Chaos, Solitons & Fractals: X 2 (2019) 100012

Contents lists available at ScienceDirect

Chaos, Solitons & Fractals: X

journal homepage: www.elsevier.com/locate/csfx

A dynamical system as the source of augmentation in a deep learning problem

P.L. Tubaro [a], G.B. Mindlin [b,*]

[a] División Ornitología, Museo Argentino de Ciencias Naturales Bernardino Rivadavia, Argentina
[b] IFIBA, CONICET and Departamento de Física, FCEyN, UBA, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 16 September 2019
Accepted 2 October 2019
Available online 3 October 2019

Keywords:
Birdsong
Deep learning
Individual recognition

ABSTRACT

In this work we build a convolutional neural network capable of identifying individual birds by their songs. Since the actual data available from each individual is very limited, we use a dynamical system capable of synthesizing realistic songs, to generate surrogate-training data. The different synthetic songs are the result of integrating the dynamical system with slightly varied parameters. We show that a data set built in this way allows us to train the network to successfully identify the different individuals in our study. In this way, we present a novel way to perform data augmentation using dynamical systems.

1. Introduction

Classification is a task that consists in assigning labels to the elements of a set, from a pre-designed set of categories. In the last years, the training of layered arrays of nonlinear units has proven to be one of the most effective ways to implement algorithmically a classification task [1,2]. In this technique, known as deep learning, the weights of the connections between the units are progressively adjusted by exposing the network, several times, to the labeled elements of a training set. Ideally, the procedure is repeated until the networkś output properly represents the labels. The adjustment procedure is known as back-propagation [3], and it is a technique known since the late eighties. It is the availability of fast computers, and large amount of data files for training, what allows now deep learning to successfully compete with other machine learning techniques.

In fact, the need of large amount of labeled data is, in many cases, a limitation to use deep learning. In the absence of a sufficiently large data set for training, the network can be trained to associate the elements in the training set with their labels, but it will not generalize. This means that new data cannot be correctly classified. In those cases, one strategy consists of generating more training samples by manipulating the existing ones through some set of random transformations, giving rise to acceptable, surrogate data. This strategy is known as augmentation [2].

One field where deep learning has been very successful is bird species identification. Xeno-canto is a website where naturalists, birders and scientists upload their labeled recordings, providing the kind of massive data necessary for training a network to perform the classification task. In fact, a yearly competition called Birdclef [4] challenges programmers with tasks that consist in the identification of birds within a selected set, from their songs. Even with the large amount of data available for the problem, augmentation is necessary to successfully train the algorithms necessary to perform the task. The most successful algorithms in the Birdclef competition use layered networks of units that receive as inputs images representing the temporal evolution of acoustic features (either Mel coefficients, spectral content of windowed data, etc. [5])

The problem presents new challenges if one is interested in identifying individual birds of a given species. The amount of data per individual that is available for training is typically much sparser, and in principle, the sounds used by different individuals within a species are very similar. Moreover, the typical augmentation operations in the field of data science are inspired in the processing of visual images of objects in physical space. Therefore, they might not enrich sufficiently the training data set for our task.

In order to overcome this issue, we generate surrogate data by integrating a dynamical system capable of synthesizing realistic birdsong [6,7]. The different files in the training set are obtained varying the parameters representing each birdś physiology and anatomy, within biologically acceptable boundaries. We train a layered convolutional network completely with surrogate data, and show that a network trained in this way is capable of

* Corresponding author.
E-mail address: gabo@df.uba.ar (G.B. Mindlin).

https://doi.org/10.1016/j.csfx.2019.100012

recognizing the subjects whose parameters guided the construction of the artificial training data.

The paper is organized as follows. We describe the species under study, and the characteristics of our problem in Section 2. In Section 3 we describe the model used to generate the surrogate data that will be used to train the network, which will be described in Section 4. The performance of the network will be described in Section 5. We close with our conclusions and discussions in Section 6.

2. The problem

Approximately forty percent of the known bird species are songbirds, a group of bird species that require some degree of learning from a tutor in order to properly sing typical species songs. The Rufous-collared Sparrow (*Zonotrichia capensis*) is a songbird, which requires exposure to a tutor in order to learn its song. In template and subtropical regions, this is a unique combination of sounds that will sing during all of its life. Its song, which lasts between one to three seconds, is built from syllables: continuous sounds with a modulated fundamental frequency. These can be grouped in two parts [8]. The first one consists of a few (one to five) introductory syllables, and is known as the theme. The second part is a rapid trill, i.e. a repetition of several copies of a downsweep syllable. In fact, this species has been studied in the framework of what is known as the "acoustic adaptation hypothesis", which postulates that the structure of the song is the one that minimizes its degradation in the birdś environment [9]. According to this hypothesis, for example, dense vegetation leads to slower trills, so that reverberations do not affect the perception of the syllables. In this way, the trills reveal features of the birdś environment, while the theme plays the role of an identity-bearing signal. Fig. 1 displays a spectrogram of a Rufous-collared Sparrow song.

We worked with six individuals recorded in Parque Miguel Lillo, Necochea, Argentina (38°33′44″S 58°44′43″O). The sounds were registered at 44.1 kHz with a digital Tascam HD-P2 recorder, connected to a directional Senheiser K6ME67 microphone. The recordings included a variable number of songs, from one to 10 songs. Fig. 2 shows a set of images, where each one corresponds to the spectrogram of a song produced by a different bird. These spectrograms were computed with Gaussian windows (standard deviations of 128 points), processing segments of 512 samples with successive overlaps of 256 points. The display of the spectrograms considers a clipping below 1/1000 the maximum value of the spectrogram.

The parameters characterizing the song (the initial and final values of the fundamental frequency for each syllable, the duration of each syllable, and the timing between syllables) varied very little across different repetitions of the song; never more than 3 percent. In this way, if we need to build surrogate data starting from the spectrograms of our scarce experimental data, the usual operations of augmentation would create very artificial images. In fact, data augmentation was conceived as a set of random operations to create believable looking images, but in the field of vision. Therefore, rotations, width shift, height shift and flipping, which are sensible parameters for vision, are of little use for generating a sensible training set for our problem, as these modified images would not be spectrograms of songs the bird could ever perform. It is for this reason that we have taken a different approach. We use a dynamical system describing the physics of birdsong production, and synthesize realistic replica of the experimental data with slightly changed parameters. Then, the spectrograms of these songs will be used to train and validate a layered network.

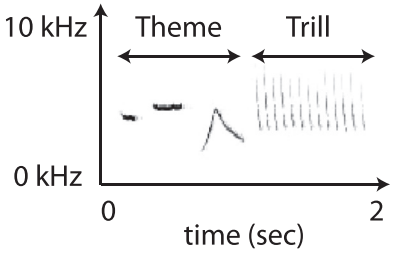

Fig. 1. The song of a Rufous-collared sparrow. The first syllables constitute a theme, and are an identity-bearing signal. The trill is a set of rapid downsweep syllables. The rate of syllable production in the trill indicates the birdś geographical origin.

3. The model

Birdsong is generated in a way that resembles how human voiced sounds are produced. A bipartite structure called the syrinx holds two pairs of labia at the juncture between the bronchi and the trachea. Each pair of labia is set into an oscillatory mode when a sufficiently strong airflow passes between them, just as the human vocal folds are when a voiced sound is uttered. These oscillations modulate the airflow, generating sound. In the last years, theoretical and experimental works have identified the biomechanical and dynamical mechanisms that rule the behavior of the oscillating labia as birdsong is produced. The basic physiological parameters that the bird needs to control are the air sac pressure, which controls the strength of the airflow through the labia, and a set of physiological instructions sent to the muscles controlling the configuration of the syrinx. The configuration of this somewhat elastic substrate affects the stretching of the labia, and therefore, the fundamental frequency of the oscillations at which these can oscillate [10].

The labia are assumed to be in a stationary position when the bird is silent. As the parameter representing the air sac pressure builds up, it eventually reaches a threshold for the oscillatory motion. While the parameters of the problem remain in the phonating region of the parameter space, the airflow is modulated and sound is produced. As the pressure is decreased, the sound will eventually stop (i.e., the syllable will end). The qualitative change in dynamics when the parameters are varied is known as a bifurcation. Close to the parameter values where a bifurcation occurs, the model can be transformed into paradigmatic, simple equations.

For this species [11], it was shown that the system of equations describing the labial dynamics could be written as:

$$\frac{dx}{dt} = y$$

$$\frac{dy}{dt} = \kappa\gamma^2 x - \gamma x^2 y + \beta\gamma y,$$

where x stands for the midpoint labial position, κ, β are system parameters, and γ is the temporal scaling of the problem. To generate sound with this labial dynamics, the pressure at the tracheal input p_i is computed as:

$$p_i(t) = \mathrm{A}\, x(t) + p_{back}\left(t - \frac{L}{c}\right)$$

$$p_{back}(t) = -r\; p_i\left(t - \frac{L}{c}\right)$$

where A is a coefficient that depends on the airflow strength, L is the tracheal length, c is the speed of sound and r, the reflection coefficient. Finally, the pressure at the output of the trachea ($p_{out} = (1-r)p_i(t - \frac{L}{c})$) forces a Helmholtz oscillator representing the oroesopharingeal cavity (OEC), which behaves as a last filter for the signal [12]. The equations ruling the filtering by the OEC are:

$$\frac{di_1}{dt} = i_2,$$

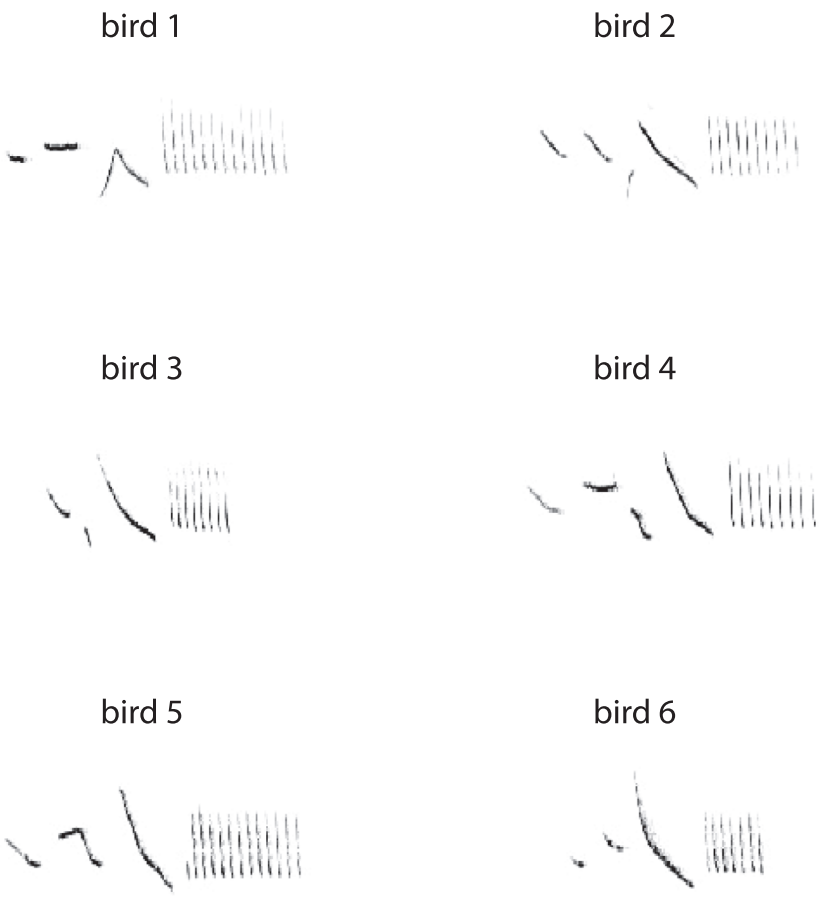

Fig. 2. Images obtained from the spectrograms of six Rufous-collared sparrows, within the same time and frequency limits of Fig. 1. The main difference between the spectrograms for different birds is the frequency modulation of the syllables in the theme. In order to generate these images, the spectrograms were computed with the same temporal and spectral ranges used in Fig. 1.

$$\frac{di_2}{dt} = -\frac{i_1}{CL_1} - \left(\frac{r_d}{L_2} + \frac{r_d}{L_1}\right) i_2 + \left(\frac{1}{CL_1} + \frac{r_2 r_d}{L_1 L_2}\right) i_3 + \frac{dp_{out}/dt}{L_1} + \frac{r_2 r_d}{L_1 L_2} p_{out}$$

$$\frac{di_3}{dt} = -\left(\frac{L_1}{L_2}\right) i_2 - \frac{r_d}{L_2} i_3 + \frac{1}{L_2} p_{out}$$

where following a deep rooted tradition in acoustic, the equations ruling the dynamics of a Helmholtz oscillator with an opening are written as those of an equivalent circuit. These equations are derived in [12], and the final sound is proportional to the variable i_3. Following previous work [12] the parameters used in the simulations are $(L_1, L_2, r_2, r_d, C) = (1/20,\ 1/10^4,\ 0.5\ 10^7, 24,000, 5/355\ x10^8)$.

These simple dynamical models for birdsong production can generate synthetic sounds with spectrograms and timbre very similar to the actual songs, just by fitting the syllables' fundamental frequencies. In fact, neurons highly selective to a bird's own song spike when the bird is exposed to synthetic copies generated by these models [6,13]. Six examples of spectrograms obtained by the integration of this model are shown in Fig. 3. Let us discuss how to unveil the time dependent parameters needed in each case in order to be able to reproduce the original songs.

In many species, the variety of acoustic modulations was found to be the result of a set of basic physiological instructions called "gestures" [14]. In the case of the Rufous-collared Sparrow, we defined a set of three frequency modulation patterns, namely exponential downsweep, linear modulations and sinusoidal modulations. The parameters for each modulation pattern are listed in Table 1.

To synthesize the song of a bird with the model, we identify the modulation pattern for each syllable of the song, and compute the parameters needed to reproduce them. Then, for each syllable we generate a list of frequencies. The values of κ necessary for the system's solutions to display solutions with fundamental frequencies ω satisfy:

$\kappa = 6.5\ 10^{-8}\omega^2 + 4.2\ 10^{-5}\omega + 2.610^{-2}$, a relationship that was reconstructed from a series of simulations on the model. In this way, the list of fundamental frequencies gets transformed into the parameters the model needs to synthesize a realistic copy of the song. By using the model, the spectral content of the sound source, properly filtered by the tracheal tube and the OEC, are automatically reproduced.

The model is now integrated a large number of times, varying the values of the parameters used to reproduce the basic gestures. In this way, a large number of surrogate spectrograms are generated, all of them differing in random parameters that are consistent with the biological variability that exists between different the songs produced by a unique individual. Training a deep neural network requires both a training set (used to adjust the parameters of the network) and a validation set. The last one is used to check that the system does not merely associate the inputs with the labels, but that it is capable of generalizing and classifying new sets as well. We generated, for each bird, 2500 sets of surrogate data to train the model, and 1500 to validate it. Notice that the system is

Table 1
Elementary frequency patterns in the song of collared-sparrows.

Modulation pattern	Frequency	Parameters
Sinusoidal	$w(t) = w_f + (w_i - w_f)\frac{(t-t_i)}{(t_f-t_i)}$	w_f, w_i, $t_f t_i$
Exponential	$w(t) = w_f + (w_i - w_f)e^{-\frac{3(t-t_i)}{(t_f-t_i)}}$	w_f, w_i, $t_f t_i$
Linear	$w(t) = w_{av} + A\ sin(\alpha_i + (\alpha_f - \alpha_i)\frac{(t-t_i)}{(t_f-t_i)})$	ω_{av}, A, α_f, α_i, $t_f t_i$

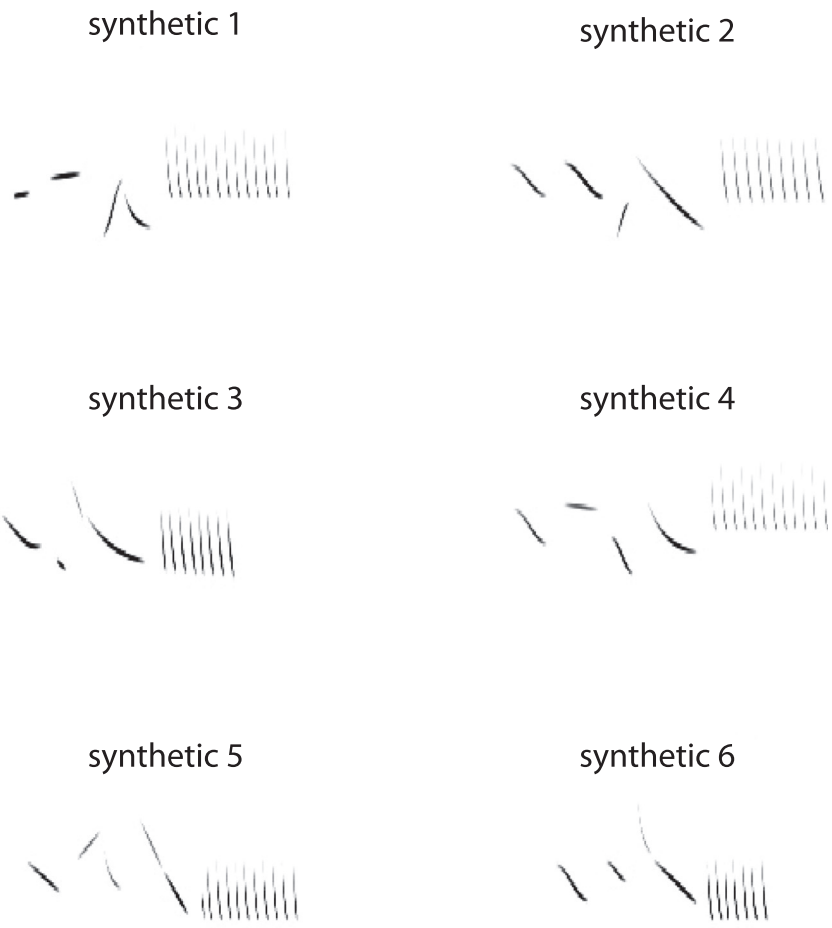

Fig. 3. Synthetic songs generated by a dynamical system, where the parameters we selected so that they would be good copies of the recorded songs. Yet, the parameters used in these simulations, as it is the case for all surrogate data, were slightly varied from the parameters that would be required to obtain optimal copies. These images are, as in Fig. 2, obtained from spectrograms of sounds within the same temporal window and frequency range shown in Fig. 1.

entirely trained with surrogate data: it is not exposed to a single real spectrogram computed from actual data.

4. The network

We fitted the parameters of a network that consists of a series of alternating convolutional 2D layers (4), and MaxPooling layers (4), with a final pair of densely connected layers. The convolutional layers had sizes 8,16,16 and 32 respectively, obtained from the respective inputs after convoluting with 3×3 windows. All the pooling layers aggressively down-sample the features maps by a factor of 2. The final two densely connected layers consist of 1024 and 6 units respectively. The last layer has as many units as categories in the problem. That means, in our case, that we will try to train the network to classify sounds as being sung by one of six recorded birds.

A common way to avoid over-fitting is by setting constraints on the connection (weights) values so these take small values. In this way, the network is more regular. This procedure is called regularization, and it is implemented by adding a cost to the loss function of the network, whenever the weights take large values. In our network, the regularizing parameter was set as $l2 = 0.001$. Another technique for avoiding over-fitting is to randomly drop some weights (setting their values to zero). Our dropout factor was set as 0.5. Finally, the learning rate was set to 10^{-4}.

The images used to train the network used a gray scale, of 300*x*200 pixels. Batches of 10 units were used, and the training took place in 15 epochs, with 200 steps per epoch. The conversion of these images into actual grids of pixels and ultimately floating point matrices was carried out using utilities of the Keras library. In particular, the class ImageDataGenerator automatically takes images and turns them into batches of tensors. We normalized the values in each image to 255.

5. The results

The evolution of the accuracy (the fraction of the images that were correctly classified) and loss (mismatch between a target

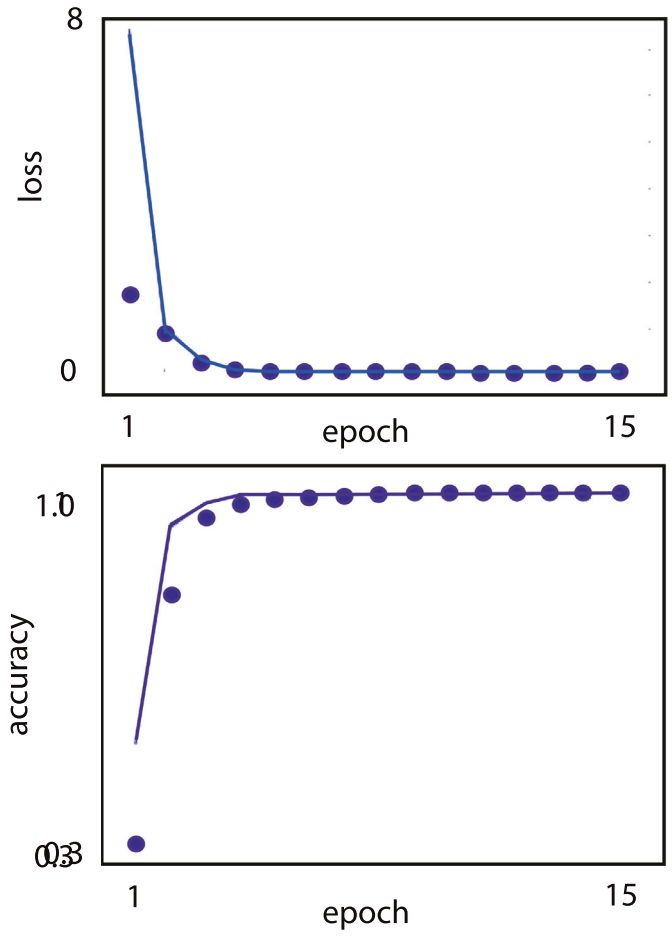

Fig. 4. We show the training (points) and validation (lines) accuracy and loss, computed during the training of the network.

candidate and the actual target) were computed in the different epochs, and they are displayed in Fig. 4. Notice that as fast as in six epochs, the validation accuracy is at 1.00, and the validation loss is as low as 0.3.

This network was trained, and validated, with a set of random images that corresponded to the spectrograms of synthetic sounds generated by a dynamical system. The sets did not include a single spectrogram from data. We then used this network to classify a small set of songs that we recorded from six individual birds. The number of songs that we could record from them varied in number. For the six birds we could record (1, 10, 9, 6, 9, 9) songs respectively. Notice that unless one bands the birds, the only parsimonious way to associate a set of songs with a bird is to extract them from a continuous recording, performed while the bird is in sight. We computed the spectrogram of these recorded song with the same parameters used to compute the spectrograms of the synthetic sounds, and used an Application Programming Interface (API) from Keras (img_to_array) to convert the images to a numerical arrays. Then, we used those data as inputs on our network, and estimated the predicted class by inspecting the output values in the last layer of our network. We assigned the class as the order of the unit presenting the largest value.

To quantify the success of the procedure we compute the confusion matrix for one numerical experiment. This is a typical layout that allows quantifying the success of the classification procedure. Each row represents an actual class, while the column represent the predicted class. For one representative numerical experiment, this reads as shown in Table 2.

Table 2
Confusion matrix, computed for one numerical experiment.

Actual\predicted	1	2	3	4	5	6
1	1	0	0	0	0	0
2	0	9	0	0	0	1
3	0	0	8	0	0	1
4	0	0	2	3	1	0
5	0	0	1	0	8	0
6	1	0	0	0	0	8

Table 3
Precision, recall and f1_score, computed from a numerical experiment.

Class	*P*	*R*	*f1_score*	Support
1	1.00	1.00	1.00	1
2	1.00	1.00	1.00	10
3	0.57	0.89	0.70	9
4	1.00	0.5	0.67	6
5	1.00	0.67	0.80	9
6	0.9	1.00	0.95	9
Avg./tot	0.89	0.84	0.84	44

From the confusion matrix, it is possible to compute a set of numbers that summarizes the performance of the network in the classification task. These are the recall, the precision and the f1_score. The precision for a class is defined as the ratio between the true positives of the class, divided by the total number of times that the class was predicted. The recall of a class is defined as the ratio between the true positives of the class and the total number of times that elements of that class were tested on the network. In other words, the precision (P) indicates, given all the predicted labels (for a given class), how many instances were correctly predicted. The recall (R), on the other hand, indicates for all instances that should have a label X, how many of these were correctly labeled. The *f1_score* measures a balance between these two indices. The indices computed from the confusion matrix in Table 2 are shown in Table 3.

Notice that not all the birds were classified equally well. The short song of the third individual has the lowest precision. Two times a song of the fourth individual, and once a song of the fifth individual were wrongfully mistaken by a song produced by the third individual. On the other hand, the fourth bird had the smallest recall. Half of the times the network was exposed to a song of this individual, it classified it wrongfully, probably due to partial matches between segments of this long song with the shorter songs of other birds. Despite these outliers, the average f1_score reaches a value of 0.84, a reasonable value, particularly since the network was never exposed to the spectrograms of real songs during its training. This numerical experiment led to a particular precise network. With the same training set, the fitting procedure for the network was performed eight times, and the confusion matrices were built for the classification of the six birds in our study. The average values and standard deviations for the precision, recall and f1_score were $P = 0.85 \pm 0.02$, $R = 0.81 \pm 0.03$, $f\_1 = 0.81 \pm 0.03$.

6. Conclusions

Deep learning is one of the most powerful techniques for classification tasks. Yet, the need of large data sets to fit the parameters of convolutional-layered networks, poses a challenge when the data available for training is scarce. This is the case in our problem, which consists of identifying individual birds by their songs.

We addressed the issue by training the network with surrogate solutions generated by a dynamical system. In this way, the network was entirely trained on synthetic data.

In general, the necessity of large data for training exists even for problems where the data is not that scarce. This has led to a strategy called augmentation, which consists of creating surrogate data through a set of operations on the data. Since the whole program of layered networks has been inspired by how the visual neural system works, the default operations considered in augmentation (and implemented in free distributed libraries as Keras) typically consist of transformations that would be natural for generating the different images of an object observed from different perspectives. In this way, many of the operations usually considered in augmentation might not be consistent with the actual variations meaningful to a particular problem, unless it is visual in nature. In our case, for example, a vertical shift in a spectrogram does not cover a realistic fluctuation (birds do not shift, generically, the frequencies of whole vocalizations upwards). Our dynamical system approach can be programmed to generate realistic fluctuations in actual physiological parameters of the problem. As a general observation, augmentation needs to be adapted to the question and its context, a problem that is capturing growing attention [15,16].

Dynamical systems, a most simplified description of how a system behaves (aiming at capturing the utmost minimal ingredients needed to understand the processes involved), and deep learning (a procedure that ostentatiously gives up on unveiling mechanisms) seem two irreconcilable extremes in the spectrum of ways to address our understanding of nature. In this work, both approaches have been used together to solve an identification problem. A dynamical system that has been derived from the understanding of the phenomenon being described is capable of capturing a large amount of coherent features of the problem. For example, a model for labial oscillations operating at parameter values close to where a saddle node in a limit cycle occurs, will generate solutions with a robust relationship between fundamental frequency and spectral content [17]. This implies that the effort of modeling the soundś fundamental frequency will automatically pay in sounds with the right harmonic content. In this way, spectrograms that are computed from sounds generated with this model, with slight variations in their fundamental frequencies, will present harmonics that need not be independently modeled. This consistency allows reducing the size of training and validation sets.

Developing voiceprints for individual animals will allow addressing important ethological questions, as well as the development of tools for monitoring ecological populations. It would not be surprising in the near future to find the tools of deep learning playing a key role in the development of new tools for neuroethology and ecology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.G. B. Mindlin on behalf of both authors.

CRediT authorship contribution statement

P.L. Tubaro: Conceptualization, Formal analysis. **G.B. Mindlin:** Conceptualization, Formal analysis.

Acknowledgments

The authors acknowledge a grant UBACYT 20020170100220BA, a grant PICT 1802-E2 from ANCyT, and a grant from the Richard Lounsbery Foundation.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning.. Nature 2015;521(7553):436.
[2] Chollet F. Deep learning with python. New York: Manning Publications and Co. Shelter Island; 2018.
[3] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Cognit Model 1988;5(3):1.
[4] H. Goeau, S. Kahl, H. Glotin, R. Planqué, W.P. Vellinga, & A. Joly, (2018). Overview of Birdclef 2018: monospecies vs. soundscape bird identification.
[5] Martinsson J. Bird species identification using convolutional neural networks Ms. Thesis. Sweden: University of Gothenburg; 2017.
[6] Mindlin GB. Nonlinear dynamics in the study of birdsong. Chaos: Interdiscip J Nonlinear Sci 2017;27(9):092101.
[7] Amador A, Perl YS, Mindlin GB, Margoliash D. Elemental gesture dynamics are encoded by song premotor cortical neurons. Nature 2013;495(7439):59.
[8] Kopuchian C, Lijtmaer DA, Tubaro PL, Handford P. Temporal stability and change in a microgeographical pattern of song variation in the Rufous-collared sparrow. Anim Behav 2004;68(3):551–9.
[9] Handford P. Trill rate dialects in the Rufous-collared sparrow, *Zonotrichia capensis*, in northwestern Argentina. Can J Zool 1988;66(12):2658–70.
[10] Goller F, Suthers RA. Role of syringeal muscles in controlling the phonology of bird song. J Neurophysiol 1996;76(1):287–300.
[11] Laje R, Gardner TJ, Mindlin GB. Neuromuscular control of vocalizations in birdsong: a model. Phys Rev E 2002;65(5):051921.
[12] Perl YS, Arneodo EM, Amador A, Goller F, Mindlin GB. Reconstruction of physiological instructions from Zebra finch song. Phys Rev E 2011;84(5):051909.
[13] Bush A, Döppler JF, Goller F, Mindlin GB. Syringeal EMGs and synthetic stimuli reveal a switch-like activation of the songbird's vocal motor program. Proc Natl Acad Sci 2018;115(33):8436–41.
[14] Gardner T, Cecchi G, Magnasco M, Laje R, Mindlin GB. Simple motor gestures for birdsongs. Phys Rev Lett 2001;87(20):208101.
[15] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 2017;24(3):279–83.
[16] Ko T, Peddinti V, Povey D, Khudanpur S. Audio augmentation for speech recognition. Sixteenth annual conference of the international speech communication association; 2015.
[17] Amador A, Mindlin GB. Beyond harmonic sounds in a simple model for birdsong production. Chaos: Interdiscip J Nonlinear Sci 2008;18(4):043123.