output of a reservoir-based classifier [14] to model the sensory response function. The learning of the inverse model between the perceptual space and the motor space is driven by a simple Hebbian learning rule and an motor random uniform exploration. We tested how the learning is influenced by (1) different learning rates, (2) different motor space dimensions and (3) the structure of the motor latent space. A higher motor space dimension allows the learning of a higher number of classes of the repertoire (Figure 8). Similar performance can be observed from the models using the 3- and 6-dimensional GAN generators, suggesting that adding more than 3 dimensions does not help to enhance the learning. Interestingly, the structure of the motor latent space can reflect which syllables are more challenging for the model. Moreover, some syllables show a synchronous behavior when learned using different latent space dimensions (see syllable *M* in Figure 8 for dimension 3 - yellow line and 6 - black line): one can hypothesize that such a syllable has a similar distance from the origin in both latent spaces. Another hypothesis could be that such syllable is contained in a subspace of the latent space that co-exists

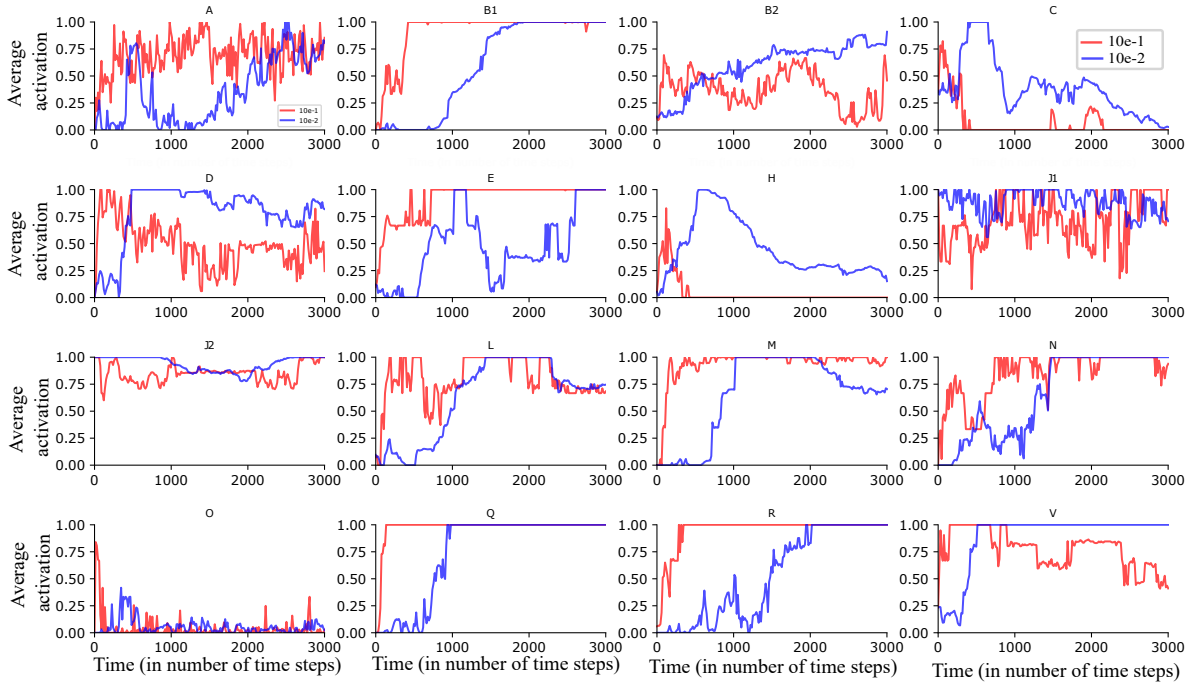


Fig. 7: **Evolution of learning: influence of the learning rate for** $n_M = 3$. Evolution of the perceptual activation for $\eta = 0.1$ (red lines) and $\eta = 0.01$ (blue lines). Each line represents the average activation obtained from 3 different instances of training. For all the classes except *O* and *B2*, the perceptual activation increases (more or less sharply) and reaches, even if does not stabilizes at, a value of 1. A higher learning rate results in faster but more unstable learning dynamics. Parameter values: $n_P = 16$, $n_M = 3$, $W_{t_0} \in U[-0.001, 0.001]$, $t_f = 3000$.

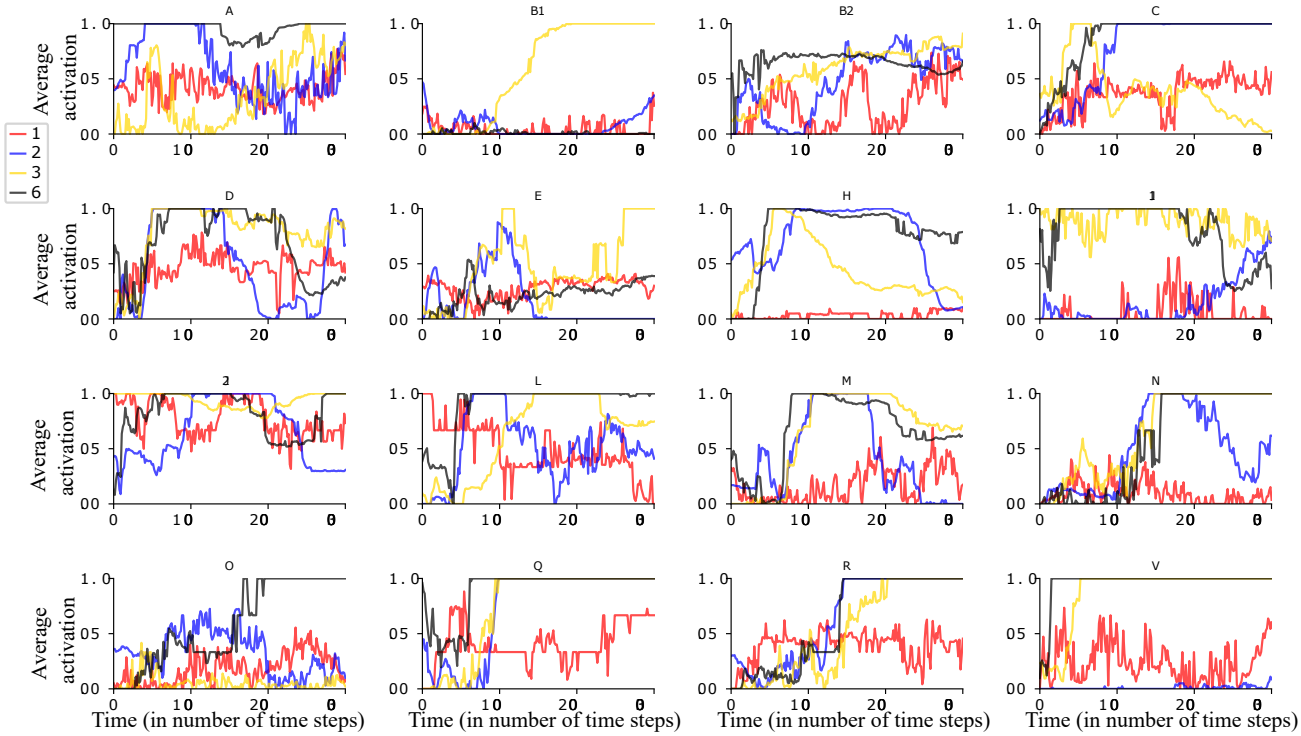


Fig. 8: **Evolution of learning: influence of motor space dimension**. Evolution of the perceptual activation for $n_M = 1$ (red lines), $n_M = 2$ (blue lines), $n_M = 3$ (yellow lines) and $n_M = 6$ (black lines). Each line represents the average activation obtained from 3 different instances of training. The perceptual activation remains generally low for $n_M = 1$ (red lines): it never reaches 1 but for syllables *J2* and syllable *L* (starting position). Alternatively, the perceptual activation almost always enables learning for $n_M = 2$ (blue lines), $n_M = 3$ (yellow lines), and $n_M = 6$ (black lines). Eventually, the stability drops after a certain time. Parameter values: $n_P = 16$, $\eta = 0.01$, $W_{t_0} \in U[-0.001, 0.001]$, $t_f = 3000$.

both for dimension 3 and 6. Further studies are needed to investigate the structure of higher dimensional latent spaces and compare within different space dimensions. The sparsity and probable non-convexity of the syllable clusters *A* and *O* are probably what is inducing an approximate learning (for A) or no-learning (for O) (see Section III-D).

Several modelers have proposed dynamical systems to model the motor control function in songbirds: in such case, the motor space describes the time-dependent motor articulations parameters which control the dynamics of the syrinx (e.g., air pressure, syringeal labial tension) [11], [19], [10], [20]. Instead of using a dynamical system that would produce sounds of approximate realism, we implemented the motor function using the generator part of low-dimensional GAN. Biologically, the GAN would represent a premotor layer rather than the control parameters of a vocal organ. Such a model learns well how to produce syllables resembling the training data (see Figure 2) from a low-dimensional latent space. We focused on a low-dimensional space, because we showed previously that a high dimensional motor space could result in slower learning convergence [21]. Although the training and the evaluation of a generator model are not trivial, once the model has been validated, it represents a powerful computational tool to produce realistic sounds. At the same time, the latent space (i.e., motor space) of GANs is redundant by construction: very similar sounds (i.e., syllables belonging to the same class of the repertoire and impossible to distinguish acoustically) can be produced by several latent vectors (i.e., motor patterns). As a consequence of the motor space redundancy, the target of the model is not a motor target but rather a perceptual target. For this reason, we did not introduce a normalization in the learning rule like we did in [21]. This choice is motivated by the fact that the motor space of the GAN cannot be normalized.

The sensory response function models the encoding of the sound in the birds' brain. The categorical classification provided by the classifier qualitatively describes the response that young birds develop in highly auditory area when they are memorizing the song [22].

A simple Hebbian learning rule allows the models to learn but does not prevent divergence after a critical time t_{critic}. The value of t_{critic} is syllable-specific (Figure 7) and depends on the learning rate. A higher learning rate results in faster learning dynamics and, thus, in an earlier t_{critic}. A simple stopping criteria could solve the decay problem. The introduction of a reinforcement signal could (1) speed the learning and (2) help the learning to stabilize after having reached the optimal plateau (i.e. the region of the motor space that enables the production of the correct perceptual goal). One could test how the introduction of a *neutral position* influences the learning. The model could be forced to use a syllable belonging to class *X* as initial position. The influence on the learning of the initial condition could then be tested. Moreover, a goal-directed strategy (e.g. goal babbling) could be used to enable the learning of the more challenging syllables (e.g. *O* or *A*).

In future work, we believe this vocal model could be extended to learn full songs (i.e. sequences of syllables) instead of single syllables. Relying on reservoirs for the sensory response function is a good choice for such extension, as it was shown that reservoirs can handle different levels of abstraction for language-like inputs [23]. A hierarchical architecture processing raw sounds but also syntactic representations [24] could be necessary to learn full songs.

AKNOWLEDGMENTS

We would like to thank Catherine Del Negro, Aurore Cazala and Juliette Giraudon for the canary recordings, and Nathan Trouvain for the classifier. We also thank Inria for the CORDI-S PhD fellowship, the ANR for grant ANR-16-CE37-0020-01, and the LabEx BRAIN for the PhD extension.

REFERENCES

[1] PK Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831, 2004.

[2] P Kuhl. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857, 2000.

[3] AJ Doupe and PK Kuhl. Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, 22(1):567–631, 1999.

[4] DK Oller. *The emergence of the speech capacity*. Psychology Press, 2000.

[5] A. Warlamount. *The Cambridge Handbook of Infant Development: Brain, Behavior, and Cultural Context*, chapter Infant vocal learning and speech production., pages 602–631. Camb. Uni. Press, 2020.

[6] MS Brainard and AJ Doupe. What songbirds teach us about learning. *Nature*, 417(6886):351, 2002.

[7] M Chakraborty and ED Jarvis. Brain evolution by brain pathway duplication. *Phil Trans Roy Soc B*, 370(1684):20150056, 2015.

[8] PY Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449, 2005.

[9] S Pagliarini, A Leblois, and X Hinaut. Vocal imitation in sensorimotor learning models: a comparative review. *IEEE TDCS*, 2020.

[10] A Amador, YS Perl, GB Mindlin, and D Margoliash. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature*, 495(7439):59, 2013.

[11] K Doya and TJ Sejnowski. A computational model of birdsong learning by auditory experience and auditory feedback. In *Central auditory processing and neural modeling*, pages 77–88. Springer, 1998.

[12] CR Ponce et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[13] S Pagliarini, N Trouvain, A Leblois, and X Hinaut. What does the canary say? WaveGAN applied to birdsong. *HAL preprint hal-03244723*, 2021.

[14] N. Trouvain and X. Hinaut. Canary song decoder: Transduction and implicit segmentation with ESNs and LTSMs. In *ICDL-EpiRob*, 2021.

[15] N Trouvain et al. Reservoirpy: an efficient and user-friendly library to design echo state networks. In *ICANN*, pages 494–505. Springer, 2020.

[16] L McInnes et al. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[17] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[18] G Giraudon, N Trouvain, A Cazala, C Del Negro, and X Hinaut. Labeled songs of domestic canary m1-2016-spring (serinus canaria) (version 0.0.1) [data set]. *Zenodo*, http://doi.org/10.5281/zenodo.4736597, 2021.

[19] IR Fiete, MS Fee, and HS Seung. Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of neurophysiology*, 98(4):2038–2057, 2007.

[20] RG Alonso et al. A circular model for song motor control in serinus canaria. *Frontiers in computational neuroscience*, 9:41, 2015.

[21] S Pagliarini, X Hinaut, and A Leblois. A bio-inspired model towards vocal gesture learning in songbird. In *ICDL Epirob, 2018*. IEEE, 2018.

[22] MM Solis and AJ Doupe. Anterior forebrain neurons develop selectivity by an intermediate stage of birdsong learning. *J Neuro*, 17(16):6447–6462, 1997.

[23] X Hinaut. Which input abstraction is better for a robot syntax acquisition model? phonemes, words or grammatical constructions? In *ICDL-EpiRob*, pages 281–286. IEEE, 2018.
[24] L Pedrelli and X Hinaut. Hierarchical-task reservoir for online semantic analysis from continuous speech. *HAL preprint hal-03031413*, 2020.