Bioacoustic fundamental frequency estimation: a cross-species dataset and deep learning baseline

Paul Best*[1], Marcelo Araya-Salas[2], Axel G. Ekström[3,4], Bárbara Freitas[5,6,7], Frants H. Jensen[8], Arik Kershenbaum[9], Adriano R. Lameira[10], Kenna D. S. Lehmann[11], Pavel Linhart[12], Robert C. Liu[13], Malavika Madhavan[12], Andrew Markham[14] Marie A. Roch[15], Holly Root-Gutteridge[16], Martin Šálek[17,18,19], Grace Smith-Vidaurre[20,21,22], Ariana Strandburg-Peshkin[23,24], Megan R. Warren[13], Matthew Wijers[25], Ricard Marxer*[1]

1 Université de Toulon, Aix Marseille Univ. CNRS, LIS, Toulon, France
2 Escuela de Biologıía & Centro de Investigación en Neurociencias, Universidad de Costa Rica
3 Speech, Music & Hearing, KTH Royal Institute of Technology
4 Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
5 National Museum of Natural Sciences, Spanish National Research Council (CSIC), Madrid, Spain
6 Centre de Recherche sur la Biodiversité et l'Environnement (UMR 5300 CNRS-IRD-TINPT-UPS), Université Paul Sabatier, 31062 Toulouse Cedex 9, France
7 Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain
8 Department of Ecoscience, Aarhus University, Frederiksborgvej 399, 4000 Roskilde, Denmark
9 Girton College, and Department of Zoology, University of Cambridge, Cambridge, UK
10 Department of Psychology, University of Warwick
11 Human Biology Program, Michigan State University
12 Department of Zoology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic
13 Department of Biology and Emory National Primate Research Center, Emory University
14 Department of Computer Science, Oxford University
15 Department of Computer Science, San Diego State University
16 School of Life and Environmental Sciences, University of Lincoln, Lincoln, United Kingdom
17 Czech Academy of Sciences, Institute of Vertebrate Biology, Brno, Czech Republic
18 Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic
19 Forestry and Game Management Research Institute, v.v.i, Jíloviště, Czech Republic
20 Department of Integrative Biology, Michigan State University, East Lansing, MI USA
21 Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI USA
22 Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI USA
23 Biology Department, University of Konstanz
24 Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany
25 Wildlife Conservation Research Unit, Recanati-Kaplan Centre, Department of Biology, University of Oxford, Oxford, UK

* paul.best@univ-amu.fr
* ricard.marxer@lis-lab.fr

Abstract

The fundamental frequency (F0) is a key parameter for characterising structures in vertebrate vocalisations, for instance defining vocal repertoires and their variations at different biological scales (*e.g.* population dialects, individual signatures). However, the task is too laborious to perform manually, and its automation is complex. Despite significant advancements in the fields of speech and music for automatic F0 estimation, similar progress in bioacoustics has been limited.

To address this gap, we compile and publish a benchmark dataset of over 250,000 calls from 14 taxa, each paired with ground truth F0 values. These vocalisations range from infra-sounds to ultra-sounds, from high to low harmonicity, and some include non-linear phenomena.

Testing different algorithms on these signals, we demonstrate the potential of neural networks for F0 estimation, even for taxa not seen in training, or when trained without labels. Also, to inform on the applicability of algorithms to analyse signals, we propose spectral measurements of F0 quality which correlate well with performance.

While current performance results are not satisfying for all studied taxa, they suggest that deep learning could bring a more generic and reliable bioacoustic F0 tracker, helping the community to analyse vocalisations via their F0 contours.

Keywords : Fundamental frequency (F0), vocalisation analysis, cross-species dataset, deep learning

Introduction

To produce acoustic signals, vertebrates typically vibrate soft tissue structures within their vocal apparatus (*e.g.* the laryngeal tissue for mammals, or the syringeal membrane for birds). The frequency at which vocal organs oscillate, measured in Hertz, is called the fundamental frequency or F0 [1, 2]. It is linked to the notion of pitch in human psychoacoustics, which relates to the perception of frequency. However, not all vocalisations result in vibration of the vocal apparatus, and such vocalisations are said to be unvoiced.

Fundamental frequency (F0) is a principal feature in the description of acoustic signals. In speech, F0 serves multiple purposes [3], from signalling speaker sex [4], to providing a cue to conversational turn-taking [5], and has been studied by phoneticians for its influence on interpretability of speech and song [6–8]. F0 is also widely used in music applications [9–11], as the physical measurement of which note is being sung or played by an instrument.

In bioacoustics, F0 can carry biologically meaningful information such as a cue for body size [12, 13] or age [14]. Additionally, within vocalisation, how the F0 evolves through time (*i.e.* the F0 contour) is a widespread feature used in defining units of a vocal repertoire [15], and can also hold community markers [16, 17] or even information on individuals' identities [18–22].

It should be noted that the concept of F0 is only applicable to approximately periodic sounds. Like in speech or music, some bioacoustic signals are imperfectly periodic, in which case F0 may be hard to define or estimate, or aperiodic, in which case there is no F0 to estimate (spectral metrics such as the centroid frequency may then be more relevant to their description). For instance, these unvoiced vocalisations can be the /sh/ or /S/ phonemes in human speech, calls produced with the tongue and lips by non-human primates [23], cetacean echolocation clicks [24], or ‘chaotic’ calls produced via vocal organ vibration [25, 26].

Task’s challenges

Biologically produced signals, even if periodic, do not always have an unambiguous fundamental frequency [27]: influences such as nonlinear phenomena can occur such as sub-harmonics or biphonation [25, 26, 28], yielding variations between and/or within the signal’s cycles. This, coupled with changes in the recorded signal due to propagation or interference with background noise (especially in outdoor far-field settings), make F0 estimation a challenging problem. Common mistakes include harmonic jumps (*e.g.* confusion between the first overtone and the fundamental), or false positives (detecting an F0 when its actually absent, either because the vocal organ isn’t active or because the vocalisation is non-periodic).

Related works

F0 estimation in speech and music

Many ‘traditional’ methods for F0 estimation rely on auto-correlating an assumed stationary signal segment to identify a period. More recently, deep neural networks have been proposed instead. In this study, we compare a set of classic and recent algorithms used in the Music Information Retrieval (MIR) and speech communities.

- PRAAT [29] (speech): after applying an auto-correlation to the signal waveform, this algorithm assumes the first peak should correspond to the main cycle period and indicate the F0. In PRAAT, the size of the auto-correlation window is three periods of the pitch floor parameter, which was set to 27.5 Hz in our experiments (to match that of CREPE[1]). The chosen window size would be problematic to deal with rapid frequency sweeps or trills, but these were not encountered in the present dataset.

- p-YIN [30] (speech and music): the original YIN alogrithm is also based on the auto-correlation method, but with several modifications such as a parabolic interpolation [31]. Then, p-YIN improves performance by storing multiple F0 candidates at each time frame, taking their probability into account to yield a smoothed F0 contour.

- CREPE [32] (music): a neural network that convolves over waveforms. With its classifier architecture, each output bin corresponds to a specific frequency, predicting whether or not it corresponds to an active F0. The original model was trained on 22 hours of synthesised and re-synthesised monophonic music of known pitch and from varying instruments.

- PESTO [33] (music): a neural network that convolves over Constant-Q Transforms (CQT), *i.e.* spectral representation with varying kernel sizes. This model is self-supervised (trained without ground truth labels) based on objectives of equivariance with respect to pitch shifts and invariance to noise addition. The original model used here was trained on two hours of people singing Chinese pop songs.

[1] infra-sonic sounds were pre-processed to be detectable despite this pitch floor (see Signal slow down / acceleration section)

- BASIC-pitch [34] (music): is a convolutional neural network trained to detect multiple active pitches (*e.g.* to deal with multiple instruments playing simultaneously). From a CQT representation, the model yields three matrices with the same number of time frames as the input. The first matrix predicts if a note is starting at a time frame (note onset), the second denotes if a note is being active (with a resolution of one bin per semi-tone), and the third denotes if a pitch is active (similarly to the latter but this time with a resolution of three bins per semi-tone). To use BASIC-pitch as a monophonic F0 estimator, for each frame, we took the frequency bin of the third matrix with the highest confidence as the predicted F0.

F0 estimation in bioacoustics

Bioacoustic analysis software (*e.g.* Raven [35], Luscinia [36]) or packages (*e.g.* Seewave [37], warbleR [38], Parselmouth [39]) integrate spectral peak finding and/or F0 estimation methods, using algorithms such as short-term cepstral transforms or PRAAT's auto-correlation. These ready-made F0 estimation tools are used in many bioacoustic studies [40–44], often using PRAAT but also sometimes combined with manual procedures [45].

For specific bioacoustic purposes, new approaches to F0 estimation were also developed. This includes training convolutional neural networks to recognise tonal energy using real or synthetic targets [46]; training with a modified loss function that enables learning from noisy pseudo-labels [47]; or tuning the YIN algorithm to bird vocalisations [48]. Another study also reported on a benchmark of numerous F0 estimation algorithms on electro-glottographic signals for bioacoustic applications [49]. However, all of them tested algorithms on a single type of signal or a single taxon, as opposed to MIR F0 estimation studies that often benchmark performance on diverse signals to get a sense of an algorithm's versatility. Working with datasets focused on a single taxon might result in algorithms being over-specialised, necessitating re-tailoring or development for each new taxon.

Overall, whether manual or automatic, the widely adopted approach to estimate the F0 of non-human vocalisations relates to finding the lowest frequency spectral peak and/or the inter-harmonic distance at each time frame. While in some cases, this leads to an imperfect measure of vocal organ vibratory speed, it still significantly correlates with, and is virtually always the same as, F0. Moreover, a large body of literature has successfully found ecologically relevant acoustic structures using F0 estimation approaches [12–14, 16, 17], supporting the idea that investing time to apply and validate deep learning tools in order to automate F0 estimation will be greatly beneficial to the scientific communities that rely on bioacoustics data.

Objectives

In bioacoustics, deep learning already strongly contributes to tasks such as vocalisation detection/classification [50] or clustering [51], but this technique is not yet widely used for F0 estimation. In speech and music F0 estimation however, deep learning has demonstrated both versatility [33, 34, 52] (*i.e.* handling a wide diversity of signals) and robustness to noise [32, 53], two important challenges in bioacoustics, as species emit diverse vocalisations in sometimes very noisy settings.

Thus, it appears that deep learning could improve bioacoustic F0 estimation, but datasets to both train and evaluate models in this specific domain are lacking. Here we compile and publish a cross-species dataset of non-human vocalisations with ground truth F0 contours from previously annotated vocalizations. They come from studies that were conducted independently from this one, most of the time including a form of annotation quality control, and some already published in peer-reviewed journals (see Supplementary text 1). We also report how both traditional and deep learning algorithms perform on these data.

With this work, we provide an analysis on different taxa so that practitioners can make informed decisions of which methods to use, how to apply them, and requirements in terms of vocalisation characteristics and annotation availability. Additionally, we hope to i) foster research and development of automatic F0 estimation on signals other than speech and music, and ii) significantly reduce the time investment required to track F0 contours of a new taxon.

Multi-F0 versus mono-F0 estimation

In many acoustic scenes, it is possible for more than one F0 to be active simultaneously. In MIR, pitch estimation is thus divided into two tasks, multi-pitch for which the goal is to identify all audible (or annotated) F0, potentially multiple in one time frame, and mono-pitch or melody estimation aiming to produce a single sequence of frequency values for a given input signal [10].

In bioacoustics, multi-F0 estimation is needed for cases with overlapping calls (which are common in natural conditions), or for species that are capable of biphony (generating two independent tones simultaneously, often referred to as F0 and G0 [54–58]). In order to limit the scope of this work, we benchmark monophonic algorithms only (BASIC-pitch was designed for multi-pitch estimation but we use it in a mono-pitch fashion here). The consideration of multi-F0 estimation is left for future work. For datasets that originally included overlapping calls, we discarded these sections to keep only those with a single active F0 according to the ground truth annotations.

Terminology

For the data introduced here, one may argue that the term ‘pitch’ would be more appropriate than ‘F0’, since annotations were conducted by humans and/or machines, and do not necessarily match vocal fold vibration speed. Nonetheless, the term ‘pitch’ might suggest that annotations describe human acoustic perception, collected from listening experiments, and the machine- or spectrogram-based annotations might give different results than perceptual tests. For this reason, we refer to the presented ground-truth as F0.

To foster transdisciplinarity and since many methods used here originated within the MIR community, we borrow many terms from the field which are not common in bioacoustics. Hence, we introduce them in the following, along with a visual illustration in Fig 1:

- *Frame*: A short time interval over which the signal is assumed to be stationary and from which we estimate the spectrum; that is one temporal bin of a spectrogram.
- *Voiced frame*: A frame containing a voiced sound (*e.g.* an animal is producing a periodic sound with its vocal apparatus).
- *Voiced section*: A temporal window of multiple voiced frames.
- *Octave*: interval of a factor of two in frequency.
- *Semitone*: interval of a twelfth of an octave.
- *Pitch accuracy*: proportion of frames with a predicted F0 that is close to the reference contour (using a fixed frequency interval threshold such as half a semitone).

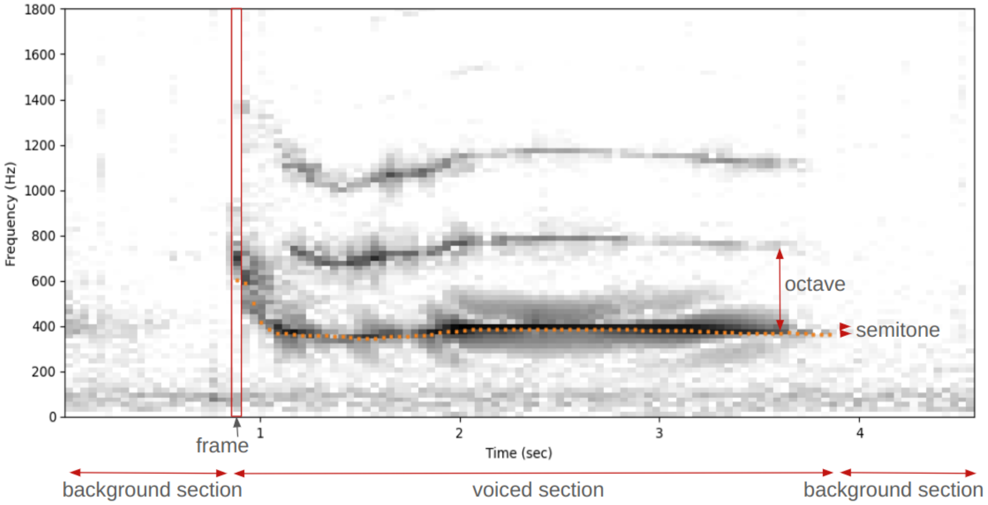


Fig 1. Example of a vocalisation spectrogram indicating different terms used in F0 estimation. This vocalisation was emitted by an Arctic grey wolf (*Canis lupus arctos* ssp.). Orange markers denote annotated ground truth F0 values.

- *Chroma accuracy*: similar to pitch accuracy, but ignoring octave shifts (*e.g.* 500 Hz ± 1 semitone is considered accurate for a 1000 Hz F0 in terms of chroma accuracy).
- *Recall*: proportion of frames that were predicted as voiced among the frames annotated as such.
- *Vocalisation recall*: proportion of voiced calls that were correctly detected. Unlike recall, this metric is based on correctly detecting at least a third of the voiced frames within the call [59].
- *Specificity*: the proportion of frames predicted as not voiced among the frames annotated as such. Specificity can be chosen here instead of precision as proportion of silent frames in all datasets remains relatively small.
- *Sub-harmonic*: A variation that occurs between consecutive cycles of a signal, leading to a dissimilarity between consecutive periods and a similarity between non-consecutive periods. In the frequency domain, this phenomenon typically generates energy at half the F0.

Materials and methods

For this study, we set out to gather, describe and publish a cross-species dataset of audio vocalisations with corresponding annotated F0 contours, as well as to report on how different algorithms perform on the task of single F0 estimation. This section starts by describing the published dataset, both in broad numbers and with fine-scale acoustic features of its components. Then, we report on implementation details for the comparison of state-of-the-art F0 estimation algorithms, especially regarding the training of models under different degrees of supervision.

Dataset

We contacted researchers who had published studies on measuring the F0 in non-human vertebrate vocalisations. In addition, we reached out to members of the International Bioacoustics Council (IBAC) via their mailing list and at the 2023 congress (Oct 27 – Nov 1, Hokkaido, Japan). The numerous answers allowed us to assemble a corpus of labeled acoustic data across 14 taxa of mammals and birds, which is described in Table 1.

Table 1. Specifications for the 14 datasets gathered in this study. Stars (*) indicate datasets that include multiple species or subspecies.

Taxon	**# Vocalisations**	**Sample Rate (kHz)**	**SNR (dB) mean ± std**	**Annotation method**
canids* (Canis spp.)	2,282	16	5.7 ± 7.2	semi-automatic [60]
spotted hyenas (*Crocuta crocuta*)	571	8	6.3 ± 6.9	semi-automatic [18]
little owls (*Athene noctua*)	1,283	4	7.2 ± 4.5	Raven PFC [22]
bottlenose dolphins (*Tursiops truncatus*)	669	96	5.4 ± 5.8	semi-automatic [21, 61]
rodents* (*Microtus ochrogaster & Mus* spp.)	224,705	250	-5.0 ± 7.0	automatic [62, 63]
hummingbirds* (Trochilidae spp.)	13,680	44	5.5 ± 5.7	Raven PFC [64]
Spix's disk-winged bats (*Thyroptera tricolor*)	340	400	1.6 ± 3.7	Raven PFC [65]
Reunion grey white eyes (*Zosterops borbonicus*)	1,174	44	10.6 ± 6.5	Raven PFC
monk parakeets (*Myiopsitta monachus*)	233	44	7.9 ± 3.2	semi-automatic [66–69]
lions (*Panthera leo*)	164	16	13.5 ± 2.9	automatic [19]
orangutans (*Pongo pygmaeus*)	1,548	44	3.0 ± 5.4	Raven PFC [70, 71]
long-billed hermits (*Phaethornis longirostris*)	160	44	7.0 ± 2.6	Raven PFC [72]
dolphins* (Delphinidae spp.)	1,113	192	-7.6 ± 5.8	manual [59]
La Palma chaffinches (*Fringilla canariensis palmae*)	347	44	7.9 ± 2.6	Raven PFC
Total	250,670	[4; 400]		

This corpus combines the results of previous works on bioacoustic signals, each of which used specific methods to generate F0 ground truths. Some were traced by hand with custom graphical interfaces (manual), others used automatically estimated F0 contours but corrected them by hand (semi-automatic) and others used fully automated procedures, either custom or out-of-the-box such as Raven Peak Frequency Contour (PFC). Note that in the latter case, the operator still annotates the spectrogram with time × frequency bounding boxes around the F0, which highly limits the potential for errors. Further details on potential quality controls over annotations are described in the data description in Supplementary text 1.

This corpus integrates diverse taxa, across mammals and birds, and with diverse vocalisation properties (Fig 2). It is not an exhaustive set of all sound-producing species, and the associated results might not be representative for taxa not included such as frogs or insects for instance. Nevertheless, vocalisations of this corpus range from infra- to ultra-sound, some are shorter than 0.1 sec and others last several seconds, some appear to reflect non-linear phenomena (*e.g.* spotted hyenas vocalisation often contain sub-harmonics) and others are close to pure tones (*e.g.* Reunion grey white eyes vocalisations do not have harmonics). Moreover, across datasets, different signal acquisition methods were used, including collar-mounted and hand-held directional recorders for spotted hyenas and hummingbirds, in-lab recording chambers for rodents, or outdoor far-field recorders for dolphins (see Supplementary text 1 for a complete description of recording protocols). This strongly affects the resulting signal quality and consequently how easily the F0 can be estimated.

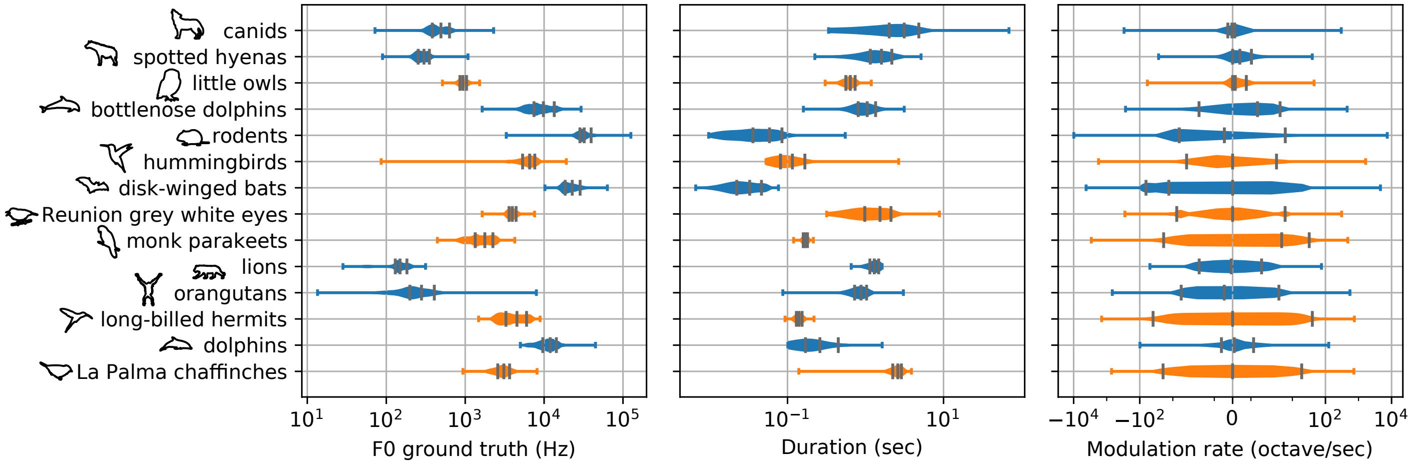


Fig 2. Distribution of F0 annotations per taxa. Blue and orange denote mammals and birds respectively, and black bars show the three quartiles of each distributions. The distribution of modulation rates report on linear F0 slopes between each annotation points.

The dataset introduced in this study is a conglomerate of previous works from different researchers. Thus, recording and ground truth characteristics vary. For some taxa, F0 labels result from a global tuning of automatic procedures, which can introduce errors, whereas labels for other taxa are the result of fine-grained manual interventions, which rely on human bias. Nonetheless, publishing such data is beneficial to the community, especially with a repository that is open to refinements from future works.

Vocalisations characterisation

Prior to estimating F0 values and comparing algorithms, we wish to obtain a fine-grained description of the dataset, enabling the formulation of hypotheses regarding signal characteristics that may influence F0 estimation accuracy. Specifically, spectral properties of vocalisations might help us identify challenges faced by algorithms to estimate F0 values. In this section, we describe the four metrics chosen for this purpose, some of which are specific to this study because they are based on annotated F0 contours: the signal-to-noise ratio (SNR), the F0 salience, the overtone-to-fundamental ratio (OFR), and the sub-harmonic ratio (SHR). Thus, we measure the energy of a vocalisation both in relation to background noise and in terms of how well it matches its corresponding annotation.

Note that except for the SNR, all metrics are computed from spectral frames (S), for which we use Hann windows without padding (window sizes and hop sizes are reported in Table 3). To reduce effects of background noise and frequency response, similarly to previous works [73, 74], we normalise spectra prior to their analysis (*i.e.* measuring salience, OFR and sub-harmonic ratio). The normalisation consists of subtracting the median of each frequency bin over background segments (not annotated as voiced) and dividing by the standard deviation. Note that this process is only used for the dataset analysis, F0 estimation methods are applied to the original audio. Finally, preliminary experiments have shown that OFR and SHR measurements were only reliable for salient F0 ground truths: unexpected values such as SHR higher than 1 were yielded if we did not target salient vocalisations. For this reason, we report only the values of frames with a salience above 0.6.

Signal-to-Noise Ratio

A commonly used descriptor of acoustic signals is their Signal-to-Noise Ratio (SNR). For each vocalisation, we high-pass filter the signal with a Butterworth filter of order three with a low-frequency cutoff set at the minimum F0 measured for the signal. The signal's power is then estimated as the root mean square (RMS) measurement of the filtered call, which is compared between sections annotated as voiced (E_{voiced}) and its surroundings ($\mathrm{E}_{background}$) to yield the SNR. Since during voiced sections vocalisation signals are mixed to background noise, to isolate their power, we subtract the power of surrounding sections before computing the ratio (Eq 1). Note that we cannot apply the logarithm for vocalisations with $\mathrm{E}_{voiced} < \mathrm{E}_{background}$, and therefore we drop these out of the measurement (an alternative to include them could have been to compute the signal to noise and signal ratio). Low SNR values are typically expected if the recorder was placed far away from the vocalising animal, in environments with high background noise, and/or if the vocalisation is produced softly. We report on SNR modes in Table 1.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathrm{E}_{voiced} - \mathrm{E}_{background}}{\mathrm{E}_{background}}\right) \tag{1}$$

F0 salience

Similarly to previous work [75], we characterise the salience of F0 contours relative to background noise. However, for this study, we aim to disentangle contributions of the fundamental frequency from harmonics. This motivated the design of two separate metrics, namely salience and OFR. The salience indicates by how much an F0 contour stands out from its surrounding spectrum. Low salience values are expected in vocalisations at low SNR and for wide-band / non-tonal vocalisations. We propose to compute the salience of an F0 annotation as the ratio of the energy in its close frequency band (set from one semitone below to one semitone above) and the energy of its surrounding octave. Eq 2 formalises this given a spectrum S, a F0 ground truth $f0$, and with numerical values in semitone. If the distribution of spectral energy were to be uniform, salience would be $\frac{1}{6}$, and if all the energy is contained in the tone surrounding the F0 contour, salience would be 1.

$$Salience = \frac{\sum_{f=f_0-1}^{f_0+1} S(f)}{\sum_{f=f_0-6}^{f_0+6} S(f)} \tag{2}$$

Overtone-to-fundamental ratio (OFR)

We use the term overtone-to-fundamental ratio (OFR) to describe the amount of energy present in the harmonics relative to the energy of the fundamental. To measure it, we chose a normalised formulation, namely the proportion of energy contained in the harmonics within the energy of the harmonics and the fundamental combined (Eq 3). Typically, a pure tone would have an OFR close to 0 whereas vocalisations with strong harmonics like human speech will have a value close to 1. In this study, we refer to vocalisations with a high OFR value as 'harmonic' and those with a low OFR value as 'non-harmonic' (harmonics can still be present but they have less energy than for other signals).

$$\mathrm{OFR} = \frac{\sum_{i=2}^{N} S(if_0)}{\sum_{i=1}^{N} S(if_0)} \tag{3}$$

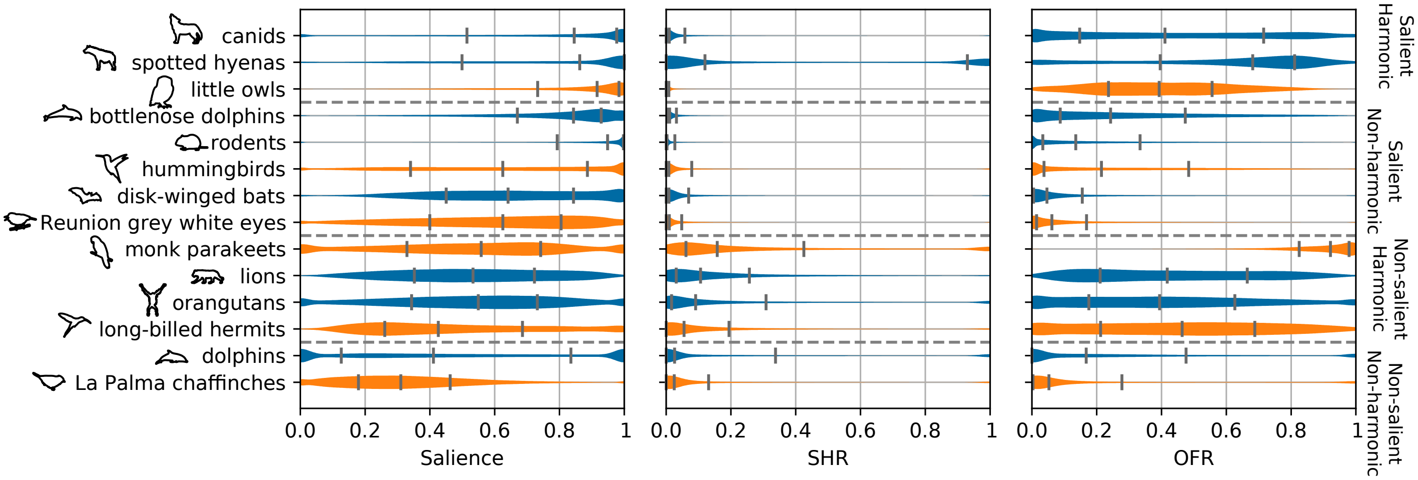
Sub-Harmonic Ratio

The Sub-Harmonic Ratio (SHR) proposed by Sun [76] can be used to detect sub-harmonics in acoustic signals (see also Herbst 2021 [77] for empirical results with this metric). Given an F0 value, it is computed by taking the ratio between the sub-harmonic amplitude (at half the F0 value) and the harmonic amplitude (Eq 4). Following Sun [76], we set N to five. SHR values are expected to approach one for signals with strong sub-harmonics (having the same amount of energy as the fundamental), or 0 for signals without any.

$$\text{SHR} = \frac{\sum_{i=1}^{N} S((i-\frac{1}{2})f_0)}{\sum_{i=1}^{N} S(if_0)} \tag{4}$$

Distributions of salience, SHR and OFR values are given in Fig 3, and help understand potential factors that might hinder F0 prediction.

Fig 3. Characterisation of F0 contours with their salience, SHR, and OFR. For more reliable measurements, SHR and OFR values are reported only for frames with a *Salience* > 0.6. Horizontal dashed lines delimit dataset groups which are labelled on the right side. Blue and orange denote mammals and birds respectively, and black bars show the three quartiles of each distributions.



Dataset grouping

To ease the reading of results across this dataset of 14 taxa, we grouped them by trends of contour characteristics (salience and OFR appeared to impact performance the most). We thus split taxa into four groups based on the median values of salience (0.6) and OFR (0.3). The resulting groups are presented in Table 2

Experimental framework

The variety of signal characteristics present in this corpus required specific pre-processing parameters to be set for each dataset, which are described in Table 3.

Signal slow down / acceleration

Slowing down or accelerating a signal by a given factor (*i.e.* artificially changing the sampling frequency) is a straightforward way to scale all frequencies by that same factor.

Table 2. Dataset groups attributed to each taxa based on their signal characteristics. Taxa with a median salience higher than 0.6 are considered salient, and taxa with a median OFR higher than 0.3 are considered harmonic. Dataset groups are delimited by dotted lines in Fig. 3.

	Low salience	**High salience**
Low OFR	dolphins, La Palma chaffinches	bottlenose dolphins, rodents, hummingbirds, disk-winged bats, Reunion grey white eyes
High OFR	monk-parakeets, lions, orangutans, long-billed hermits	canids, spotted hyenas, little owls

Table 3. Processing parameters chosen for each dataset. The window size, defined before signal slow down or acceleration, is used without padding, only for spectrogram visualisation, vocalisation characterisation, and for the p-YIN algorithm which operates on a user-set frame length. Time steps are set as a fraction of the window size, and used for spectrogram generation, vocalisation characterisation, annotation resampling, and also F0 prediction (F0 time series are resampled via linear interpolations using the `mir_eval` package [78]). A slow down factor of one is neutral, and a factor below one is an acceleration.

Taxon	Window size (ms)	Time step	Slow down factor
canids	64	1/8	1
spotted hyenas	256	1/8	1
little owls	13	1/8	1
bottlenose dolphins	11	1/8	20
rodents	2	1/8	50
hummingbirds	12	1/16	5
disk-winged bats	1	1/16	20
Reunion grey white eyes	23	1/16	5
monk parakeets	12	1/16	3
lions	128	1/8	0.5
orangutans	47	1/8	1
long-billed hermits	12	1/16	5
dolphins	8	1/8	20
La Palma chaffinches	23	1/16	5

For instance, if a 440 Hz tone was recorded at 44 kHz and we play it at 22 kHz, its F0 will shift to 220 Hz. This comes in useful for humans to listen to ultra-sonic sounds, or when using models that operate within a fixed frequency range like in this study.

Indeed, many of the methods that we evaluate were designed for signals within the frequency range of human production and perception. For instance, the CREPE and PESTO neural networks are trained as classifiers (as opposed to regression models), with a fixed output dimensionality, and with each output bin corresponding to a specific frequency (for PESTO: three bins per semitone from 27.5 Hz to 8 kHz; for CREPE: 5 bins per semitone from 32.7 Hz to 2 kHz). This presents challenges for detecting F0 with these algorithms for vocalisations that fall outside of this range. For instance, rodents or dolphins emit ultra-sonic vocalisations (above the human hearing range), and others such as lions emit close-to-infra-sonic vocalisations (Fig 2).

To be able to use the pre-trained CREPE, BASIC-pitch and PESTO models for ultra- and infra- sonic vocalisations, we slow down or accelerate signals to shift them into a human perceptual frequency range. The signal samples remain unaltered, but the sample rate is modified by dividing it by a fixed factor. As an example, the rodent corpus was sampled at 250 kHz with vocalisations between 10 and 100 kHz. With a slow down factor of 50, we set the sampling frequency to 5 kHz for vocalisations to lie between 0.2 and 2 kHz. Chosen slow down factors for each taxa are specified in Table 3, with a factor of one being neutral, and a factor below one being an acceleration.

After slowing down or accelerating the signal, for algorithms such as CREPE that work with a fixed sample rate, the signal is resampled using the bandlimited sinc interpolation method [79]. As a post-processing step, we multiply predicted frequencies by the slow down factor before evaluating them against ground truths.

Benchmarked algorithms and model training

In this study, we compare different algorithms and deep learning models used in speech, music and/or bioacoustic F0 estimation: PRAAT, p-YIN, CREPE, PESTO, and BASIC-pitch; which were introduced in the introduction.

The CREPE and PESTO deep neural networks have been trained for pitch estimation in music, but we wish to investigate how their performance might evolve if we train them for bioacoustic F0 estimation. Wishing to highlight the effect of training data on model performance, we follow the published training procedures to control for performance variation due to other factors:

- *CREPE* [32] is a supervised model with a classifier architecture. It takes the raw waveform as input, on which six convolutional layers are applied, before a fully-connected layer outputs confidence values predicting if frequency bins correspond to the F0 ground truth (there are 360 frequency bins between 32.70 Hz and 1975.5 Hz, each covering 20 cents). Following the original publication, we train the same model architecture, iteratively minimising the binary cross entropy between predictions and ground truths, using a ADAM optimiser and a learning rate of 0.0002. We use CREPE's pytorch implementation [80], initialising weights with that of the published model trained on musical signals (this was motivated by the observation of better performance when doing so).

- *PESTO* [33] is a self-supervised model that learns without ground truth labels. It does so by pitch-shifting training examples, and predicts F0 values from both the original and shifted versions (based on CQT representations of signals). During training, the model optimises a specific loss function which expects F0 predictions to have the same difference as the known shift (*i.e.* equivariance objective). As the algorithm does not learn actual F0 values, after training, synthetic signals of known frequency are used to produce a calibration that permits F0 recovery. Again to minimise confounding factors, we used PESTO's public implementation, using the same architecture and ADAM optimiser configuration. Only a few modifications to the original settings were necessary to achieve a functional learning, namely increasing the minimum CQT frequency to deal with small files (to represent low frequencies, the CQT needs large temporal windows), and increasing the range of the frequencies used in the post-training calibration).

Given these two model architectures and training protocols, each state of the art in either supervised or self-supervised F0 estimation in music, we test how training them with bioacoustic data might improve their performance in this domain. For this, we emulate different scenarios of data availability described in the following (by 'target' we refer to the taxon that a given model will be evaluated on):

- *Self-supervised*: In this scenario, a model is trained without the need of annotated F0 contours, which is the most common case when engaging in bioacoustic analysis. Here, we train PESTO on the vocalisations of the target taxon without using their associated F0 ground truth. Therefore, the size of the training set is the number of available vocalisations reported in Table 1. Since no labels are used in training, vocalisations from the same taxon are used to evaluate model performance post-training.

- *Supervised on other taxa*: In this scenario, we test the generalisation capacity of a supervised model trained on many taxa, by measuring how it behaves on a new one. With this, we assess the feasibility of a generic bioacoustic F0 estimation model that doesn't need retraining. Here, we train CREPE on a dataset combining all taxa except the target. To mitigate risks of over-representing taxa with many vocalisations (*e.g.* rodents), we limited the number of vocalisations per taxon to 1,000. Once the model is trained, its performance is evaluated on the target taxon that was retained from the training set.
- *Supervised on the target taxon*: In this scenario, we test how well a supervised model performs if it was trained on its target taxon. It is expected to attain the best performance, but is only applicable in a limited number of use cases: when researchers have access to F0 annotations for the taxon they want to analyse. We thus train CREPE on the target taxon, splitting the data in a 5-fold manner to dissociate training from evaluation data. Therefore, here training set sizes are 80% of the number of available vocalisations reported in Table 1.

Performance computation

We use the `mir_eval` package [78] to compute recall, specificity, pitch accuracy and chroma accuracy for each vocalisation independently, before averaging them per taxa. For pitch accuracy and chroma accuracy, we consider an F0 prediction to be correct if closer than half a semitone from the ground truth [10] (preliminary experiments with more permissive thresholds did not significantly change results).

The frequency resolution of F0 annotations varies across datasets, with particularly small sizes of the Fourier windows or specific manual label procedures that may lead to coarse quantisation of F0. To avoid biased results we ensured that the threshold used for evaluating pitch and chroma accuracy is larger than any of the F0 quantisation.

The recall and specificity metrics reflect an algorithm's behaviour in terms of voicing detection (*i.e.* the algorithm's capacity to differentiate between voiced and non-voiced frames). Some algorithms such as p-YIN compute a voicing probability, and others a F0 confidence value. Therefore, to generate a binary voicing prediction, we apply a threshold on these values. For each taxon and algorithm combination, we set this threshold to the balance point of the Receiver Operating Characteristic (ROC) curve (*i.e.* the point with equal recall and specificity). Threshold values are reported in Supplementary Figure 1.

For some detection tasks, the specificity metric can be over-optimistic as compared to the precision. This occurs for imbalanced datasets that have many more negative labels than positive ones (*e.g.* for voicing detections, having much more background sections than voiced sections). Since the specificity normalises the proportion of true negatives by negative labels, specificity scores might be high even with a significant proportion of detection errors. The precision however, normalises by the number of positive predictions, and does not suffer from this bias. In our case, amounts of positive and negative labels are similar, hence we report on the specificity metric which is more commonly found in the F0 estimation literature.

Results

After analysing per-taxon vocalisation characteristics, and having trained models under varying levels of supervision, we run F0 estimation on the whole corpus. We visually demonstrate this with randomly sampled examples of predictions in Fig 4.

In this section, we detail how the different algorithms and neural network models behave on each taxon, and this for different performance metrics. Throughout the text and figures, we refer to the out-of-the-box CREPE and PESTO models trained on musical signals as '`crepe-music`' and '`pesto-music`' respectively, PESTO models self-supervised on their target taxon as '`pesto-bio`', CREPE models supervised on their target taxon as '`crepe-consp.`' (for conspecific), and CREPE models supervised on other taxa than their target as '`crepe-heterosp.`' (for heterospecific).

F0 estimation accuracy

We report on the accuracy of estimating the F0 for each algorithm in Fig 5 with the pitch accuracy and chroma accuracy (tolerating octave errors).

Overall, the dataset grouping categories seem to explain most of the variations in performance. For taxa with a salient F0, neural networks trained on bioacoustic signals perform well (pitch accuracy > 0.69 and chroma accuracy > 0.74 for `pesto-bio`, `crepe-heterosp.` and `crepe-consp.`). Interestingly, for taxa with salient contours, the type of supervision (being supervised on the target taxon or being self-supervised) has a relatively small impact on performance. This is shown in Fig. 6, comparing `crepe-consp.` and `pesto-bio`. Also, for canids and little owls, the self-supervised model performs slightly better than the model supervised on the target taxon. This could indicate label noise (*i.e.* annotation errors) and/or overfitting. Typically for the latter, the model finds an over-specialised relationship between inputs and correct predictions that works well on the training data, but does not generalise to new examples. However, for taxa with non-salient vocalisations, the self-supervised training procedure becomes counter productive (`pesto-music` outperforms `pesto-bio`, Fig. 5), suggesting that the equivariance objective relies on contours with strong energy to function correctly.

As for the other algorithms, on salient contours, PRAAT seems to be subject to octave errors for vocalisations with strong harmonics (gap between pitch and chroma accuracies), but still shows the most reliable performance as compared to p-YIN or some neural networks trained on musical signals (BASIC and `pesto-music`). Comparisons between `crepe-music` and `crepe-consp.` (Fig. 6), or between `crepe-music` and `crepe-heterosp.` (Fig. 5), show that training neural networks on bioacoustic data mostly improves performance in F0 estimation. This is true even for relatively small datasets such as for disk-winged bats (272 vocalisations used in training). However, it is worth noting that the only taxon for which `crepe-consp.` has lower performance than `crepe-music` (although by a small margin) is the taxon with the lowest amount of annotated vocalisations (128 vocalisations in the training set).

For the less salient vocalisations, performance variability increases. Supervised training on the target taxon (`crepe-consp.`) still leads to the best results, but except for the long-billed hermits, models trained on other taxa (`crepe-heterosp.`) remain relatively close (their median pitch accuracy are 0.67 and 0.63 respectively). We show this relationship in Fig. 6, in which we compare the pitch accuracy of `crepe-consp.` with other methods. Despite being relatively close in performance, the superiority of `crepe-consp.` over `crepe-heterosp.` demonstrates that in general, for training, data proximity (training with data that is similar to the application domain) is more effective than data quantity.

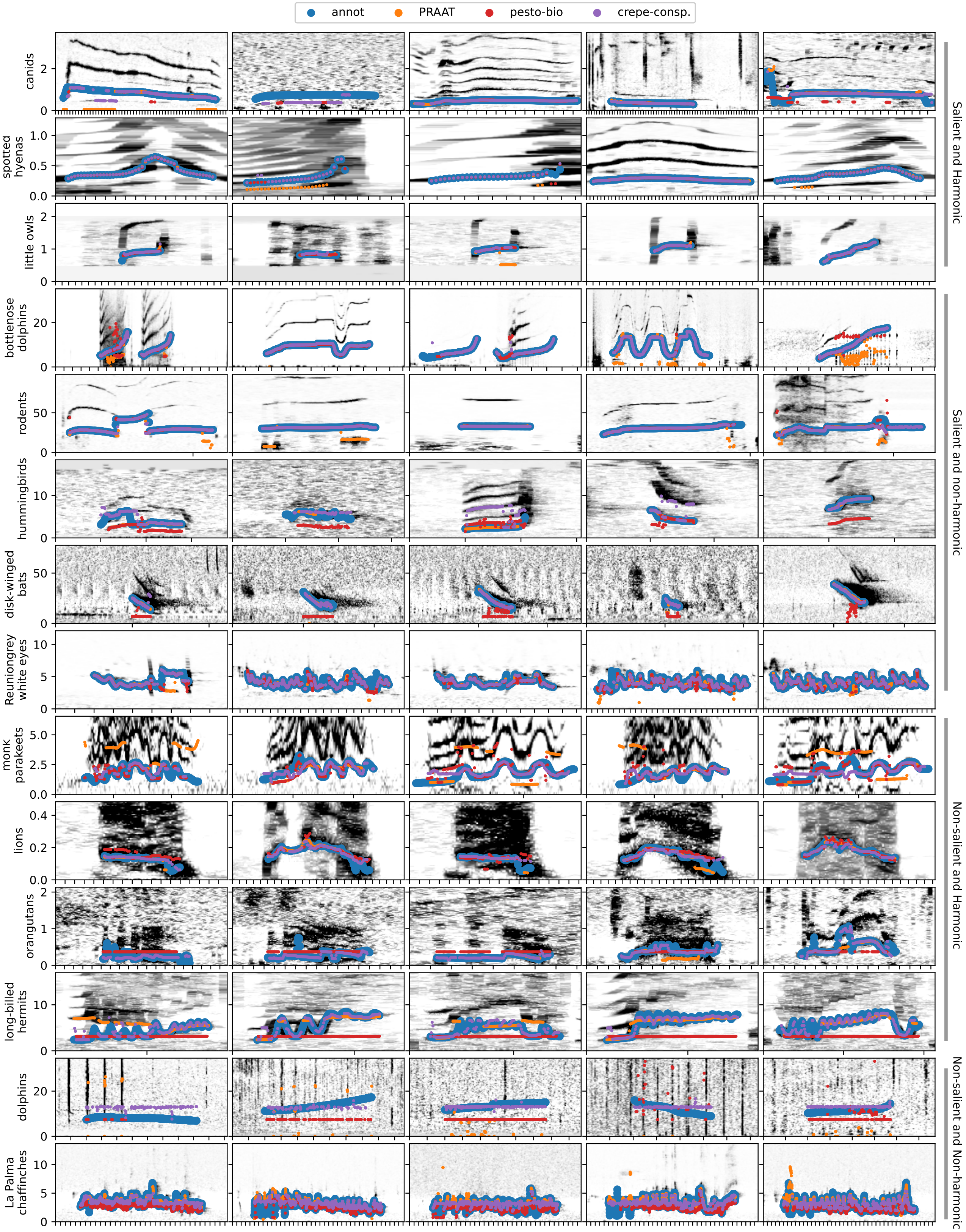


Fig 4. Spectrograms of randomly sampled vocalisations for each of the dataset's taxa, along with F0 predictions from different algorithms. Frequency values are given in kHz, and ticks on the abscissa are placed every 0.1 sec. Dataset grouping categories are also indicated on the right side of the figure.

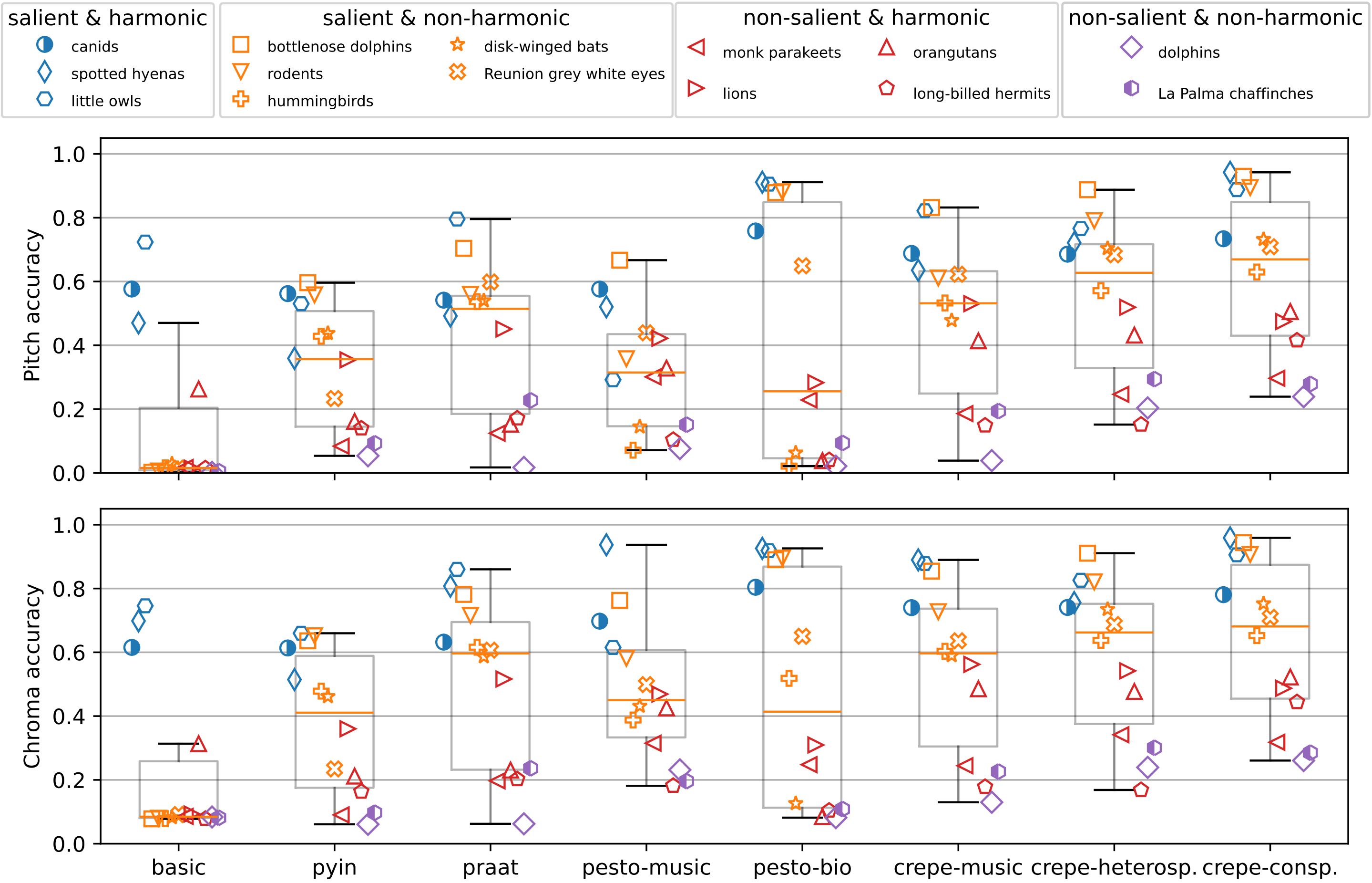


Fig 5. F0 estimation performance for each algorithm and taxon. Categories follow the dataset grouping described in Methods (the abscissa within a category varies for readability but does not carry information). Boxes extend from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median, and whiskers extend from the box to the farthest data point lying within 1.5x the inter-quartile range (IQR) from the box.

Accuracy as a function of salience

As salience appeared to be the most impacting vocalisation characteristic on F0 estimation accuracy, we report on pitch accuracy as a function of salience in Fig 7. To characterise each vocalisation, we used its average salience and average OFR. Results here are reported across all taxa, for vocalisations with an average OFR above and below 0.5 separately.

The salience metric appears to be a strong indicator of how well an F0 can be estimated. Indeed, the two almost follow a perfect identity relationship. This representation also confirms that for salient vocalisations `pesto-bio` performs well, similarly to supervised models, but a performance gap appears for the fainter contours with low OFR.

At low salience values, a tendency appears for lower performance with non-harmonic vocalisations as compared to harmonic ones, especially for p-YIN, BASIC and PESTO. This phenomenon suggests some reliance on harmonic structures to correctly estimate F0 values. Furthermore, Fig. 7 shows that supervised models trained on bioacoustic data (`crepe-consp.` and `crepe-heterosp.`) generally perform better than other

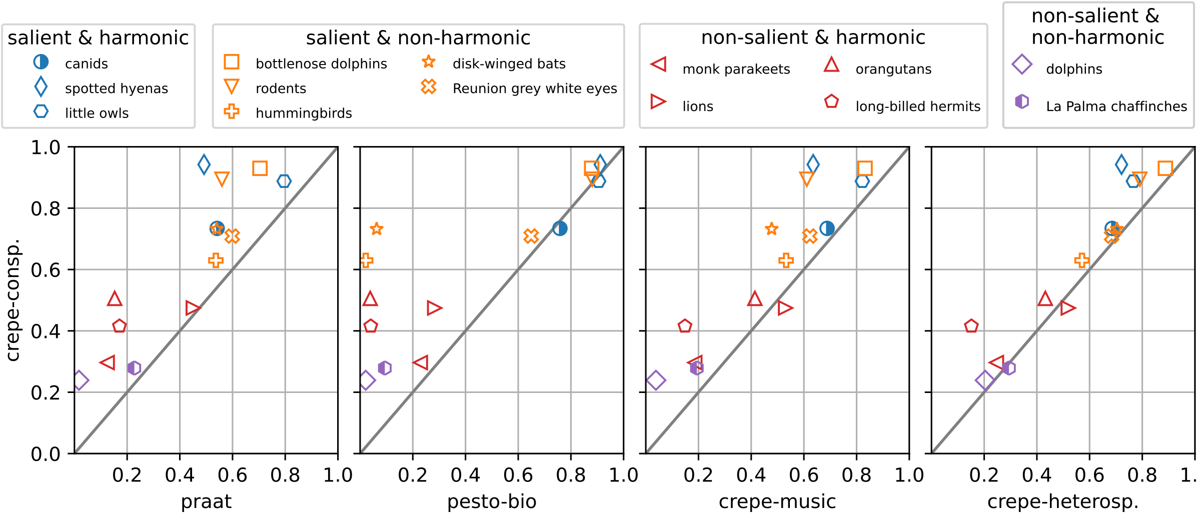


Fig 6. Pitch accuracy comparison. The performance of the top performing model (`crepe-consp.`) is plotted against the performance of other algorithms. Points on the diagonal ($y = x$) show equally performing models, whereas points far apart from it show large performance gaps.

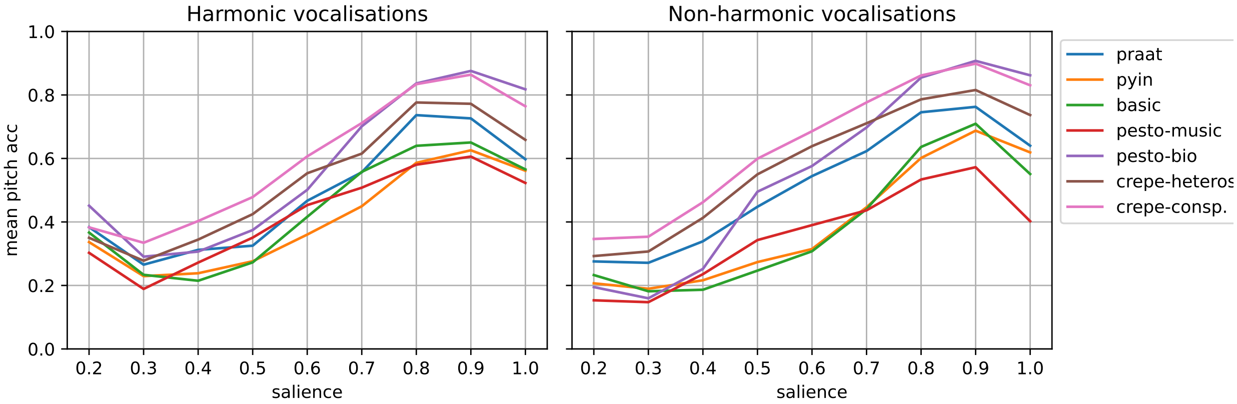


Fig 7. F0 estimation performance of each algorithm as a function of salience. Results are reported for vocalisations considered harmonic and non-harmonic separately (see Eq 3, with a threshold of 0.5).

algorithms regardless of their harmonicity or F0 salience.

Voicing detection

We report on the capacity of the different algorithms to correctly discriminate between voiced and background frames in Fig 8. Across algorithms, vocalisation OFR does not appear to impact performance, but their salience does.

For the most part, all algorithms perform similarly in terms of voicing detection, except BASIC-pitch. This is probably due to the fact that an onset matrix is normally used to predict note activation, thus the distribution of confidence values used here might not allow a good discrimination between voiced and background frames.

In their development of automated whistle contour extraction, Roch et al. [59] also reported on vocalisation-wise recall (the proportion of vocalisations with a recall of at

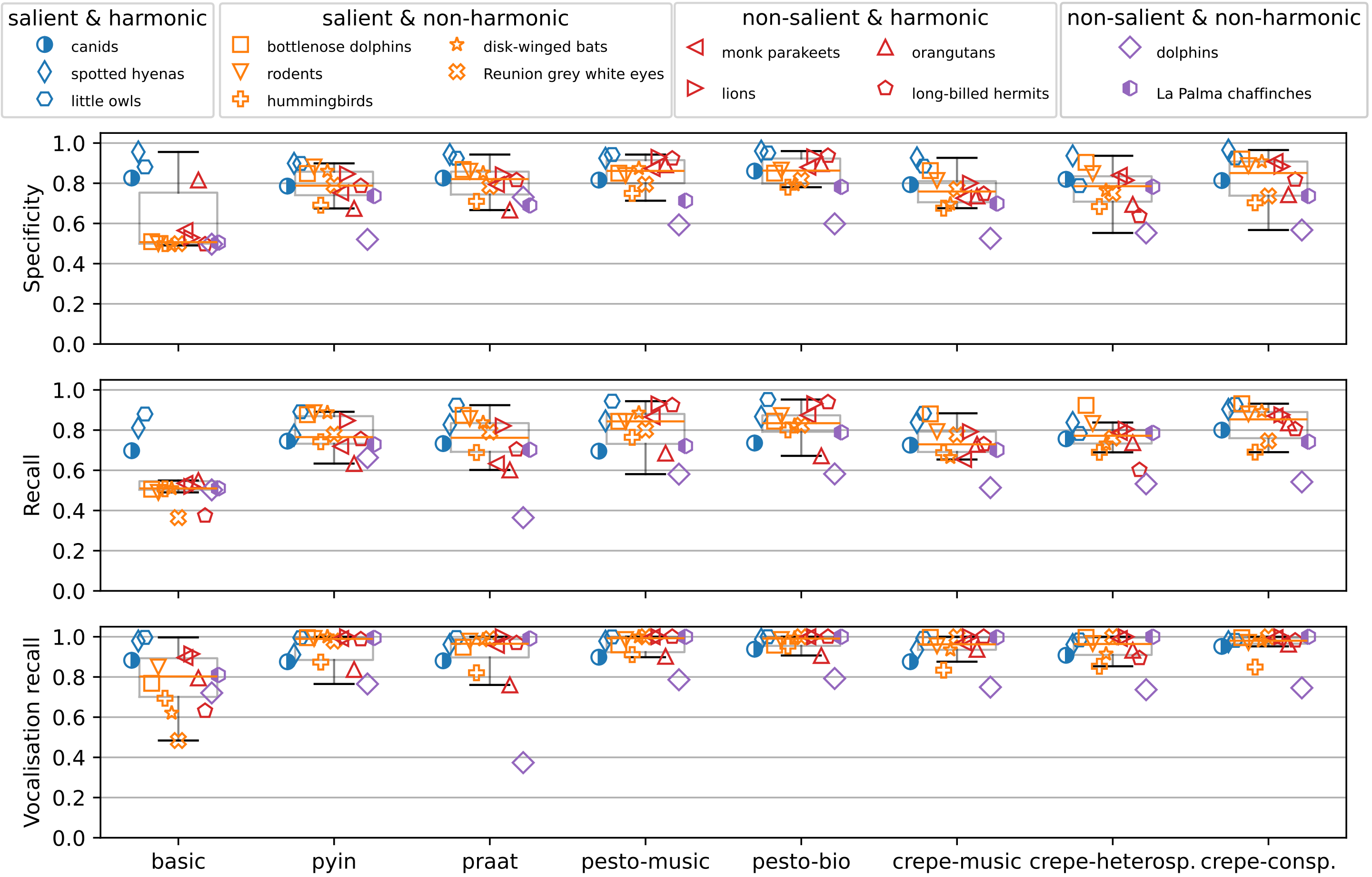


Fig 8. Voicing detection performance for each algorithm and taxa (following the dataset grouping). The recall metric is averaged across all temporal frames, whereas the vocalisation recall gives the proportion of vocalisations with at least a third of its frames detected as voiced. Boxes extend from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median, and whiskers extend from the box to the farthest data point lying within 1.5x the inter-quartile range (IQR) from the box.

least 33%) to inform on the potential use of such algorithm for vocalisation detection. As shown in Fig 8 for this metrics, to the exception of BASIC-pitch, all scores are relatively high, with `crepe-consp.` being close to the top-line performance for almost all taxa.

We report optimal threshold values that lead to these voicing detection performance in Supplementary Figure 2 An important insight brought by this visualisation is the amount of variability of this threshold depending on the taxon. Specifically comparing `crepe-heterosp.` and `crepe-consp.`, we see that training on a variety of taxa makes that the model yields a more stable confidence prediction, which implies less taxon-specific tuning to find the optimal voicing confidence threshold.

Temporal smoothing of F0 predictions

In the performance comparison conducted so far, we focused on instantaneous F0 estimation without temporal smoothing. Temporal smoothing such as the Viterbi algorithm [53, 81] is commonly used to track F0 values, as using priors from salient parts of a vocalisation might help in gaining accuracy for its fainter parts. We ran the `crepe-consp.` models applying a public implementation of the Viterbi algorithm [32], and compare scores in Table 4. Overall, temporal smoothing lead to similar result or was slightly detrimental. This might be due to the public Viterbi implementation used being specifically tuned for musical signals [32], for which typical frequency modulation rates differ from that of non-human vocalisations.

Table 4. Pitch accuracy for the `crepe-consp.` model with and without temporal smoothing.

Taxon	**Without Viterbi**	**With Viterbi**	**Gain by using Viterbi**
canids	0.73	0.74	0.004
spotted hyenas	0.94	0.91	-0.028
little owls	0.89	0.87	-0.019
bottlenose dolphins	0.93	0.91	-0.019
rodents	0.89	0.90	0.002
hummingbirds	0.63	0.63	0
disk-winged bats	0.73	0.73	0
Reunion grey white eyes	0.71	0.71	0
monk parakeets	0.30	0.30	0
lions	0.47	0.48	0.004
orangutans	0.51	0.51	0
long-billed hermits	0.42	0.40	-0.011
dolphins	0.24	0.24	0
La Palma chaffinches	0.28	0.28	0

Discussion and Conclusion

Estimating the F0 of non-human vertebrate vocalisations is crucial for bioacousticians to unveil structures in these signals, helping to address ecological and evolutionary questions. Automating this task would help to reach comparative scales for these measures (*e.g.* at individual or population-level) given how prohibitively time-intensive it can be to manually trace frequency contours.

With this study, we propose to take advantage of deep learning models in this regard, after they provided significant advances in the speech and music communities. Overall, performances are rather low as compared to what is common in the speech or music communities, in which pitch accuracies are most often above 95% [32, 33]. Several facts might explain this observation. Speech and music F0 estimation benchmarks are often with data recorded indoors if not fully synthetic, with a relatively high SNR (microphones being placed in proximity to sound sources and in quiet environments), and it took significant research efforts for algorithms to reach such scores despite an extensive knowledge of production mechanisms and great experimental control. The algorithms tested here result from this effort but were not designed to work in bioacoustic conditions, which include vocalisations that might lack harmonics, often recorded at a distance and outdoors, in the presence of other noise sources.

Scores are low as compared to tests on indoors near-field speech or music, but they are in most cases above twice the random baseline performance (0.8 % for the chroma accuracy), and `crepe-consp.` performs above three times the random baseline for all taxa. In this sense, we propose these automatic algorithms as a baseline for further developments, without which they are only reliable for signals with a relatively high F0 salience / SNR.

Nonetheless, our results demonstrate that deep learning models systematically outperform traditional methods in bioacoustic F0 estimation (Fig. 6), even for taxa not seen during training. Using this technology would therefore be advantageous to the field, similar to its benefits in other tasks such as vocalisation detection or classification. One of the advantages with training models is for instances where the first overtone has a stronger energy than the F0, which typically triggers harmonic jumps with traditional algorithms (*e.g.* this can occur in the effect of formants, a major component of speech [82]). We see from the median scores in Fig. 5 that supervised models (`crepe-consp.` and `crepe-heterosp.`) have the smallest difference between pitch and chroma accuracies. This suggests less confusion between the first overtone and the fundamental, and thus that explicitly guiding models to predict the F0 even when the first overtone has a higher energy is effective.

Moreover, our experiments with self-supervised models show that knowledge can be gained even without labels, with scores being comparable to that of supervised models for species with salient vocalisations (Fig. 6). This is especially relevant since the lack of annotated data is a major obstacle in using deep learning for bioacoustics [50]. We expect that a semi-supervised training procedure, with only few labelled examples, would allow to improve training with less salient vocalisations, which so far pose challenges as compared to more salient ones. In this sense, combining training paradigms or algorithms outputs could be a lead for further developments [11].

In our corpus characterisation, we propose a F0 salience metric which, based on simple spectral measurements, informs on the potential reliability of algorithms at estimating the F0. The visualisation of performance as a function of salience (Fig. 7) demonstrates that for vocalisations that are highly tonal and with low background noise, algorithms can reach an accuracy of 90%, but this performance drops down to 35% with the more 'noisy' vocalisations (whether they are less tonal, exposed to more background noise or both). The lion dataset for instance has the highest mean SNR (Table 1), unsurprisingly since microphones were collar-mounted, but calls seem to contain deterministic chaos [26], which makes their salience distribution relatively low (Fig. 3). For these data, scores are around 50%, and the performance gap between PRAAT and deep learning models is relatively small (about 5% depending on the model). We encourage bioacousticians to evaluate their data in this regard, in order to anticipate the potential viability of automated F0 estimation for their specific use case. Future methodological research on bioacoustic F0 estimation should focus on vocalisations with low salience, as they are the most challenging to track.

For some taxa, the BASIC-pitch model gave reasonable performances, without having been trained on non-human signals. With such an architecture designed for polyphonic music [34], there is a potential for analysing biphonic and overlapping calls, which this study does not tackle. It should be noted that pitch classifier architectures such as CREPE and PESTO could also be modified for multi-pitch tasks [33], and that their last layer's activation can already inform on multiple F0 candidates. Regardless of the chosen approach, F0 trackers will only be fully ready for real-world applications when they are capable of managing multiple simultaneous F0s, as many bioacoustic applications require. Despite the fact that the proposed dataset contains only mono-F0 annotations, mixing up signals could easily emulate multiple-F0 scenarios, and thus this corpus could still be suited to develop and evaluate multiple-F0 trackers.

With the hope of fostering further methodological developments, we publish all acoustic signals and their associated F0 ground truths in an open repository. Being aware that some of the dataset's ground truths might not perfectly match the actual oscillation frequency of vocal organs, since both manual and semi-automatic F0 annotations can be biased [49, 59], we believe this corpus can still help in automating what a bioacoustician would annotate as F0 in a signal, and hence are worth investigating. Indeed, if one generic algorithm can yield the same F0 values as one that was specifically tuned for some signal, and if it can replicate what an annotator would have considered as F0, it will help researchers save time and extend their analysis to more vocalisations. Nonetheless, users of such methods should be aware that F0 predictions can be corrupted by numerous phenomena such as non-linearities or background noise, and do not guarantee an accurate measurement of vocal organ vibration speed.

Nonetheless, before benchmarking more algorithmic procedures, future work using this dataset should focus on refining heuristics to filter label noise (especially resulting from automatic annotation procedures). Otherwise, fine-grained evaluations may be unreliable. To facilitate the application of our current experiments in other studies, we provide the Python code necessary to train and utilise pre-trained F0 estimators. The published Python interface allows to infer F0 values using a CREPE model trained on the whole dataset published here. Depending on the signals they wish to analyse, users can easily choose another model trained on a specific taxon, or set slow down / acceleration factors (to deal with infra- or ultra-sounds, or with rapid frequency sweeps such as bird trills), prediction time step, or prediction post-processing (among argmax, weighted argmax, or viterbi). Specifically, such tool can facilitate a large range of studies on non-human vocal behaviour, including to characterise frequency contours at the scale of species or communities (vocal repertoires), across individuals (individual signatures), or within individuals (across behaviours or during development).

Acknowledgments

We thank Alain Riou for taking the time to discuss his PESTO model and how to overcome training difficulties, and Christian Herbst for enlightening discussions preliminary to this work.

We thank Laela S. Sayigh and Vincent M. Janik for contributing to this work with the bottlenose dolphin data.

We thank contributors to the https://www.phylopic.org/ website for the taxonomic icons in Figures 2 and 3, all under the CC0 1.0 license, except for the disk-winged bat (under the CC0 3.0 license https://creativecommons.org/licenses/by/3.0/, credits to Roberto Díaz Sibaja), orangutans (under the CC0 3.0 license https://creativecommons.org/licenses/by/3.0/, credits to Gareth Monger), and for the La Palma chaffinch (under the CC0 3.0 license https://creativecommons.org/licenses/by/3.0/, credits to Maxime Dahirel).

Financial disclosure statement

We thank the Foundation for Science and Technology (FCT, Portugal) for funding B.F. through a PhD grant (2020.04569.BD).

We thank the Swedish Research Council for funding A.G.E. through the grant 2017–00626.

We thank the UK Research & Innovation, Future Leaders Fellowship for funding A.R.L. through the grant MR/T04229X/1; ARL.

We thank the Czech Academy of Sciences for funding M.S. through the grant RVO 68081766.

We thank the National Institute on Deafness and Other Communication Disorders (NIDCD) for supporting the mouse recordings through the grant R01DC008343 (R.C.L.).

We thank the National Institute of Mental Health (NIMH) for supporting the vole recordings through the grants P50MH100023 (R.C.L. and M.R.W.) and R01MH115831 (R.C.L. and M.R.W.).

Collection of hyena vocalizations was supported by National Science Foundation grant nos OISE1853934 and IOS1755089, Carlsberg Foundation grant no. CF 15-0915 and Human Frontier Science Program grant no. RGP0051/2019.

We thank the National Science Foundation for supporting K.D.S.L. through a Graduate Research Fellowship.

We thank the Gips-Schüle Stiftung, the Max Planck Institute of Animal Behaviour, and the Zukunftskolleg at the University of Konstanz for supporting A.S.P..

We thank the Aarhus Institute of Advanced Studies for funding F.H.J. through a AIAS-COFUND fellowship under the FP7-PEOPLE programme of the EU (agreement no. 609033).

We thank the Czech Science Foundation for funding P.L. and M.M. through the grant 21-04023K.

Monk parakeet recording and frequency contour tracing was supported by a Fulbright Study/Research grant to G.S.V., an American Ornithologists' Union Carnes Award to G.S.V., Experiment.com crowdfunding to G.S.V., a donation to G.S.V. from Drs. Michael and Susan Achey, and an NMSU Whaley Field Award to G.S.V..

We thank the Agence Nationale de la Recherche for supporting P.B. through the grant ANR-20-CE23-0012-01 (R.M.), as well as to the Institute ofConvergence ILCB, which is supported by grants from France 2030 (ANR-16-CONV-0002) andthe Excellence Initiative of Aix-Marseille University (A*MIDEX).

Funders did not play any role in the study design , data analysis, decision to publish or preparation of the manuscript.

Data availability

The Python code used to reproduce experiments are available in an open source repository[2]. It shows all packages used, in which version, and gives implementation details to evaluate existing algorithms or train deep learning models. Besides, the weights of models trained for this study, along with a ready-to-use Python interface, are also made available for researchers to use them in their own applications.

The data are accessible through this repository `https://doi.org/10.5061/dryad.prr4xgxw8`. The whole dataset is structured in a uniform way, with a sound file cut around each vocalisations (with some non-voiced padding), and a text file containing a list of time × frequency annotated values.

References

1. Huang X, Acero A, Hon HW, Reddy R. Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR; 2001.

2. Herbst CT. Biophysics of vocal production in mammals. Vertebrate sound production and acoustic communication. 2016; p. 159–189.

3. Hirst DJ, de Looze C. Measuring Speech. Fundamental frequency and pitch.; 2021.

[2] `https://github.com/mim-team/bioacoustic_F0_estimation`

4. Honorof DN, Whalen D. Identification of speaker sex from one vowel across a range of fundamental frequencies. The Journal of the Acoustical Society of America. 2010;128(5):3095–3104.

5. Skantze G. Turn-taking in conversational systems and human-robot interaction: a review. Computer Speech & Language. 2021;67:101178.

6. Lindblom B. Accuracy and limitations of sona-graph measurements. In: Proceedings of the fourth international congress of phonetic sciences. vol. 1. Mouton The Hague; 1962. p. 188–202.

7. Sundberg J, Titze I, Scherer R. Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. Journal of Voice. 1993;7(1):15–29.

8. Ekström AG. Ape vowel-like sounds remain elusive: a comment on Grawunder et al.(2022). International Journal of Primatology. 2023;44:237–239.

9. Orio N, et al. Music retrieval: A tutorial and review. Foundations and Trends in Information Retrieval. 2006;1(1):1–90.

10. Salamon J, Gómez E, Ellis DP, Richard G. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. IEEE Signal Processing Magazine. 2014;31(2):118–134.

11. Bosch JJ, Marxer R, Gómez E. Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. Journal of New Music Research. 2016;45(2):101–117.

12. Bowling DL, Garcia M, Dunn JC, Ruprecht R, Stewart A, Frommolt KH, et al. Body size and vocalization in primates and carnivores. Scientific reports. 2017;7(1):41070.

13. Fitch WT, Hauser MD. Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. In: Acoustic communication. Springer; 2003. p. 65–137.

14. Stoeger AS, Zeppelzauer M, Baotic A. Age-group estimation in free-ranging African elephants based on acoustic cues of low-frequency rumbles. Bioacoustics. 2014;23(3):231–246.

15. Kershenbaum A, Blumstein DT, Roch MA, Akçay Ç, Backus G, Bee MA, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. Biological Reviews. 2016;91(1):13–52.

16. Garland EC, Castellote M, Berchok CL. Beluga whale (*Delphinapterus leucas*) vocalizations and call classification from the eastern Beaufort Sea population. The Journal of the Acoustical Society of America. 2015;137(6):3054–3067.

17. Henry L, Barbu S, Lemasson A, Hausberger M. Dialects in animals: Evidence, development and potential functions. Animal Behavior and Cognition. 2015;2(2):132–155.

18. Lehmann KD, Jensen FH, Gersick AS, Strandburg-Peshkin A, Holekamp KE. Long-distance vocalizations of spotted hyenas contain individual, but not group, signatures. Proceedings of the Royal Society B. 2022;289(1979):20220548.

19. Wijers M, Trethowan P, Du Preez B, Chamaillé-Jammes S, Loveridge AJ, Macdonald DW, et al. Vocal discrimination of African lions and its potential for collar-free tracking. Bioacoustics. 2021;30(5):575–593.

20. Deecke VB, Janik VM. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. The Journal of the Acoustical Society of America. 2006;119(1):645–653.

21. Sayigh LS, Janik VM, Jensen FH, Scott MD, Tyack PL, Wells RS. The Sarasota Dolphin Whistle Database: A unique long-term resource for understanding dolphin communication. Frontiers in Marine Science. 2022;9:923046.

22. Linhart P, Šálek M. The assessment of biases in the acoustic discrimination of individuals. Plos One. 2017;12(5):e0177206.

23. Lameira AR, Hardus ME, Bartlett AM, Shumaker RW, Wich SA, Menken SB. Speech-like rhythm in a voiced and voiceless orangutan call. PloS one. 2015;10(1):e116136.

24. Tyack PL, Miller EH. Vocal anatomy, acoustic communication and echolocation. Marine mammal biology: An evolutionary approach. 2002;59:142–84.

25. Wilden I, Herzel H, Peters G, Tembrock G. Subharmonics, biphonation, and deterministic chaos in mammal vocalization. Bioacoustics. 1998;9(3):171–196.

26. Fitch WT, Neubauer J, Herzel H. Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. Animal behaviour. 2002;63(3):407–418.

27. Titze IR, for Voice NC, Speech. Workshop on Acoustic Voice Analysis: Summary Statement. National Center for Voice and Speech; 1995. Available from: `https://books.google.es/books?id=POk2HQAACAAJ`.

28. Riede T, Tokuda IT, Munger JB, Thomson SL. Mammalian laryngseal air sacs add variability to the vocal tract impedance: Physical and computational modeling. The Journal of the Acoustical Society of America. 2008;124(1):634–647.

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

30. Mauch M, Dixon S. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In: International conference on acoustics, speech and signal processing (ICASSP). IEEE; 2014. p. 659–663.

31. De Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America. 2002;111(4):1917–1930.

32. Kim JW, Salamon J, Li P, Bello JP. CREPE: A Convolutional Representation for Pitch Estimation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 161–165.

33. Riou A, Lattner S, Hadjeres G, Peeters G. PESTO: Pitch Estimation with Self-supervised Transposition-equivariant Objective. In: 24th International Society for Music Information Retrieval Conference (ISMIR); 2023.

34. Bittner RM, Bosch JJ, Rubinstein D, Meseguer-Brocal G, Ewert S. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2022. p. 781–785.

35. K Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. Raven Pro: Interactive Sound Analysis Software; 2023. Available from: `https://www.ravensoundsoftware.com/`.

36. Lachlan R. Luscinia; 2022. Available from: `http://rflachlan.github.io/Luscinia/`.

37. Sueur J, Aubin T, Simonis C. Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics. 2008;18:213–226.

38. Araya-Salas M, Smith-Vidaurre G. warbleR: an R package to streamline analysis of animal acoustic signals. Methods in Ecology and Evolution. 2017;8(2):184–191. doi:https://doi.org/10.1111/2041-210X.12624.

39. Jadoul Y, De Boer B, Ravignani A. Parselmouth for bioacoustics: automated acoustic analysis in Python. Bioacoustics. 2024; p. 1–19.

40. Röper K, Scheumann M, Wiechert A, Nathan S, Goossens B, Owren M, et al. Vocal acoustics in the endangered proboscis monkey (*Nasalis larvatus*). American Journal of Primatology. 2014;76(2):192–201.

41. Hagiwara M, Miron M, Liu JY. ISPA: Inter-Species Phonetic Alphabet for Transcribing Animal Sounds. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024;.

42. Gamba M, Favaro L, Torti V, Sorrentino V, Giacoma C. Vocal tract flexibility and variation in the vocal output in wild indris. Bioacoustics. 2011;20(3):251–265.

43. Poupard M, Best P, Schlüter J, Symonds H, Spong P, Lengagne T, et al. Large-scale unsupervised clustering of orca vocalizations: a model for describing orca communication systems. In: 2nd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR); 2019. p. 82–87.

44. Garcia N, Macias-Toro E, Vargas-Bonilla J, Daza J, López J. Segmentation of bio-signals in field recordings using fundamental frequency detection. In: 3rd International Work-Conference on Bioinspired Intelligence. IEEE; 2014. p. 86–92.

45. Torti V, Bonadonna G, De Gregorio C, Valente D, Randrianarison RM, Friard O, et al. An intra-population analysis of the indris' song dissimilarity in the light of genetic distance. Scientific Reports. 2017;7(1):10140.

46. Li P, Liu X, Palmer K, Fleishman E, Gillespie D, Nosal EM, et al. Learning deep models from synthetic data for extracting dolphin whistle contours. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. p. 1–10.

47. Li P, Liu X, Klinck H, Gruden P, Roch MA. Using deep learning to track time × frequency whistle contours of toothed whales without human-annotated training data. The Journal of the Acoustical Society of America. 2023;154(1):502–517.

48. O'Reilly C, Harte N. Pitch tracking of bird vocalizations and an automated process using YIN-bird. Cogent Biology. 2017;3(1):1322025.

49. Herbst CT, Dunn JC. Fundamental frequency estimation of low-quality electroglottographic signals. Journal of Voice. 2019;33(4):401–411.

50. Stowell D. Computational bioacoustics with deep learning: a review and roadmap. PeerJ. 2022;10:e13152.

51. Best P, Paris S, Glotin H, Marxer R. Deep audio embeddings for vocalisation clustering. Plos one. 2023;18(7):e0283396.

52. Gfeller B, Frank C, Roblek D, Sharifi M, Tagliasacchi M, Velimirović M. SPICE: Self-supervised pitch estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020;28:1118–1128.

53. Han K, Wang D. Neural network based pitch tracking in very noisy speech. IEEE/ACM transactions on audio, speech, and language processing. 2014;22(12):2158–2168.

54. Papale E, Buffa G, Filiciotto F, Maccarrone V, Mazzola S, Ceraulo M, et al. Biphonic calls as signature whistles in a free-ranging bottlenose dolphin. Bioacoustics. 2015;24(3):223–231.

55. Filatova O, Fedutin I, Nagaylik M, Burdin A, Hoyt E. Usage of monophonic and biphonic calls by free-ranging resident killer whales (Orcinus orca) in Kamchatka, Russian Far East. Acta ethologica. 2009;12:37–44.

56. Brown CH, Alipour F, Berry DA, Montequin D. Laryngeal biomechanics and vocal communication in the squirrel monkey (Saimiri boliviensis). The Journal of the Acoustical Society of America. 2003;113(4):2114–2126.

57. Zollinger SA, Riede T, Suthers RA. Two-voice complexity from a single side of the syrinx in northern mockingbird Mimus polyglottos vocalizations. Journal of Experimental Biology. 2008;211(12):1978–1991.

58. Suthers RA, Zollinger SA. Producing song: the vocal apparatus. Annals of the new York Academy of Sciences. 2004;1016(1):109–129.

59. Roch MA, Scott Brandes T, Patel B, Barkley Y, Baumann-Pickering S, Soldevilla MS. Automated extraction of odontocete whistle contours. The Journal of the Acoustical Society of America. 2011;130(4):2212–2223.

60. Kershenbaum A, Root-Gutteridge H, Habib B, Koler-Matznick J, Mitchell B, Palacios V, et al. Disentangling canid howls across multiple species and subspecies: structure in a complex communication channel. Behavioural processes. 2016;124:149–157.

61. Sayigh LS, Esch HC, Wells RS, Janik VM. Facts about signature whistles of bottlenose dolphins, Tursiops truncatus. Animal Behaviour. 2007;74(6):1631–1642.

62. Warren MR, Campbell D, Borie AM, Ford IV CL, Dharani AM, Young LJ, et al. Maturation of Social-Vocal Communication in Prairie Vole (*Microtus ochrogaster*) Pups. Frontiers in Behavioral Neuroscience. 2022;15:814200.

63. Liu RC, Miller KD, Merzenich MM, Schreiner CE. Acoustic variability and distinguishability among mouse ultrasound vocalizations. The Journal of the Acoustical Society of America. 2003;114(6):3412–3422.

64. Beltrán DF, Araya-Salas M, Parra JL, Stiles FG, Rico-Guevara A. The evolution of sexually dimorphic traits in ecological gradients: an interplay between natural and sexual selection in hummingbirds. Proceedings of the Royal Society B. 2022;289(1989):20221783.

65. Araya-Salas M, Hernández-Pinsón HA, Rojas N, Chaverri G. Ontogeny of an interactive call-and-response system in Spix's disc-winged bats. Animal Behaviour. 2020;166:233–245.

66. Smith-Vidaurre G, Perez-Marrufo V, Wright TF. Individual vocal signatures show reduced complexity following invasion. Animal Behaviour. 2021;179:15–39.

67. Smith-Vidaurre G, Pérez-Marrufo V, Hobson EA, Salinas-Melgoza A, Wright TF. Individual identity information persists in learned calls of introduced parrot populations. PLoS Computational Biology. 2023;19(7):e1011231.

68. Smith-Vidaurre G, Perez-Marrufo V, Wright TF. Simpler signatures post-invasion; 2021. Available from: `https://figshare.com/articles/dataset/Simpler_signatures_post-invasion/14811636`.

69. Smith-Vidaurre G, Perez-Marrufo V, Hobson EA, Salinas-Melgoza A, Wright TF. Smith-Vidaurre_et_al_2023_IdentityInformationEncoding; 2023. Available from: `https://figshare.com/articles/dataset/Smith-Vidaurre_et_al_2023_IdentityInformationEncoding/22582099`.

70. Lameira AR, Wich SA. Orangutan long call degradation and individuality over distance: a playback approach. International Journal of Primatology. 2008;29:615–625.

71. Ekström AG, Moran S, Sundberg J, Lameira A. PREQUEL: Supervised phonetic approaches to analyses of great ape quasi-vowels; 2023. Available from: `osf.io/preprints/psyarxiv/8aeh4`.

72. Araya-Salas M, Wright T. Open-ended song learning in a hummingbird. Biology letters. 2013;9(5):20130625.

73. Chen CP, Bilmes JA. MVA processing of speech features. IEEE Transactions on Audio, Speech, and Language Processing. 2006;15(1):257–270.

74. Xie J, Colonna JG, Zhang J. Bioacoustic signal denoising: a review. Artificial Intelligence Review. 2021;54:3575–3597.

75. Salamon J, Gómez E, Bonada J. Sinusoid extraction and salience function design for predominant melody estimation. In: Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11), Paris, France; 2011. p. 73–80.

76. Sun X. A pitch determination algorithm based on subharmonic-to-harmonic ratio. In: Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000); 2000. p. vol. 4, 676–679.

77. Herbst CT. Performance evaluation of subharmonic-to-harmonic ratio (SHR) computation. Journal of Voice. 2021;35(3):365–375.

78. Raffel C, McFee B, Humphrey EJ, Salamon J, Nieto O, Liang D, et al. MIR_EVAL: A Transparent Implementation of Common MIR Metrics. In: ISMIR. vol. 10; 2014.

79. McFee B. resampy: efficient sample rate conversion in Python. Journal of Open Source Software. 2016;1(8):125.

80. Morrison M. torchcrepe; 2023. Available from: https://github.com/maxrmorrison/torchcrepe.

81. Fujihara H, Kitahara T, Goto M, Komatani K, Ogata T, Okuno HG. F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search. In: International Conference on Acoustics Speech and Signal Processing (ICASSP). vol. 5. IEEE; 2006. p. V–V.

82. Fitch WT. The evolution of speech: a comparative review. Trends in cognitive sciences. 2000;4(7):258–267.