Feature and structural learning of memory sequences with recurrent and gated spiking neural networks using free-energy: application to speech perception and production I

Alexandre Pitti, Mathias Quoy, Catherine Lavandier, Sofiane Boucenna

Feature and structural learning of memory sequences with recurrent and gated spiking neural networks using free-energy: application to speech perception and production I

Alexandre Pitti[1*], Mathias Quoy[1], Catherine Lavandier[1] and Sofiane Boucenna[1]

***Abstract*— We propose a unified framework for modeling the cortico-basal system (CX-BG) and the fronto-striatal system (PFC-BG) for the generation and recall of audio memory sequences; ie, sound perception and speech production. Our genuine model is based on the neural architecture called INFERNO standing for Iterative Free-Energy Optimization of Recurrent Neural Networks. Free-energy (noise) minimization is used for exploring, selecting and learning in PFC the optimal choices of actions to perform in the BG network (eg sound production) in order to reproduce and control the most accurately possible the spike trains representing sounds in CX. The difference between the two working memories relies in the neural coding itself, which is based on temporal ordering in the CX-BG networks (Spike Timing-Dependent Plasticity) and on the rank ordering in the sequence in the PFC-BG networks (gating or gain-modulation). We detail in this paper only the CX-BG system responsible to encode the audio primitives at few milliseconds order, while the PFC-BG system responsible for the learning of temporal structure in sequences will be presented in a complementary paper. Two experiments done with a small and a big audio database show the capabilities of exploration, generalization and robustness to noise of the neural architecture to retrieve audio primitives.**

I. Introduction

In order to make optimal choices out of many sensory or motor options, brain networks have to be organized hierarchically, but also flexibly, for retrieving and categorizing memory sequences from milliseconds order to seconds order [1], [2]. Depending on the degree of novelty and variability in incoming signals, top-down expectations help to recognize any familiar patterns or to detect unfamiliar ones. So, what are the neural mechanisms behind these bottom-up and top-down processes to construct robust and coherent behaviors?

A. Neural fundations

In different brain areas, working memories (WMs) are hypothesized to embed neural processes with forward and inverse models that can encode, anticipate and eventually control incoming signals to be more robust and to overcome their variability [3], [4], [5]. Two brain areas namely the Basal Ganglia (BG) that selects actions with respect to current states [6] and the Prefrontal Cortex (PFC) that represents forthcoming actions with respect to current contexts [7], [2], [8], are important for embedding these WMs. Being part of two different loops but connected at the BG level, they realize reactive (BG) and proactive (PFC) control, processing information differently and at different speed.

On the one hand, some evidences indicate that the striatum in BG has a principal function in learning-related plasticity associated with selecting one set of actions from many, resulting in the acquisition of habitual behavior [9], [10]. On the other hand, PFC achieves behavioral planning in terms of the end result, rather than in terms of the movement required to perform the task [11], [12].

Graybiel and Grafton suggest in [13] that proactive control is associated with sustained and/or anticipatory activation of lateral PFC, which reflects the active maintenance of task goals. By contrast, reactive control should be reflected in transient activation, along with a wider network of additional brain regions such as the BG. Therefore, these two control mechanisms differ in terms of the involvement during learning and retrieving tasks or sequences, with the BG dynamics working at a faster pace than the PFC.

In Machine Learning, reactive and proactive control relate to what is called model-free and model-based systems in Reinforcement Learning (RL) [14], [15], [16], [6], having one system for stimulus-response tasks doing greedy-like optimization and the other learning distinct policies for prediction, which serves for planning goal-directed behaviors.

These features are linked therefore to what is called now the Bayesian theory of the brain [17], [18] and to the paradigm of predictive coding for cognition [19], [20], [21]. These general theories describe how our expectations (as well as our errors) on sensory inputs are used as attention signals to adjust the prior expectations for the next events. Brain areas are hypothesized to use error prediction as a core information to *control* their dynamics from each others, not just for binding them mutually.

Under this framework, two or more brain networks can interact dynamically (eg Cortex CX with Basal Ganglia BG) so that we have always one network (eg the controller) that infers the reliability of another (eg the observer) with respect to a specific context. Along with Bayes theory, predictive coding has also its link with optimal control theory [22], which we think interesting in terms of perspective for modeling the corticostriatal system as it moves the problem of learning and retrieving memory sequences into a control problem.

Problem-solving tasks are good examples for understanding the involvement of the BG-PFC loops in goal-directed behaviors under uncertainty especially during infancy. These

[1] University Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS, France surname.name@u-cergy.fr

goal-directed behaviors are also called *task-sets* in cognitive and developmental sciences [23], [24], [25]. Task sets relate to the novel capabilities acquired by twelve-months-old babies such as tool-use, sustained attention, spatial memory, asymmetric imitation and rule-based learning and are argued to be linked to the maturation of the PFC [26], [27], [28], [29], [30]. Other crucial examples during infancy are speech production and the sequential organization of actions [31], [32], [33]. These two important cognitive tasks presumably involve the BG and PFC loops to adjust timely and orderly motor primitives [34], [11] and [35], [36], [37].

This neural process has been particularly studied for speech and language sequences because auditory modality is the sense especially sensitive to temporal structure. In the case of speech production, Romanski and colleagues propose that the phonotopical level requires the implementation of high-order models for encoding words or sentences as articulatory vocal tracks [38]. In other experiments with 3 months-old children [31], [32], [33], the stronger activation of the PFC has been observed for detecting temporal dissonance in regular temporal structures of spoken sequences of totally random syllabes such as the ABA structure in "tomato" or "mifumi", independent of the syllabes pronounced for the A or B items [39], [40].

B. Proposal framework for feature extraction and sequence learning

In line with these finding, we propose a neural architecture to model the CX, BG and PFC systems that combines model-free and model-based learning for retrieving and controlling long-range memory sequences hierarchically at the signal level and at the abstract level, see Figs. 1. The two working memories are developped within the same framework of predictive coding and reinforcement learning [20], [41], [42] but each system is working differently to code information and to minimize online error and external noise. The models use spiking neural networks (SNN) in order to learn temporal delays between pre- and post-synaptic firing neurons with the mechanism of Spike Timing-Dependent Plasticity (STDP) [43], [44], [45]. In line with the framework of free-energy minimization [19], we exploit also intrinsic noise within the system in order to realize stochastic descent gradient and novelty detection.

We propose that these neural mechanisms can serve for the learning of temporal delays between neurons in a self-organizing manner and makes possible the discovery of causes and effects necessary for active inference and predictive coding. This extends previous researches in which we developped several models of WMs using SNNs corresponding to different brain areas. These models exploited noise or novelty to iteratively infer a model and minimize error prediction either to control one system's dynamics (eg the hippocampus or BG-like model-free networks [46], [47]) or to select dynamically a better controller (eg a PFC-like model-based network [48]).

In our BG-like network modeled in [47], we showed that it is possible to control long-range memory sequences of spikes –, above 1000 iterations without loss,– and to solve the so-called credit assignment problem by infering causes and effects, even with long-range delays. Because of its property to optimize and control dynamics iteratively using noise or free-energy, we named our network INFERNO, which stands for Iterative Free-Energy Optimization for Recurrent Neural Networks [47].

Our framework will be applied to speech learning (perception and production). The global framework combines the corticostriatal and prefrontal systems for the recognition and generation of audio memory sequences, see Fig. 1. But in this paper, only the model of the cortico-striatal (CX- BG) system is developped in order to better describe the process for retrieving audio primitives for a short time scale. The model named INFERNO network is then used to solve the credit assignment problem for retrieving the motor primitives (articulatory motions) that cause specific sound signal (vocal tracks). In a complementary paper, the combination with the prefrontal system (PFC) will be presented, with the use of Gain-Modulated neurons for learning the temporal organization within memory sequences and for predicting the next ones; the sensitivity of Gain-Modulated or gated neurons to the items' order within a sequence will serve for finding structure within signals. Then, the use of this gated version of the INFERNO network will make possible to retrieve long range sequences through iterative optimization for long time scale.

C. Neural model for corticostriatal system

In our comprehension of the free-energy optimization strategy proposed by Friston [], it is similar to a reinforcement learning process done between two or more learning structures that attempt to minimize error prediction. To us, it conveys the learning problem into the ones of optimal control and predictive coding. We can apprehend the cortico-striatal loop as two learning systems that attempt to perform an optimal control and resolve error prediction among their dynamics. In Fig. 2, we display our framework with the Primary Auditory Cortex (PAC) system and the Intra-Parietal Lateral (IPL) layer modeled with SNNs to encode incoming inputs, the Striatum layer that categorizes the state of the IPL dynamics and the Globus Pallidus that attempt to control back the input dynamics of the PAC and IPL with a reentrant loop. The error prediction is evaluated and minimized over time by supervision of the STR units (critic) and by noise generation and stochastic search done on the GP output layer (actor).

This paper is organised as follows. In the section II, the neural architecture and the learning mechanisms of INFERNO network are presented. Two experimental setups for sound sequences are presented in section III, respectively for a limited learning database (only one speaker, 3 minutes length) and for a larger database (several speakers of different genders, 30 minutes length). The results of these two experiments are developped and discussed in section IV.

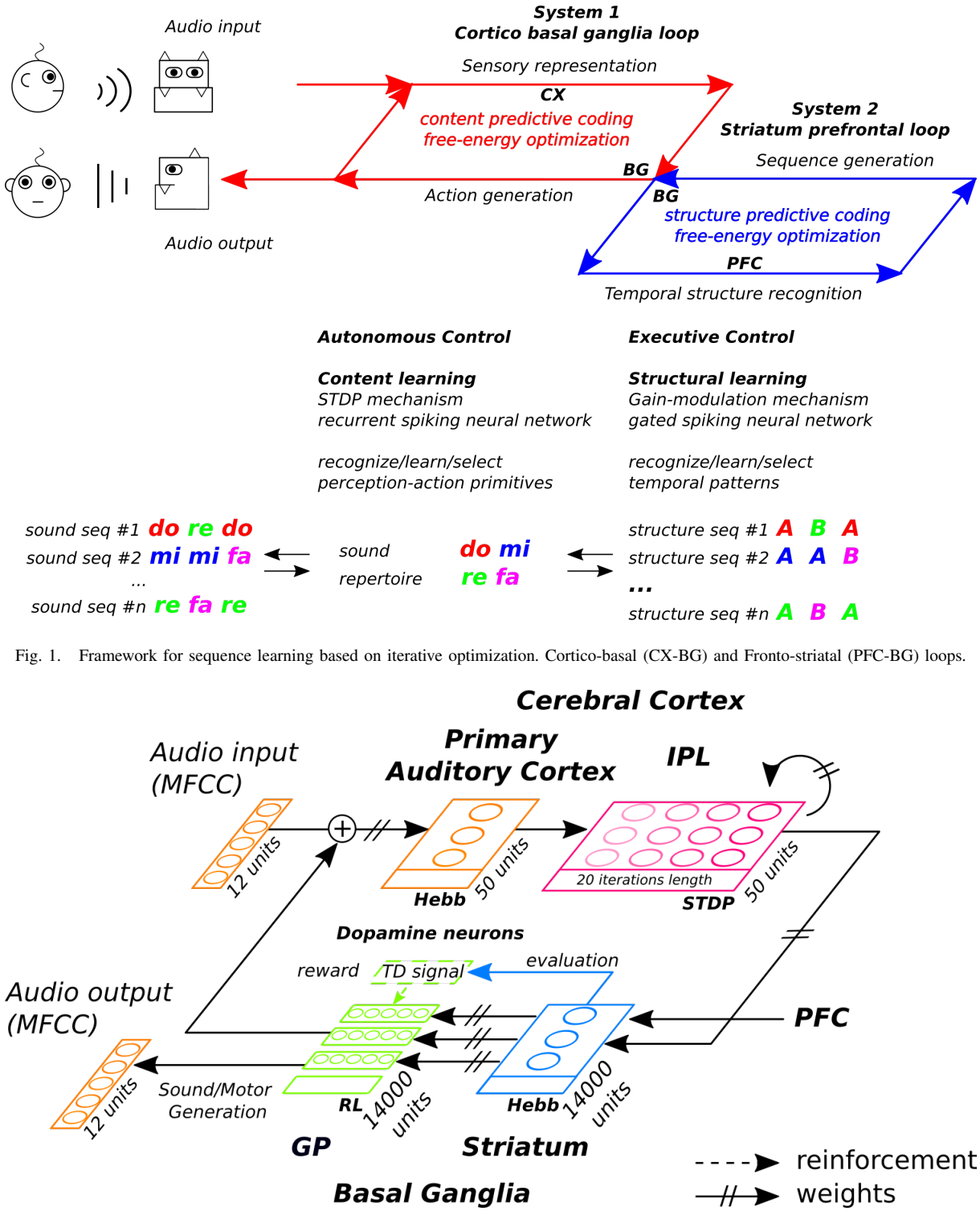

Fig. 1. Framework for sequence learning based on iterative optimization. Cortico-basal (CX-BG) and Fronto-striatal (PFC-BG) loops.

Fig. 2. Framework of the INFERNO architecture for audio primitive retrieving based on iterative optimization through cortico-basal ganglia loop (CX-BG). The Primary Auditory Cortex (PAC) receives and categorizes the audio vectors as a first stage, the Intra-Parietal Lateral cortex (IPL) integrates over time its output that are categorized at the end by the Striatum (STR) in the basal ganglia. The Globus Pallidus (GP) searches and retrieves the audio vectors that best matches the IPL dynamics recognized by the striatal units. The iterative optimization process is done by minimizing noise with a temporal difference reinforcement signal.

II. Methods

We present here the neural architecture INFERNO used for predictive coding associated with CX and BG. We describe then the coding mechanism used for modeling the spiking neurons and the learning mechanisms associated with temporal order and rank coding. We define then the

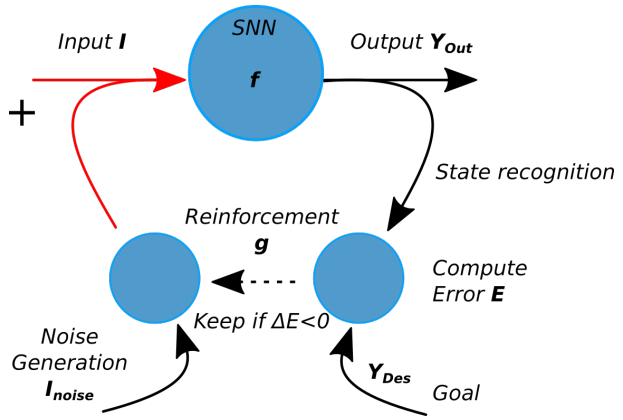

Fig. 3. Stochastic descent gradient optimization used to control the neural dynamics. Free-energy (noise) is injected as Input in the network. After a period of time, the Output vector is read to recognize the state and its value is compared to a goal vector. If the variationnal error E is decreasing, the stochastic descent gradient keeps the current Input. After several cycles, the Input converges to its optimal values that minimizes error and maximizes the state recognition stage.

experimental setup and the parameters used in the context of audio primitives retrieval for encoding the audio signals.

A. The recurrent network INFERNO

The neural architecture INFERNO [47] consists of two coupled learning systems arranged as in Fig. 2. The first network corresponds to one recurrent neural network of spiking neurons (SNNs) and the second network consists on one associative map. The SNN implements a forward model of the incoming signals whereas the associative map implements an inverse model aimed at retrieving and controlling those signals. The inverse-forward controller can be modeled with the function $Y_{out} = f(I)$ for the SNN and with the function $I = g(Y_{out})$ for the associative map, in which I is the input vector and Y_{out} are the output dynamics.

In order to minimize error, the second network generates intrinsic noise I_{noise} to control the dynamics of the first one following a RL mechanism. The activity of the SNN Y_{out} is compared to one desired goal vector Y_{des} to compute the error E between Y_{des} and Y_{out} and the current input is kept for the next step $I(t+1) = I(t) + I_{noise}$, if and only if it diminishes the gradient ΔE. Over time, I converges to I_{opt} its optimum value, and Y_{out} converges to Y_{des} the desired vector. This scheme is in line with actor-critic algorithms and predictive coding. Its organization is similar to novel architectures combining two or more competitive neural networks such as auto-encoders or the generative adversarial networks.

We showed in [47] that this variational process is similar to a stochastic descent gradient algorithm performed iteratively and can solve the temporal credit assignment problem for delays above dizains of iterations. For instance, the convergence to the desired goal after a certain delay can be viewed as the retrieval of a memory sequence for such duration. Furthermore, the free-energy minimization is generative in the sense that it can retrieve novel solutions I for the same output Y. This can be viewed as a synchronization process toward attractor memories [49].

B. Neuron model – Rank-Order Coding algorithm

We use the rank-order coding (ROC) algorithm to model integrate-and-fire neurons [50]. For instance, ROC neurons can translate ordered spatio-temporal patterns into ranked weights, see Fig. 4. The more similar the sequence order of the incoming signals, the higher the amplitude level of the ROC neurons. Reversely, the less similar the sequence order of the incoming signals, the lower the amplitude level of the ROC neurons.

If we adopt an ordinal ranking sensitive to the amplitude level of incoming units as displayed in Fig. 4, this coding strategy retranscribes well the Hebbian rule of "neurons that fire together wire together". These units can model well the properties of common neural populations in the neocortex.

$$Y_i^{IPL}(t) = w_i Y_i^{PSA}(t) + \sum_{j=1}^{50}\sum_{k=1}^{20} w_{jk}^{IPL} rank(Y_k^{IPL}(t-1)) \tag{1}$$

The equations of the rank-order coding algorithm that we used are as follows. The neurons' output Y is computed by doing the dot product between the function *rank()* sensitive to a specific rank ordering within the input signal vector I and the synaptic weights w; $w \in [0,1]$. As an example, one possible rank function can be $rank(i) = \frac{1}{1+i}$ that decreases monotonically with respect to the i^{th} rank of one item. For a vector signal of dimension $M = 50$ and for a population of $N = 14000$ neurons (M afferent synapses), we have:

$$Y_n^{Str} = \sum_{m}^{M} rank(Y_m^{IPL})\, w_{nm}^{IPL-Str}, \forall n \in N \tag{2}$$

The rank function *rank()* can be implemented classically as a power law of the `argsort()` function normalized between $[0,1]$ for modeling the STDP. This warranties that the density distribution is bounded and that the weight matrix is sparse, which makes the rank-order coding neurons similar to radial basis functions. This attribute permits to use them as receptive fields so that the more distant the input signal is to the receptive field, the lower is its activity level. The updating rule of the weights is similar to the winner-takes-all strategy in Kohonen networks [51] with an adaptive learning rate $\alpha_n, \forall n \in N$. For the best neuron Y_b, we have:

$$\Delta w_{bm}^{IPL-Str} = \alpha_b (rank(Y_m^{IPL}) - w_{bm}^{IPL-Str}), \forall m \in M \tag{3}$$

$$\alpha_b = 0.9\,\alpha_b \tag{4}$$

$$\Delta w^{Str-GP} = \beta (Y^{Str} - w^{Str-GP}).\delta_1 \tag{5}$$

where $\delta_1 = 1$ if reinforcement, and 0 otherwise.

$$Y^{GP}(t+1) = Y^{GP}(t) + noise.\delta_{\Delta E} \tag{6}$$

where $\delta_{\Delta E} = 1$ if $\Delta E > 0$, and 0 otherwise.

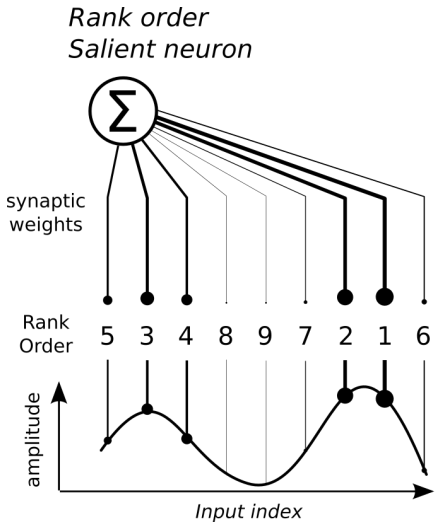

Fig. 4. Rank-Order Coding principle [50]. This type of neuron encodes the rank code of an input signal. Its amplitude is translated into an ordered sequence and the neuron's synaptic weights are associated to this sequence. In our example, the neural activity is salient to this particular order, which is seen in the line widths of the synaptic weights.

C. Experimental setup

The experimental setup for Experiment 1 in section III-A consists on a small audio dataset of 2 minutes length of a native french woman speaker repeating three times five sentences. The audio .wav file is translated into MFCC vectors (dimension 12) sampled at 25ms each and tested either with a stride of 10ms and no stride. The whole sequence represents 14.000 MFCC vectors for the case with strides and 10.000 MFCC vectors for the case with no strides.

The number of Striatal and GP units are chosen so that they correspond to the number of MFCC vectors, which means 14000 units (10.000 units) for each layer. We do so in order to test the reliability of our architecture to retrieve input data with an orthogonal representation. The compression rate is however low (1:1).

In contrast, Experiment 2 in section III-B will use a bigger audio dataset of 27 minutes length from six native french speakers, the same speaker as in Experiment 1 plus two other women and three men, repeating the same sentences as in the previous experiment. The audio .wav file is translated into MFCC vectors (dimension 12) sampled at 25ms each, which corresponds to 140.000 MFCC vectors for the case with 10ms stride. The number of Striatal and GP units are kept the same as for the first experiment (14.000 units), which means that the size for the BG layers are now ten times lower than the total number of MFCC to retrieve in the sequence. The compression rate is this time high (1:10). This second experiment will serve to test the generalization capabilities of our architecture and its robustness to high variabilities with respect to the inputs.

We provide a link to .wav files samples at `https://promethe.u-cergy.fr/alexpitt/inferno`.

III. Results

We perform two experiments in sections III-A and III-B with both consisting on learning and retrieving audio primitives considering the cortico-basal ganglia system, see the CX-BG system in Fig. 2. The two use the same network with the same number of units, the first experiment is performed with a small audio dataset of 2 minutes length whereas the second one is performed with a bigger audio dataset of 27 minutes length, see section II-C for more details.

In section III-A.1, we make to learn the Primary Auditory Cortex (PAC), IPL and Striatum layers in an unsupervised manner so that the three structures self-organize to sparse distributions using Hebb law for the PAC and the Striatum whereas the IPL learns the temporal dependencies across time using the STDP learning mechanism; the direction of the information flow is PAC→IPL→STR. In section III-A.2, the GP layer learns audio primitives (the MFCC vectors) through free-energy optimization; the direction of the information flow is STR→GP→PAC→IPL→STR. We study the two cases when we let the system unsupervised (self-organized regime) and when we control its dynamics (forced regime), resp. section III-A.3 and III-A.4. We analyze the performance of the inferno architecture in section III-A.5.

A. Experiment 1 – specialization capabilities in small audio database

1) Striatal categorization of IPL states: In order to understand the behavior of the system during the learning stage, we display the raster plots of the different dynamics for the PAC, IPL and Striatum layers for 1000 iterations respectively in Fig. 5 a-c). While the PAC first receives at each iteration the MFCC vectors, the IPL integrates with a temporal horizon of 20 iterations the different dynamics. Then, a third layer, the Striatum, categorizes the current state of the IPL network in a higher dimension. We justify the need to have a Striatum network of dimension as big as the audio database in order to separate orthogonally the MFCC vectors.

In so far, the learning stage is feed-forward from PAC→IPL→STR and the categorization is done in an unsupervised manner. The plasticity coefficient added to the learning mechanism of the Striatal units in eq. 4 permits to avoid any catastrophic forgetting after several weights updating.

2) CX-BG Iterative free-energy exploration-optimization: Once several passes are done over the complete audio sequence, the neurons stabilize to certain representations. It is possible then to perform an active exploration stage in the other direction – which means STR→GP→PAC→IPL→STR for retrieving the corresponding audio entries in GP through reinforcement learning.

This stage corresponds to a motor babbling in which the audio inputs are generated in GP and evaluated after a delay in STR. The prediction error in STR is used to drive the dynamics in GP using free-energy and to control the PAC layer and IPL dynamics via an iterative optimization process.

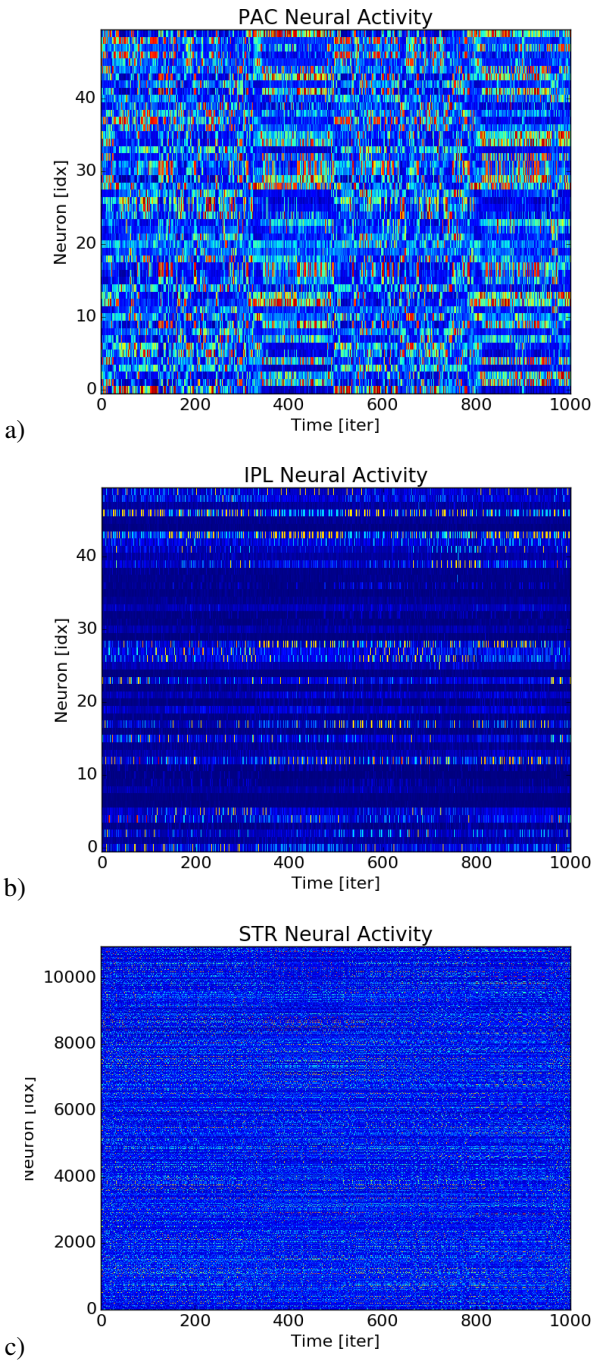

Fig. 5. Dynamics of the different structures during the learning stage. In a), the PAC layer categorizes the MFCC vectors in a higher representation. In b), this information is passed to the IPL layer that integrates over time (20 iterations) the incoming information. In c), the final layer, the STR, categorizes a second time the filtered information in a bigger neural population.

Over time, each audio vector is reinforced for each GP-Striatal pair whenever the GP auditory pattern makes to fire its corresponding Striatal unit. The audio pattern converges to an optimal MFCC vector for which the Striatal unit was the most active. As proposed by several neuroscientists, the GP layer may control indirectly the Striatal layer through the cortical dynamics [9], [10], [42]. The prediction error may drive the amount of noise within the system and the ratio between exploration and exploitation. This scheme corresponds to a predictive coding mechanism, which can solve the temporal credit assignment problem between causes (in GP) and delayed effects (in IPL).

We display in Fig. 6 three examples of retrieved GP dynamics (middle chart) for which the prediction error in Striatum is diminished over time (top chart) with respect to the spatio-temporal patterns of the IPL layer (bottom chart). The dashed line corresponds to a reset performed on the GP dynamics in order to observe dynamically the error minimization mechanism at work. The three samples correspond to the optimization process for three different Striatal units and for three GP vectors. During the free-energy descent gradient, each GP vector converges to one audio pattern for which the IPL activity is the most recognized by the corresponding Striatal unit. As showed in the graphs, the optimization process does not necessarily converge to the same minima after the reset done on the GP vector but can be stacked to another one. This means that different patterns of activity in the GP layer can influence in a similar way the activity in the IPL layer. Therefore, the categorization done in STR is not perfectly orthogonal (sparse) and different solutions coexist to retrieve the IPL spatio-temporal dynamics.

We analyze in Fig. 7 the learning performance of the free-energy optimization stage on the GP-STR dynamics. Fig. 7 a) presents the density distribution of the prediction error minimization for all the Striatal units and Fig. 7 b) presents the reconstruction error in the GP units with respect to the MFCC vectors. In Fig. 7 a), the prediction error is computed as the difference between the maximal activity of neurons when triggered and their upper limit, which means that for an error equals to zero, the STR neuron is firing maximally whereas for an error equals to 1, the STR neuron is not firing at all. The result in this graph shows that for a majority of the GP-STR units (80% of the population), the optimization process permits to minimize the prediction error below a value of 0.3, which means that most the GP neurons retrieved the optimal input vector that cause the STR to fire. Instead, for a small proportion of them (20% of the population), the error is above 0.4, which means that the the optimization process was not effective. in this case, the inferno architecture did not find the relationship between auditory input and the striatal category.

In Fig. 7 b), the reconstruction error is computed as the euclidean distance between the MFCC vectors presented in the audio database with the nearest GP vectors retrieved through free-energy optimization after normalization. The density probability distribution normalized between $[0, 1]$ shows that the reconstruction process is good with an approximation error centered at 4%. The GP layer has found most of the MFCC vectors.

We present in Fig. 8 further statistical analysis on the retrieved sound signals. In Fig. 8 a), we show a histogram about the MFCCs reconstruction error over 4 periods pro-

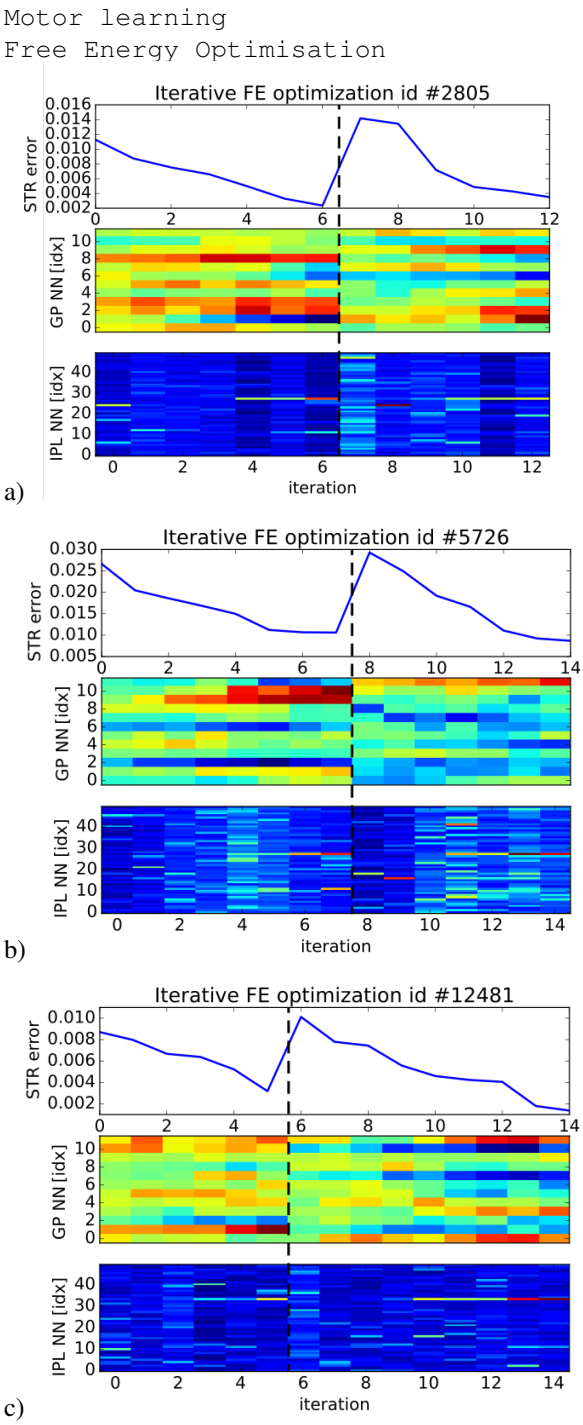

Fig. 6. Free-energy optimization. a-c), error minimization of three Striatal units (top chart) using noise to retrieve GP vectors for which the Striatal units fire maximally (middle chart). The IPL units display different spike trains for which a solution is found (bottom charts). The dashed lines correspond to a reset of the GP dynamics in order to show that the minimization process is always present and that different solutions can be retrieved dynamically.

cessing all over the audio sequence. The error is computed with the euclidean distance between each GP vector with the nearest MFCCs from the audio samples. The error is not normalized between $[0, 1]$ as in Fig. 7 b), the MFCCs vary

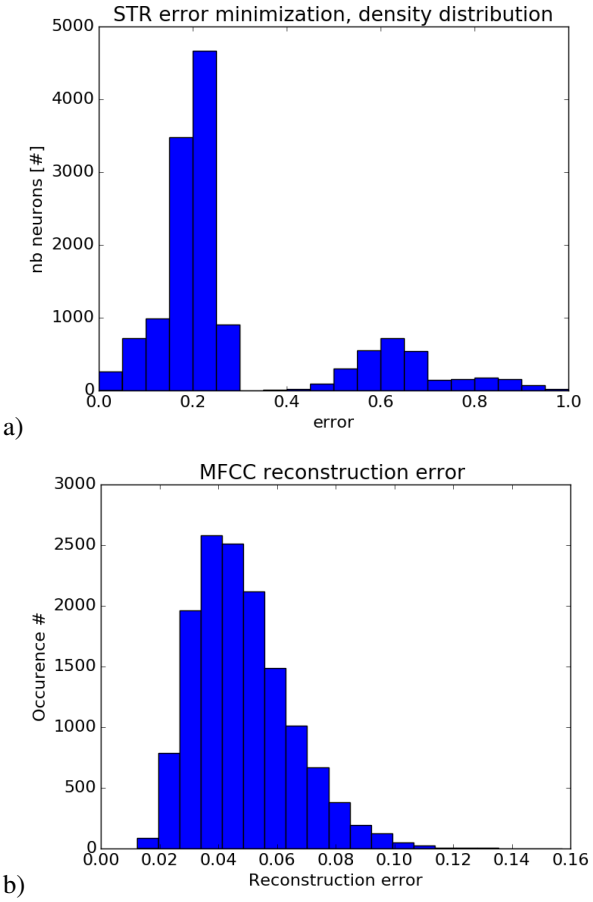

Fig. 7. Reconstruction analysis after free-energy optimization. In a), density probability distribution of the Striatal units with respect to their prediction error level . In b), density probability distribution of reconstruction error of MFCC vectors by the GP layer. For most of the neurons within the STR layer, the optimization process permits to construct MFCC vectors close the real ones from the audio database. The error reconstruction follows a central field distribution centered at 0.05 and standard deviation $\pm$ 0.05.

between $[0 + 1200]$. After each period, the error on each sample follows a distribution with lower error mean and narrower variance. The iterative optimization process goes from a 12% error to a 2% error on average on the samples. This shows the efficiency of the reinforcement learning stage to reconstruct the input dynamics.

A different curve is plotted in Fig. 8 b) obtained from an euclidean measure of the identity mismatch between the retrieved MFCC indices and the correct one (ground truth) and displayed ordered in time within the sequence. A low level indicates that the index of retrieved MFCC vector expected is near the real one and a high level indicates that the indices do not match. As similar to the previous figure, the error distribution diminishes gradually after each pass on the sequence. We can observe also that at the beginning and at the end of the sequence, the relative error is rather small corresponding to background noise when the person did not start speaking and when she ended up in advance.

When reconstituting the .wav file in Fig. 8 c) from the retrieved MFCC vectors, we can observe a gradual refining

of the audio waveform from the four periods with respect to the ground truth displayed at the bottom chart. The sequence is showed for 11 seconds although the global test was performed over two minutes length of the audio database.

After four exposures of the neural architecture to the audio sequence, the retrieved signals are gradually converging to the correct waveform. At period #0, the waveform is very discrete with square-like pattern and the amplitude and the wavelength are not respected. Gradually from period #1 to #3, we can observe a refinement of the waveform matching the ground truth curve[1].

3) Self-supervised learning: The learning of the MFCCs does not need to be done in a specific order. It can be done in an unsupervised manner by testing dynamically different sounds through cortico-basal recursion. This learning strategy may be seen as a motor babbling stage with random exploration. The resulting sequence is not necessarily coherent but at each iteration, the optimization process is at work to explore and improve the MFCC vectors found in GP. We present in Fig. 9 a) the unsupervised learning of the GP units combined with the information processing done in the STR and IPL layers for two thousand iterations. Below a certain error level (1st chart), the Striatal neurons have decharged maximally and another exploration cycle is engaged with the selection of a different Striatal unit (2nd chart). This second cycle will modify the dynamics in the GP (3rd chart), the PAC and the IPL layer till (4th chart) maximization of the STR units. The recall is not instantaneous in the beginning of the cycle and several iterations are necessary to make the different layers to converge. The process is similar to a greedy hill-climbing strategy although it is more visible in Fig. 6.

4) Forced learning: As opposed to the unsupervised learning strategy presented previously, we can force the recall of the Striatal neurons in a specific serial order, see Fig. 9 b). This control is normally done by another structure, the PFC, to retrieve an ordinal sequence.

The error minimization stage takes longer time to converge to the optimum IPL dynamics in comparison to the unsupervised learning strategy. However, the errors are smaller as we can expect.

Comparing the two learning strategies, we found that the unsupervised learning with self-organization could achieve error minimization and control on the IPL dynamics but the retrieving of longer sequences was not completely effective. These results are similar to what we found previously in [47]. Using unsupervised learning, the search space is not fully explored if the dimensionality is too big and the neural architecture can be trapped into local minima even if we use noise for descent gradient.

The learning stage can be very long and sub-optimal in comparison to the forcing method performed in a supervised manner. Over time, the supervised learning appeared more efficient to tutor the INFERNO network by providing goals and forcing the minimizing of errors till a certain threshold.

[1]We provide the link of the different .wav files at `https://promethe.u-cergy.fr/alexpitt/inferno`.

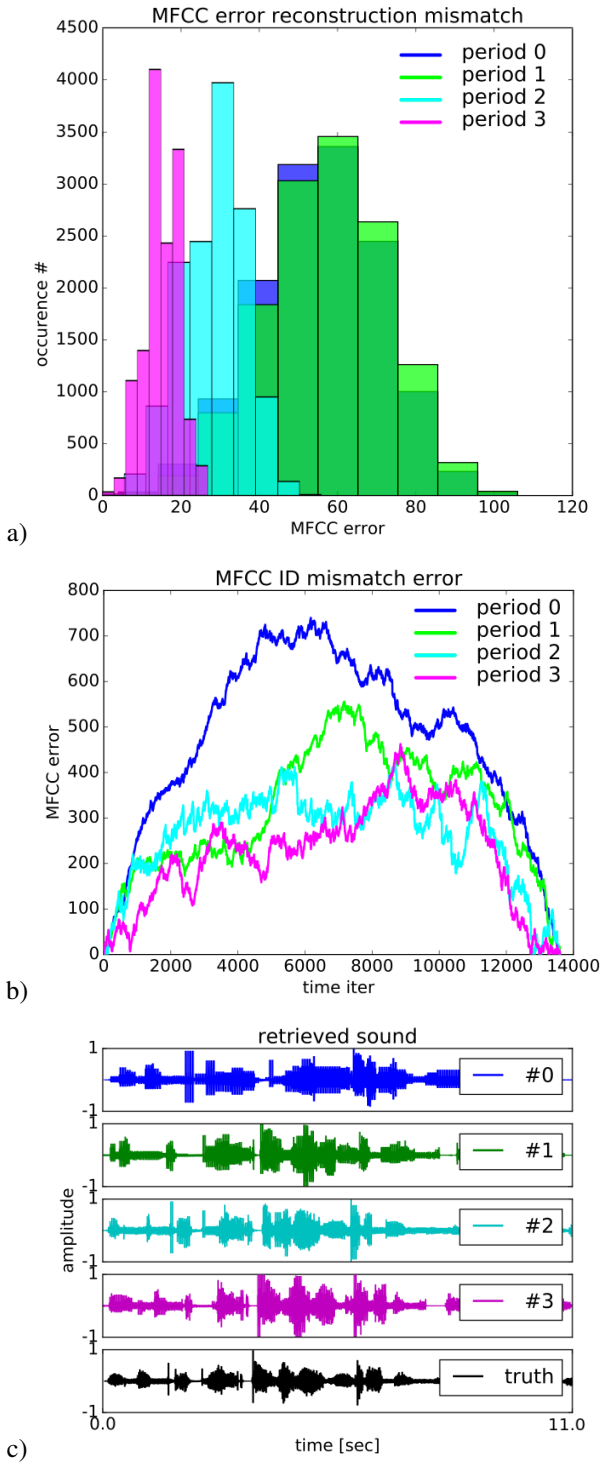

Fig. 8. Perforance analyse after several exposures and reconstruction analysis of the audio signals. In a), euclidean distance between the MFCCs retrieved and the ones from the audio database. In b), identity mismatch between the predicted MFCCs and the correct one for the whole audio sequence. in c), waveform reconstruction for the four learning periods.

This is in line with the idea that a goal-based approach plays a structuring role in comparison with a random-based approach, which will not take off if the dimension space is too big. Such structuring role is maybe played by the PFC and Hippocampus on the whole cortex during

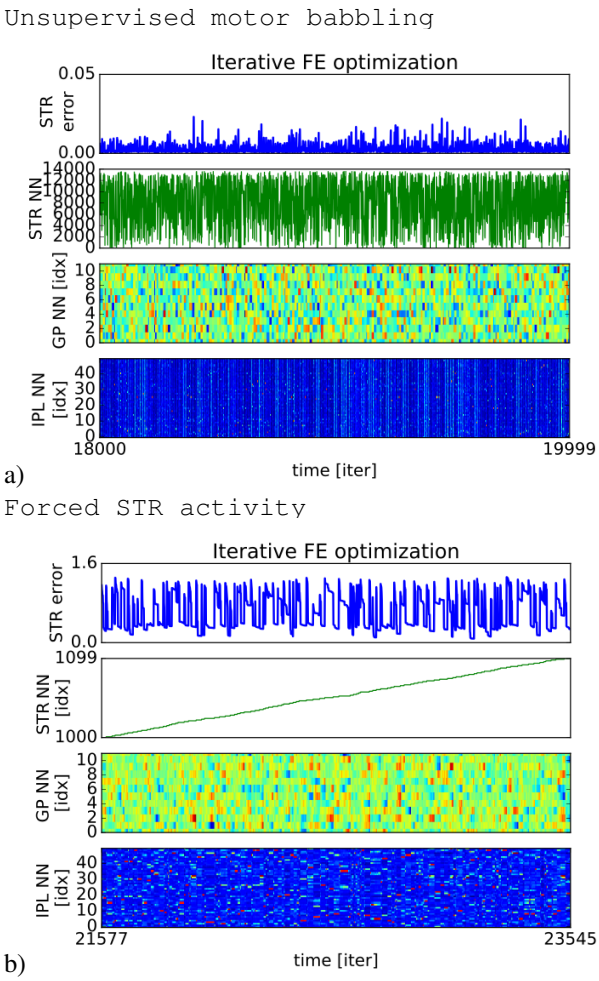

Fig. 9. Self-supervised VS forced learning. We compare the two learning strategies resp. in a) and b), in terms of convergence and dynamics. the self-supervising strategy might correspond to a babbling stage in which each audio unit is selected and tested at each cycle in a random fashion. Instead, the forcing strategy permits to control the learning of each unit separately till convergence.

development [52].

5) Retrieved MFCCs & audio primitives: We display in Fig. 10 a) the reconstructed .wav signal (in red) with respect to the real signal (blue) (2 minutes length) from the MFCC retrieved in GP and realigned in the correct order, Fig. 10 b). The MFCC coefficient errors between the real signal and the one reconstructed are displayed in Fig. 10 c).

We can observe that the overall waveform of the sound signal is correctly reconstructed although some errors and some delays are visible and hearable. The MFCC coefficient errors in Fig. 10 c) show that error is bigger for the high MFCC coefficients (high pitch) than for the small MFCC coefficients (low pitch). As the smaller coefficients correspond to low frequencies, it makes sense that the important part of the signal, which is in the high frequencies, is harder to retrieve.

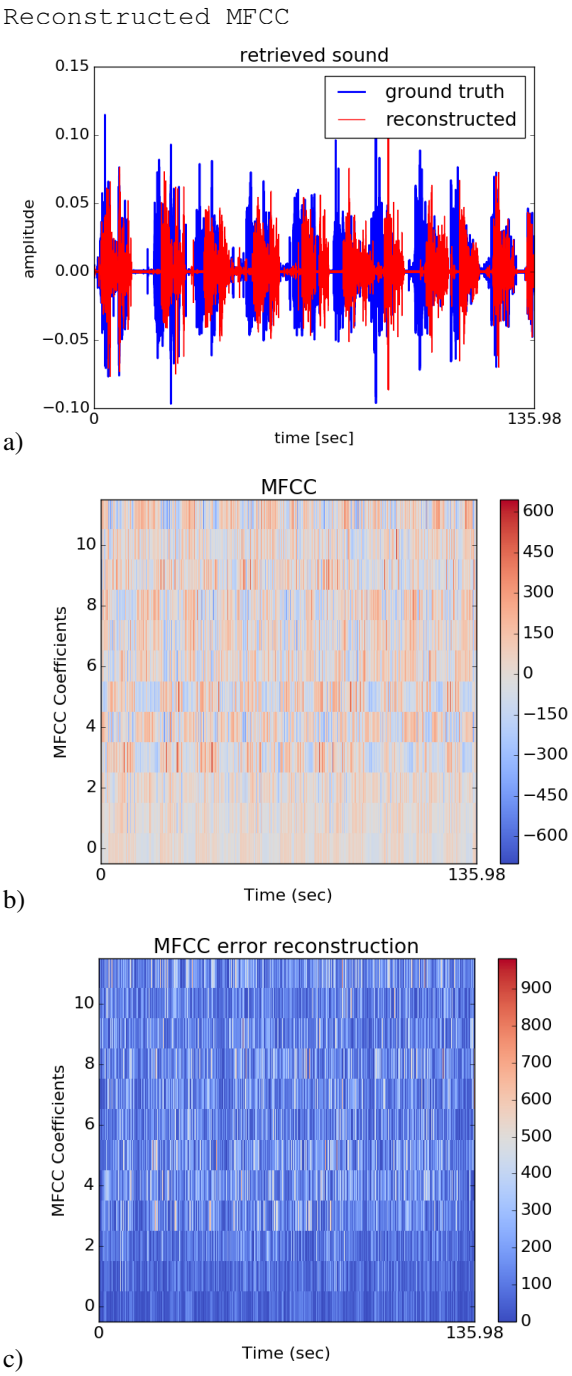

Fig. 10. Reconstructed Waveform and MFCC comparison. In a), the original waveform in blue and the reconstructed one in red. in b), the reconstructed MFCC raster plot. In c), the raster plot of the MFCC error between the original sequence and the retrieved one.

B. Experiment 2 – generalization capabilities in big audio database

In this section, we present the experiment done on a bigger audio database with an architecture of the same size as in the previous section. As the ratio between STR units and MFFCC to encode is now 1:10, we investigate here the generalization and redundancy capabilities of the network. The audio database of 27 minutes is also more difficult as

it consists of sentences pronounced by six different speakers with equal number of gender.

We present in Figure 11 a-c) different analysis done after the learning stage, resp. in a) the correspondance matrix between STR units and MFCCs vector within the audio database, in b) the correspondance matrix between the real MFCC vectors and the ones retrieved from the STR units. Fig. 11 d) displays a sample of the retrieved waveform.

The graph plotted in Fig. 11 a) corresponds to the mapping between the STR units and the MFCC vectors in the audio database. The euclidean distance is computed from the GP vectors retrieved to have correspondance between the MFCC and STR indices. The histogram in top chart indicates the generalization within the STR network: the same STR unit codes for several MFCC vectors. It shows the generalization capabilities of certain STR units within the network. For instance, certain STR units are clustering more than 100 MFCC vectors. At reverse, the histogram in the right chart indicates that the redundancy within the STR network: several STR units code for the same MFCC vector. The population coding is seen by the horizontal stripes. For instance, certain MFCC vectors are coded by more than 100 STR units.

Using this mapping, it is possible to construct in Fig. 11 b) the correspondance matrix between the retrieved MFCC vectors from the STR units and the ground truth MFCC vectors. The diagonal indicates that the mapping is bijective and that the network has retrieved the MFCC vectors from the STR units. The horizontal stripes indicate the redundancy and population coding within the network as well as the reconstruction errors due to the big audio database.

These results describe how the network performs on a large audio dataset, the discrepancy indicates that the number of vectors to retrieve is high in comparison to the number of units within the network. The reconstructed waveform in Fig. 11 c) plotted in red in comparison of the real waveform plotted in blue is one illustration of it: although the wave enveloppe is mostly preserved, the sound details are degraded. This is how the inferno network imposes a dimensionality reduction and has attempted to limit discrepancy and reconstruction errors.

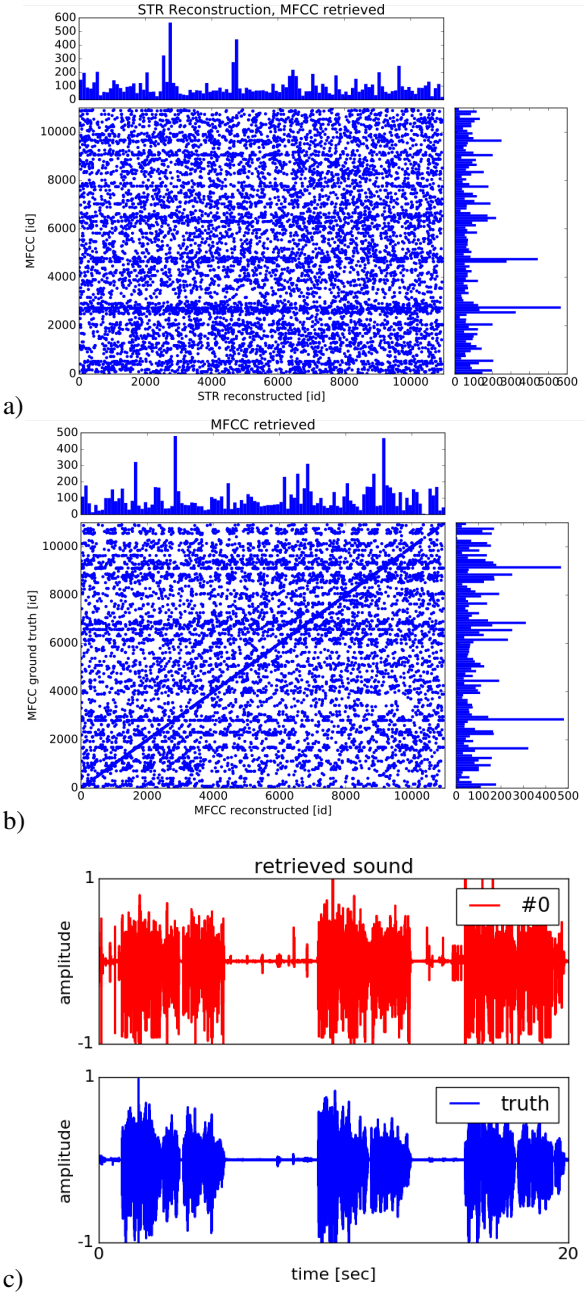

Fig. 11. Different analysis done on STR reconstruction and MFCC mapping. In a), the correspondance matrix between STR units and MFCCs vector within the audio database. In b), the correspondance matrix between the real MFCC vectors and the reconstructed ones based the correspondance matrix in a). In c), an example of a retrieved waveform is provided.

IV. Discussion

We have presented the neural architecture INFERNO based on free-energy minimization using recurrent spiking neural networks for modeling the CX-BG loop. This neural architecture is used for learning temporal sequences and for retrieving motor primitives by evaluating sensory feedback.

In [47], we have described this architecture for modeling the CX-BG structure with random examples. Here, we have showed its effectiveness in the more challenging tasks of audio primitives recognition and generation. The BG network explores and retrieves MFCC sound vectors by testing them stochastically through the CX layer. The more the Striatal units recognize and predict the CX output, the stronger it reinforces its link with the discovered GP units. At the end of this minimization process, the GP layer constitutes a sound repertoire of MFCCs.

The INFERNO network has two features, namely generalization and robustness to temporal delays. On the one hand, the number of units in the Striatum layer imposes a dimensionality reduction depending on the number of sound primitives to be learned (eg the number of MFCC vectors). On the other hand, the temporal chains in the CX layer permits to solve the temporal credit assignment problem and to link causes and effects thanks to STDP.

In the first experiments we have designed the network with the same number of STR units as MFCCs to retrieve (14.000

units) in order to have an orthogonal representation with few overlapping. These experiments were necessary to assess the robustness of the network particularly in high dimensions.

Although we have showed that the CX-BG network was capable to retrieve audio primitives in a self-organized manner, its exploration phase takes longer time than in a supervised manner. The exploration of the audio primitives in a self-organized manner is similar to a motor babbling, testing different sounds till convergence to the correct ones. In comparison to [47], the precise recovery of the temporal sequence was not possible due to the redundant sound repertoire in GP which possessed too much similar MFCC vectors. At reverse, it is aknowledged that the Basal Ganglia possesses a limited number of motor primitives. This result makes senses as we reconstruct audio MFCC vectors in the GP layer and not motor primitives as we should have with a robot or with a vocoder. Despite the dimensionality problem, the BG-CX loop is known to encode conditioning responses and its role is not devoluted to the control of the precise serial recall of sequences. Instead, the PFC is known to perform such executive control on the cortico-basal ganglia system to realize a precise control of temporal sequences. This second PFC-BG system will be presented in a complementary article.

In the last experiment, we have performed the learning of an audio sequence ten times longer than the previous ones (30 minutes .wav) in order to assess the generalization capabilities of INFERNO to higher dimensions. However, the number of sound primitives was the same as in the first experiment (14.000 units). Although the reconstruction error was important in comparison to the first case, the network was still capable to generalize correctly to this larger temporal sequence. This underlies the capabilities of inference of the architecture despite the large variability in the database.

These attributes for generalization and inference appear in line with what is happening during development. For instance, infants appear to learn a dictionnary of prototypic sounds and to know how to adjust in their mother tongue different voices, different context [53]. One difficulty is to know how speech is decomposed into distinct units to be analyzed. At the end of the developmental stage, a large number of sounds will seem similar to infants althought they are different; eg “r” and “l” in japanese. This phenomenon occuring from 6 month to 18 months is known as perceptual categorization, in which the discriminating capabilities are narrowing. during this period, infants appear to organize a repertoire of prototypic sounds with which he can compare and infer any sound he thinks to be the closest as sort of ’perceptual magnet’ [54], [55]. This repertoire is either perceptual, motor, or sensorimotor and the decision making done seems to correspond to Bayesian inference in speech [56], [57], [58].

In our present research, the sound repertoire encoded is only perceptual as audio primitives as encoded in the GP layer in the form of MFCC vectors. In future researches, we can think to use a vocoder with an audio speaker in place of the MFCCs in order to generate a real sound with a microphone to retrieve the sound information from another channel. That is, we think that having a robot that can speak and listen will help to learn from itself and from its social environment in a more ecological fashion following a developmental process [59], [60]. We envision to extend our framework also to visual information for audio-speech recognition [61].

ACKNOWLEDGEMENTS

We would like to thank Mathieu Lagrange (LS2N, UMR 6004 CNS, Ecole Centrale de Nantes) for providing the audio database. This work was partially funded by EQUIPEX-ROBOTEX (CNRS), chaire dexcellence CNRS-UCP and project Labex MME-DII (ANR11-LBX-0023-01).

REFERENCES

[1] G. Buzsaki, *Rhythms of the Brain*. Oxford University Press, 2006.
[2] E. Miller, “The ”working” of working memory,” *Dialogues Clin Neurosci.*, vol. 15, no. 4, pp. 411–418, 2015.
[3] H. Orban and D. Wolpert, “Representations of uncertainty in sensorimotor control,” *Current Opinion in Neurobiology*, vol. 21, pp. 1–7, 2011.
[4] D. Wolpert and M. Kawato, “Multiple paired forward and inverse models for motor control,” *Neural Networks*, vol. 11, no. 6, pp. 1317–1329, 1998.
[5] D. Wolpert, K. Doya, and M. Kawato, “A unifying computational framework for motor control and social interaction,” *Philosophical Transactions of the Royal Society*, vol. 358, pp. 593–602, 2003.
[6] K. Doya, “Metalearning and neuromodulation,” *Neural Networks*, vol. 15, p. 495506, 2002.
[7] T. Buschman and E. Miller, “Goal-direction and top-down control,” *Phil. Trans. R. Soc. B*, vol. 369, p. 20130471, 2014.
[8] E. Koechlin, “Prefrontal executive function and adaptive behavior in complex environments,” *Current Opinion in Neurobiology*, vol. 37, pp. 1–6, 2016.
[9] A. Graybiel, “The basal ganglia and chunking of action repertoires,” *Neurobiol Learn Mem*, vol. 70, p. 119136, 1998.
[10] K. Doya, “Metalearning, neuromodulation, and emotion,” *G. Hatano, N. Okada, H. Tanabe (Eds.), Affective Minds*, pp. 101–104, 2000.
[11] J. Tanji and E. Hoshi, “Behavioral planning in the prefrontal cortex,” *Curr. Opin. Neurobiol.*, vol. 11, p. 164170, 2001.
[12] J. Tanji, K. Shima, and H. Mushiake, “Concept-based behavioral planning and the lateral prefrontal cortex,” *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 528–534, 2007.
[13] A. Graybiel and S. Grafton, “The Striatum: Where Skills and Habits Meet,” *Cold Spring Harb Perspect Biol*, vol. 7, p. a021691, 2015.
[14] A. Barto, “Adaptive critics and the basal ganglia,” *In J. Houk, J. Davis, D. Beiser (Eds.), Models of information processing in the basal ganglia. Cambridge, MA MIT Press.*, pp. 215–232, 1995.
[15] A. Barto and R. Sutton, “Reinforcement learning in artificial intelligence,” *Advances in Psychology*, vol. 121, pp. , 358–386, 1997.
[16] W. Schultz, P. Dayan, and P. Montague, “A neural substrate of prediction and reward,” *Annu. Rev. Neurosci.*, vol. 275, pp. 1593–1599, 1997.
[17] W. Ma, J. Beck, P. Latham, and A. Pouget, “Bayesian inference with probabilistic population codes,” *Nat Neurosci*, vol. 9, no. 11, p. 14321438, 2006.
[18] J. Tenenbaum, C. Kemp, T. Griffiths, and N. Goodman, “How to grow a mind statistics, structure, and abstraction,” *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.
[19] K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *Journal of Physiology-Paris*, vol. 100, no. 1-3, p. 7087, 2006.
[20] A. Clark, *Surfing Uncertainty Prediction, Action, and the Embodied Mind*. Oxford University Press, 2015.
[21] M. Spratling, “Predictive coding as a model of cognition,” *Cognitive Processing*, vol. 17, no. 3, p. 279305, 2016.
[22] T. E and M. Jordan, “Optimal feedback control as a theory of motor coordination,” *Nat Neurosci*, vol. 5, p. 12261235, 2002.

[23] H. Harlow, "The formation of learning sets," *Psychological Review*, vol. 56, no. 1, pp. 51–65, 1949.

[24] K. Adolph and A. Joh, *Multiple learning mechanisms in the development of action*. New York Oxford University Press, 2009.

[25] M. Rouault and E. Koechlin, "Prefrontal function and cognitive control: from action to language," *Current Opinion in Behavioral Sciences*, vol. 21, p. 106111, 2018.

[26] A. Diamond, "Development of the ability to use recall to guide action, as indicated by infants' performance on a-not-b," *Child Development*, vol. 74, pp. 24–40, 1985.

[27] A. Diamond and P. Goldman-Rakic, "Comparison of human infants and rhesus monkeys on piaget's a-not-b task evidence for dependence on dorsolateral prefrontal cortex," *Experimental Brain Research*, vol. 74, pp. 24–40, 1989.

[28] A. Diamond, "A model system for studying the role of dopamine in the prefrontal cortex during early development in humans early and continuously treated phenylketonuria," *Handbook of Developmental Cognitive Neuroscience. Edited by Charles A. Nelson and Monica Luciana*, pp. 433–472, 1998.

[29] Y. Munakata, "Infant perseveration and implications for object permanence theories a pdp model of the a-not-b task," *Developmental Science*, vol. 1, no. 2, pp. 161–211, 1998.

[30] ——, "Computational cognitive neuroscience of early memory development," *Developmental Review*, vol. 24, pp. 133–153, 2004.

[31] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, p. 19261928, 1996.

[32] T. Nazzi, J. Bertoncini, and J. Mehler, "Language discrimination by newborns: toward an understanding of the role of rhythm," *J. Exp. Psychol. Hum. Percept. Perform*, vol. 24, p. 756766, 1998.

[33] G. Marcus, S. Vijayan, S. Bandi Rao, and P. Vishton, "Rule learning by seven-month-old infants," *Science*, vol. 283, p. 7780, 1999.

[34] P. Barone and J. J.P., "Prefrontal cortex and spatial sequencing in macaque monkey," *Exp Brain Res*, vol. 78, p. 447 464, 2018.

[35] M. Botvinick, Y. Niv, and A. Barto, "Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective," *Cognition*, vol. 113, no. 3, p. 262280, 2009.

[36] B. Averbeck, M. Chafee, D. Crowe, and G. A.P., "Neural activity in prefrontal cortex during copying geometrical shapes. i. single cells encode shape, sequence, and metric parameters," *Exp Brain Res.*, vol. 150, no. 2, pp. 127–41, 2003.

[37] B. Averbeck, D. Crowe, M. Chafee, and G. A.P., "Neural activity in prefrontal cortex during copying geometrical shapes. ii. decoding shape segments from neural ensembles," *Exp Brain Res.*, vol. 150, no. 2, p. 143153, 2003.

[38] L. Romanski, B. Averbeck, and D. M., "Neural representation of vocalizations in the primate ventrolateral prefrontal cortex," *J Neurophysiol*, vol. 93, p. 734747, 2005.

[39] C. Wacongne, E. Labyt, V. van Wassenhove, T. Bekinschtein, L. Naccache, , and S. Dehaene, "Evidence for a hierarchy of predictions and prediction errors in human cortex," *Proc. Natl. Acad. Sci. USA*, vol. 108, p. 2075420759, 2011.

[40] S. Dehaene, F. Meyniel, C. Wacongne, L. Wang, and C. Pallier, "The neural representation of sequences from transition probabilities to algebraic patterns and linguistic trees," *Neuron*, vol. 88, pp. 2–19, 2015.

[41] R. Rao and D. Ballard, "Predictive coding in the visual cortex a functional interpretation of some extra-classical receptive-field effects," *Nat Neurosci*, vol. 2, pp. 79–87, 1999.

[42] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 364, pp. 1211–21, 2009.

[43] G. Bi and M. Poo, "Activity-induced synaptic modifications in hippocampal culture, dependence of spike timing, synaptic strength and cell type," *J. Neurscience*, vol. 18, pp. 10 464–10 472, 1998.

[44] E. M. Izhikevich, J. Gally A., and M. Edelman, G., "Spike-timing dynamics of neuronal groups," *Cerebral Cortex*, vol. 14, pp. 933–944, 2004.

[45] E. Izhikevich, "Polychronization computation with spikes," *Neural Computation*, vol. 18, pp. 245–282, 2006.

[46] A. Pitti and Y. Kuniyoshi, "Modeling the cholinergic innervation in the infant cortico-hippocampal system and its contribution to early memory development and attention," *Proc. of the International Joint Conference on Neural Networks (IJCNN11)*, pp. 1 – 8, 2011.

[47] A. Pitti, P. Gaussier, and M. Quoy, "Iterative free-energy optimization for recurrent neural networks (inferno)," *PLoS ONE*, vol. 12, no. 3, p. e0173684, 2017.

[48] A. Pitti, R. Braud, S. Mah, M. Quoy, and P. Gaussier, "Neural model for learning-to-learn of novel task sets in the motor domain," *Frontiers in Psychology*, vol. 4, no. 771, 2013.

[49] I. Tsuda, "Chaotic itinerancy and its roles in cognitive neurodynamics," *Current Opinion in Neurobiology*, vol. 31, p. 6771, 2015.

[50] S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," *Neural Networks*, vol. 14, pp. 715–725, 2001.

[51] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[52] J. McClelland, M. Botvinick, D. Noelle, D. Plaut, M. Rogers, T.T. Seidenberg, and L. Smith, "Letting structure emerge connectionist and dynamical systems approaches to cognition," *Trends in Cognitive Science*, vol. 14, no. 5, pp. 348–356, 2010.

[53] P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

[54] P. Kuhl, "Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not," *Percept. Psychophys.*, vol. 50, no. 2, p. 93107, 1991.

[55] P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, no. 5044, pp. 606–608, 1992.

[56] R. Laurent, M. Barnaud, J. Schwartz, P. Bessire, and J. Diard, "The complementary roles of auditory and motor information evaluated in a bayesian perceptuo-motor model of speech perception," *Psychological Review, American Psychological Association*, vol. 14, no. 1, p. e0210302, 2017.

[57] K. Kording and D. Wolpert, "Bayesian decision theory in sensorimotor control," *Trends Cogn. Sci.*, vol. 10, pp. 319–326, 2006.

[58] M. Barnaud, J. Schwartz, P. Bessire, and J. Diard, "Computer simulations of coupled idiosyncrasies in speech perception and speech production with cosmo, a perceptuo-motor bayesian model of speech communication," *PLoS ONE, Public Library of Science*, vol. 14, no. 1, p. e0210302, 2019.

[59] M. Asada, "Modeling early vocal development through infantcaregiver interaction: A review," *IEEE TCDS*, vol. 8, no. 2, pp. 128–138, 2016.

[60] A. Cangelosi and T. Ogata, "Speech and language in humanoid robots," *A. Goswami, P. Vadakkepat (eds.), Humanoid Robotics: A Reference, Springer Nature B.V. 2019*, pp. 2261–2292, 2018.

[61] A. Pitti, A. Blanchard, M. Cardinaux, and P. Gaussier, "Gain-field modulation mechanism in multimodal networks for spatial perception," *12th IEEE-RAS International Conference on Humanoid Robots Nov.29-Dec.1, 2012. Business Innovation Center Osaka, Japan*, pp. 297–302, 2012.