



## REVIEW

# The science of birdsong and the spectrogram, the technique that changed it all [version 1; peer review: 2 approved with reservations]

Ana Amador <sup>1,2</sup>, Gabriel B. Mindlin<sup>1-3</sup>

<sup>1</sup>Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Ciudad Universitaria, Buenos Aires, 1428, Argentina

<sup>2</sup>CONICET - Universidad de Buenos Aires, Instituto de Física Interdisciplinaria y Aplicada (INFINA), Ciudad Universitaria, Buenos Aires, 1428, Argentina

<sup>3</sup>Departamento de Matematica Aplicada, Universidad Rey Juan Carlos, Móstoles, Community of Madrid, 28933, Spain

**V1** First published: 17 May 2023, 2:9  
<https://doi.org/10.12688/molpsychol.17520.1>  
Latest published: 17 May 2023, 2:9  
<https://doi.org/10.12688/molpsychol.17520.1>

## Abstract

It is common for significant advances in science to come hand in hand with innovative techniques. The study of birdsong was able to immensely expand incorporating novel rigorous investigations when a graphic and informative representation of sound was achieved. The result of this technique, the spectrogram, allows describing the pitch modulations and timbre properties of sounds as a function of time. In this paper we review its history and some of its applications.

## Keywords



Songbirds, vocal production, spectrogram





This article is included in the [Foundations of Neuroethology](#) collection.

## Open Peer Review

Approval Status  

	1	2
version 1		
17 May 2023	<a href="#">view</a>	<a href="#">view</a>

1. **Eduardo Mercado** , University at Buffalo, Buffalo, USA
2. **Dan Stowell** , Tilburg University, Tilburg, The Netherlands  
Naturalis Biodiversity Center, Leiden, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Gabriel B. Mindlin ([gabo@df.uba.ar](mailto:gabo@df.uba.ar))

**Author roles:** **Amador A:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mindlin GB:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The work was partially funded by the University of Buenos Aires (UBA, grant 20020130100094BA), the National Scientific and Technical Research Council (CONICET) and the National Agency for the Promotion of Research, Technological Development and Innovation (ANPCyT, grants PICT-2018-0619 and PICT-2017-4681), Argentina.

**Copyright:** © 2023 Amador A and Mindlin GB. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Amador A and Mindlin GB. **The science of birdsong and the spectrogram, the technique that changed it all [version 1; peer review: 2 approved with reservations]** Molecular Psychology: Brain, Behavior, and Society 2023, 2:9 <https://doi.org/10.12688/molpsychol.17520.1>

**First published:** 17 May 2023, 2:9 <https://doi.org/10.12688/molpsychol.17520.1>

## Birdsong before the spectrogram

William Hudson (1842-1922) was an Argentine-British naturalist. Born in Argentina, he emigrated to England at the age of 33. Upon his arrival in his adoptive homeland, he wrote about the experience of hearing the song of a Eurasian black-cap (*Sylvia atricapilla*) for the first time: “In my childhood, this bird warbled to me through the lines of a poem I read, and now, many years later, I hear the song for the first time. It is beautiful, but how unlike what I had imagined!” This anecdote [Wilson, 2016] illustrates the historical difficulty of incorporating birdsong into scientific research: how to transmit an accurate, informative description of a song.

There is a field of human culture in which we faced a challenge similar to the problem of transmitting a birdsong’s description: music. Both human music and birdsong involve the production of sounds with varying pitch, timbre and duration. However, there are also differences. Human music evolved as a collective activity, leading different cultures to develop a consensus on the discretization of the pitch (*i.e.*, the definition of discrete notes to be used) and the measurement of duration times based on which to build a musical discourse [Tsuji & Muller, 2021]. In contrast, the vocalizations of birds typically consist of a succession of acoustic elements during which the pitch of the sound is continuously modulated. While some bird species do exhibit a predominant rhythm in their songs [Norton & Scharff, 2016], there is generally no adherence to prescribed rules for the duration of the acoustic elements in birdsong production. In addition, the possible timbres of birdsong are diverse, with over 10,000 known bird species capable of producing a wide range of sound types, from very harsh to surprisingly tonal ones.

If we focus on the similarities between birdsong and music, we can find precedents for strategies used to generate a visual record of a given succession of acoustic elements. In the 9th century, a notation for Gregorian choirs began to spread in European monasteries, called “cheironomic neumes” [Tsuji & Muller, 2021]. These consisted of inflectional marks indicating the general form of the music, but not necessarily the exact notes or rhythms to be sung. The marks were written from left to right, and an ascending mark corresponded to a modulation of the pitch of the sound from low to high frequencies. The problem with this notation was that the system was not very accurate and it was immensely complicated. Beginning in the 14th century, this representation in space of frequencies against time (in which frequencies are indicated by the height of the symbols, and time is read from left to right) was ordered by writing the symbols on a set of lines (staves), which regulated the heights of the acoustic elements. Also, the relative duration of the notes was organized in fixed symbols (whole note, half note, *etc*) [Tsuji & Muller, 2021].

Attempts have been made to use musical notation to capture the characteristics of the songs of different bird species. One of the most successful and known examples is the work of composer and ornithologist Olivier Messiaen (1908–1992), who incorporated evocative fragments of various birdsongs

into compositions like *La Nativité*, *Quatuor*, and *Vingt regards* [Schultz, 2008]. Another example is music inspired in songbirds, as the *Symphony No. 6 “Pastoral”* by Ludwig van Beethoven, or the 1952 work “*Merle noire*” by Messiaen, in which the flute plays melodies identifiable with the song of a blackbird [Bowden, 2008]. However, using musical notation to accurately transcribe bird song requires a high level of musical training and may not be accessible to those without it. An alternative approach is the use of cheironomic neumes (see previous paragraph and [Tsuji & Muller, 2021]). This approach, used by Saunders in 1935 to illustrate his description of Swamp sparrows (*Melospiza georgiana*) [Saunders, 1935], allows for the representation of modulations in pitch and duration in a qualitative way, but does not capture timbre or other aspects of the song in a quantitative manner. Even with this simplified approach, transcribing birdsong effectively requires a certain level of skill and talent.

Another approach, used extensively in Bird Field Guides for birders is the use of onomatopoeias to mimic the birds’ songs (see for example, [Narosky & Yzurieta, 2010]). This could be helpful when being in the field but not accurate to capture acoustic details of birdsongs.

## The appearance of the spectrograph

The spectrograph was a machine capable of generating a pictorial representation of the sound, informative of its pitch modulations and timbre. The operation of a spectrograph is based on the mathematical principle of Fourier’s theorem, which states that any periodic function  $f(t)$  can be broken down into a sum of terms (in principle, an infinite number of them), each of which is a trigonometric function [Pipes, 1946]:

$$f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{T} t\right) + b_n \sin\left(\frac{2\pi n}{T} t\right), \quad (1)$$

with the coefficients  $a_n$  and  $b_n$  given by integrals of the product between these trigonometric functions by the function to be analyzed, that is:

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos\left(\frac{2\pi n}{T} t\right) dt, \quad (2)$$

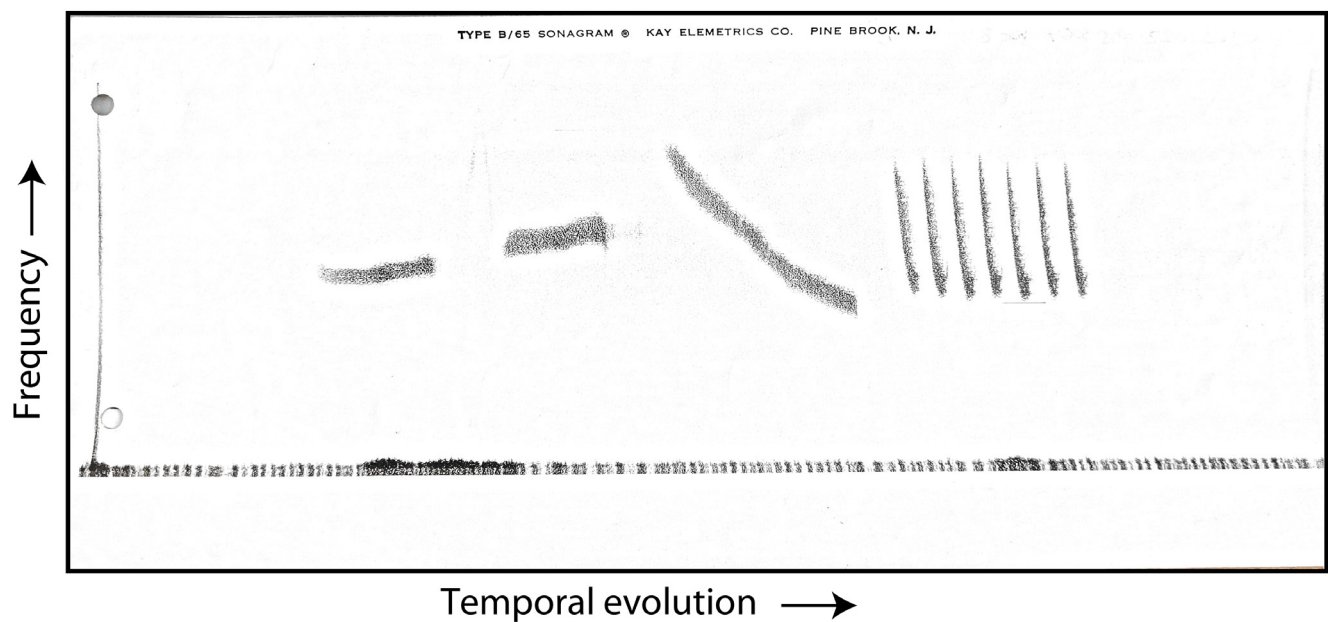
$$b_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \sin\left(\frac{2\pi n}{T} t\right) dt \quad (3)$$

With the exception of one term in the sum ( $a_0$ ) in Equation (1), all the others terms are periodic functions with periods  $T_n$  that are different submultiples of the original period ( $T_n \equiv T/n$ ). Therefore, these functions are also periodic functions with the same period  $T$  as the original function. If the function of interest is not periodic, we can consider the function as a fragment of a hypothetical periodic function with a period of  $T$ . This means we extend the original signal indefinitely, adding copies of the original signal in both directions in time, and treat new signal as a  $T$ -periodic function. In this way, we can apply Fourier’s theorem to analyze our original non-periodic time series. The key to writing this series is the ability to compute the weight of each term

in the expansion. The set of all the weights necessary to reproduce the original sound ultimately constitutes its “fingerprint”. Some terms will weigh more than others, determining the fundamental properties of the sound (such as its timbre) [Calus & Fairley, 1970; Crawford 1968]. Additionally, we can analyze small fragments of the sound centered on a series of consecutive times and create a three-dimensional diagram to describe the sound. One axis of this diagram, would be for the frequencies of the different terms (which are simply the inverses of the periods). The other axis would be for time. Finally, for each time and frequency, the third axis would contain points at heights proportional to the weights of each term in our expansion. If the weights are encoded in a gray scale (for example, high values of the weights would correspond to dark grays, and small values to light grays), a two-dimensional graphical representation of the sound can be created. The good news is that there is a simple mathematical formula that allows us to calculate these weights for a given sound signal using Equation (2) and Equation (3). The bad news is that the process involves (infinitely) many integrals. Even calculating a relatively small number of terms was an enormous challenge in the 1950s [Cooley *et al.*, 1967]. The solution to this challenge was to build a physical device, a piece of hardware, that could generate a diagram like the one we have just described without needing to perform the calculations of the terms’ weights in the Fourier series. Around 1940, a group of researchers at the Bell Telephone Laboratories decided that it was time to develop methods for making the details of speech more visible and intelligible [Potter *et al.*, 1947]. Also, it was wartime so a driving force was the need to monitor movements of ships and submarines. As a consequence, much of the information remained classified until World War II was over

(see historic details in [Marler, 2004]). Soon after hostilities ended a company was created to build and market a machine for visible speech. This device was the spectrograph, described in 1946 by Koenig, Dunn and Lacy [Koenig *et al.*, 1946].

A spectrograph consists of several components. The first one was a support material for the sound being analyzed [Koenig *et al.*, 1946]. Then, there was electronic hardware that transformed the recorded signal into an electrical one, which was subsequently filtered by a bandpass circuit (an electronic circuit that selected only those components of the signal whose frequencies were within a pre-established range). The process was repeated multiple times, with the parameters of a modulator being changed in order to bring different regions of the spectrum into the filter’s frequency range. A thermal paper was used to record the output of this repetitive process. This paper was placed on a rotating drum, synchronized to complete a full turn in the time it took to analyze the sound once. The printing mechanism was a hot tip that left a mark on the thermal paper, the density of which was proportional to the output of the filter. With each repetition of the process, the modulation was slightly modified to increase the frequencies selected by the filter, and the position of the tip moved upwards horizontally. By repeating this process multiple times, the spectrograph generated a figure that could be read from left to right, with marks indicating the presence of sound in a frequency band (at a vertical position proportional to the frequency), whose duration was inferred from the horizontal length of the stain. An example of thermal paper marked by a spectrograph is shown in Figure 1. Hard to describe is “the delicate smell of ozone left when the procedure was over” (D. Margoliash, personal communication).



**Figure 1.** Scanned thermal paper generated by a spectrograph showing the analogical spectrogram of a rufous-collared sparrow (*Zonotrichia capensis*) song. Each continuous line defines a syllable. The initial three syllables compose the introductory theme, characteristic of each individual. The following seven short syllables constitute the trill. This figure has been adapted and reproduced with permission from [Tubaro, 1990].

The use of the spectrograph revolutionized the study of bird-song, providing a highly informative graphic representation of sound. Prior to this technique, it was difficult to validate basic features of birdsong. An illuminating example is the use of two independent sound sources. These sounds are generated simultaneously, so that the only other graphic representation existing to date (the graph of sound wave pressure fluctuations) could not shed light on the phenomenon. Peter Marler describes the first use of spectrography in the study of birdsong [Marler, 2004], carried out by Donald Borror at Ohio State University in 1953 [Borror & Reese, 1953]. Those studies allowed a deeper understanding of the songs of sparrows, thrushes, and wrens. However, perhaps the most impressive results of this technique in the field of birdsong were those that began to emerge from Thorpe's group, who in 1950 purchased one of the first spectrographs imported into Britain [Marler, 2004]. Spectrography allowed for the study of micro and macro variations in the songs of individuals within a species, providing a new perspective on the cultural evolution of vocal communication and its influence on speciation. Focusing on chaffinches (*Fringilla coelebs*), they applied spectrographic analysis to the problem of vocal learning, giving oscine birds a preferential place as animal models. Peter Marler had already written about dialects in 1952, stating that the geographical variations in the song of chaffinches are phenotypic, due to vocal learning [Marler, 1952]. But it was only after incorporating spectrographic tools that Marler was able to study the subtleties of song variation in white-crowned sparrows (*Zonotrichia leucophrys*) in San Francisco Bay [Baptista & Wells, 1975; Marler & Tamura, 1962]. These micro and macro variations present in the songs of individuals within a species, interpretable within the framework of vocal learning, opened a new and quantitative perspective on the problem of cultural evolution in vocal communication and its influence on speciation. As Peter Marler states, "spectrographic analysis elevated dialect studies from the level of the enthusiastic observer to that of a scientific subject of investigation" [Marler, 2004].

The use of spectrograms quickly spread throughout the academic world. However, the equipment was large, expensive, and predated commercial portable tape recorders making it difficult to obtain sound recordings of field work during the 1960s. An example of this transition can be seen in the work of F. Nottebohm, who in the mid-sixties, spent five months in Argentina without access to portable recorders [Nottebohm, 1969]. These months were, however, ideal for studying the song of the rufous-collared sparrow (*Zonotrichia capensis*) in different parts of the country. This species' song consists of an introductory theme of two to four syllables followed by a trill (see Figure 1). The rate of syllable generation in the trill is a characteristic of the country region inhabited by the bird. On the other hand, the structure of the introductory theme is an identifying trait of the individual. These themes are relatively tonal sequences of syllables in which the pitch modulates upwards or downwards. Deprived of better tools, Nottebohm annotated the various themes found in different parts of the country using "Saunders"-style notes [Nottebohm, 1969]. Back at Rockefeller University, he *calibrated himself* doing the following: he took a song, described it with his

schematic notation, and then compared it to the result of analyzing the song with a spectrograph. This was a resourceful way to do science in 1966. As a side note, Nottebohm turned out to be quite accurate in his descriptions, since his schematic notes were similar to the pictures generated by the spectrograph [Bistel *et al.*, 2022; Nottebohm, 1969].

## The history of digitalization

The existence of hardware capable of spectrally analyzing sound and creating an informative image of it marked a significant change in the sound analysis in general and in birdsong in particular [Marler, 2004]. However, the digitization of sound signals allowed researchers to examine the details of sound signals at various scales. In the context of audio signals, a significant moment in the history of digitization was 1937, when engineer Alec Reeves developed the use of pulse code modulation (PCM) in telecommunications [Reeves, 1968]. PCM is an efficient method for digitizing analog signals displayed at discrete time intervals. In this process, the analog signal is sampled at regular intervals, and the recorded amplitudes are approximated to the closest value in a set of possible discrete values. This procedure, along with the expression of these numbers in binary language, made it possible to computationally process sound, allowing for the application of programmable operations on the signal without the need for specific physical equipment as the spectrograph. However, the first commercial digital recorder was not released until 1977, with Sony's PCM-1, which combined digital analog processors with a Betamax tape recorder to allow for digital recording and playback (see Sony History Chapter for further details). These innovations greatly enhanced our ability to analyze and understand the singing of birds and of sounds in general.

A digital format for the recording and storage of acoustic data opened the doors to its computational processing. Furthermore, there was a numerical algorithm (listed among the top 10 algorithms of the 20th century [Dongarra & Sullivan, 2000]) that allowed acoustic data processing with low computational costs: the fast Fourier transform (FFT).

To understand why this algorithm was so revolutionary in the field of acoustic signal analysis, let's remember some basic elements that enter into the description of a sound phenomenon. As we discuss in the previous section, every periodic function can be written as an infinite sum of trigonometric functions as shown in Equation (1). Based on this rigorous mathematical result, when working with real experimental signals certain compromises need to be done. We have already discussed that in order to work with non-periodic functions (*e.g.* signals of an acoustic phenomenon, such as the song of a bird), we can assume that the signal fragment to be analyzed repeats infinitely, so that the total duration time of the sample becomes the period. Then it is calculated the weight of the components of the trigonometric functions used to reconstruct the "periodic" function ( $a_n$  in Equation (2) and  $b_n$  in Equation (3)). Now, this restriction puts a lower bound on the frequencies of the problem. The inverse of the total duration time of the signal becomes a minimum frequency of the



problem, and each term will be associated with a frequency that will be a multiple of that minimum frequency [Calus & Fairley, 1970; Crawford, 1968]. In other words, unlike the mathematical description with which we started the discussion, in the case of a real problem, the possible frequencies are multiples of one, very small, associated with the size of the signal. On the other hand, when digitizing the signal, a new concept comes into play: the sampling time, which is the time interval between those successive records taken at regular time intervals. Since it is not possible to elucidate what happens between two of these successive times, the sampling frequency imposes a maximum on the frequencies that enter in the description. Unlike the mathematical problem in which time is a continuous variable, a digitized signal is composed by a finite number of samples (taken at a finite number of times), and the inverse of the difference between successive times sets a limit to the number of terms that it makes sense to include in a sum. In this way, Equation (1) will sum from  $n = 1$  to  $n = N$ . This sum will be a mathematical approximation of the signal to be analyzed. There is an additional element that we must take into account when discussing how to carry out a graphic representation of a sound signal. We said that we were interested in being able to describe how pitch, for example, is modulated along a syllable. This means that when describing the weights of a Fourier series decomposition of a signal, we will not be interested in analyzing the entire function. On the contrary, we would be interested in analyzing successive fragments. To do that, filters  $f_i$  can be defined, that is, functions that hold the identity in a neighborhood of an instant  $t_p$ , and zero away from that neighbourhood. In this way, the following function can be generated:  $g_i(t) \equiv f(t)f_i(t)$ , so that the task is set to spectrally analyze those  $i$  segments [Boersma & Van Heuven, 2001].

In the previous paragraph we discuss that the digitalization of the signal will give as a result a discrete signal and this will result in  $N$  terms to calculate in Equation (1). The computation of the coefficients is a problem that scales as  $N^2$ , since it takes the order of  $N^2$  operations to spectrally describe the function ( $N$  coefficients, each one requiring  $N$  sums). Having a problem of order  $N^2$  makes it impractical for computational calculations as  $N$  is generally a big number. The FFT method takes note of a multiplicity of operations whose result can be inferred from others. This leads the problem of calculating the FFT to be  $O(N \log N)$ , resulting in significantly less terms to calculate than  $N^2$  [Cooley & Tukey, 1965]. It is difficult to emphasize enough the impact of this technology in the field. For example, the applications that perform spectrograms in real time, within the framework of smart phone applications, are based in calculation of FFT analysis. These applications brought powerful analytical tools to thousands of naturalists and researchers, and could not be run in a small computational device as a smart phone without implementing an FFT analysis. Useful and revolutionary as this algorithm is for birdsong research, it is interesting to mention that John Tukey (co-responsible with James Cooley for its present implementation and realizing that the algorithm was  $O(N \log N)$ ), came up with the idea during a meeting of a presidential advisory committee. The topic of discussion was not birdsong,

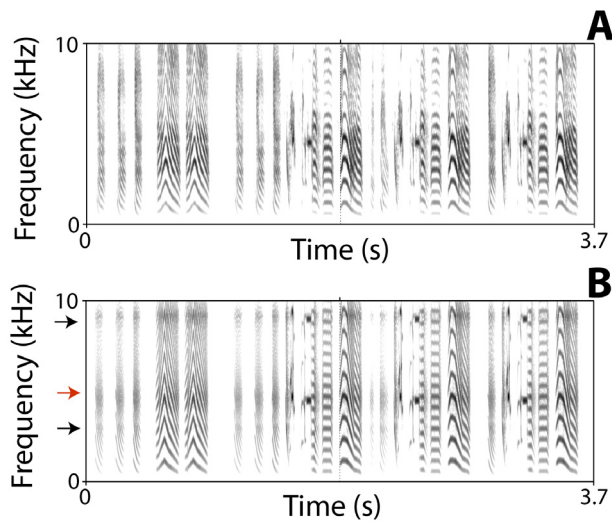
but how to detect nuclear tests by the Soviet Union, analyzing seismological time series obtained from off-shore seismometers [Rockmore, 2000].

## Analyzing spectrograms

The spectrogram, then, is a graph in which we display, for different times  $t_p$ , the spectral analysis of the product of two functions: the sound signal  $S(t)$ , and a window  $w_i(t)$ . Some aspects of the resulting graph will be due to the parameters used in the procedure, and others will reflect the properties of the signals to be analyzed. As an example, let's consider a sequence of signals, which are essentially zero, except in the neighborhood of a given time  $t_p$ , in which they resemble the original acoustic signal. So even if the original signal were a sinusoidal function, none of the signals analyzed when computing a spectrogram is actually a sinusoidal function. To represent a function that is similar to a sine function in the neighborhood of a given instant and decays to zero when moving away from that value, many sinusoidal functions of similar frequencies should be added. An intuitive explanation of the phenomenon is the following. By adding many sinusoidal functions of similar frequencies, in phase at a time  $t_p$ , it is achieved that, around that time, the functions are approximately in phase, giving rise to a signal resembling a sine [Calus & Fairley, 1970; Crawford, 1968]. By times away from  $t_p$ , however, the different functions that were added are going to be out of phase, since they all had arguments that incremented at slightly different rates. For this reason, even if the original sound is tonal (single fundamental frequency), the windowing implies a broadening of the frequencies present.

How do we interpret a spectrogram? If we have, as in Figure 1, a series of lines, the interpretation is straightforward: each continuous line represents a continuous sound. The height of the line indicates the pitch of the sound, and if there are variations in the vertical position of the line from left to right, they are interpreted as pitch variations across time. Figure 2, however, presents us with a slightly greater challenge, as multiple frequencies are present at the same time. This means that the sound when analyzed in the FFT framework has a spectral decomposition with multiple terms. In this case, the timbre of the sound has more richness than in cases where a single frequency carries the majority of the representation for each beat.

Why are some birdsong sounds spectrally rich, while others are not? This is associated with the mechanism of avian phonation and the region of physiological parameters in which the animal operates. It has been known since the 1990s that birds use a phonation mechanism similar to the one used by humans to produce voiced sounds: the modulation of airflow through the creation of self-sustaining oscillations in flexible tissue [Elemans *et al.*, 2015; Goller & Larsen, 1997a; Mindlin & Laje, 2006; Suthers *et al.*, 1999]. In the case of oscine birds (a suborder containing over 4000 species within the passerine order), the vocal apparatus (called the syrinx) is a highly conserved structure [Ames, 1971]. It consists of opposing pairs of membranes called labia (medial and ventral) at each junction between the bronchi and trachea. Now, a pair of labia



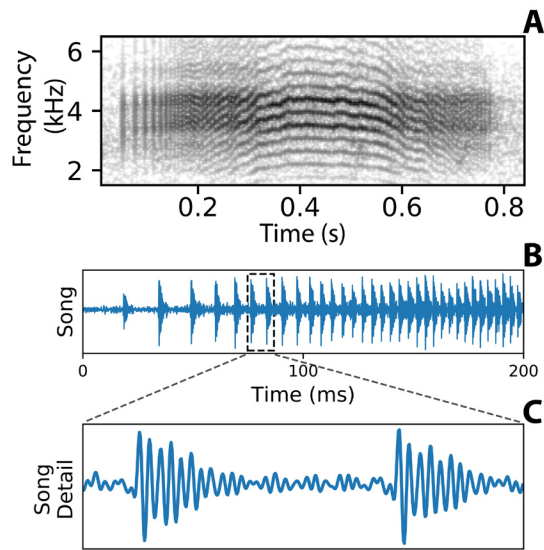
**Figure 2. Digital spectrogram of a zebra finch (*Taeniopygia guttata*) song and its synthetic copy.** The recorded birdsong (A) can be copied by a synthetic song (B) generated by a dynamical systems model for the vocal apparatus. Each panel shows the recorded sound wave and the digital spectrogram generated from it. The red arrow in (B) indicates a frequency range where resonances exist and therefore the magnitude of the frequencies are enhanced. In this case, the resonance of the oroesopharyngeal cavity (4 kHz) is the most salient feature. This figure has been adapted and reproduced with permission from [Amador & Mindlin, 2014].

can begin to gently oscillate when the flow of air passing between them is fast enough, similar to how two pieces of paper held very close to each other will begin to oscillate when air is blown between them [Goller & Larsen, 2002; Mindlin & Laje, 2006]. These oscillations are regular and harmonic, resulting in a tonal sound [Larsen & Goller, 1999]. However, labia can also begin to oscillate in another way. Imagine the labia pressed against each other. During the exhalation process, if the air reaches a sufficient level of pressure, it will eventually force its way through them. But by temporarily opening a channel for the air to pass through, the pressure drops and the labia close again. This type of labia movement is “explosive” (*i.e.* resulting from a short opening of the labia) [Goller & Larsen, 2002; Jensen *et al.*, 2007] and the oscillations can be generated with arbitrarily low frequencies (that is, a long time elapses between explosive releases of air). In this case, the sound signals will be spectrally richer [Amador & Mindlin, 2008; Sitt *et al.*, 2008].

The zebra finch is an example of a bird that generates sounds with a high spectral content, as well as more tonal ones (see Figure 2.A). It is interesting that there is a simple functional relationship between the spectral content and the fundamental frequency in the sounds in this species. This relationship is a consequence of the dynamical mechanism behind the labial oscillations at the sound source [Amador *et al.*, 2013; Sitt *et al.*, 2008]. Upon closer examination, we can see that some of the frequencies present in spectrally rich sounds are more prominent (*i.e.*, the coefficients corresponding to the

respective terms in a Fourier expansion will have higher values). For example, in Figure 2, there are prominent frequencies around 4 kHz. These frequencies are emphasized by the filters of the avian vocal apparatus. After being generated in the sound sources by the effects of flow modulation induced by the labia, the sound passes through the trachea (which behaves like a closed-end tube) and excites the oroesopharyngeal cavity (which behaves like a Helmholtz resonator). In fact, Figure 2.B shows a synthetic sound generated by a mathematical model that implements these physical processes. One of the advantages of using a dynamical model to replicate a birdsong is that its success builds confidence on the hypotheses on which the model was constructed. But beyond that, the model's parameters can be explored in order to understand how the different physical features of the animal's anatomy affect the acoustic properties of the song [Amador *et al.*, 2013; Mindlin, 2017]. More specifically, modifying the parameters that characterize the different filters of the system, allows to emphasize portions of the spectrum corresponding to a given filter [Fletcher *et al.*, 2006; Gardner *et al.*, 2001; Sanz Perl *et al.*, 2011]. The 4 kHz frequency (red arrow in Figure 2.B) corresponds to the resonance of the oroesopharyngeal cavity. This was verified by changing the volume of the Helmholtz resonator (that represents the oroesopharyngeal cavity in the model) and finding that the enhanced frequencies varied following the model's predictions [Amador *et al.*, 2013; Sanz Perl *et al.*, 2011]. To generate a spectrally rich sound source, ultimately filtered by the passive vocal tract, the labial dynamics has to be ruled by nonlinear equations. Therefore, the subtle structure of harmonics that enriches the spectrogram vertically is the combination of nonlinear phenomena ruling the sound source, and a passive vocal tract that filters the spectrally rich sounds emitted by the sound source.

A mechanism similar to that used by the zebra finch to generate low-frequency sounds is employed by the suboscine *Phytotoma rutila*, a South American bird whose unusual vocalizations resemble the sound of a rusty hinge [Uribarri *et al.*, 2020]. A spectrogram of the vocalization is shown in Figure 3.A. In [Uribarri *et al.*, 2020] the model of a source generating explosive sounds that are filtered by an oroesopharyngeal cavity was tested by studying vocalizations of this species (whose song is not learned) recorded at different altitudes. The sizes of specimens from those altitudes were estimated from tarsus measurements. After verifying that the mean sizes of the specimens correlated with altitude using scaling laws, it was found that the frequencies emphasized in the song spectrograms decreased with altitude [Uribarri *et al.*, 2020]. This example is interesting because it provides an example of bird anatomy being encoded in song, which is not common because many species have the ability to adapt their filters and mask this effect [Riede *et al.*, 2006]. In Figure 3 we show the sound signal, with its low frequency bursts, and a magnified segment illustrating that each sound bursts consists of a damped oscillation. The model tested in [Uribarri *et al.*, 2020] proposes that a pulse tone sound filtered by the oroesopharyngeal cavity can account for this unique bird sound. In Figure 3 we display the time series data corresponding



**Figure 3. *Phytotoma rutila* song.** Digital spectrogram of the song (A), the waveform of the initial segment (B); and a soundwave detail of two sound segments (C). This figure has been adapted and reproduced with permission from [Uribarri *et al.*, 2020].

to the sound uttered by a *Phytotoma rutila*, that consists of bursts generated at approximately 100 Hz. Each burst consists of damped oscillations of approximately 4 kHz: these are interpreted as the result of filtering “explosive” sounds generated at approximately 100 Hz. The spectrogram in Figure 3.A shows a darker region around 4kHz during the whole vocalization, indicating the presence of a filter that enhances frequencies at about 4 kHz.

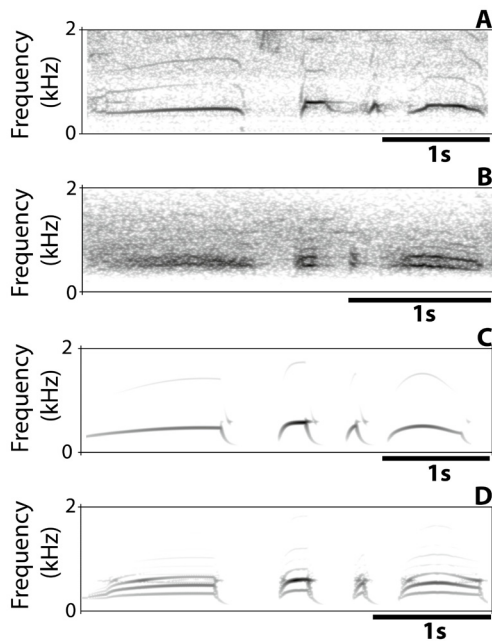
In these examples, we recognize similarities to human vocal phonation of voiced sounds, such as vowels, in which a low-frequency oscillation generated by the modulation of airflow by the vocal cords generates a sound that is subsequently filtered by the vocal tract. Subtle articulations change the emphasis of different harmonics, generating what we recognize as different vowels [Titze, 1994]. However, this is not the norm in bird song, where it is common to find tonal sounds. Through X-ray cinematography, it was found that in northern cardinals (*Cardinalis cardinalis*) song is accompanied by movements of the hyoid skeleton and changes in the diameter of the esophagus, which allow for the adaptation of the volume of the oroesopharyngeal cavity in order to enhance the changing frequency modulations of the sound generated by the sound source [Riede *et al.*, 2006]. In this way, the components of the sound signal corresponding to frequencies higher than the fundamental have little energy. This behavior is believed to be present in many oscine birds displaying tonal sounds in their songs.

All the cases mentioned so far involved a source generating a sound whose pitch was given by a single frequency, and its timbre by the weight of its different harmonic components. However, we have already discussed that many birds are

capable of generating sounds using their two sound sources independently. These dissonant sounds are achieved when the sources generate two pitches that are not harmonically linked. Greenewalt [Greenewalt, 1968] and Stein [Stein, 1968] were the first to report this effect, further suggesting that its origin was the activation of both sides of the syrinx. This observation was experimentally validated by Suthers and collaborators [Goller & Suthers, 1996], who measured airflow across each of the junctions between the trachea and bronchi. The observation of spectrograms corresponding to sounds in which two sources are active makes it clear that the motor controls of each side of the syrinx are independent. Furthermore, the two sound sources can display a wide variety of dynamic behaviors. Since the lateral labia of both sources are supported on a shared physical substrate, the two sources have a mechanical coupling, which allows their dynamics (governed by non-linear equations) to interact [Laje *et al.*, 2008]. Nowicki and Capranica conducted a delicate study of the frequencies present in vocalizations from two sources, finding not only the frequencies associated with the vocalizations generated on each side of the syrinx, but also the combinations that result from their nonlinear interaction [Nowicki & Capranica, 1986]. To do this, they analyzed a frequent vocalization in the black-capped chickadee (*Parus atricapillus*) and showed that the multiple frequencies are heterodyne products resulting from the interaction of two signals.

Surprisingly, there is evidence of non-linear interaction between two oscillators when we carefully explore the spectrogram corresponding to a sound produced by a tracheal syrinx. In a tracheal syrinx there is only one pair of labia responsible for airflow modulation [Goller & Larsen, 1997b]. A non-linear interaction between oscillating labia is possible if there is a significant anatomical difference between the two opposing membranes [Alonso *et al.*, 2016]. Each one will begin to oscillate when the airflow exceeds a certain threshold, but if the syrinx is extremely asymmetric, each membrane will respond differently to the same airflow, and their oscillations may be significantly different. If the asymmetry is not very significant, the oscillations will be synchronous. Otherwise, the oscillations will not lock, and the spectrogram will show a set of bands instead of a single well-defined frequency. The mathematics of this phenomenon was discussed in the context of a comparative analysis between the vocalizations of two phylogenetically close species of pigeons (*Patagioenas maculosa* and *Patagioenas picazuro*) [Alonso *et al.*, 2016]. The temporal modulation of the fundamental frequencies in both species is very similar, but the difference in timbre is significant; in *P. maculosa* there are bands around the fundamental frequency (see Figure 4 and [Alonso *et al.*, 2016]). In addition to mathematically modeling the phenomenon, anatomical exploration of the syrinxes confirmed the hypothesis of the anatomical difference. Figure 4.A and 4.B shows the spectrograms of the vocalizations of both species. Simulations of a computational model implementing the hypothesis were used to account for the structure of the spectra. Figure 4.C shows a spectrogram of soundwaves generated by simulations of a model assuming symmetric membranes and Figure 4.D assuming





**Figure 4. Recorded and synthesized vocalizations of *Patagioenas picazuro* and *Patagioenas maculosa*.** (A) Typical *P. picazuro* vocalizations, (B) *P. maculosa* vocalizations, (C) synthetic songs of *P. picazuro* and (D) *P. maculosa*. This figure has been adapted and reproduced with permission from [Alonso *et al.*, 2016].

asymmetric membranes. Again, a subtle anatomical hypothesis suggested by a precise signature in the spectrogram.

### Spectrograms and artificial intelligence

In 1997, Podos reported his analysis of the calls of 34 species of songbirds [Podos, 1997]. He focused on two specific variables: the rate of syllabic production and the range of frequencies covered during vocalization. In most of the species analyzed, Podos found that the bandwidth of frequencies covered during each syllable decreased as the rate of syllabic production increased. It is plausible to interpret that this restriction reflects a physiological limitation. Since there is a maximum speed at which a muscle can contract, it is parsimonious to expect little modulation in vocalization frequencies when the time available to perform it is short. Based on this hypothesis, it was proposed that generating syllables at the highest limit of this physiological restriction (that is, with the greatest possible frequency range given a syllable production rate) could be an element evaluated when choosing a mate. This idea of *performance* has been highly influential (as well as debated [Kroodsma, 2017]) in the field. Its impact is largely due to its conceptual attractiveness. The relative computational simplicity of the observable, which can be inferred from the vertical and horizontal lengths of the continuous sections present in a spectrogram, probably contributed to its popularity.

In the application that we just discussed, there are specific characteristics that are calculated from the spectrogram, that is,

from the image that represents the sound. It was intuited that these two properties of the image, the length and height of each continuous stretch, could not be varied independently within the limits of physiology, and Podos proceeded to link them by means of a “law” that would be expressed as a limit to the animal’s ability [Podos, 1997].

Recently there has been a growing interest in obtaining information directly from images, without the need to extract a predetermined set of properties [Chollet, 2021]. In charge of these new processes are a set of algorithms known generically as deep learning networks, which are capable of performing a task after being provided not with *rules* but with *examples* [LeCun *et al.*, 2015]. For example, to solve the task of image classification, these algorithms are provided not with a set of pre-established operations, but simply with a set of examples [Chollet, 2021]. In the process known as training, the algorithm has access to the result of the classification obtained when the algorithm used a certain set of parameters, and it is allowed to modify them based on that result. After training a successful algorithm will be capable of correctly classifying images that are different than those used during training [Rawat & Wang, 2017]. It is natural that this type of procedure can be adapted to the analysis of images that visually expresses acoustic properties: spectrograms [Tubaro & Mindlin, 2019].

A classification algorithm based on a neural network, in its simplest version, consists of building a network of processing units arranged in layers [Chollet, 2021]. Each processing unit has an input and an output. In the simplest algorithm of this class, the input consists of a weighted sum of the outputs of the previous layers. The output of a unit is the result of evaluating a non-linear function at the value of the input. The first layer has as input the numerical translation of the object to be classified. For example, in the case of an image, this would be a set of numbers, each one indicating the color or shade of gray for each pixel of the image. The last layer consists of a number of units that matches the total count of classes in terms of which the data is grouped. The nonlinear function in this last layer is usually chosen as a sigmoid function, whose arguments can vary between zero and one. The parameters of the network are the weights with which the outputs of the *i*-th units of each layer enter the *j*-th unit of the following layer. The goal is to find a set of parameters that guarantee that all the images of a class, when processed by the network, result in a single unit of the last layer with values close to one, while the other units give zero [Chollet, 2021].

There are many variations of this algorithm. To classify images, the most widely used neural network architecture is known as a convolutional neural network [Chollet, 2021; LeCun *et al.*, 2015]. This architecture begins with a first two-dimensional layer, in which the numerical translation of the image is introduced. The following stages of the network are each constituted by a set of layers, and the values of the units at those layers result from applying filters to the layers of the previous stage. The adjustment of the model parameters (which are the ones that define the filters) was made manageable

from the computational point of view by introducing a learning method based on back propagation [LeCun *et al.*, 1998]. The applications of this algorithm in science and industry are innumerable. The level of efficiency is such that many classification problems are resolved by initially converting them into an image classification problem. For example, in the field of data science, the annual BirdCLEF competition is an open challenge to the community for the automatic identification of bird species through their song. Since 2017 (more information can be found [here](#)), the winning computational scheme has consisted of processing images that encode the spectral temporal evolution of the song, using neural networks [Fazeka *et al.*, 2018]. Today, there are a wide variety of applications for smart phones that are capable of recording and processing spectrograms in real time. Not only that, but the information is sent to servers that run previously trained neural networks that can identify a bird by its song with significant precision [Kahl *et al.*, 2021].

In a previous section we described Nottebohm's manual recording of the frequency modulations of the syllables that constituted the *Zonotrichia capensis* song themes he heard in Argentina during 1966. Recently, an avian vocal production model was used to generate synthetic songs whose syllables presented modulations similar to those reported in 1966. These songs were then processed with a neural network that was trained to identify the various themes that individuals of the same species sing to each other at one of the sites that Nottebohm visited in the 1960s (Parque Pereyra Iraola, Buenos Aires, Argentina). The network used in this study was not designed to classify songs, but to measure a "distance" between images. It consisted of two identical copies of a network, the last layer of which was made up of a number of units. The values that these units have constitute its representation coordinates in an abstract space. The network is trained in such a way that if the inputs are two images corresponding to the same class, the images have similar representation coordinates. On the other hand, if the network inputs are two images corresponding to different classes, these are represented by two points that are far apart in the abstract representation space. This neural network architecture, called Siamese, can in general give us a metric, a distance, between images. In this specific application, Mindlin and collaborators show that two of the songs that were reported in 1966 were still being sung in the same region in 2022 [Bistel *et al.*, 2022].

## Conclusions

There have not been many leaps in the understanding of nature that have not been accompanied, or even preceded, by the

appearance of some revolutionary technique. The spectrogram, used for the study of birdsong, was the technique that brought the question fully into focus as an object of scientific interest. A spectrogram allows us to generate a visual record of sound, freeing us from the temporality that characterizes our auditory experience. It not only captures vocal timing, like a sound time series, but it also captures pitch modulation and timbre variability, which are essential properties in the description of birdsong.

The history of this technique, on the other hand, has followed the very interesting dynamics of interaction between intellectual curiosity, defense research, and markets. Born from classified projects during World War II, the spectrographs were later appropriated by researchers interested in birdsong, human phonation and seismology, among others. It is difficult to find a higher contrast between a war effort and birdsong research. The market, on the other hand, quickly made portable and later digital recording modes available. The volumes involved in the entertainment industry made these devices rapidly accessible to researchers around the world, with widely varying degrees of funding, but with close proximity to diverse avian species. This allowed our global understanding of some aspects of avian neuroethology to be enriched by the study of a huge variety of species. It is difficult to imagine these studies without a spectro-temporal description of the song carried out at some point.

From the historical description of dialects, the elucidation of the most subtle biomechanical mechanisms used in phonation, or population-scale issues such as monitoring species through song, there is no area of application in avian neuroethology that has not taken advantage of this technique. The market, on the other hand, has made massive automated image analysis simple and efficient. Combining these tools with the data being gathered in global repositories such as *Xeno-canto*, it is possible to foresee, without much imaginative effort, that we will witness in the near future new qualitative leaps in our understanding of avian vocal behavior.

## Data availability

No data are associated with this article.

## Acknowledgements

GBM thanks URJC for its hospitality during his sabbatical stays. The authors thank Dr. Pablo Tubaro for providing a copy of an original spectrograph output (Figure 1) that was part of his Ph. D. Thesis [Tubaro, 1990].

## References

Alonso RG, Kopuchian C, Amador A, *et al.*: **Difference between the vocalizations of two sister species of pigeons explained in dynamical terms.** *J Comp Physiol A Neuroethol Sens Neural Behav Physiol.* 2016; **202**(5): 361–70. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Amador A, Perl YS, Mindlin GB, *et al.*: **Elemental gesture dynamics are encoded by song premotor cortical neurons.** *Nature.* 2013; **495**(7439): 59–64. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Amador A, Mindlin GB: **Beyond harmonic sounds in a simple model for**

**birdsong production.** *Chaos.* 2008; **18**(4): 043123.

[PubMed Abstract](#) | [Publisher Full Text](#)

Amador A, Mindlin GB: **Low dimensional dynamics in birdsong production.** *Eur Phys J B.* 2014; **87**: 300.

[Publisher Full Text](#)

Ames PL: **The Morphology of the Syrinx in Passerine Birds.** (Peabody Museum of Natural History, Yale University), 1971.

Baptista LF, Wells H: **Additional evidence of song-misprinting in the white-crowned sparrow.** *Bird-Banding.* 1975; **46**(4): 269–272.

[Publisher Full Text](#)

Bistel R, Martinez A, Mindlin GB: **An analysis of the persistence of *Zonotrichia capensis* themes using dynamical systems and machine learning tools.** *Chaos Solitons Fractals.* 2022; **165**: 112803.

[Publisher Full Text](#)

Boersma P, Van Heuven VJ: **Speak and unSpeak with PRAAT.** *Glot International.* 2001; **5**(9/10): 341–347.

[Reference Source](#)

Borror DJ, Reese CR: **The analysis of bird songs by means of a vibralyzer.** *Wilson Bull.* 1953; **65**(4): 271–276.

[Reference Source](#)

Bowden S: **The theming magpie: the influence of birdsong on Beethoven motifs.** *The Musical Times.* 2008; **149**(1903): 17–35.

[Publisher Full Text](#)

Calus IM, Fairley JA: **Fourier series and partial differential equations: a programmed course for students of science and technology.** John Wiley & Sons, 1970; 111.

[Reference Source](#)

Chollet F: **Deep learning with Python.** Manning publications, 2021.

[Reference Source](#)

Cooley JW, Tukey JW: **An algorithm for the machine calculation of complex Fourier series.** *Math Comp.* 1965; **19**(90): 297–301.

[Publisher Full Text](#)

Cooley JW, Lewis PAW, Welch PD: **Historical notes on the fast Fourier transform.** *Proceedings of the IEEE.* 1967; **55**(10): 1675–1677.

[Publisher Full Text](#)

Crawford FS: **Waves.** Berkeley Physics Course-Volume 3. McGraw-Hill, 1968; 600.

[Reference Source](#)

Dongarra J, Sullivan F: **Guest Editors Introduction to the top 10 algorithms.** *Computing in Science & Engineering.* 2000; **2**(01): 22–23.

[Publisher Full Text](#)

Elemans CPH, Rasmussen JH, Herbst CT, *et al.*: **Universal mechanisms of sound production and control in birds and mammals.** *Nat Commun.* 2015; **6**(1): 8978.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Fazeka B, Schindler A, Lidy T, *et al.*: **A multi-modal deep neural network approach to bird-song identification.** *arXiv preprint arXiv:1811.04448.* 2018.

[Publisher Full Text](#)

Fletcher NH, Riede T, Suthers RA: **Model for vocalization by a bird with distensible vocal cavity and open beak.** *J Acoust Soc Am.* 2006; **119**(2): 1005–11.

[PubMed Abstract](#) | [Publisher Full Text](#)

Gardner T, Cecchi G, Magnasco M, *et al.*: **Simple motor gestures for birdsongs.** *Phys Rev Lett.* 2001; **87**(20): 208101.

[PubMed Abstract](#) | [Publisher Full Text](#)

Goller F, Suthers RA: **Role of syringeal muscles in controlling the phonology of bird song.** *J Neurophysiol.* 1996; **76**(1): 287–300.

[PubMed Abstract](#) | [Publisher Full Text](#)

Goller F, Larsen ON: **A new mechanism of sound generation in songbirds.** *Proc Natl Acad Sci U S A.* 1997a; **94**(26): 14787–91.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Goller F, Larsen ON: **In situ biomechanics of the syrinx and sound generation in pigeons.** *J Exp Biol.* 1997b; **200**(Pt 16): 2165–76.

[PubMed Abstract](#) | [Publisher Full Text](#)

Goller F, Larsen ON: **New perspectives on mechanisms of sound generation in songbirds.** *J Comp Physiol A Neuroethol Sens Neural Behav Physiol.* 2002; **188**(11–12): 841–50.

[PubMed Abstract](#) | [Publisher Full Text](#)

Greenewalt CH: **Bird song: Acoustics and physiology.** Smithsonian Institution Press, 1968.

[Reference Source](#)

Jensen KK, Cooper BG, Larsen ON, *et al.*: **Songbirds use pulse tone register in two voices to generate low-frequency sound.** *Proc Biol Sci.* 2007; **274**(1626): 2703–2710.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kahl S, Wood CM, Eibl M, *et al.*: **BirdNET: A deep learning solution for avian diversity monitoring.** *Ecol Inform.* 2021; **61**: 101236.

[Publisher Full Text](#)

Koenig W, Dunn HK, Lacy LY: **The sound spectrograph.** *J Acoust Soc Am.* 1946; **18**(1): 19–49.

[Publisher Full Text](#)

Kroodtsma D: **Birdsong performance studies: a contrary view.** *Animal Behaviour.* 2017; **125**: e1–e16.

[Publisher Full Text](#)

Laje R, Sciamarella D, Zanella J, *et al.*: **Bilateral source acoustic interaction in a syrinx model of an oscine bird.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2008; **77**(1 Pt 1): 011912.

[PubMed Abstract](#) | [Publisher Full Text](#)

Larsen ON, Goller F: **Role of syringeal vibrations in bird vocalizations.** *Proc R Soc Lond B Biol Sci.* 1999; **266**(1429): 1609–1615.

[Publisher Full Text](#) | [Free Full Text](#)

LeCun Y, Bottou L, Bengio Y, *et al.*: **Gradient-based learning applied to document recognition.** *Proceedings of the IEEE.* 1998; **86**(11): 2278–2324.

[Publisher Full Text](#)

LeCun Y, Bengio Y, Hinton G: **Deep learning.** *Nature.* 2015; **521**(7553): 436–444.

[PubMed Abstract](#) | [Publisher Full Text](#)

Marler P, Tamura M: **Song “dialects” in three populations of White-crowned Sparrows.** *The Condor.* 1962; **64**(5): 368–377.

[Publisher Full Text](#)

Marler P: **Variation in the song of the Chaffinch *Fringilla coelebs*.** *Ibis.* 1952; **94**(3): 458–472.

[Publisher Full Text](#)

Marler PR: **Science and birdsong: the good old days.** *Nature's music: the science of birdsong.* Elsevier, Ed: Marler, P. R., & Slabbekoorn, H, 2004.

Mindlin GB, Laje R: **The physics of birdsong.** Springer Science & Business Media, 2006.

Mindlin GB: **Nonlinear dynamics in the study of birdsong.** *Chaos.* 2017; **27**(9): 092101.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Narosky T, Yzurieta D: **Aves de Argentina y Uruguay. Guía de identificación/Birds of Argentina and Uruguay. A field guide.** Editora Vazquez Mazzini, Buenos Aires, Argentina, 2010.

Norton P, Scharff C: **“Bird song metronomics”: Isochronous organization of zebra finch song rhythm.** *Front Neurosci.* 2016; **10**: 309.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Nottebohm F: **The song of the chingolo, *Zonotrichia capensis*, in Argentina: Description and evaluation of a system of dialects.** *The Condor.* 1969; **71**(3): 299–315.

[Publisher Full Text](#)

Nowicki S, Capranica RR: **Bilateral syringeal coupling during phonation of a songbird.** *J Neurosci.* 1986; **6**(12): 3595–3610.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Pipes LA: **Applied mathematics for engineers and physicists.** Mc Graw-Hill Book Company, Madison, 1946.

[Reference Source](#)

Podos J: **A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae).** *Evolution.* 1997; **51**(2): 537–551.

[PubMed Abstract](#) | [Publisher Full Text](#)

Potter RK, Kopp GA, Green HC: **Visible Speech.** Dover Publications Inc., New York, 1947.

[Reference Source](#)

Reeves AH: **Pulse code modulation: Past, present and future.** *J Franklin Inst.* 1968; **285**(4): 243–250.

Riede T, Suthers RA, Fletcher NH, *et al.*: **Songbirds tune their vocal tract to the fundamental frequency of their song.** *Proc Natl Acad Sci U S A.* 2006; **103**(14): 5543–5548.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rockmore DN: **The FFT: an algorithm the whole family can use.** *Comput Sci Eng.* 2000; **2**(1): 60–64.

[Publisher Full Text](#)

Rawat W, Wang Z: **Deep convolutional neural networks for image classification: A comprehensive review.** *Neural Comput.* 2017; **29**(9): 2352–2449.

[PubMed Abstract](#) | [Publisher Full Text](#)

Sanz Perl Y, Arneodo EM, Amador A, *et al.*: **Reconstruction of physiological instructions from Zebra finch song.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2011; **84**(5 Pt 1): 051909.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Saunders AA: **A guide to bird songs.** D Appleton. 1935.

Schultz R: **Melodic contour and nonretrogradable structure in the birdsong of Olivier Messiaen.** *Music Theory Spectrum.* 2008; **30**(1): 89–137.

[Publisher Full Text](#)

Sitt JD, Amador A, Goller F, *et al.*: **Dynamical origin of spectrally rich vocalizations in birdsong.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2008; **78**(1 Pt 1): 011905.

[PubMed Abstract](#) | [Publisher Full Text](#)

Stein RC: **Modulation in bird sounds.** *Auk*. 1968; **85**(2): 229–243.

[Publisher Full Text](#)

Suthers RA, Goller F, Pytte C: **The neuromuscular control of birdsong.** *Philos Trans R Soc Lond B Biol Sci*. 1999; **354**(1385): 927–39.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Titze IR: **Principles of voice production.** Prentice Hall, 1994; 354.

[Reference Source](#)

Tsuji K, Müller SC: **Physics and Music: Essential Connections and Illuminating Excursions.** Springer Nature, 2021; 424.

[Reference Source](#)

Tubaro PL: **Aspectos causales y funcionales de los patrones de variación del canto del chingolo (*Zonotrichia capensis*).** Tesis doctoral. Facultad de Ciencias

Exactas y Naturales. Universidad de Buenos Aires, 1990.

[Reference Source](#)

Tubaro PL, Mindlin GB: **A dynamical system as the source of augmentation in a deep learning problem.** *Chaos Solitons Fractals: X*. 2019; **2**: 100012.

[Publisher Full Text](#)

Uribarri G, Rodríguez-Cajarville MJ, Tubaro PL, *et al.*: **Unusual avian vocal mechanism facilitates encoding of body size.** *Phys Rev Lett*. 2020; **124**(9): 098101.

[PubMed Abstract](#) | [Publisher Full Text](#)

Wilson J: **Living in the sound of the wind: personal quest for W. H. Hudson, naturalist and writer from the River Plate.** Constable Ed., London, 2016.

[Reference Source](#)



# Open Peer Review

Current Peer Review Status: ? ?

---

## Version 1

Reviewer Report 06 February 2024

<https://doi.org/10.21956/molpsychol.18795.r27392>

© 2024 Stowell D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Dan Stowell** 

<sup>1</sup> Tilburg University, Tilburg, North Brabant, The Netherlands

<sup>2</sup> Naturalis Biodiversity Center, Leiden, South Holland, The Netherlands

This is a review paper with a strongly history-focused perspective on frequency analysis and time-frequency analysis of birdsong. Overall, the article is good and well-written, providing an interesting background that will especially benefit researchers who haven't lived through these different epochs.

However, I found the review a little too heavy on self-citation, and in some other respects its view is not quite broad enough. The strongest aspect is the section describing how subtleties of physical sound production relate to what is seen on a spectrogram. The weakest aspects are the lack of mathematical precision in the description of spectrogram-related concepts and terms, and the narrow and unclear section on machine learning.

Further, I was hoping to read about variants of time-frequency analysis that have been used in birdsong analysis, such as the reassigned spectrogram, the multitaper spectrogram, or even wavelet spectrograms. Although these are not mainstream, they each have theoretical justifications which may make them useful for high-precision time-frequency analysis or for robust analysis in high-noise scenarios (see e.g. Xiao *et al.*, 2007<sup>1</sup>). Having read the article, I now see that the focus is purely on the "mainstream" Fourier magnitude spectrogram, but it seems a shame to leave the reader with the impression that it's the only type of analysis possible.

To return to the mathematical issues: since the short-term Fourier transform (STFT) is the core tool being discussed here, the reader must be given a precise and complete portrait of the method. A crucial step that is completely ignored is that the "sin" and "cos" elements of the Fourier analysis must typically be reduced from their complex form to magnitudes, or else there is a gap of understanding before "the weights are encoded in a gray scale". It's not clear how these weights should map on to the formulae. The authors must explicitly include this step. Some sources also work with the full (complex) spectrogram data rather than the magnitudes (for machine learning, though not for visualisation).

The section "Spectrograms and artificial intelligence" is quite narrow and limited. (I point out also that "machine learning" would be a better term than "artificial intelligence" here.) This topic has seen enormous growth in the past ten or twenty years. It is perhaps not surprising that the authors don't give an exhaustive coverage. However, this section does not really treat the key issues in the topic. Instead it gives an unsatisfactory description of neural networks followed by some examples that lead, all too soon, to a self-citation when the authors describe their own use of machine learning. It would be better to reduce the attempted explanation of deep learning, and instead cite more of the foundational works that demonstrated the strength of machine learning for birdsong analysis, and/or recent scientific work that reviews this field or makes strong contributions. The authors correctly refer to BirdCLEF and BirdNET as influential.

Specific comments (detailed, or minor) follow below.

The abstract is a little bit vague - I'd like more of a hint of what the reader will learn from this review. Please expand a little on the sentence "In this paper we review its history and some of its applications."

#### **Section "Birdsong before the spectrogram":**

- "Human music evolved as a collective activity" - are you asserting that birdsong didn't? Consider slightly rephrasing.
- In this paragraph I'm a little dubious of the implied claim that human music has predominant rhythm "because" humans have discussed and negotiated it. Isn't it true that many human musics (but not all) follow patterns of metre and rhythm, even non-formalised music cultures? All human cultures discuss and negotiate, but it's not clear how far that shapes all human musics.
- "However, using musical notation to accurately transcribe bird song requires..." -- Here, the authors assert that musical notation can accurately transcribe bird song, which is not a well-supported view. In my opinion it would be better to observe that Western musical notation is not really up to the task because not well-matched to the variations and idiosyncracies of birdsong.

#### **Section "The appearance of the spectrograph":**

- After equation 1: I think the description can be slightly simplified since  $a_0$  can technically be considered a periodic function too.
- "Additionally..." - the introduction of STFT is a little cryptically-written here. It is here motivated by "creat[ing] a three-dimensional diagram", when really it should be motivated by analysing a sound whose frequency characteristics (its "weights") are not stationary but vary over time.
- "If the weights are encoded in a gray scale" - you haven't dealt with the conversion from sin and cos to the complex magnitudes, so it's not quite clear how these weights should map on to the formulae.

- "An illuminating example is the use of two independent sound sources." - This is unexplained. I presume they're referring to the two sides of the syrnix. We don't need detail here but the unexplained "example" will be incomprehensible to some readers.

#### **Section "The history of digitalization":**

- "the sampling frequency imposes a maximum on the frequencies that enter in the description" - It would be pedagogical to include a mention of the Nyquist sampling theorem here, especially since "the Nyquist frequency" is such a common concept in time-frequency analysis.
- "To do that, filters  $f_i$  can be defined" - these are commonly referred to as "window" or "windowing" functions. Again, including the common terminology would be pedagogic.

#### **Section "Analyzing spectrograms":**

- "let's consider a sequence of signals" - this should be described as a set of signals, not a sequence.
- "except in the neighborhood of a given time  $t_i$ , in which they resemble the original acoustic signal" - what original signal? Unclear formulation should be rephrased.
- Extended discussion of vocal production details and how it relates to the observed spectrogram data - this section is very good, and connects together plenty of literature which is not commonly considered when doing spectrographic work. It includes an extended example from the authors' own prior work, but also sufficient citations from elsewhere.

#### **Section "Spectrograms and artificial intelligence":**

- "a set of algorithms known generically as deep learning networks" - in fact this is not generic enough. The correct name (corresponding to the description in this text) is "machine learning", not "deep learning". Since the authors here wish mainly to discuss deep learning, they should probably make brief reference to machine learning and then a slightly tighter definition of deep learning.
- The self-citation of "[Tubaro & Mindlin, 2019]" at the end of the paragraph describing machine learning is not appropriate. There are many, many papers applying machine learning and deep learning to spectrogram images, and even focusing on this concept, prior to 2019, as well as good recent reviews.
- "A classification algorithm based on a neural network..." - This description of neural networks is a little clumsy (because the authors are aiming to be brief). Is the description of layers needed at all? I suggest it is a distraction in this paper - the reader does not need any discussion of "layers" at all - and these 2 paragraphs could be simplified down to just a few sentences which take us smoothly to BirdCLEF and BirdNET.
- It is not clear to me whether spectrograms will continue to be the most important representation used for machine learning on birdsong (see e.g. WaveNet), so it might also be beneficial to have a passing comment on that consideration. Is the magnitude

spectrogram purely pre-eminent? Are other types of spectrogram, or even the raw waveform, well-suited to modern machine learning?

## References

1. Xiao J, Flandrin P: Multitaper Time-Frequency Reassignment for Nonstationary Spectrum Estimation and Chirp Enhancement. *IEEE Transactions on Signal Processing*. 2007; **55** (6): 2851-2860  
[Publisher Full Text](#)

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Partly

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioacoustics, machine learning, deep learning, signal processing.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Reviewer Report 23 January 2024

<https://doi.org/10.21956/molpsychol.18795.r27388>

© 2024 Mercado E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Eduardo Mercado** 

University at Buffalo, Buffalo, New York, USA

Researchers studying animal bioacoustics rely heavily on spectrographic analyses to describe, compare, and classify vocalizations, and to relate vocal actions to mating, learning, echolocation, and communication. This approach to measuring animals' sounds has become so ubiquitous that many experts are likely to view a scientific report on animal vocalizations that does not include a spectrogram to be incomplete. Amador and Mindlin provide a chronological review of how



spectrographic analyses have impacted studies of bird song, including details on specific ways in which spectrograms can be used to clarify mechanisms of sound production. The authors also highlight the advantages spectrograms provide when combined with automated classifiers such as neural networks.

The manuscript provides a useful and interesting overview of the technical and mathematical development of spectrograms in the context of avian bioacoustics. The authors' examples of how spectrograms can contribute to the development and testing of theoretical models of sound production are clear and compelling. The paper would be an excellent way to introduce students to bioacoustic analysis approaches, and as an example of how science and technology interact to create new advances. I do not have any major concerns with the content of the paper, but do have some questions and suggestions for ways to supplement the information that is provided.

In the first section of the paper, the authors describe early attempts to characterize bird songs prior to the invention of the spectrograph. The spectrograph was not the first mechanical approach to analyzing sounds, and I was left wondering why none of these other approaches advanced analyses of bird songs in the same ways as the spectrograph did. Surveying the history of acoustic measuring devices, I came across the sonometer, phonautograph, kymograph, tonometer, vibrometer, resonators, spectrometer, stroboscope, and oscilloscope, any of which could have potentially been applied to the analysis of bird songs (some were). What did the spectrograph offer that made it special? I think maybe these other devices focused more on time-domain analyses or narrow frequency bands, while the spectrograph revealed both dimensions together, and had the capacity to create more easily publishable records. Was it the discovery that time-frequency domain measures presented in concert were more informative that made the difference? Or did something about the graphical nature of the images make them more appealing to scientists?

The abstract by (Koenig W et al, 1946) [Ref1] cited by the authors (on the early development of the spectrograph) includes the statement that "Spectrograms are shown for a wide variety of sounds, including voice sounds, animal and bird sounds, music, frequency modulations, and miscellaneous familiar sounds," meaning that the earliest talk on spectrographic analyses included examples of bird calls/songs. This might be worth mentioning since the actual use of spectrograms for more scientific purposes came a bit later.

In the section on digitalization and throughout the paper the term "filters" is used in a variety of ways by the authors. This could potentially be confusing. First, filtering is mentioned in the context of electronic bandpass filtering by spectrographs. Next, filters are described as a mathematical function. Later, they are described as a resonance feature of a vocal tract/cavity in birds. Finally, they are described as parameters within a neural network layer. I'm not sure exactly how the authors might revise the text to avoid or explain this overlap. Perhaps the term could be dealt with explicitly early on and its multiple versions explained. Or maybe different terms could be used in different contexts (e.g., there are other ways of describing how neural networks process inputs).

The section on "analyzing spectrograms" has lots of good tutorial examples. One aspect that I thought could have been presented in a more uniform way is the discussion of signals and the effects of resonators. While reading this section I was thinking of source-filter models and how they could be mapped onto the syrinx vs cavity resonance that is being described. In that context, I also wondered if birds are producing any noisy, consonant-like sounds in their songs (unvoiced);

if so, it might be worth noting how these noisy sounds show up in spectrograms as diffuse non-lines. For instance, the first sound in Figure 4B looks like it might have some sort of chaotic/noisy aspects to it which are not fully captured by the model in 4D. Given that the authors are talking about non-linear models of production and “explosive” elements, noisy/chaotic elements of sounds might be worth mentioning – they often are not in bioacoustic analyses. I also thought it might be useful to note that sound elements can be both enhanced and effectively damped by surrounding structures. This is mentioned on p. 8, “In this way, the components of the sound signal corresponding to frequencies higher than the fundamental have little energy.” Those components might have had more if the source were sitting outside of the bird’s body. The authors could potentially contrast resonance frequencies with anti-resonance frequencies.

Figure 3A provides a good example of how continuous gradations in pulse rate can lead to discrete shifts in image properties in spectrograms. I think this is worth highlighting because it relates to a statement on p. 6 that, “Some aspects of the resulting graph will be due to the parameters used in the procedure, and others will reflect the properties of the signals to be analyzed.” Changing the analysis parameters for this vocalization could dramatically affect how it looks in the spectrogram (e.g., the whole series could be shown as either vertical or horizontal lines). This is a potential limitation of the spectrographic approach: the same sound can generate a variety of different images and gradual changes in sound can lead to abrupt changes in image features. How does one decide what is the “right” image for a vocalization?

I think this sentence, “the multiple frequencies are heterodyne products resulting from the interaction of two signals” could be fleshed out a bit to explain what heterodyne products are and how they might be detected in a spectrogram (and maybe whether this is something birds do to extend their vocal repertoire).

On p. 9, I think the statement “without the need to extract a predetermined set of properties” is a bit misleading. It’s true that there is no need to specify and measure a specific set of properties, but in choosing which input sounds are examples, by selecting how they are represented numerically, and by deciding what they are examples of, one is implicitly selecting a predetermined set of rules/properties/dimensions, often ones that can be verbalized. Also on this page, the phrase “the numerical translation” seems to imply that there is only one translation, but there are lots of different ways the sound could be numerically represented, so maybe “a numerical translation” or “some numerical translation” would be clearer.

In the final conclusions section, or somewhere earlier in the paper, I think it would be useful to point out some of the limitations of the spectrogram as an analysis tool. For instance, it discards spatial information and signal phase information. Spectrograms can obscure differences in the relative intensity of different signals. Depending on how parameters are selected, the resulting images can highlight features that are biologically irrelevant and obscure features that may be very relevant to animals. In freeing observers from temporality, spectrograms may also provide a misleading impression of the kinds of information that listening birds extract from songs. This is not to say that spectrograms aren’t useful, but it would be more balanced to summarize both their pros and cons, and this more balanced assessment would build on the theme that new technologies that show signals in new ways can be scientifically advantageous.

## References

1. Koenig W, Dunn H, Lacy L: The Sound Spectrograph. *The Journal of the Acoustical Society of*

*America*. 1946; **18** (1): 19-49 [Publisher Full Text](#)

**Is the topic of the review discussed comprehensively in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Is the review written in accessible language?**

Yes

**Are the conclusions drawn appropriate in the context of the current research literature?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Animal bioacoustics, behavioral neuroscience, cognitive psychology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

-----