A Hierarchical Spatio-Temporal Model for Bio-Acoustic Analysis

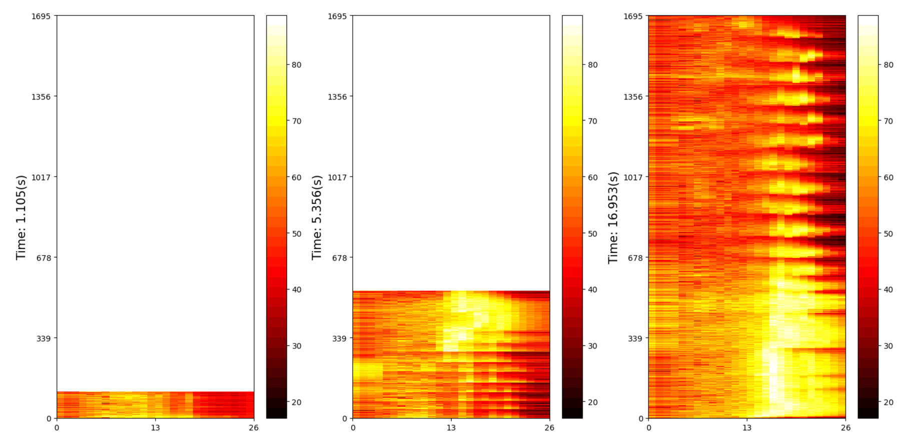
Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Marco Gamba, Daria Valente

Politecnico di Torino

Università degli Studi di Torino

Bio-acoustic Data

The available data is a set of vocal signals of multiple species of lemurs that are native to Madagascar. Below is the spectrogram representation of 3 recorded signals on a regular time-frequency grid. Each signal has a call-type label and a species label, which are characterized by the behaviour of the lemur during the vocal emission and the species to which the lemur belongs to. Note that: (i.) the noticeable distortions of the observed time domains with respect to each other; (ii). the oscillations along the time axis that arise from time discretization and signal reconstruction.



The goal of the proposed model is to obtain the representative acoustic structure of a behavioural call type of a single species

The Model

Each $i-$th recorded signal is assumed to be a realization of a two-dimensional process $\mathcal{Y}_i(t,h) \in \mathbb{R}$, where $(t,h) \in \mathbb{R}_{\geq 0} \times \mathbb{R}$, over an observed regular time-frequency grid that is denoted by $\mathcal{T}_i = \{0.01(k-1) \mid k = 1,\dots,n_{t,i}\}$ and $\mathcal{H} = \{0.23k + \log 63 \mid k = 1,\dots,n_h\}$ with $n_{t,i}$ being the number of time coordinates on $\mathcal{T}_i$ and n_h being the number of frequency coordinates on $\mathcal{H}$, respectively. Let $l_i = \max\{\mathcal{T}_i\}$ be the unique duration of the $i-$th recorded signal. The model is :

$$\begin{aligned}\mathcal{Y}_i(t,h) &= \mu_i + \mathcal{W}(t,h,\psi(t|\boldsymbol{\chi}_i)) + \epsilon_i(t,h)\\ \mathcal{W}(t,h,d) &\sim \mathrm{GP}(0, C(\cdot,\cdot,\cdot|\boldsymbol{\theta}))\\ \epsilon_i(t,h) &\sim \mathrm{GP}(0,\tau_i^2)\end{aligned}$$

where

- $\mu_i \in \mathbb{R}$ is the scalar mean sound intensity and τ_i^2 is the nugget effect
- $\mathcal{W}(t,h,d) \in \mathbb{R}$ is a latent Gaussian process of zero mean that specifies the latent spectral shape of the representative acoustic structure and is defined over a tri-dimensional space that consists of the real-time dimension $t \in \mathbb{R}_{\geq 0}$, the log-frequency dimension $h \in \mathbb{R}$ and the dimension for the warped time $d \in \mathbb{R}_{\geq 0}$
- $\psi(\cdot|\boldsymbol{\chi}_i) : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is the non-linear temporal stretching function dependent on a vector of parameters $\boldsymbol{\chi}_i$ that quantifies the time distortion of the $i-$th recorded signal with respect to the latent process
- $C(\cdot,\cdot,\cdot|\boldsymbol{\theta}) : \mathbb{R}^3_{\geq 0} \longrightarrow \mathbb{R}$ is the stationary cross-covariance function for the latent process

Non-linear temporal stretching & Circular time

The covariance function $C(\cdot,\cdot,\cdot|\boldsymbol{\theta})$ for the latent spectral shape is :

$$\begin{aligned}C\big(|t-t'|,|h-h'|,|d-d'|\ \big|\ \boldsymbol{\theta}\big) &= \sigma^2\lambda\, C_g\big(|h-h'|,|d-d'|\big) + \sigma^2(1-\lambda)\, C_c\big(|t-t'|,|h-h'|\big)\\ &= \frac{\sigma^2\lambda}{\phi_d|d-d'|+1}\exp\left(-\frac{\phi_h|h-h'|}{(\phi_d|d-d'|+1)^{\rho/2}}\right) + \frac{\sigma^2(1-\lambda)}{\phi_c\Delta_c(t,t'|\gamma)+1}\exp\left(-\frac{\phi_h|h-h'|}{(\phi_c\Delta_c(t,t'|\gamma)+1)^{\rho/2}}\right)\end{aligned}$$

where

- $d = \psi(t|\boldsymbol{\chi}_i) \in \mathbb{R}_{\geq 0}$ is the warped time coordinate given by the non-linear temporal-stretching function
- $\Delta_c(\cdot,\cdot|\gamma) : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is the periodic distance function that gives the circular distance between two real-time coordinates

Non-linear temporal-stretching: The first component $C_g(\cdot,\cdot)$ is the Gneiting correlation function that describes how the representative acoustic structure changes across the warped-time dimension and the frequency dimension. The relative relationship between the unique times of the data and the latent spectral shape of $\mathcal{W}(t,h,d)$ is described by the function $\psi(t|\boldsymbol{\chi}_i)$ which maps the observed time coordinates in the real-time dimension $t \in \mathcal{T}_i$ onto the dimension for the warped time $d \in \mathbb{R}_{\geq 0}$ that is given by :

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

where

- $\alpha_i \in \mathbb{R}_{\geq 0}$ is the misalignment parameter
- $\beta_i \in \mathbb{R}_{>0}$ is the parameter that linearly stretches the non-linearly warped real-time coordinate
- $\Omega(q|\boldsymbol{\xi}_i) = \dfrac{\Gamma(\exp\zeta_i + \exp\delta_i)}{\Gamma(\exp\zeta_i)\Gamma(\exp\delta_i)}\displaystyle\int_0^q x^{\exp\zeta_i - 1}(1-x)^{\exp\delta_i}\,dx$

  is the Beta cumulative distribution function with $\Omega(0|\boldsymbol{\xi}_i) = 0$ and $\Omega(1|\boldsymbol{\xi}_i) = 1$ that non-linearly warps the real-time coordinate within the $[0,1]$ scale

Circular distance: Due to the fact that the sampling artefacts do not exist in the dimension for the warped time, the effects of the periodicity of the real-time coordinates that arises from the presence of the artefacts needs to be accounted for separately in the real-time dimension. The second component $C_c(\cdot,\cdot)$ addresses this circular nature of the data by treating the circular distances between two real-time coordinates as two angles on a circle of circumference γ . The periodic distance between two real-time coordinates is :

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

The choice of the periodic distance function implies that the circular distance between any two real-time coordinates is restricted to a circular scale with period $\gamma/2$ such that $\Delta_c(t,t'|\gamma) \in [0,\gamma/2]\ \forall\, t,t'$.

Hierarchical model: The idea is to infer from the data the latent representative acoustic structure $\mathcal{W}(t,h,d)$ which quantifies the distortion as well as the periodicity in time as shown in the figures of data. Let $\boldsymbol{y}_i = \{y_{i,t,h}\}_{t\in\mathcal{T}_i, h\in\mathcal{H}}$ be the $i-$th recorded signal, $\boldsymbol{\theta} = \{\sigma,\lambda,\phi_d,\phi_h,\phi_c,\gamma,\rho\}$ and $\boldsymbol{\chi}_i = \{\alpha_i,\beta_i,\boldsymbol{\xi}_i\}$.