End-to-End Probabilistic Inference for Nonstationary Audio Analysis

William Wilkinson[1,2] Michael Riis Andersen[2,3] Josh Reiss[1] Dan Stowell[1] Arno Solin[2]

[1]Queen Mary University of London [2]Aalto University [3]Technical University of Denmark

Overview

- Utilise **prior knowledge** to learn from a single audio recording.
- **Time-frequency (TF) analysis** and **nonnegative matrix factorisation (NMF)** are ubiquitious in signal processing, but are always treated as disjoint, deterministic methods.
- We treat them probabilisticly in a joint **Gaussian process (GP)** model, the **GTF-NMF** [2].
- A **spectral mixture GP** (1) models covariance as a sum of quasi-periodic components [1]. We model the amplitude / variance with another GP projected through NMF-like mapping.
- Results in a nonstationary version of the spectral mixture GP.
- We formulate the **stochastic differential equation (SDE)** representation.
- Inference via expectation propagation (EP) in the Kalman filter. **Scales linearly** in the number of time steps.
- Applied to **multiple signal processing tasks** vs. Extended Kalman filter (EKF) and baseline methods.

Inference

- Construct **SDE form** of the GTF-NMF model:

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \mathbf{F}\mathbf{f}(t) + \mathbf{L}\mathbf{w}(t),$$
$$y_k = \mathcal{H}(\mathbf{f}(t_k)) + \sigma_y \varepsilon_k.$$

- Inference via assumed density filtering (ADF) in the nonlinear Kalman filter [4].
  - The trick is to use the Kalman predictions, $p(\mathbf{f}(t_k) \,|\, \mathbf{f}(t_{k-1}))$, as the cavity distributions.
- ADF does not perform well for this highly nonlinear likelihood model, so we implement **full EP**.
  - Must calculate true marginal update at each time step for nonlinear likelihood $\mathcal{H}(\cdot)$ via sigma-point integration – scales poorly with dimensionality.
- **Infinite-horizon** (steady state) GP solution reduces computation to **$\mathcal{O}(M^2 T)$ complexity and $\mathcal{O}(MT)$ memory** (T = time steps, M = state dimensionality).

Gaussian Time-Frequency + NMF

$$g_n(t) \sim \mathrm{GP}(0, \kappa_g^{(n)}(t,t')), \quad n = 1,2,\ldots,N,$$
$$z_d(t) \sim \mathrm{GP}(0, \kappa_z^{(d)}(t,t')), \quad d = 1,2,\ldots,D,$$

$g_n(t)$ are temporal NMF components and $z_d(t)$ the frequency channels. Kernel $\kappa_z^{(d)}$ is quasi-periodic. Amplitude kernel $\kappa_g^{(n)}$ typically from Matérn class.

The *likelihood* model:

$$y_k = \sum_d a_d(t_k)\, z_d(t_k) + \sigma_y \varepsilon_k,$$

for square amplitudes (the magnitude spectrogram):

$$a_d^2(t_k) = \sum_n W_{d,n}\, \psi(g_n(t_k)).$$

$W_{d,n}$ = NMF weights,
$\psi(\cdot)$ = softplus mapping to enforce positivity.

=

Nonstationary Spectral Mixture GP

Hierarchical model with hyper-GP prior

$$g_n(t) \sim \mathrm{GP}(0, \kappa_g^{(n)}(t,t'))$$

for each component with an NMF-like positivity mapping, $\alpha_d^2(t) = \sum_n W_{d,n}\, \psi(g_n(t))$, such that:

$$z(t) \sim \mathrm{GP}\left(0, \sum_{d=1}^{D} \alpha_d(t)\alpha_d(t') \cos(\omega_d(t-t'))\kappa_d(t,t')\right),$$
$$y_k = z(t_k) + \sigma_y \varepsilon_k.$$

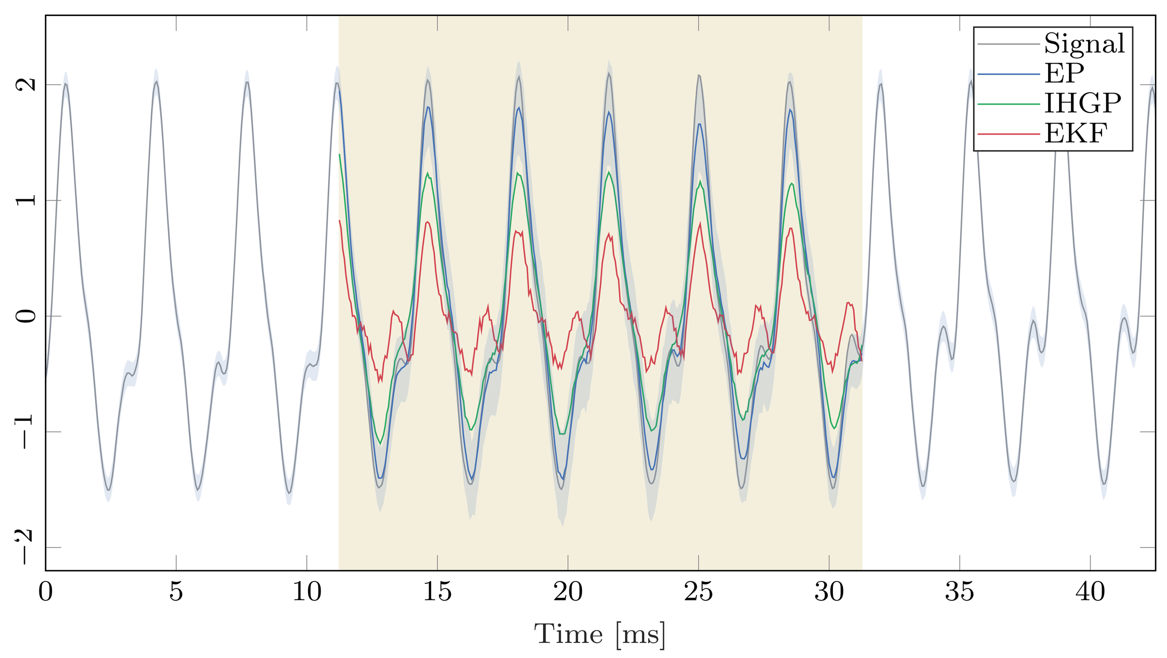
A GP model whose kernel is a sum of quasi-periodic functions with time-dependent variance [3].

ℓ_d = lengthscale,
ω_d = frequency.

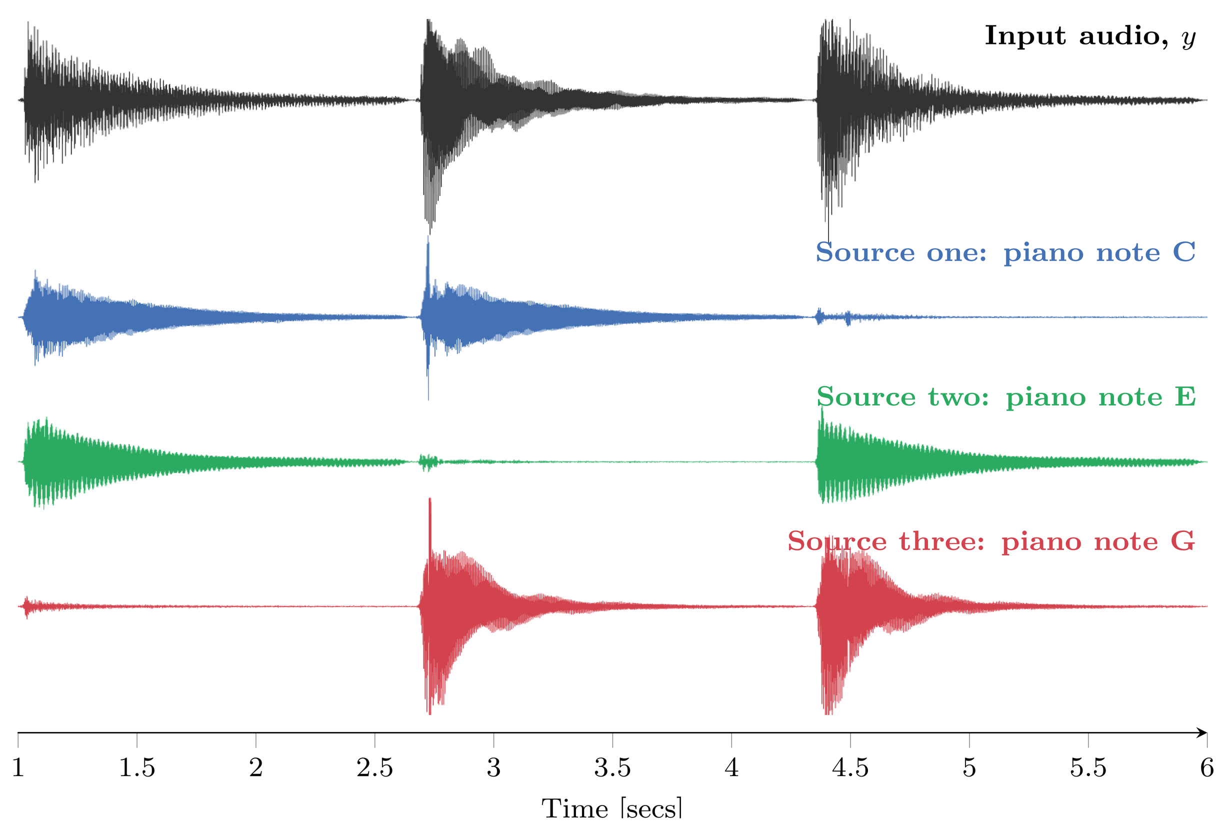
Results

- Same method applied to **missing data synthesis**, **denoising** and **source separation** without modification.
- Full EP consistently outperforms EKF, ADF and IHGP.
- However, memory saving in IHGP allows us to process audio signals of 6 seconds ($T = 96{,}000$, $M = 123$) which is not possible with other methods.
- Outperforms baseline on missing data synthesis, but less competitive on denoising.
- Still work to be done scaling to longer time series and larger models.

Missing Data Synthesis

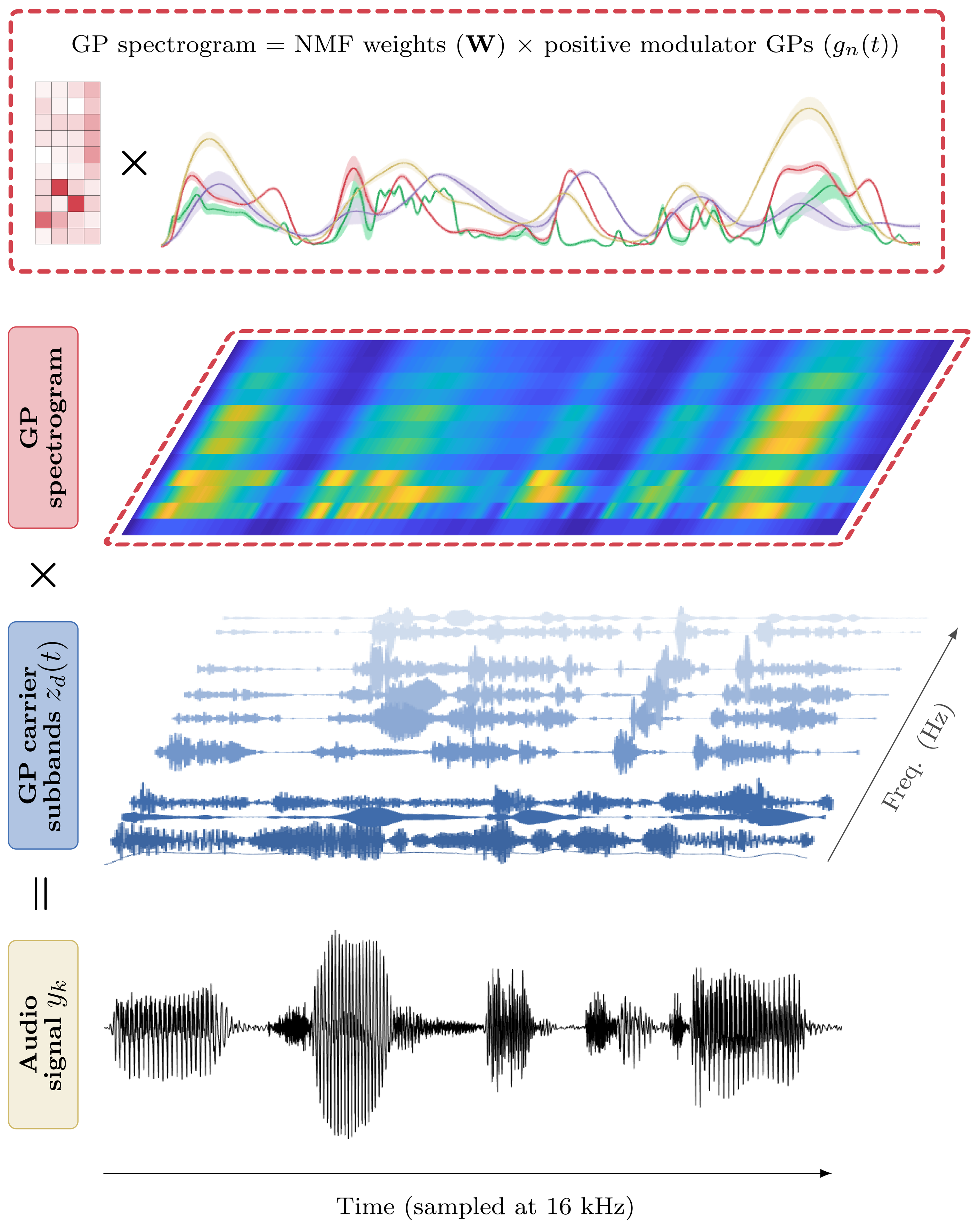


Source Separation



Code and resources:

https://github.com/AaltoML/nonstationary-audio-gp



Overview: Nonstationary modelling of audio data. Input (**bottom**) is a recording of female speech. We decompose the signal into Gaussian process carrier waveforms (blue block) multiplied by a spectrogram (red block). The spectrogram is learned from the data as a nonnegative matrix of weights times positive modulators (**top**).

References

[1] A. Wilson and R. Adams (2013). Gaussian process kernels for pattern discovery and extrapolation. *Proceedings of ICML.*

[2] R. E. Turner and M. Sahani (2014). Time-frequency analysis as probabilistic inference. *IEEE Trans. on Signal Processing.*

[3] S. Remes, M. Heinonen and S. Kaski (2017). Non-stationary spectral kernels. *Advances in NIPS.*

[4] H. Nickisch, Hannes, A. Solin and A. Grigorievskiy (2018). State space Gaussian processes with non-Gaussian likelihood *Proceedings of ICML.*