Bayesian inference for latent spectral shapes

Journal of the Royal Statistical Society Series C: Applied Statistics, 2026, **75**, 564–582
https://doi.org/10.1093/jrsssc/qlaf057
Advance access publication 5 November 2025
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Bayesian inference for latent spectral shapes

Daria Valente[1], Hiu Ching Yip[2,3], Gianluca Mastrantonio[3], Enrico Bibbona[3], Olivier Friard[1] and Marco Gamba[1]

[1]Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy
[2]HKJC Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong SAR, China
[3]Department of Mathematical Science, Polytechnique of Turin, Turin, Italy

Address for correspondence: Gianluca Mastrantonio, Department of Mathematical Science, Polytechnique of Turin, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy. Email: gianluca.mastrantonio@polito.it

Abstract

This paper proposes a hierarchical model for the analysis of spectrograms of animal calls. The motivation stems from analysing recordings of the so-called grunt calls emitted by various lemur species. Our goal is to identify a latent spectral shape that characterizes each species and facilitates measuring dissimilarities between them. The model addresses the synchronization of animal vocalizations, due to varying time-lengths and speeds, with nonstationary temporal patterns and accounts for periodic sampling artifacts produced by the time discretization of analogue signals. The former is achieved through a synchronization function, and the latter is modelled using a circular representation of time. To overcome the curse of dimensionality inherent in the model's implementation, we employ the Nearest Neighbour Gaussian Process, and posterior samples are obtained using the Markov chain Monte Carlo method. We apply the model to a real dataset comprising sounds of eight different species. We define a representative sound for each species and compare them using a distance measure. Cross-validation is used to evaluate the predictive capability of our proposal and explore special cases. Additionally, a simulation study is used to demonstrate how effectively the Markov chain Monte Carlo algorithm can identify the parameters used to generate the data.

Keywords: bioacoustics, circular time, Nearest Neighbour Gaussian Process, nonstationary covariance function, nonlinear warping

1 Introduction

The study of vocal repertoires is an essential step in understanding animal behaviour. Communication plays a crucial role in individual interactions and can provide crucial information about social relations and hierarchical organization within a group, such as mating behaviour, territoriality, group dynamics, and even predator-prey interactions. Understanding these behaviours is fundamental for conservation efforts and ecological research. Vocalizations are often collected as sound recordings and their analysis is most typically performed using spectrograms, making bioacoustic analysis a form of time-frequency analysis. Processing and learning from such high-dimensional data require statistical methods or automated algorithms. An overview of the contemporary computational methods in bioacoustic analysis is summarized in the recent surveys by Sainburg et al. (2020) and Tavakoli et al. (2024). The most common practice is the application of feature engineering methods, which generally involve manual or automatic selection of a set of basic-features for quantitative comparison. There are numerous types of such basic-features in bioacoustic such as pitch, amplitude envelope, Wiener entropy, and others. Kershenbaum et al. (2016) provided additional background on several paradigms of features and suggested a protocol to analyse them with the appropriate methodologies. Studies conducted by Gamba and Giacoma (2007) and Gamba et al. (2016) are examples of bioacoustic analyses that manually select features

Received: March 4, 2025. **Revised:** September 29, 2025. **Accepted:** October 7, 2025

such as pitches, behaviours and sex to be independent predictors or response variables for linear models. Analysing the full acoustic dataset without feature selection often involves dimensionality reduction and clustering; for instance, Valente et al. (2019) implemented a dimensionality reduction method to construct a probabilistic weighted graph using the distances between sounds.

The motivating data of this work are a set of lemur vocal signals, represented as spectrograms, that were recorded in Madagascar over 12 years. The lemurs of Madagascar are particularly fascinating because they evolved in a diverse and isolated environment from the extensive continental shelves. Moreover, among lemurs there is one of the few 'singing primates', Indri indri, a species possessing rhythmic capabilities similar to that of human musicality (De Gregorio et al., 2021) and a rich vocal repertoire (Valente et al., 2019). Lemurs can emit different signals with different aims (e.g. alarm, territorial defence, group cohesion and coordination, anti-predator mobbing) and their classification and interpretation have been subject to several investigations (Pozzi et al., 2010; Valente et al., 2022). Among the various vocalizations observed in lemurs, we focus on the 'grunt', a nasal vocalization mostly ubiquitous among lemurs (Macedonia & Stanger, 1994) and emitted in various contexts, including group movements in forests (Sperber et al., 2017). Interestingly, while the grunt is shared across several lemur species, each species produces its own distinct version due to differences in vocal tract morphology (Gamba et al., 2012). Such acoustic divergence has been hypothesized to allow discrimination across species, and to have played a role in speciation. Therefore, investigating grunting behaviour could provide insights into the unique evolutionary history of these primates (Zimmermann, 2017). Moreover, grunts belong to the so-called close calls, i.e. low-intensity, low-frequency calls emitted when individuals are in close spatial proximity with one another, common in many primate species. Being lemurs among the most basal primates, investigating their close-calling behaviour could shed light on the evolutionary roots of primate gregarious behaviour, as well as social and cognitive abilities (Pflüger & Fichtel, 2012).

The primary aim of this paper is to model the 'typical' spectrogram of the grunt sound, nonparametrically, by means of Gaussian processes (GPs) that allow for quantitative comparison across different species. A similar approach has been taken in Pigoli et al. (2018), where the statistical analysis of human languages was conducted. The authors encoded the differences between languages in the covariance operator of a GP, which is less affected by different spoken words with respect to the mean level, and more characteristic of the specific language. In our case, since we are modelling animal vocalization, which is intrinsically less complex, we assume that most of the information is instead contained in the mean of the (aligned) spectrogram, which we describe with GPs. To achieve our goal, two important factors must be considered. Firstly, each grunt call has its own specific duration and different samples of the same call may be emitted at different speeds, with the speed possibly changing during the call itself, resulting in accelerated vocal signals. Secondly, the spectrograms are affected by periodic artifacts that arise when processing the original audio signal with the short-time Fourier transform (STFT) using a fixed time window (Graps, 1995; Kumar & Foufoula-Georgiou, 1997). It should be noted that the artifacts are an integral part of spectrogram data that cannot be removed. We address these two effects by taking a weighted average of two independent stationary GPs, each defined to account for one of the two aforementioned factors. Inspired by the functional data analysis literature, see, for example, Ramsay and Silverman (2005), Ma et al. (2024), or Bryner and Srivastava (2022), where functions need to be aligned, we introduce a *synchronization function* in the first GP to align and stretch the signal, while a correlation function based on a cyclic representation of time models the cyclic artifact, with its cyclic period included in the set of parameters to be estimated. It should be noted that, even if a correlation function based on a cyclic component has already been used, see, for example, Mastrantonio et al. (2017) or Shirota and Gelfand (2017), to the best of our knowledge, this is the first time that the cyclic period is considered a parameter. By formalizing the model under a Bayesian framework, we are then able to obtain a realization of a spectrogram, that we call *latent spectral shape* of the sound or *representative* sound, which can be considered representative of the grunt calls emitted by a species. The modelling of the cyclic artifact, the use of periodicity as random variables, the way we synchronized the sound, and the model we used to describe the latent shapes are all new contributions of this work.

There are a few other papers where GPs have been proposed for analysing bioacoustic signals and modelling spectrograms. In Tavakoli et al. (2019), GPs are used in linguistics to model the variation of the British accent across Great Britain. In Wilkinson (2019), a spectral mixture of

GPs is utilized for denoising and source separation tasks of audio signals. In Pigoli et al. (2018), GPs describe differences between romance languages. Finally, in Turner and Sahani (2014), a GP is used to capture the latent spectral shape. Compared to our proposal, all these methods lack a cyclic component. In works such as Pigoli et al. (2018) and Tavakoli et al. (2019), data alignment is treated as a preprocessing step, with its uncertainty excluded from the main procedure, likely leading to an underestimation of the uncertainty. Additionally, Wilkinson (2019) and Turner and Sahani (2014) do not incorporate alignment or warping.

The dataset comprises recorded sounds varying significantly in size across species, ranging from 145 to 6,594 sounds per species (see Table 1). Each recording consists of 26 frequency points, with most calls containing more than 10 time points. This large data size makes inference with GPs computationally challenging: this problem is often referred to as the 'Big n problem' (Jona Lasinio et al., 2013). The first step we employed to solve this issue is to use the Nearest Neighbour Gaussian Process (NNGP) method introduced by Datta, Banerjee, Finley, Gelfand (2016), which is well-suited for Bayesian analysis of large spatial and spatio-temporal datasets. Even though our data are not spatio-temporal in the traditional sense, we can consider them as a realization of a bivariate process over the Cartesian product of time and (log-)frequency coordinates, where frequency plays the role of a spatial coordinate. The NNGP requires setting the maximum number of neighbours to be considered in the approximation, and although it is generally assumed that 10–15 give reasonable results (Datta, Banerjee, Finley, Gelfand, 2016), our model is more complex than the usual GP. Therefore, we prefer to have a larger set of neighbours, but, on the other hand, computational costs increase. For this reason, instead of using the entire dataset, we decided to work with a subset of the most informative observations, which are the longer ones. The model is estimated using real data, and for each species, we present the estimate of the latent sound. Additionally, we provide a method to compare and measure distances between these sounds using a quadratic distance. We discuss the results and use this distance to construct a tree, which is then compared with established phylogenetic trees from the literature. To assess whether the complexity of our model is justified, we perform cross-validation against simpler submodels in which certain features are omitted. We evaluate their performance relative to the full model using the Continuous Ranked Probability Score (CRPS) (Gneiting & Raftery, 2007; Gneiting et al., 2008). While the results vary across species, in most cases, introducing our features improves predictive performance. Using a simulated example, we also demonstrate that all parameters of our model can be learned from the data.

The paper is organized as follows. Section 2 describes the available dataset. Section 3 introduces the proposed model. Section 4 discusses computational issues, including the marginal model, the NNGP implementation, identifiability, and the sampling of the latent spectral shape. Section 5 presents the analysis of the results obtained from the application of the model to the lemur dataset, the cross-validation experiment, the construction of the distance-based tree, and the results based solely on the prealignment of the sounds. Section 6 concludes the paper and outlines directions for future work. The appendix, available on the journal's website, includes: a description of how and where the data were collected (online supplementary material Appendix A), an explanation of how prior distributions were chosen (online supplementary material Appendix B), simulations (online supplementary material Appendix C), additional results for the real data examples (online supplementary material Appendix D), a discussion of the posterior predictive checks (online supplementary material Appendix E), and the sample means of the prealigned data (online supplementary material Appendix F).

2 The data set

The motivating dataset consists of vocal signals from eight lemur species representing two taxonomic families: Lemuridae and Indriidae. All Lemuridae species were congeneric, including Eulemur coronatus (EC), Eulemur rubriventer (ER), Eulemur flavifrons (FL), Eulemur fulvus (FU), Eulemur macaco (MA), and Eulemur mongoz (MO). The Indriidae species included Indri indri (II) and Propithecus diadema (PD). We collected recordings from numerous individuals over 19 years in Madagascar and various institutions maintaining lemurs ex situ. A detailed description of the recording locations and procedures is provided in online supplementary material Appendix A. Recordings were made using Sennheiser shotgun ME 66, ME 67, and

Table 1. The number of recordings, the minimum and maximum recorded time-length, and the number of recordings that last longer than 0.1, 0.2, and 0.3 s, respectively

	EC	ER	FL	FU	II	MA	MO	PD
# recordings	1,966	6,594	1,174	1,041	1,144	3,615	1,492	145
min length	0.025	0.020	0.020	0.021	0.011	0.020	0.031	0.027
max length	4.016	0.374	0.474	0.513	0.238	2.057	1.135	0.105
# length > 0.1	1311	4531	542	464	25	1599	1277	2
# length > 0.2	412	180	15	33	4	204	268	0
# length > 0.3	87	12	2	13	0	47	73	0

Note. EC = Eulemur coronatus; ER = Eulemur rubriventer; FL = Eulemur flavifrons; FU = Eulemur fulvus; MA = Eulemur macaco; MO = Eulemur mongoz; II = Indri indri; PD = Propithecus diadema.

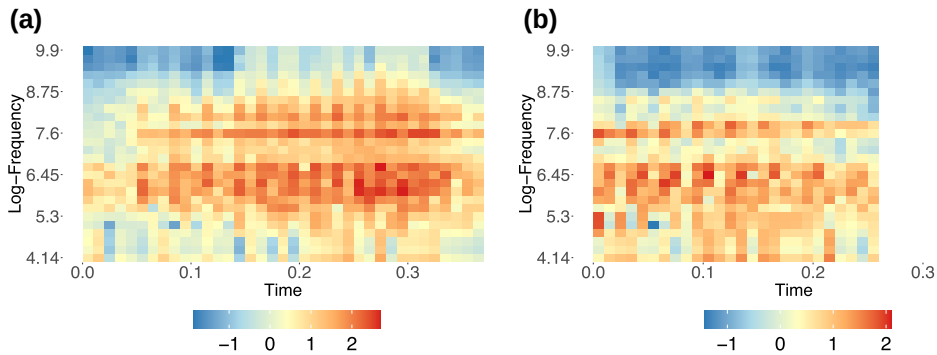

Figure 1. Spectrograms of two recorded signals of species Eulemur rubriventer (ER). The *x*-axis and *y*-axis correspond to the discretized time and frequency domains, respectively. The two figures have the same *x*-axes. (a) ER—sound 2, (b) ER—sound 26.

AKG CK 98 microphones. The signals were captured at a sampling frequency of 44.1 kHz with solid-state digital audio recorders, including the Marantz PMD671, Sound Devices 702, Olympus S100, and Tascam DR-100MKII, with a 16-bit amplitude resolution. Vocalizations were recorded from a distance of 2–10 m, with the microphone oriented toward the vocalizing animal. Each sound was categorized into one of the vocal types identified in Maretti et al. (2010), e.g. clacsons, hums, roars, based on its acoustic characteristics and evolution. Among these types there is the 'grunt', which is the main focus of this work.

Each grunt emitted by an individual lemur has been converted into a spectrogram, using the STFT implemented in the Praat 6.0.28 software (Boersma, 2001). The frequency is divided into one-third octave bands ranging from 63 to 20,000 Hertz, evenly segmented into 26 bins on a log-scale with a step size of 0.23 in log-frequency, and the temporal axis is discretized with a constant time-step of 0.01 s. While all spectrograms have the same number of log-frequency bins, the number of time bins may vary due to differences in recording lengths, as summarized in Table 1. Given that shorter sounds are more challenging to analyse and convey less information, the analysis in the paper focuses solely on the 100 longest sounds available for each species, striking a balance between precision and efficiency. Figures 1 and 2 depict examples of these spectrograms. Since we are focused on understanding the differences between spectrograms across species in terms of their frequency content and time evolution, we decided to disregard information about sound length. Although sound length can help differentiate species, it does not provide insight into the characteristics of the spectrogram itself. The sounds of different species are assumed independent; therefore, a separate instance of the model is defined for each species, and all introduced notation refers to that specific species. Let $N = 100$ denote the total number of recorded

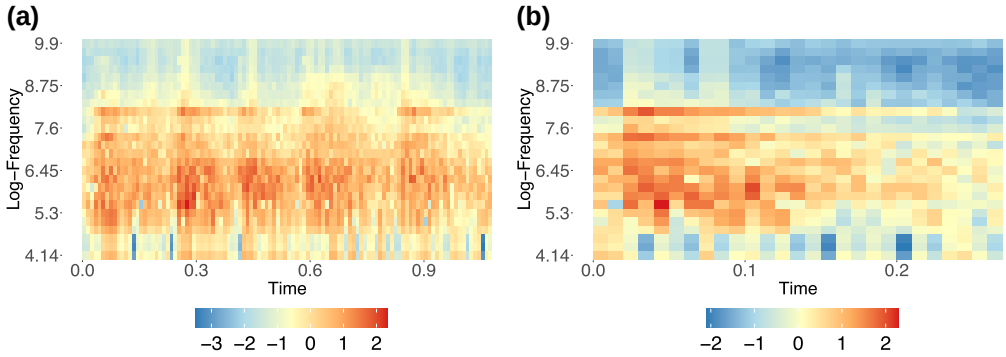

Figure 2. Spectrograms of two recorded signals of species Eulemur mongoz (MO). The *x*-axis and *y*-axis correspond to the discretized time and frequency domains, respectively. The figures have different *x*-axes. (a) MO—sound 2, (b) MO—sound 100.

signals, $H = 26$ denote the number of log-frequencies, which remains constant across all signals, and T_i denote the number of observed time points in the ith recorded signal, where $i = 1, \ldots, N$. Define $\mathcal{T}_i$ and $\mathcal{H}$ as the sets of time log-frequency points for the ith spectrogram, which, in our case, are defined as

$$\mathcal{T}_i = \{0.01(j-1) \mid j = 1, \ldots, T_i\}, \quad \mathcal{H} = \{0.23k + \log 63 \mid k = 1, \ldots, H\},$$

and let $l_i = \max \mathcal{T}_i$ represent the length of signal i. Each point in the ith spectrogram is measured in dB SPL, and can be seen as the realization of a two-dimensional process $Y_i(\cdot, \cdot) \in \mathbb{R}$, existing in continuous time and continuous log-frequency, evaluated at (t, h), where $t \in \mathcal{T}_i$ and $h \in \mathcal{H}$. Additionally, let $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_N^\top)^\top$ denote the collection of all recorded signals, where the elements of $\mathbf{y}_i$ are ordered first by ascending time and, within each time point, by increasing log-frequency values.

3 The model

The two-dimensional process $Y_i(t, h)$ is assumed to be a noisy version of a process $\mathcal{A}_i(t, h)$, modelled as follows:

$$\begin{aligned} Y_i(t, h) &= \mu_i + \mathcal{A}_i(t, h) + \epsilon_i(t, h), \\ \mathcal{A}_i(t, h) &= \sigma\Big(\sqrt{\lambda}\mathcal{W}_1\big(\psi(t; \boldsymbol{\chi}_i), h\big) + \sqrt{1-\lambda}\mathcal{W}_2(t, h)\Big), \end{aligned} \tag{1}$$

where $\mathcal{W}_1(\cdot, \cdot)$ and $\mathcal{W}_2(\cdot, \cdot)$ are two unknown functions $\mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$, which we estimate by assuming they are realizations of GPs, $\psi(t; \boldsymbol{\chi}_i)$ is a function that we will introduce in Section 3.1, and $\epsilon_i(t, h)$ is an error term assumed to follow a Gaussian distribution. Hence, we have

$$\begin{aligned} \mathcal{W}_1(d, h) &\sim \mathrm{GP}\big(0, \mathcal{C}^g(\cdot, \cdot; \boldsymbol{\theta})\big), \\ \mathcal{W}_2(t, h) &\sim \mathrm{GP}\big(0, \mathcal{C}^c(\cdot, \cdot; \boldsymbol{\theta})\big), \\ \epsilon_i(t, h) &\overset{ind}{\sim} \mathrm{GP}(0, \tau_i), \end{aligned} \tag{2}$$

where $\mathcal{C}^g(\cdot, \cdot; \boldsymbol{\theta})$ and $\mathcal{C}^c(\cdot, \cdot; \boldsymbol{\theta})$ are correlation functions. Here, $\mu_i \in \mathbb{R}$ represents the mean sound intensity level and $\tau_i \in \mathbb{R}^+$ represents the variance, and they account for factors such as environmental noise, individual variability, and variations in the global intensity of the sound due to unknown distances from the receiver. It should be noted that the i.i.d. assumption for the noise term $\epsilon_i(t, h)$ is a simplification, as in reality the noise may exhibit heteroscedasticity across frequencies. However, this assumption is introduced to avoid overcomplicating the model.

From (1) and (2), we can see that the process $\mathcal{A}_i(t, h)$ is defined as the weighted average of two GPs, which are introduced to serve distinct purposes. The component $\mathcal{W}_1(d, h)$ operates as a stationary process defined over a latent time dimension d and log-frequency h. Here, d represents a dimension to which the actual time of each recording is nonlinearly mapped by a signal-specific transformation encoded in the *synchronization function* $\psi(t; \boldsymbol{\chi}_i)$ (see Section 3.1). Time d is termed the *shared-time* and $\mathcal{W}_1(d, h)$ is referred to as the *shared-time process* or *shared-time component*. The parametric construction of the synchronization function and the correlation function $\mathcal{C}^g(\cdot, \cdot; \boldsymbol{\theta})$ will be discussed in Section 3.1. On the other hand, the component $\mathcal{W}_2(t, h)$ is an additional GP utilized exclusively for modelling periodic sampling artifacts, as detailed in Section 3.2 along with the correlation function $\mathcal{C}^c(\cdot, \cdot; \boldsymbol{\theta})$. We refer to $\mathcal{W}_2(t, h)$ as the *artifact-process* or *artifact-component*.

It is noteworthy that although $\mathcal{A}_i(t, h)$ is call-specific, $\mathcal{W}_1(d, h)$ and $\mathcal{W}_2(t, h)$ are shared across all spectrograms. Consequently, the distinction between two latent sounds, $\mathcal{A}_i(t, h)$ and $\mathcal{A}_{i'}(t, h)$, relies solely on different synchronization functions. Hence, a representative sound for the entire set of spectrograms emitted by a species can be encapsulated in a new realization $\mathcal{A}_\ell(t, h)$, with $\ell = N + 1$, such that its synchronization function $\psi(t; \boldsymbol{\chi}_\ell)$ is representative of all other synchronization functions. We refer to $\mathcal{A}_\ell(t, h)$ as the *latent spectral shape* of the sound. Further details on how to define $\psi(t; \boldsymbol{\chi}_\ell)$ and obtain an estimate of $\mathcal{A}_\ell(t, h)$ are provided in Section 3.3.

3.1 The shared-time component $\mathcal{W}_1(d, h)$ and the synchronization function $\psi(t; \boldsymbol{\chi}_i)$

From Figure 1, several notable features of the data become apparent. Regarding the time dimension, it is evident that the recordings exhibit varying durations, and they may not be perfectly aligned. For instance, Figure 1b could correspond to a shifted portion of Figure 1a, or it might represent the same sound emitted at a higher speed. Alternatively, the actual sound in Figure 1a might commence around 0.05 s instead of 0. A similar analysis can be applied to Figure 2. Hence, it becomes necessary to introduce a family of functions capable of synchronizing the sounds along the time dimension, which is defined as follows:

$$\psi(t; \boldsymbol{\chi}_i) = \alpha_i + \beta_i \Omega\left(\frac{t}{l_i}; \boldsymbol{\xi}_i\right) l_i,$$

where $\boldsymbol{\chi}_i = (\alpha_i, \beta_i, \boldsymbol{\xi}_i)$, and $\Omega(q; \boldsymbol{\xi}_i)$ represents the nonlinear *time-warping function*, defined as the Beta cumulative distribution function (CDF):

$$\Omega(q; \boldsymbol{\xi}_i) = \frac{\Gamma(\exp\zeta_i + \exp\delta_i)}{\Gamma(\exp\zeta_i)\Gamma(\exp\delta_i)} \int_0^q x^{\exp\zeta_i - 1}(1-x)^{\exp\delta_i - 1}\, dx, \qquad q \in [0, 1], \tag{3}$$

with warping parameters $\boldsymbol{\xi}_i = (\zeta_i, \delta_i)$. This synchronization function maps the real-time interval $[0, l_i]$ of the ith spectrogram to the interval $[\alpha_i, \alpha_i + \beta_i l_i]$ of the shared-time. In principle, various choices for the warping function can be made, as long as it remains continuous and strictly increasing in $[0, 1]$, satisfying the boundary conditions $\Omega(0; \boldsymbol{\xi}_i) = 0$ and $\Omega(1; \boldsymbol{\xi}_i) = 1$. We selected the Beta CDF due to its ease of interpretation, flexibility, and because, having few parameters, it avoids over-parametrization.

Figure 3 presents several examples of the time-warping function under different parametrizations. Notably, a special case of equation (3) arises when $\delta_i = \zeta_i = 0$, resulting in $\Omega(q; \boldsymbol{\xi}_i) = q$ being a linear function where the Beta reduces to the Uniform CDF. It is evident from Figure 3 that if $|\zeta_i|$ or $|\delta_i|$ are very large, e.g. > 0.75, the time-warping becomes so severe that a large portion of the real time is compressed into a very small portion of the warped time. In this case, this portion of the call would be emitted at an implausibly high speed. Therefore, constraints on the values of the warping parameters are necessary, and this will be discussed in Section 4.3.

The correlation function $\mathcal{C}^g(|d - d'|, |h - h'|; \boldsymbol{\theta})$ for the shared-time component is defined as follows:

$$\mathcal{C}^g(|d - d'|, |h - h'|; \boldsymbol{\theta}) = \frac{1}{\phi_d|d - d'| + 1} \exp\left(-\frac{\phi_h|h - h'|}{(\phi_d|d - d'| + 1)^{\rho/2}}\right). \tag{4}$$

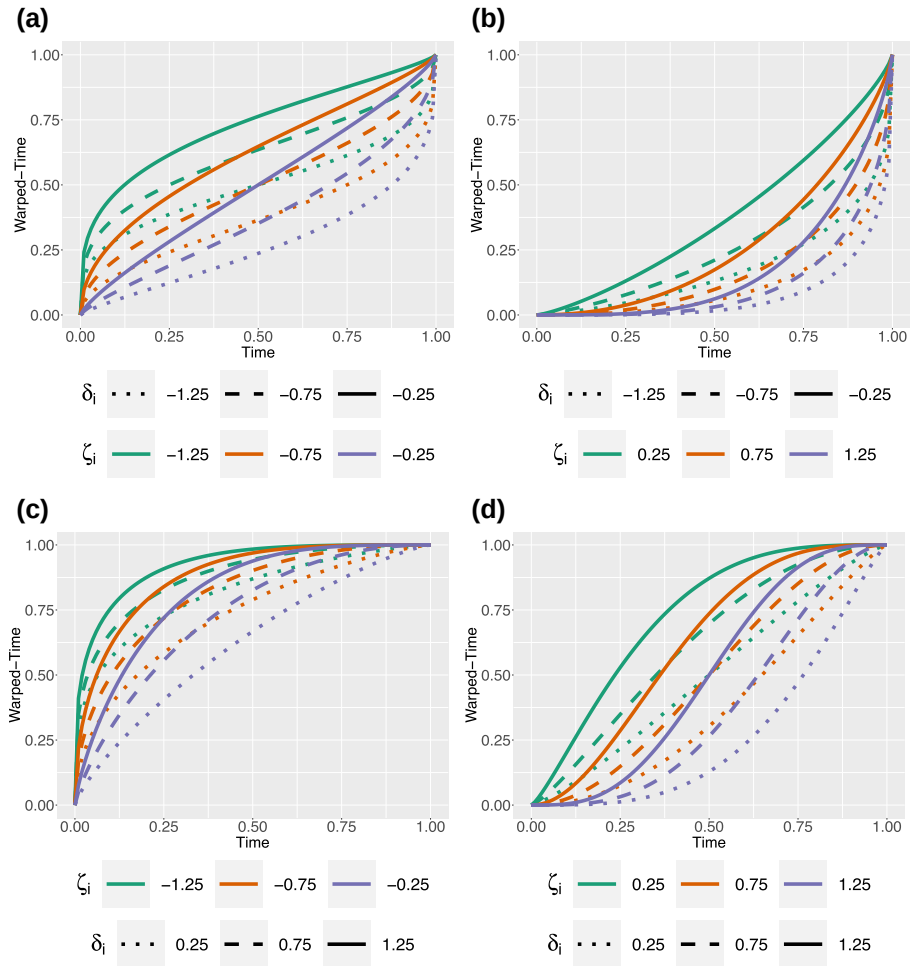

Figure 3. The time-warping function $\Omega(q \mid \boldsymbol{\xi}_i)$ under different parametric values. Line colours refer to the value of ζ_i, whereas the line types (solid, dashed, dotted …) refer to the value of δ_i. The most severe warpings are present when parameters are large in absolute value, e.g. $|\zeta_i|0.75$ or $|\delta_i|0.75$. If $\delta_i = \zeta_i = 0$, the warped time coincides with the real time.

Here, $\phi_h > 0$, $\phi_d > 0$, and $\rho \in [0, 1]$ represent the entries of the vector $\boldsymbol{\theta}$. This is a parametrization of the correlation function introduced by Gneiting (2002), selected here for its nonseparability and flexibility. The nonseparable parameter ρ represents the time-frequency interaction, while ϕ_d and ϕ_h denote the time and frequency decays, respectively. It should be noted that the model above presents identifiability issues, which will be addressed in Section 4.2.

3.2 The artifact-component $\mathcal{W}_2(t, h)$

As shown in Figure 4, the selection of the time window for the STFT can induce the appearance of artifacts, particularly when the time window is relatively small compared to the waveform cycles (Graps, 1995; Kumar & Foufoula-Georgiou, 1997). These artifacts manifest as approximately cyclical variations in intensity along the real-time axis. Upon examining Figures 1 and 2, it appears that both spectrograms exhibit temporal patterns, with a period of 0.02–0.04 s, displaying oscillations along the time domain potentially attributable to these artifacts.

The artifact component $\mathcal{W}_2(t, h)$, defined on the real time t rather than the shared-time, is designed to model this phenomenon. Consequently, it is specified by a correlation function $\mathcal{C}^c(\cdot, \cdot; \boldsymbol{\theta})$ based on circular time-distance:

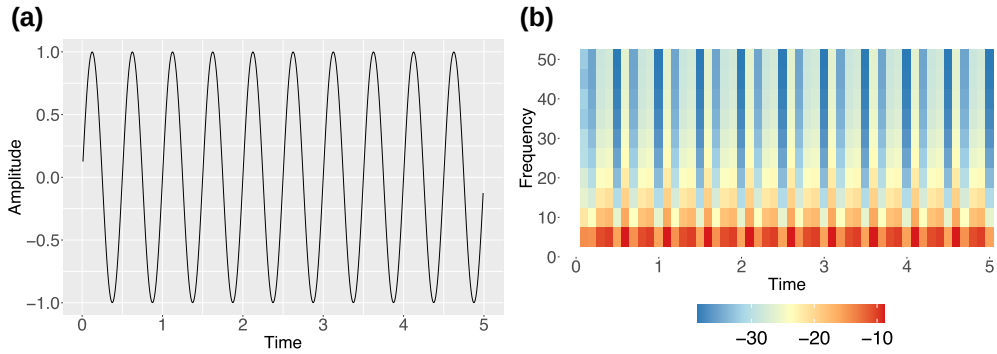

Figure 4. Example of the periodic sampling artifacts caused by the windowing effect. (a) A periodic signal generated by $y(t) = \sin(2\pi f_o t)$ with frequency $f_o = 2$. (b) The spectrogram obtained by applying the short-time Fourier transform with a time-step of 0.1 s to the periodic signal.

$$\Delta_c(|t-t'|;\gamma) = \min\{|t-t'| \bmod \gamma,\ \ \gamma - |t-t'| \bmod \gamma\}.$$

This choice of periodic distance function restricts the circular distance between any two time coordinates to a circular scale with period γ, ensuring $\Delta_c(|t-t'|;\gamma) \in [0, \gamma/2]\ \forall\ t, t'$. Subsequently, this circular distance is employed to define the Gneiting circular correlation function, akin to the approaches of Shirota and Gelfand (2017) and Mastrantonio et al. (2017), resulting in

$$\mathcal{C}^c(|t-t'|, |h-h'|; \boldsymbol{\theta}) = \frac{1}{\phi_c \Delta_c(|t-t'|;\gamma)+1} \exp\left(-\frac{\phi_b |h-h'|}{(\phi_c \Delta_c(|t-t'|;\gamma)+1)^{\rho/2}}\right), \tag{5}$$

where $\phi_c > 0$, and $\gamma > 0$ represent additional entries of the vector $\boldsymbol{\theta}$. To prevent overparametrization, the frequency decay parameter ϕ_b and the nonseparability parameter ρ in equation (5) are shared with those in (4). It should be noted that γ is a parameter that will be inferred from the data. To the best of our knowledge, this is the first time that estimation of this parameter has been proposed in the literature.

3.3 The latent spectral shape $\mathcal{A}_\ell(t, h)$

As delineated in Section 3, we aim to derive a representative sound that summarizes the information contained in all spectrograms of the same species, and it describes its 'shape'. This representative sound is defined as the predicted latent sound of a prospective observation $\mathcal{A}_\ell(t, h)$, where $\ell = N+1$. The time duration of this representative sound is denoted by l_ℓ, and its synchronization parameters are represented by $\boldsymbol{\chi}_\ell = (\alpha_\ell, \beta_\ell, \boldsymbol{\xi}_\ell)$. For $\mathcal{A}_\ell(\cdot,\ \cdot)$ to serve as a comprehensive representation, it must capture all the information conveyed by the sample. Consequently, all points in the set of observed data must be represented, which implies that $\psi(0 \mid \boldsymbol{\chi}_\ell) = \min_{i\in 1,\ldots,N} \psi(0 \mid \boldsymbol{\chi}_i)$ and $\psi(l_\ell \mid \boldsymbol{\chi}_\ell) = \max_{i\in 1,\ldots,N} \psi(l_i \mid \boldsymbol{\chi}_i)$, and that at least one cycle of the cyclic component is discernible, i.e. $l_\ell > \gamma$. If we define the set D as

$$D = \left\{ \boldsymbol{\chi}_\ell \in \mathbb{R}_+^2 \times \mathbb{R}^2 \,\middle|\, \begin{array}{l} \alpha_\ell = \min_{i\in\{1,\ldots,N\}} \alpha_i, \\ \alpha_\ell + \beta_\ell l_\ell = \max_{i\in\{1,\ldots,N\}} (\alpha_i + \beta_i l_i), \\ l_\ell > \gamma \end{array} \right\}, \tag{6}$$

it is easy to see that if $\boldsymbol{\chi}_\ell \in D$, the required constraints are satisfied. Notably, these determine the parameters α_ℓ and β_ℓ, while $\boldsymbol{\xi}_\ell$ remains unconstrained. Hence, our primary interest lies in the predictive distribution of

$$\mathcal{A}_\ell(t, h) \mid \mathbf{y}, \boldsymbol{\chi}_\ell \in D, \tag{7}$$

which represents a distribution over the latent sound, conditioned on the data and subject to the constraint (6) being satisfied.

4 Computational details

Due to the high complexity of the model, we adopt a Bayesian approach and obtain samples from the posterior distribution of the model parameters using an MCMC algorithm. In the next section, we provide details on the implementation, identifiability issues, prior specification, the methodology for sampling from the predictive distribution of $\mathcal{A}_\ell(t, h)$, and a description of the NNGP approximation, which is employed to reduce the computational time required for inference.

4.1 Marginal model

The evaluation of $\mathcal{W}_1(\cdot, \cdot)$ at different time points, determined by the individual synchronization function $\psi(t; \boldsymbol{\chi}_i)$, would necessitate sampling an impractically large number of random variables due to the dataset size. Consequently, we opt to marginalize the model with respect to $\mathcal{W}_1(\cdot, \cdot)$ and $\mathcal{W}_2(\cdot, \cdot)$, a common practice in geo-statistical models (see, for example, Gelfand et al., 2010; Mastrantonio et al., 2015).

While the observed processes $Y_i(t, h)$ and $Y_{i'}(t', h')$, where $i \neq i'$, are conditionally independent given the latent processes $\mathcal{W}_1(\cdot, \cdot)$ and $\mathcal{W}_2(\cdot, \cdot)$, the marginalization introduces dependence between the signals. Consequently, the entire set of observations forms a multivariate GP with a covariance function given by

$$\mathcal{C}^y_{i,i'}\big((t, h), (t', h'); \boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}, \boldsymbol{\theta}\big) = \mathcal{C}^a_{i,i'}\big((t, h), (t', h'); \boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}, \boldsymbol{\theta}\big) + \tau_i \mathbb{I}\big((i, t, h) = (i', t', h')\big), \tag{8}$$

where $\mathbb{I}((i, t, h) = (i', t', h'))$ is an indicator function that is equal to 1 if the condition of its argument is true, and 0 otherwise, while $\mathcal{C}^a_{i,i'}((t, h), (t', h'); \boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}, \boldsymbol{\theta})$ is equal to

$$\sigma^2\big(\lambda \mathcal{C}^g(\Delta_d(t, t'; \boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}), |h - h'|; \boldsymbol{\theta}) + (1 - \lambda)\mathcal{C}^c(|t - t'|, |h - h'|; \boldsymbol{\theta})\big) \tag{9}$$

with

$$\Delta_d(t, t'; \boldsymbol{\chi}_i, \boldsymbol{\chi}_{i'}) = |\psi(t; \boldsymbol{\chi}_i) - \psi(t'; \boldsymbol{\chi}_{i'})| = |\alpha_i + \beta_i \Omega(t/l_i; \boldsymbol{\xi}_i) l_i - \alpha_{i'} - \beta_{i'} \Omega(t'/l_{i'}; \boldsymbol{\xi}_{i'}) l_{i'}|.$$

This formulation reveals that while the stationary processes $\mathcal{W}_1(\cdot, \cdot)$ and $\mathcal{W}_2(\cdot, \cdot)$ are instrumentally employed to formulate the model in equation (1), the spectrograms $Y_i(t, h)$ are not assumed to be stationary. This fact is reflected in the covariance (8), which explicitly depends on t and t', not solely on $|t - t'|$. Furthermore, an important observation is that the warping parameters $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_{i'}$ exclusively contribute to the nonstationary part of the covariance function (8), and thus they can be interpreted as parameters of nonstationarity.

4.1.1 Nearest neighbours Gaussian process

The most obvious problem in the implementation of the marginalized model is the so-called *Big* n *Problem* (Jona Lasinio et al., 2013) caused by the computation of the multivariate Gaussian likelihood, which requires the inversion of the covariance matrix of dimension $M = \prod_{i=1}^N T_i H$. To be able to estimate the model, we adopt the NNGP approximation (Datta, Banerjee, Finley, Gelfand, 2016), which is one of the best methods for efficiently and accurately approximating the density of a GP. The idea behind the NNGP is to write the joint density as a product of univariate conditional distributions, and then to reduce the size of each conditioning set to increase the computational efficiency of the MCMC algorithm. It has been shown (Datta, Banerjee, Finley, Gelfand, 2016) that the approximation error introduced by the NNGP is negligible if the observations included in the conditioning sets are highly correlated with those for which the conditional density is expressed. The observations left in the conditioning set of each variable are called its neighbours.

To implement the NNGP, we first remind the reader that $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_N^\top)^\top$ denotes the collection of all recorded data, and that the elements of each $\mathbf{y}_i$ are ordered by ascending time first, and then by increasing values of log-frequencies within each time. To facilitate the notation, we

indicate as $\mathbf{y}^{1:j}$ the vector composed of the first j elements of $\mathbf{y}$ and as y^j the jth element of $\mathbf{y}$. Then, the joint density of the data can be approximated as

$$f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(y^1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}) \prod_{j=2}^{M} f(y^j \mid \mathbf{y}^{1:j-1}, \boldsymbol{\theta}, \boldsymbol{\eta}) \approx f(y^1 \mid \boldsymbol{\theta}, \boldsymbol{\eta}) \prod_{j=2}^{M} f(y^j \mid \mathcal{N}_j, \boldsymbol{\theta}, \boldsymbol{\eta}),$$

where $\mathcal{N}_j \subseteq \mathbf{y}^{1:j-1}$ contains the set of neighbours of y^j that are used in the NNGP approximation. To better approximate the original process, the set $\mathcal{N}_j$ must contain observations that allow the NNGP approximation to capture all types of dependencies described by the covariance (or correlation) function in (8). Hence, $\mathcal{N}_j$ should include observations from the same call as y^j, to capture within-call dependence, as well as observations from other calls, to capture between-call dependence. Moreover, the dependence structure is modelled through two additive components, $\mathcal{C}^g(\cdot, \cdot; \cdot)$ and $\mathcal{C}^c(\cdot, \cdot; \cdot)$, which must be learned from the neighbour set. Let assume that y^j belongs to call i, we define four different sets of neighbours to construct $\mathcal{N}_j$ such that

$$\mathcal{N}_j = \mathcal{E}_j^g \cup \mathcal{R}_j^g \cup \mathcal{E}_j^c \cup \mathcal{R}_j^c, \quad \mathcal{E}_j^g \cap \mathcal{E}_j^c = \mathcal{R}_j^g \cap \mathcal{R}_j^c = \emptyset, \tag{10}$$

where $\mathcal{E}_j^g$ contains neighbours from call i that maximize the dependence based on $\mathcal{C}^g(\cdot, \cdot; \cdot)$, and $\mathcal{E}_j^c$ contains neighbours from call i, not included in $\mathcal{E}_j^g$, that maximize the dependence based on $\mathcal{C}^c(\cdot, \cdot; \cdot)$. The sets $\mathcal{R}_j^g$ and $\mathcal{R}_j^c$ are defined similarly, but include only neighbours from the call $i-1$, if the call index is greater than 1, otherwise, they are empty sets. It is important to note that the neighbour sets change at each iteration of the MCMC algorithm, as the parameters in the covariance functions are updated, similarly to the approach proposed by Datta, Banerjee, Finley, Hamm, et al. (2016). Additionally, to eliminate any effect arising from the arbitrary ordering assigned to the set of calls, which determines the possible elements in $\mathcal{R}_j^g$ and $\mathcal{R}_j^c$, the permutation of the call indices is treated as an additional parameter, with a uniform prior distribution over the space of permutations. This parameter is updated using a Metropolis step to ensure thorough exploration of the parameter space and convergence of the algorithm.

4.2 Identifiability issues

The model suffers from identifiability issues. For instance, adding a constant $c \in \mathbb{R}^+$ to all α_i, or multiplying all α_i, β_i by a constant $c \in \mathbb{R}^+$ while dividing the decay rate ϕ_d by the same constant, leaves the value of the covariance $C^a_{i,i'}$ in equation (9) unchanged. To resolve this issue, a constraint can be set as follows:

$$\begin{cases} \min\{\alpha_1, \ldots, \alpha_N\} = 0, \\ \max\{\alpha_1 + \beta_1 l_1, \ldots, \alpha_N + \beta_N l_N\} = 1. \end{cases} \tag{11}$$

Equation (11) defines [0, 1] as the interval for the shared-time component: it should be noted that the choice of this interval is arbitrary, and changing it does not affect posterior inference. This condition can be imposed a priori using an appropriate prior distribution, or it can be enforced, after model fitting, by remapping posterior samples to an identifiable version, by exploiting the covariance invariance property; we decided to implement the latter. Let $\alpha^{(b)}, \beta^{(b)}$, and $\phi_d^{(b)}$ represent the bth posterior sample of the parameters. If we compute $\alpha_{\min}^{(b)} = \min\{\alpha_1^{(b)}, \ldots, \alpha_N^{(b)}\}$ and $l_{\max}^{(b)} = \max\{\alpha_1^{(b)} + \beta_1 l_1, \ldots, \alpha_N^{(b)} + \beta_N l_N\} - \alpha_{\min}^{(b)}$, then we can remap $\alpha^{(b)}, \beta^{(b)}$, and $\phi_d^{(b)}$ to

$$\begin{cases} \alpha_i^{*(b)} = \frac{\alpha_i^{(b)} - \alpha_{\min}^{(b)}}{l_{\max}^{(b)}}, \quad i = 1, \ldots, N, \\ \beta_i^{*(b)} = \frac{\beta_i^{(b)}}{l_{\max}^{(b)}}, \quad i = 1, \ldots, N, \\ \phi_d^{*(b)} = l_{\max}^{(b)} \phi_d^{(b)}. \end{cases} \tag{12}$$

Consequently, the remapped posterior samples obtained using equation (12) adhere to the identification constraints given by equation (11). Additionally, the covariance function in equation (8), evaluated with $(\alpha_i^{*(b)}, \beta_i^{*(b)}, \phi_d^{*(b)})$, is equal to the one computed with $(\alpha_i^{(b)}, \beta_i^{(b)}, \phi_d^{(b)})$.

4.3 Sampling from the latent spectral shape

As mentioned in Section 3.3, we need to obtain posterior samples from the distribution (7) of the representative sound $\mathcal{A}_\ell(t, h)$. We define $\boldsymbol{\eta}_i = (\mu_i, \tau_i, \alpha_i, \beta_i, \boldsymbol{\xi}_i)$, $\boldsymbol{\theta} = (\sigma^2, \lambda, \phi_d, \phi_c, \phi_h, \rho, \gamma)$, and we denote by $\mathbf{A}_\ell$ the random vector formed by evaluating $\mathcal{A}_\ell(t, h)$ at all $t \in \cup_i \mathcal{T}_i$ and $h \in \mathcal{H}$, with its elements ordered first by ascending time and, within each time point, by increasing log-frequency values. We can easily see that

$$\begin{pmatrix} \mathbf{A}_\ell \\ \mathbf{Y} \end{pmatrix} \Big| \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\chi}_\ell \sim N\left(\begin{pmatrix} 0 \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{a}_\ell} & \boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}} \\ \boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}}^T & \boldsymbol{\Sigma}_{\mathbf{y}} \end{pmatrix} \right), \tag{13}$$

where $\mathbf{Y}$ is the vector of random variables describing the set of all observations, $\boldsymbol{\nu} = (\boldsymbol{\nu}_1^\top, \boldsymbol{\nu}_2^\top, \ldots, \boldsymbol{\nu}_N^\top)^\top$ with $\boldsymbol{\nu}_i = \mu_i \mathbf{1}_{T_i H}$, $\boldsymbol{\Sigma}_{\mathbf{y}}$ is the covariance between the observed data that has elements computed using (8), while $\boldsymbol{\Sigma}_{\mathbf{a}_\ell}$ and $\boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}}$ are computed using (9). The conditional distribution of $\mathbf{A}_\ell$ can be derived from (13) with the standard results from the multivariate normal:

$$\mathbf{A}_\ell \mid \mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\chi}_\ell \sim N\left(\boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{Y} - \boldsymbol{\nu}), \boldsymbol{\Sigma}_{\mathbf{a}_\ell} - \boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \boldsymbol{\Sigma}_{\mathbf{a}_\ell,\mathbf{y}}^T \right). \tag{14}$$

If we assume $\alpha_\ell = 0$, $\beta_\ell l_\ell = 1$ and $l_\ell > b_\gamma$, equation (14) can be seen as the distribution of $\mathbf{A}_\ell \mid \mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\xi}_\ell, \boldsymbol{\chi}_\ell \in D$ (see equation (6)). Posterior samples from the predictive distribution $\mathcal{A}_\ell(t, h) \mid \mathbf{Y}, \boldsymbol{\chi}_\ell \in D$ can be drawn by using a standard Monte Carlo procedure, since

$$f(\mathbf{a}_\ell \mid \mathbf{y}, \boldsymbol{\chi}_\ell \in D) = \iiint f(\mathbf{a}_\ell \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{y}, \boldsymbol{\xi}_\ell, \boldsymbol{\chi}_\ell \in D) f(\boldsymbol{\xi}_\ell \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\eta} \, \mathrm{d}\boldsymbol{\xi}_\ell,$$

where $\mathbf{a}_\ell$ is a realization of $\mathbf{A}_\ell$, $f(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{y})$ is the posterior distribution of all model parameters, $f(\mathbf{a}_\ell \mid \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{y}, \boldsymbol{\xi}_\ell, \boldsymbol{\chi}_\ell \in D)$ is the density of (14), and $f(\boldsymbol{\xi}_\ell \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ is a distribution over the warping parameters, which must be specified. We assume that the time-warping parameters $\boldsymbol{\xi}_i$, with $i \in \{1, 2, \ldots, N, \ell\}$, are independent samples from a common parametric distribution, reflecting the biological assumption of individual variability in the way they produce a sound. However, since it is hardly justifiable that these values may induce severe time-warping (see Section 3.1 and Figure 3), we consider a population distribution with support in a finite domain such as $\zeta_i \in (-b_\zeta, b_\zeta)$ and $\delta_i \in (-b_\delta, b_\delta)$ for all $i = 1, \ldots, N$. The distributions are

$$\begin{aligned} \log\left(\frac{\zeta_i + b_\zeta}{b_\zeta - \zeta_i} \right) \Big| m_\zeta, v_\zeta &\sim \mathcal{N}(m_\zeta, v_\zeta), \\ \log\left(\frac{\delta_i + b_\delta}{b_\delta - \delta_i} \right) \Big| m_\delta, v_\delta &\sim \mathcal{N}(m_\delta, v_\delta). \end{aligned} \tag{15}$$

A discussion on how to specify the prior distribution for the mean and variance parameters in (15) is deferred to online supplementary material Appendix B.

5 Application to lemurs' calls

In this section, we present the results of the model applied to the data outlined in Section 2. We estimate eight models, one for each species. The models are implemented with 60,000 iterations, discarding the first 48,000 as burn-in, and retaining every sixth iteration, resulting in 2,000 posterior samples for inference. The maximum number of neighbours in the NNGP, i.e. the maximum number of elements in $\mathcal{N}_j$, is fixed at 40, equally distributed across the four sets of Equation (10). In selecting the bounds b_ζ and b_δ for the intervals of the warping parameters, we aimed to balance two opposing requirements. On the one hand, excessive time warping would imply an unrealistic

speed at which the sounds are produced, as discussed in Section 3.1. On the other hand, we want to restrict the parameter space as little as possible, to allow the algorithm to explore a broad range of configurations. The upper bound we chose for both, $b_\zeta = b_\delta = 0.75$, offers a good compromise between these two needs. As shown in Section 3.1 it allows for substantial time warping, e.g. in panel (c) where, approximately, the first 10% of the sound is mapped to the first 50% of the warped time, while the last 50% of the sound is in the last 10% of the warped time. The priors are

$$\begin{aligned} m_\zeta &\sim N_{(-5,5)}(0, 0.75), & m_\delta &\sim N_{(-5,5)}(0, 0.75), \\ v_\zeta &\sim \mathrm{IG}_{<0.75}(0.01, 0.01), & v_\delta &\sim \mathrm{IG}_{<0.75}(0.01, 0.01). \end{aligned}$$

Additionally, $\alpha_i \sim U(0, 0.2)$, $\tilde{\beta}_i \sim U(0.75, 1)$, σ^2, $\tau_i \sim \mathrm{IG}(1.0, 1.0)$, $\mu_i \sim N(0, 100000)$, $\phi_h \sim \mathrm{U}(0.521, 13.025)$,

$$\phi_c \,|\, \gamma \sim \mathrm{U}\left(\frac{1.0 - 0.05}{0.05(0.5\gamma)}, \frac{1.0 - 0.05}{0.05 \times 0.01}\right)$$

and

$$\phi_d \,|\, (\beta_1, \beta_2, \ldots, \beta_N) \sim \mathrm{U}\left(\frac{1.0 - 0.05}{0.05 \max\left(\{\beta_i l_i\}_{i=1}^N\right)}, \frac{1.0 - 0.05}{0.05 \min\left(\{\beta_i l_i/(T_i - 1)\}_{i=1}^N\right)}\right).$$

The motivation behind the choice of these prior distributions is explained in online supplementary material Appendix B. Convergence of the MCMC chain is assessed using the Gelman–Rubin R convergence diagnostic (Vats & Knudson, 2021). The total computation time, using a parallel implementation in Julia (Bezanson et al., 2017) on 32 cores, ranged from a minimum of 4 days for the PD species to 15 days for the EC species; the differences in computational time are due to variations in sound lengths across species (see Table 1). A Julia package with the model implementation, as well as the data used in this section, is available in a GitHub repository: https://github.com/GianlucaMastrantonio/Bayesian_inference_for_Latent_Spectral_Shapes

5.1 Representative sound

In terms of posterior means of $\mathcal{A}_\ell(t, h)$, shown in Figure 5, there are important differences among species, especially in the log-frequencies 4.14, which is the lowest one, and between 6.9 and 8.75. In the frequency range (5.06, 6.65), all species have strong intensity, which is generally quite constant across time, with the exception of species EC and MA, which decrease along the time axis. The log-frequency distribution reflects the formant (local maxima of the spectrum) distribution of the respective species, showing that frequency variation in ER, FU, and MA is higher than in the other species at higher frequencies. This result agrees with Gamba et al. (2012) findings concerning formant distribution and variation across lemur species, where the fourth formant in FU, MA and ER was estimated and measured at around 6 kHz (see also Figure 1 in the same paper). Variation in frequency distribution across species also reflects previous findings across Eulemur species, where distinctive species-specific traits were found in low-pitched grunts (see Gamba et al., 2012, 2016).

For all species, the variance shown in Figure 6 is approximately constant, with larger values at the initial time points. This can be attributed to the synchronization function, since these points may be estimated with fewer observations, while the apparent cyclic pattern may be caused by the cyclic process. An interesting exception is species FU, which has larger variability at frequency 8.75. The almost constant level suggests that variability across individuals is not specific to a particular frequency or time point, but rather affects the entire spectrum, while the band at frequency 8.75 for FU shows that not all animals of this species share the same pattern at this frequency.

5.2 Distances between the latent spectral shapes and corresponding tree

As a means to compare the representative sounds of each species, we decided to compute a distance between the posterior means of $\mathcal{A}_\ell(t, h)$. We can do this in different ways, but the easiest one is to

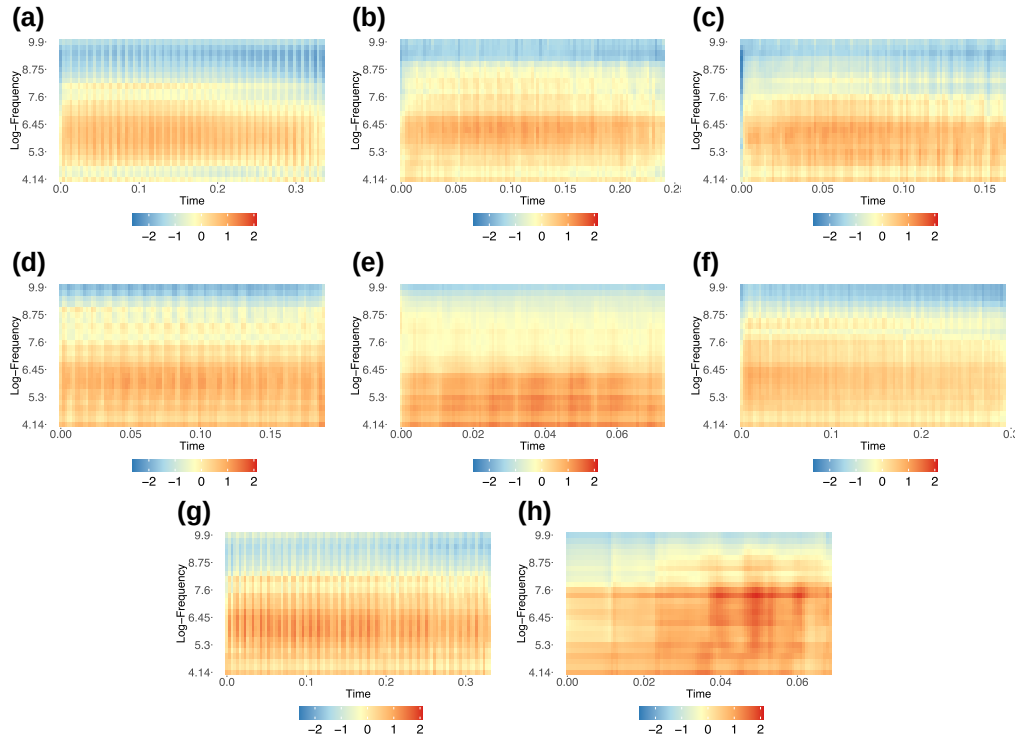

Figure 5. Posterior means of $\mathcal{A}_\ell(t, h)$, assuming l_ℓ is equal to the median observed duration for the given species, with time points spaced by 0.01 (as in the observed data), and with the number and positions of log-frequency points matching those in the data. The colour scale is shared across all images. (a) Eulemur coronatus (EC), (b) Eulemur rubriventer (ER), (c) Eulemur flavifrons (FL), (d) Eulemur fulvus (FU), (e) Indri indri (II), (f) Eulemur macaco (MA), (g) Eulemur mongoz (MO), (h) Propithecus diadema (PD).

evaluate all $\mathcal{A}_\ell(t, h)$ on a shared set of frequency and temporal points. Hence, for all species, we assume l_ℓ to be equal to the median length of the observed time across all species, with time points separated by 0.01, and the number and positions of log-frequency are given by the set $\mathcal{H}$. Then, the distance between species A and B is the sample mean over all points in the spectrogram of the square differences between the intensity of the spectrogram at the same time-frequency points of the two species. The results are shown in Figure 7a, which will be used to compute the distance-based tree. In this context, the posterior means represent the best estimates we have of the latent spectral shapes, which is why we chose to compare species using these. Using the R package phylogram (Wilkinson & Davy, 2018), we compute a tree based on Figure 7 with a hierarchical clustering approach (Paradis, 2012). To evaluate the uncertainty in the method, we compute the posterior distribution of the pairwise distances between species, shown as boxplots in Figure 8, where in each panel, one species is taken as the reference, and all distances are computed with respect to it. Another possible way to assess uncertainty over the tree is to compute, for each iteration, the corresponding distance matrix and then build a tree from it. With this method, given the variability of the results, each iteration would produce a different distance hierarchy, and consequently, a different tree would be selected. With only 2,000 samples and given the vast size of the tree space, this approach does not allow us to identify a consensus tree. In contrast, using the average posterior distances yields a clearer pattern with a meaningful interpretation that, at least partially, agrees with previous findings in the literature, as shown in the comments below.

Reliable reconstructions of the phylogenesis have been prevented by the absence of fossils for lemurs, the high degree of parallelism among lemurs, and the inconsistency of reconstructions regarding relationships within Lemuridae. Concerning this last point, understanding relationships between the Eulemur species is the most daunting task, as demonstrated by the many available reconstructions that rarely agree. According to the dissimilarity matrices in Figure 7a, the distances

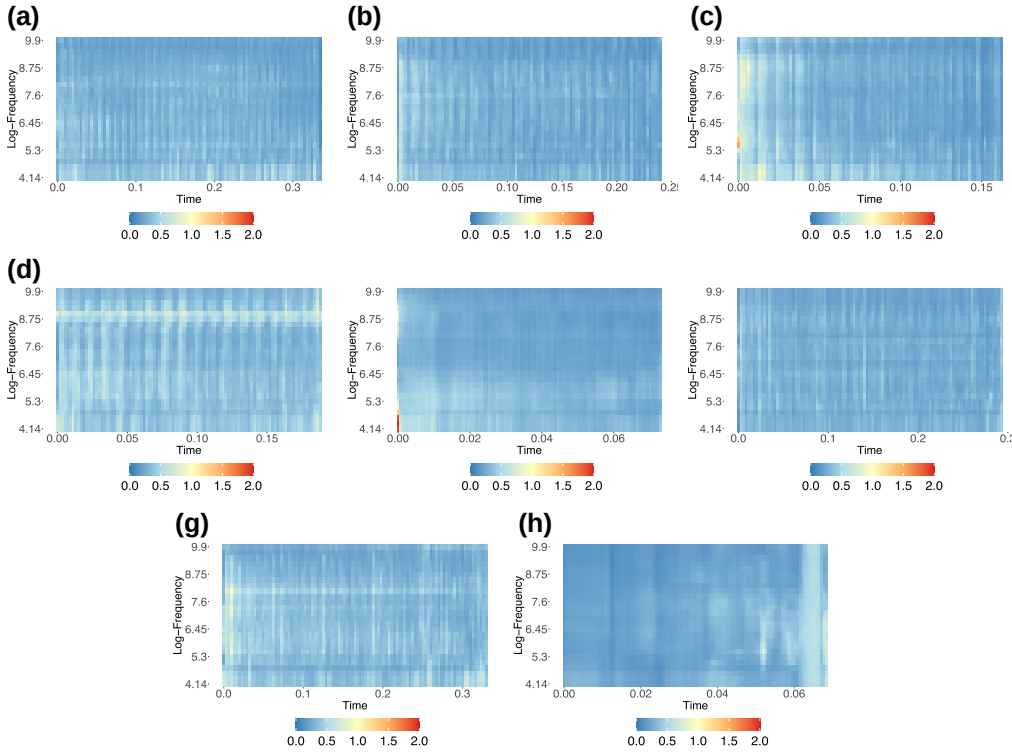

Figure 6. Posterior variances of $\mathcal{A}_\ell(t, h)$, assuming l_ℓ is equal to the median observed duration for the given species, with time points spaced by 0.01 (as in the observed data), and with the number and positions of log-frequency points matching those in the data. The colour scale is shared across all images. (a) Eulemur coronatus (EC), (b) Eulemur rubriventer (ER), (c) Eulemur flavifrons (FL), (d) Eulemur fulvus (FU), (e) Indri indri (II), (f) Eulemur macaco (MA), (g) Eulemur mongoz (MO), (h) Propithecus diadema (PD).

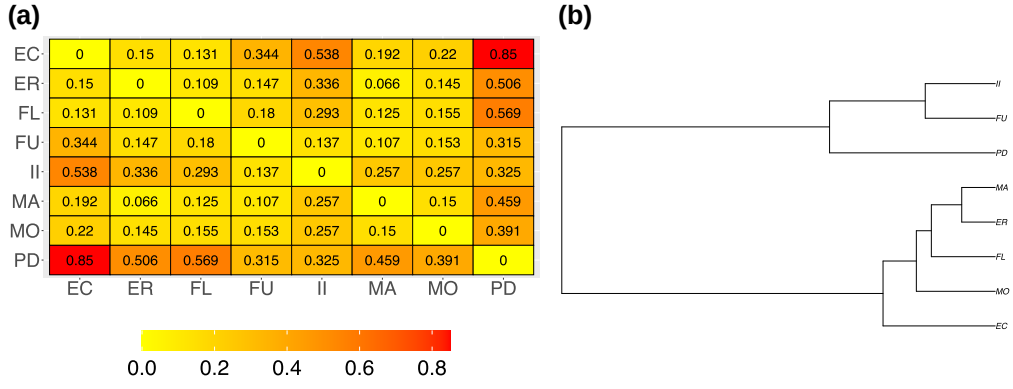

Figure 7. (a) Matrix containing the quadratic distances, for each pair of species, between the posterior means of $\mathcal{A}_\ell(t, h)$, and (b) the associated tree computed from the distance matrix. (a) Quadratic distance, (b) Tree.

reveal that MA and FL are not closer than the other taxa, which is surprising considering they were earlier considered subspecies of Eulemur macaco. However, a study by Gamba and Giacoma (2008) already showed differences in grunt vocalizations among individuals of what were then considered two subspecies of Eulemur macaco, suggesting that divergence across vocalizations may only partially map differences at the phylogenetic level (in agreement with findings of Macedonia & Stanger, 1994). These differences were mainly attributed to formants and

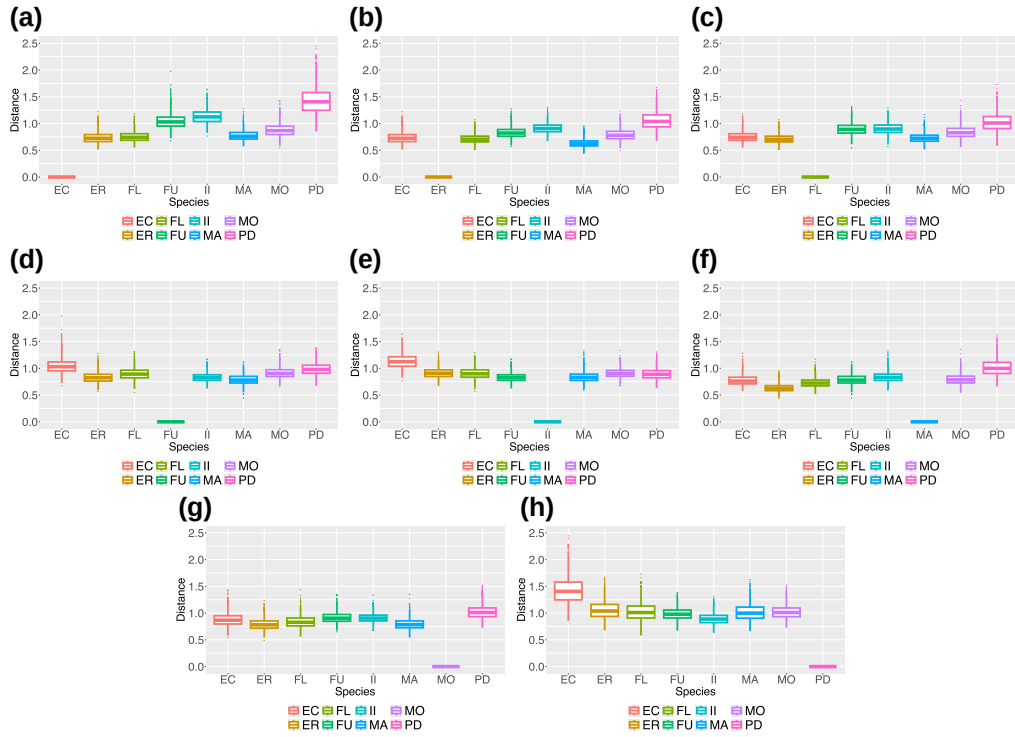

Figure 8. Graphical representation of the posterior distribution of the quadratic distances between species, whose means are shown in Figure 7. Each panel shows one species in relation to all the others. (a) Eulemur coronatus (EC), (b) Eulemur rubriventer (ER), (c) Eulemur flavifrons (FL), (d) Eulemur fulvus (FU), (e) Indri indri (II), (f) Eulemur macaco (MA), (g) Eulemur mongoz (MO), (h) Propithecus diadema (PD).

fundamental frequency and were interpreted as informative of morphological differences between the two species (Gamba & Giacoma, 2008). Vocalisations in MA and FL may have diverged to prevent hybridization, which has been observed in the wild. This is supported by research on African Cichlids, particularly of the genus Pseudotropheus, which reside in Lake Malawi. These fish have undergone adaptive divergence in genes that contribute to odour-mediated mate selection, showing how communicative signals can play a critical role in species-specific interactions. Although phylogenetic reconstructions obtained from the analyses of the lemuroid data sets are often based on very different datasets, and those on communication data are rare, the trees generated during this study show a relatively different picture from previous investigations. We must also consider that DNA research data must still establish a clear connection in some clades (DelPero et al., 2006). By looking at Figure 7b, we can see that the analysis of low-pitched calls we performed led to identification of II and PD as outgroups of the Lemuridae in this study, but FU surprising clustered together with II, possibly due to higher energy in low frequencies of their calls. On the other hand, as it has been found in our distance-based tree, MA and ER are sister groups. We found inconsistent positioning of the Indriidae species in this study, as PD and II did not show lower dissimilarity than the Lemuridae. This is probably related to an inherent problem with the sounds we considered for PD. The short duration of the PD sounds prevents proper construction of our models and mapping of differences with other species.

5.3 Cross-validation

In this section, we assess how our model describes the data compared to specific sub-cases obtained by removing certain model components. The rationale behind this approach is to test whether all model components are necessary for a satisfactory data description. Specifically,

Table 2. CRPS indices for both cross-validation experiments

	CV1				CV2			
	All	NoWarp	NoCic	NoAl	All	NoWarp	NoCic	NoAl
EC	**0.95**	0.96	1.22	0.97	1.04	**1.02**	1.42	1.20
ER	**0.91**	0.92	1.17	0.98	**1.10**	1.13	1.54	1.28
FL	**0.92**	0.96	1.31	0.94	**1.02**	1.26	1.54	1.05
FU	0.91	1.02	1.33	**0.90**	1.11	1.43	1.45	**1.04**
II	**0.93**	1.06	1.38	0.94	**1.04**	1.42	1.52	1.11
MA	0.82	0.87	1.18	**0.81**	0.96	0.12	1.28	**0.92**
MO	**0.92**	0.93	1.29	0.93	**0.99**	1.05	1.32	1.02
PD	1.17	1.35	1.43	**0.71**	1.47	1.72	1.56	**0.92**

Note. It is shown in bold the best model for each species and experiment.

we compare (i) our proposed model (All); (ii) a model with no warping (NoWarp); (iii) a model without the cyclic component (NoCic); and (iv) a model where $\alpha_i = 0$ and $\beta_i = 0$, representing a model without alignment (NoAl). The most natural way of performing cross-validation in this context would be to use the trained model to predict an entirely new spectrogram. The problem with this approach is due to the presence of the call-specific parameters μ_i, τ_i, β_i, and α_i, about which we do not make any distributional assumptions beyond a prior distribution. We therefore proceed with two different cross-validation experiments. In the first one, referred to as CV1, for each species and spectrogram, we randomly select 5% of the time-frequency points and remove them from the data. We then fit the model using only the remaining 95%. Subsequently, we compute the CRPS (Gneiting & Raftery, 2007; Gneiting et al., 2008), a proper scoring rule commonly used to evaluate predictive performance using the held-out sample. The CRPS generalizes the absolute error to probabilistic forecasts and is widely used to assess both the calibration and sharpness of predictive distributions. In our case, it provides a quantitative evaluation of how well the predictive distribution recovers the held-out samples. In the second cross-validation experiment (CV2), we fit the model using the entire dataset, which, we remind the reader, comprises the 100 longest calls for each species, and assess its predictive performance on the longest spectrogram that was not included in model fitting. Since the parameters μ_i, τ_i, α_i, and β_i are call-specific and their values are required to obtain posterior samples for the held-out spectrogram, at each iteration, an index i is sampled from the set 1, 2, …, 100, and the corresponding μ_i, τ_i, α_i, and β_i are then used. Although this is not a fully rigorous way of sampling the parameter values, it is the most practical option available without altering the model formulation. The CRPS values for both experiments are show in Table 2.

For CV1, in 5 out of the 8 species, our proposed model shows the lowest CRPS. For species FU and MA, the NoAl model has a slightly lower CRPS, but our model's CRPS is very close to it. Interestingly, removing the cyclic component results in a worsened CRPS for all species, with a significant increase. Conversely, alignment and warping generally produce smaller CRPS values. However, since the differences are not substantial, in future applications or when data are not highly informative, it may be better to estimate a model without these components. For species PD, the best model is the one without alignment. This can be attributed to the short observations present in this dataset, which do not allow for alignment. In CV2, with the exception of species EC, where the model without the warping function has a lower CRPS, the results are in line with CV1, although the values are consistently higher, as expected.

5.4 Benchmarking against existing technologies

As a further test, we compare our dissimilarity matrix and tree with those obtained using a simple prealignment of the data. To this aim, we apply the preprocessing procedure used in Pigoli et al. (2018), whose code is available in its online supplementary material, to align the spectrograms on a common time grid, conventionally set to 100 points in the interval [0,1]. The first step

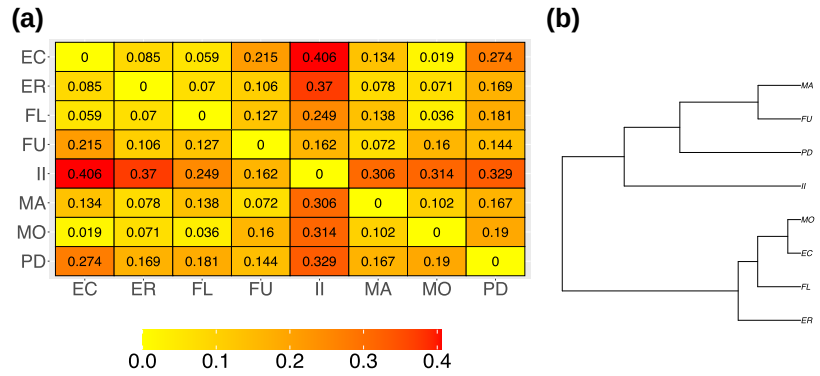

Figure 9. (a) Matrix containing the quadratic distances, for each pair of species, between the sample means of the prealigned spectrograms obtained using the code from Pigoli et al. (2018), and (b) the associated tree computed from the distance matrix. (a) Quadratic distance, (b) Tree.

of the procedure subtracts the average value from each spectrogram. Then, for each i, the code produces an aligned version of the data on a fixed grid of time-frequency points, using the same frequency points as the original data, allowing the computation of a sample mean over the spectrograms. We then take the mean at each time-frequency point, which serves as a frequentist nonparametric estimate of a quantity comparable to our latent sound; these estimates are shown in online supplementary material Figure 5 of Appendix F, while the associated distance matrix and tree are presented in Figure 9. Although discrepancies between the tree presented in Section 5.2 and DNA data were already observed in our results, Figure 9 reveals even greater differences. In particular, the tree places *Indri* as a sister group to all other species, while grouping *PD*, *MA*, and *FU* into a separate branch. Moreover, the resulting tree appears to be less consistent with the DNA-based and communication-based relationships previously described within the Lemuridae (Kumar & Foufoula-Georgiou, 1997; Macedonia & Stanger, 1994). Focusing specifically on the position of *II*, we observe that it shows greater distances from the other species than those measured among the Lemuridae.

6 Conclusion and future development

This work has presented a spatial-temporal model for bioacoustic data with nonstationary temporal patterns and periodic artifacts. The model combines novel ideas with several techniques to describe the complex structures that are intrinsic to bioacoustic data, and to handle the size of the available dataset. The model was applied to a motivating example, where a large dataset of recorded spectrograms was analysed. From the model output, we were able to estimate a representative sound for each species, which we then compared in terms of quadratic distance and distance-based tree. We also performed cross-validation to evaluate if all the model components were necessary to describe the data. Although this was not always true, with the exception of one species, our model was always the best or very close to the best.

We are currently working on the application of this methodology to a larger dataset comprising different calls. This will provide a better description of the distance between species in relation with a distance-based tree. We are also developing a model that can be used to assign a species label to newly recorded sounds, which, in this case, takes advantage of the sound length that was not utilized in this study.

Although, in principle, the methods can be applied to human speech, we believe that, given its more complex structure, including different languages, large vocabularies, and accents, the model would require a more complex covariance structure to capture the variability of human speech. Moreover, animal vocalizations are much shorter in duration than even a single sentence in human language, which would prohibitively increase the computational effort required to analyse the data. We believe that the model proposed by Pigoli et al. (2018) is more suitable for this task.

Conflicts of interest: None declared.

Funding

The work of G.M. has been partially carried out within the FAIR-Future Artificial Intelligence Research Foundation and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3-D.D.1555 11/10/2022, PE00000013).

Data availability

The data and a Julia package implementing the MCMC algorithm are available at https://github.com/GianlucaMastrantonio/Bayesian_inference_for_Latent_Spectral_Shapes.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Bezanson J., Edelman A., Karpinski S., & Shah V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Boersma P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9), 341–345.

Bryner D., & Srivastava A. (2022). Shape analysis of functional data with elastic partial matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9589–9602. https://doi.org/10.1109/TPAMI.2021.3130535

Datta A., Banerjee S., Finley A. O., & Gelfand A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800–812. PMID: 29720777. https://doi.org/10.1080/01621459.2015.1044091

Datta A., Banerjee S., Finley A. O., Hamm N. A. S., & Schaap M. (2016). Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, *10*(3), 1286–1316. https://doi.org/10.1214/16-AOAS931

De Gregorio C., Valente D., Raimondi T., Torti V., Miaretsoa L., Friard O., Giacoma C., Ravignani A., & Gamba M. (2021). Categorical rhythms in a singing primate. *Current Biology*, *31*(20), R1379–R1380. https://doi.org/10.1016/j.cub.2021.09.032

DelPero M., Pozzi L., & Masters J. C. (2006). A composite molecular phylogeny of living lemuroid primates. *Folia Primatologica*, *77*(6), 434–445. https://doi.org/10.1159/000095390

Gamba M., Friard O., & Giacoma C. (2012). Vocal tract morphology determines species-specific features in vocal signals of lemurs (Eulemur). *International Journal of Primatology*, *33*(6), 1453–1466. https://doi.org/10.1007/s10764-012-9635-y

Gamba M., & Giacoma C. (2007). Quantitative acoustic analysis of the vocal repertoire of the crowned lemur. *Ethology Ecology & Evolution*, *19*(4), 323–343. https://doi.org/10.1080/08927014.2007.9522555

Gamba M., & Giacoma C. (2008). Subspecific divergence in the black Lemur's low-pitched vocalizations. *The Open Acoustics Journal*, *1*(1), 49–53. https://doi.org/10.2174/1874837600801010049

Gamba M., Torti V., Estienne V., Randrianarison R. M., Valente D., Rovara P., Bonadonna G., Friard O., & Giacoma C. (2016). The indris have got rhythm! timing and pitch variation of a primate song examined between sexes and age classes. *Frontiers in Neuroscience*, *10*(249), 1–12. https://doi.org/10.3389/fnins.2016.00249

Gelfand A., Diggle P., Fuentes M., & Guttorp P. (2010). *Handbook of spatial statistics*. Chapman and Hall.

Gneiting T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, *97*(458), 590–600. https://doi.org/10.1198/016214502760047113

Gneiting T., & Raftery A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting T., Stanberry L. I., Grimit E. P., Held L., & Johnson N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, *17*(2), 211–235. https://doi.org/10.1007/s11749-008-0114-x

Graps A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, *2*(2), 50–61. https://doi.org/10.1109/99.388960

Jona Lasinio G., Mastrantonio G., & Pollice A. (2013). Discussing the "big n problem". *Statistical Methods and Applications*, *22*(1), 97–112. https://doi.org/10.1007/s10260-012-0207-2

Kershenbaum A., Blumstein D. T., Roch M. A., Akçay Ç., Backus G., Bee M. A., Bohn K., Cao Y., Carter G., Cäsar C., Coen M., DeRuiter S. L., Doyle L., Edelman S., Ferrer-i-Cancho R., Freeberg T. M., Garland

E. C., Gustison M., Harley H. E., …Zamora-Gutierrez V., *et al.* (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Review*, *91*(1), 13–52. https://doi.org/10.1111/brv.2016.91.issue-1

Kumar P., & Foufoula-Georgiou E. (1997). Wavelet analysis for geophysical applications. *Reviews of Geophysics*, *35*(4), 385–412. https://doi.org/10.1029/97RG00427

Ma Y., Zhou X., & Wu W. (2024). A stochastic process representation for time warping functions. *Computational Statistics and Data Analysis*, *194*, 107941. https://doi.org/10.1016/j.csda.2024.107941

Macedonia J. M., & Stanger K. F. (1994). Phylogeny of the Lemuridae revisited: Evidence from communication signals. *Folia Primatologica*, *63*(1), 1–43. https://doi.org/10.1159/000156787

Maretti G., Sorrentino V., Finomana A., Gamba M., & Giacoma C. (2010). Not just a pretty song: An overview of the vocal repertoire of indri indri. *Journal of Anthropological Sciences*, *88*, 151–165. PMID: 20834055.

Mastrantonio G., Jona Lasinio G., & Gelfand A. E. (2015). Spatio-temporal circular models with non-separable covariance structure. *TEST*, *25*(2), 331–350. https://doi.org/10.1007/s11749-015-0458-y

Mastrantonio G., Jona Lasinio G., Pollice A., Capotorti G., Teodonio L., Genova G., & Blasi C. (2017). A hierarchical multivariate spatio-temporal model for large clustered climate data with annual cycles. *Annals of Applied Statistics*, *13*(2), 797–823. 10.1214/18-AOAS1212

Paradis E. (2012). *Analysis of phylogenetics and evolution with R* (2nd ed.). Use R! Springer New York.

Pflüger F. J., & Fichtel C. (2012). On the function of redfronted lemur's close calls. *Animal Cognition*, *15*(5), 823–831. https://doi.org/10.1007/s10071-012-0507-9

Pigoli D., Hadjipantelis P. Z., Coleman J. S., & Aston J. A. D. (2018). The statistical analysis of acoustic phonetic data: Exploring differences between spoken romance languages. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *67*(5), 1103–1145. https://doi.org/10.1111/rssc.12258

Pozzi L., Gamba M., & Giacoma C. (2010). The use of artificial neural networks to classify primate vocalizations: A pilot study on black lemurs. *American Journal of Primatology*, *72*(4), 337–348. https://doi.org/10.1002/ajp.v72:4

Ramsay J. O., & Silverman B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer New York. Hardcover published: 08 June 2005; Softcover published: 10 November 2010; eBook published: 28 June 2006.

Sainburg T., Thielk M., & Gentner T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, *16*(10), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

Shirota S., & Gelfand A. E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *The Annals of Applied Statistics*, *11*(2), 481–503. https://doi.org/10.1214/16-AOAS960

Sperber A. L., Werner L. M., Kappeler P. M., & Fichtel C. (2017). Grunt to go - vocal coordinate of group movements in redfronted lemurs. *Ethology: Formerly Zeitschrift Fur Tierpsychologie*, *123*(12), 894–905. https://doi.org/10.1111/eth.2017.123.issue-12

Tavakoli S., Matteo B., Pigoli D., Chodroff E., Coleman J., Gubian M., Renwick M. E., & Sonderegger M. (2024). Statistics in phonetics. *Annual Review of Statistics and Its Application*. https://doi.org/10.1146/annurev-statistics-112723-034642

Tavakoli S., Pigoli D., Aston J. A. D., & Coleman J. S. (2019). A spatial modeling approach for linguistic object data: Analyzing dialect sound variations across Great Britain. *Journal of the American Statistical Association*, *114*(527), 1081–1096. https://doi.org/10.1080/01621459.2019.1607357

Turner R. E., & Sahani M. (2014). Time-frequency analysis as probabilistic inference. *IEEE Transactions on Signal Processing*, *62*(23), 6171–6183. https://doi.org/10.1109/TSP.78

Valente D., De Gregorio C., Torti V., Miaretsoa L., Friard O., Randrianarison R. M., Giacoma C., & Gamba M. (2019). Finding meanings in low dimensional structures: Stochastic neighbor embedding applied to the analysis of indri indri vocal repertoire. *Animals*, *9*(5), 243. https://doi.org/10.3390/ani9050243

Valente D., Miaretsoa L., Anania A., Costa F., Mascaro A., Raimondi T., De Gregorio C., Torti V., Friard O., Ratsimbazafy J., Giacoma C., & Gamba M. (2022). Comparative analysis of the vocal repertoires of the indri (Indri indri) and the diademed sifaka (Propithecus diadema). *International Journal of Primatology*, *43*(4), 733–751. https://doi.org/10.1007/s10764-022-00287-x

Vats D., & Knudson C. (2021). Revisiting the Gelman–Rubin diagnostic. *Statistical Science*, *36*(4), 518–529. https://doi.org/10.1214/20-STS812

Wilkinson S. P., & Davy S. K (2018). *phylogram: an R package for phylogenetic analysis with nested lists*.

Wilkinson W (2019). *Gaussian Process Modelling for Audio Signals* [Phd thesis]. Queen Mary University of London, Example City, CA.

Zimmermann E., & Primate Hearing and Communication. (2017). Evolutionary origins of primate vocal communication: Diversity, flexibility, and complexity of vocalizations in basal primates. In *Primate hearing and communication* (pp. 109–140). Springer International Publishing.