End-to-End Probabilistic Inference for Nonstationary Audio Analysis
(or how to apply Spectral Mixture GPs to audio)

William Wilkinson, Michael Riis Andersen, Josh Reiss, Dan Stowell, Arno Solin
June 12, 2019

Queen Mary University of London / Aalto University / Technical University of Denmark
We previously showed that a spectral mixture Gaussian process is equivalent to a probabilistic filter bank, i.e. a filter bank that adapts to the signal and can make predictions / generate new data.
We previously showed that a **spectral mixture Gaussian process is equivalent to a probabilistic filter bank**, i.e. a filter bank that adapts to the signal and can make predictions / generate new data.

**[Prior]** \( f(t) \sim \text{GP}\left(0, \sum_{d=1}^{D} \sigma_d^2 \exp(-|t - t'|/\ell_d) \cos(\omega_d (t - t'))\right), \)

**[Likelihood]** \( y_k = f(t_k) + \sigma_{y_k} \varepsilon_k, \)
The next step in the signal processing chain is often to analyse the dependencies in the spectrogram, with e.g. non-negative matrix factorisation (NMF).
End-to-End probabilistic time-frequency analysis

Audio signal $y_k$

Time (sampled at 16 kHz)
End-to-End probabilistic time-frequency analysis

Audio signal $y_k$ = GP carrier subbands $f_d(t)$ × GP spectrogram

Time (sampled at 16 kHz)
End-to-End probabilistic time-frequency analysis

$$y_k = \text{GP carrier subbands } f_d(t) \times \text{GP spectrogram}$$

$$\text{GP spectrogram} = \text{NMF weights } (W) \times \text{positive modulator GPs } (g_n(t))$$

Audio signal $y_k$

Time (sampled at 16 kHz)

Freq. (Hz)
The model

**GP prior:**

\[ f_d(t) \sim \text{GP}(0, \sigma_d^2 \exp(-|t - t'|/\ell_d) \cos(\omega_d (t - t'))), \quad d = 1, 2, \ldots, D, \]
\[ g_n(t) \sim \text{GP}(0, \kappa_g^{(n)}(t, t')), \quad n = 1, 2, \ldots, N, \]
The model

GP prior:

\[ f_d(t) \sim \text{GP}(0, \sigma_d^2 \exp(-|t - t'|/\ell_d) \cos(\omega_d (t - t')), \quad d = 1, 2, \ldots, D, \]
\[ g_n(t) \sim \text{GP}(0, \kappa_g^{(n)}(t, t')), \quad n = 1, 2, \ldots, N, \]

Likelihood model:

\[ y_k = \sum_d a_d(t_k) f_d(t_k) + \sigma_y \varepsilon_k, \]

for square amplitudes (the magnitude spectrogram):

\[ a_d^2(t_k) = \sum_n W_{d,n} \text{softplus}(g_n(t_k)), \]
The model

GP prior:

\[ f_d(t) \sim \text{GP}(0, \sigma_d^2 \exp(-|t - t'|/\ell_d) \cos(\omega_d (t - t')), \quad d = 1, 2, \ldots, D, \]
\[ g_n(t) \sim \text{GP}(0, \kappa^{(n)}_g(t, t')), \quad n = 1, 2, \ldots, N, \]

Likelihood model:

\[ y_k = \sum_d a_d(t_k) f_d(t_k) + \sigma_y \varepsilon_k, \]

for square amplitudes (the magnitude spectrogram):

\[ a^2_d(t_k) = \sum_n W_{d,n} \text{softplus}(g_n(t_k)), \]

This is a nonstationary spectral mixture GP
Inference

We show how to write the model as a stochastic differential equation:

\[
\frac{d\tilde{f}(t)}{dt} = \mathbf{F}\tilde{f}(t) + \mathbf{L}w(t),
\]

\[
y_k = \mathcal{H}(\tilde{f}(t_k)) + \sigma_y e_k,
\]

such that inference can proceed via Kalman filtering & smoothing.
We show how to write the model as a stochastic differential equation:

\[
\frac{d\tilde{f}(t)}{dt} = F\tilde{f}(t) + Lw(t),
\]

\[
y_k = \mathcal{H}(\tilde{f}(t_k)) + \sigma_y \varepsilon_k,
\]

such that inference can proceed via Kalman filtering & smoothing.

Usually the nonlinear \( \mathcal{H}(\cdot) \) is dealt with via linearisation (EKF), but we implement full Expectation Propagation (EP) in the Kalman smoother, and the infinite-horizon solution which scales as:

\[ \mathcal{O}(M^2 T) \]
The fully probabilistic model can, *without modification*, be applied to:

- Missing Data Synthesis
- Denoising
- Source Separation

---

**Figure 1:** An example of missing data imputation with the GTF-NMF model for each inference method with 20 iterations. Grey signal is the ground truth, a recording of a bamboo flute. The yellow shaded region indicates where the data is missing. Blue shaded area is the 95% confidence region for the EP method.

**Figure 1:** Denoising with various inference methods across five levels of corruption noise variance (0.01–0.5). y-axis is the signal-to-noise ratio of the recovered waveform. Mean values across 10 speech signals are shown. Shaded areas are standard error. SpecSub is the spectral subtraction baseline.

**Figure 1:** Infinite-horizon GP source separation example showing three piano notes (sources) recovered from a mixture signal (top), where two notes are played at a time in the original recording.

---

**Thanks for listening! Poster: 6:30pm Weds, Pacific Ballroom #217**

**Contact:** william.wilkinson@aalto.fi
The fully probabilistic model can, **without modification**, be applied to:

**Missing Data**

**Synthesis**

---

**Figure 1**: An example of missing data imputation with the GTF-NMF model for each inference method with 20 iterations. Grey signal is the ground truth, a recording of a bamboo flute. The yellow shaded region indicates where the data is missing. Blue shaded area is the 95% confidence region for the EP method.

---

**Figure 1**: Denoising with various inference methods across five levels of corruption noise variance (0.01–0.5). y-axis is the signal-to-noise ratio of the recovered waveform. Mean values across 10 speech signals are shown. Shaded areas are standard error. SpecSub is the spectral subtraction baseline.

---

**Figure 1**: Infinite-horizon GP source separation example showing three piano notes (sources) recovered from a mixture signal (top), where two notes are played at a time in the original recording.
Applications and Results

The fully probabilistic model can, **without modification**, be applied to:

**Missing Data**

**Synthesis**

**Denoising**

![Missing Data](image1)

![Denoising](image2)
Applications and Results

The fully probabilistic model can, without modification, be applied to:

**Missing Data**

**Synthesis**

**Denoising**

**Source Separation**

---

**Figure 1:** An example of missing data imputation with the GTF-NMF model for each inference method with 20 iterations. Grey signal is the ground truth, a recording of a bamboo flute. The yellow shaded region indicates where the data is missing. Blue shaded area is the 95% confidence region for the EP method.

**Figure 1:** Denoising with various inference methods across five levels of corruption noise variance (0.01–0.5). Y-axis is the signal-to-noise ratio of the recovered waveform. Mean values across 10 speech signals are shown. Shaded areas are standard error. SpecSub is the spectral subtraction baseline.

**Figure 1:** Infinite-horizon GP source separation example showing three piano notes (sources) recovered from a mixture signal (top), where two notes are played at a time in the original recording.
Applications and Results

The fully probabilistic model can, without modification, be applied to:

**Missing Data Synthesis**

**Denoising**

**Source Separation**

Thanks for listening! Poster: 6:30pm Weds, Pacific Ballroom #217

Contact: william.wilkinson@aalto.fi