Reduced-rank Spectral Mixtures Gaussian Processes for Probabilistic Time-frequency Representations

Anis Fradi, Khalid Daoudi

Journal Pre-proof

Reduced-rank spectral mixtures Gaussian processes for probabilistic time-frequency representations

Anis Fradi, Khalid Daoudi

PII: S0165-1684(23)00429-2
DOI: https://doi.org/10.1016/j.sigpro.2023.109355
Reference: SIGPRO 109355

To appear in: *Signal Processing*

Received date : 22 May 2023
Revised date : 30 November 2023
Accepted date : 5 December 2023

Please cite this article as: A. Fradi and K. Daoudi, Reduced-rank spectral mixtures Gaussian processes for probabilistic time-frequency representations, *Signal Processing* (2023), doi: https://doi.org/10.1016/j.sigpro.2023.109355.

Reduced-rank Spectral Mixtures Gaussian Processes for Probabilistic Time-frequency Representations

Anis Fradi*, Khalid Daoudi*

Abstract

Deterministic time-frequency representations are commonly used in signal processing, particularly in audio processing. Whilst presenting many potential advantages, their probabilistic counterparts are not widely used, essentially because of the computational load and the lack of clear interpretability of the different underlying models. However, using state space models, they have been shown recently to be equivalent to Spectral Mixtures Gaussian processes (SM-GP). This pioneer work unlocks this problem and opens the path for the development of tractable and interpretable probabilistic time-frequency analysis. In this paper, we propose a relatively simple yet a significant improvement of that work in terms of computational complexity, flexibility and practical application. To do so, we use a recent approach for covariance approximation to develop an algorithm for faster inference of SM-GP, while opting for a frequency-domain approach to hyperparameter learning. We illustrate the practical potential of our method using voiced speech data. We first show that key speech features can be accurately learned from the data. Second, we show that our method can yield better performances in denoising.

Keywords: probabilistic time-frequency analysis, Gaussian process, spectral mixtures Gaussian process, reduced-rank covariances, probabilistic filter banks.

1. Introduction

Time-Frequency (TF) analysis is in the heart of signal processing science and technology Cohen (1995). In many applications, such as audio processing, the signal is assumed to be locally stationary and the task is to uncover the (time-dependent) latent spectral structures of the signal. There exists several ways to perform such

*Corresponding author
Email address: anis.fradi@inria.fr (Anis Fradi)

an analysis, for instance the short-time Fourier transform (STFT), filter banks and wavelets, which are the most widely used tools in this field (Mallat, 2008). The problem is thus ill-posed as there exist a variety of possible representations but no principled way to infer on the "best" one. As stated in (Turner and Sahani, 2014): *there is no consensus on how to select the best time-frequency representation for a particular signal or task, nor are there robust algorithms for automatic (and potentially time-varying) signal-dependent adaptation of the representation.* The pioneer work (Turner and Sahani, 2014) proposed a new perspective on classical (deterministic) TF analysis by framing it as a probabilistic Bayesian inference problem. The observed signal is assumed to be composed of a finite superposition of unobserved band-limited components possibly corrupted by noise, with varying time-frequency concentration, according to a conditional probability distribution. The problem is then to make a Bayesian inference of the latent components from the observations by putting priors on these components. Under the assumption of Gaussian priors and noise, (Turner and Sahani, 2014) showed that several TF representations, such as STFT, spectrograms, filter banks and wavelets can be interpreted as a probabilistic inference in a Gaussian regression model. Besides providing a unified view to various TF methods, this Gaussian time-frequency (GTF) formulation has the main advantages of making natural to propagate uncertainty in the TF representation, to quantify noise and to handle missing/irregular observations. However, whilst this framework is appealing, it suffers from the significant computational complexity burden required to carry out the inference, because of the need to invert (generally) large signal covariances.

It has been clear that GTF and other approaches to probabilistic inference in TF analysis, such as Bayesian Spectrum Estimation (Qi et al., 2002) and the probabilistic phase vocoder (PPV) (Cemgil and Godsill, 2005), fall under the framework of Gaussian processes (GP) (Rasmussen and Williams, 2006). This indicates that there is a bridge between GP techniques developed in machine learning and signal processing algorithms. Such a bridge has been recently established in (Wilkinson et al., 2019; Wilkinson, 2019) based on the link between state space models (SSM) and GP inference (Solin and Särkkä, 2014), and using spectral mixtures Gaussian processes (SM-GP), first introduced in (Wilson and Adams, 2013) and then adapted in (Alvarado and Stowell, 2017). It was shown that despite being developed from a very different perspective, the PPV is a special case of the SM-GP. This in turn shows that the latter can be viewed as probabilistic filter banks. Thus, new probabilistic TF models can be drawn for which inference can be solved using classical Kalman filtering (Kalman, 1960), which scales linearly in time.

In this paper, we propose an improvement of the method proposed in (Wilkinson

et al., 2019; Wilkinson, 2019). Whilst the inference of the latter scales linearly in time, it scales cubicly in state dimension which can become rapidly intractable in practice. We propose to use a recent method for the approximation of covariance functions (Solin and Särkkä, 2020) in order to reduce the computational complexity of inference. This approximation relies on series expansion of the covariance function in terms of eigen-decomposition of the Laplace operator. The resulting approximation allows the proposed model to adapt the signal so that the eigenvalues can be expressed as simple functions of the spectral density associated with the covariance function (Akhiezer and Glazman, 2013). We show that computing the posterior mean of a fully Bayesian model has a lower linear-time complexity compared to the methods mentioned above. Besides the gain in computational complexity, this approach has the advantage of yielding an inference algorithm which is simple to understand and implement, as compared to the SSM approach which involves stochastic differential equations and dynamical systems. Moreover, it allows to consider any SM kernel, not only Matérn-SM ones.

Our approach falls within the framework of kenel approximation which has been largely studied in GP. In particular, Yin et al. (2020) introduced a sequential majorization-minimization technique for optimizing hyperparameters based on the grid spectral mixture kernel. This method yields better local minima, albeit with quadratic computational complexity in time. More recently, (Suwandi et al., 2022) harnessed the power of a multicore computing environment to optimize kernel hyperparameters in a distributed fashion. They also introduced a doubly distributed learning algorithm based on the alternating direction method of multipliers (ADMM), with an overall computational complexity scaling cubically with time. In contrast, our approach scales linearly in time, making it advantageous for applications with large sample size, such as audio processing.

To illustrate the practical potential of our method we carry out experiments where we first show, using voiced speech, that the frequency-domain hyperparameter learning technique proposed in (Turner and Sahani, 2014) is effective, as it learns important spectral cues such as the fundamental and formant frequencies. Second, denoising experiments show that the proposed method outperforms not only the one of (Wilkinson et al., 2019; Wilkinson, 2019) but also, interestingly, the "optimal" GTF inference (with full inversion of the signal covariance).

The rest of the paper is organized as follows. Section 2 provides a brief background on Gaussian processes regression. Section 3 presents the spectral mixtures Gaussian processes, their link to PPV and the time-frequency approach to hyperparameter learning. Section 4 gives a brief presentation of covariance approximation based on the expansion of Laplace operator. Our proposed method for probabilistic time-

frequency analysis based on the approximation of SM-GP is developed in Section 5. Experiments using speech data are presented and discussed in Section 6. We finally conclude the paper in Section 7.

2. Gaussian processes regression

A Gaussian process (GP) is a probabilistic model that is used to make predictions about unknown functions (Williams and Barber, 1998). It is a collection of random variables, any finite number of which have a joint Gaussian distribution. In other words, it defines a distribution over functions, where each function is a probability distribution over possible values of the function at any input. One advantage of GPs is that they provide a measure of uncertainty for each prediction. This can be useful in cases where it is important to know how confident we are in our predictions. A GP is completely specified by a mean function $m(t)$ and a covariance $k(t,t')$:

$$\begin{aligned} m(t) &= \mathbb{E}(f(t)); \quad t \in \mathbb{R} \\ k(t,t') &= \mathbb{E}((f(t)-m(t))(f(t')-m(t'))); \quad t,t' \in \mathbb{R} \end{aligned}$$

The mean function is usually assumed to be zero for simplicity $m \equiv 0$ so that the GP is said to be a zero mean (centered) GP.

GP regression is a type of probabilistic regression that uses a GP to model the relationship between input variables and output variables. The goal is to learn a function that maps input variables t_i to output variables y_i, while accounting for uncertainty in the prediction according to a training set $(\mathbf{t},\mathbf{y}) = (t_i, y_i)_{i=1}^N$. In its canonical form, the GP regression considers a nonparametric function f assumed to be a realization of a stochastic GP prior whereas the likelihood term holds from observations corrupted by a noise term $\epsilon_i \sim \mathcal{N}(0,\gamma^2)$:

$$\begin{aligned} f(t) &\sim \mathcal{GP}(m(t), k(t,t')) \\ y_i &= f(t_i) + \epsilon_i \end{aligned}$$

Based on the conjugate property of the Gaussian distribution, the posterior distribution over $\mathbf{f} = f(\mathbf{t})$ is also Gaussian: $p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$. From Bayes' rule, the posterior mean and covariance are given by:

$$\begin{aligned} \boldsymbol{\mu}_{\text{post}} &= \mathbf{K}(\mathbf{K}+\gamma^2\mathbf{I}_N)^{-1}\mathbf{y} \\ \boldsymbol{\Sigma}_{\text{post}} &= \mathbf{K} - \mathbf{K}(\mathbf{K}+\gamma^2\mathbf{I}_N)^{-1}\mathbf{K} \end{aligned}$$

where $\mathbf{K}$ is the $N \times N$ covariance matrix with entries $\mathbf{K}_{i,j} = k(t_i,t_j)$ and $\mathbf{I}_N$ refers

to the $N \times N$ identity matrix. The covariance function $k(t, t')$ usually depends on a set of hyperparameters θ_k that need to be fitted from the training data.

In general, the most widely used covariances (in theory and practice) are stationary covariances (or kernels), for which $k(t, t')$ is a function of $\tau = t - t'$. For such covariances, the Bochner's theorem (Snelson and Ghahramani, 2005) states that they can be equivalently represented in terms of their spectral density, giving rise to the Fourier duality of covariance and spectral density:

$$\begin{aligned} k(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} S(w) e^{iw\tau} dw \\ S(w) &= \int_{-\infty}^{+\infty} k(\tau) e^{-iw\tau} d\tau \end{aligned}$$

where $S(.)$ is called the spectral density associated to the kernel k.

3. Spectral mixtures Gaussian processes (SM-GP)

3.1. Background

The choice of the covariance/kernel function has a strong influence on the modeling and inference performance of a Gaussian process. However, standard canonical kernels, such as the squared exponential (SE) and Matérn kernels, are mostly used regardless of the task specificities. For instance, standard periodic kernels were used in (Liutkus et al., 2011) to model audio signals. Such kernels are generally not realistic in practice because most of audio signals, speech in particular, are not pure harmonic signals. To meet with task/application specificities, standard kernels are sometimes combined and tuned. When doing so, strong and task-specific constraints must be imposed on the combination to keep interpretability and to avoid overfitting. In fact, generally speaking, in most applications it is unclear what kernel functions should be used. Motivated by this fact, (Wilson and Adams, 2013) introduced a class of kernels which can approximate any kernel, as long as it has a spectral density. (Wilson and Adams, 2013) started by noticing that the spectral density of SE kernels, and mixtures of SE kernels, correspond only to Gaussian spectral densities centered on the origin. This naturally led to the consideration of mixture of Gaussians as they are dense in the set of probability distributions (Plataniotis, 2000). The resulting kernel, the spectral mixture kernel, was showed in (Wilson and Adams,

2013) to be:

$$c(t,t') = \sum_{d=1}^{D} c^{(d)}(t,t') \quad t,t' \in \mathbb{R}$$

where $c^{(d)}(t,t') = cos(w_d(t-t')) \times k^{(d)}(t,t')$, w_d can be interpreted as a center frequency and $k^{(d)}$ is the SE kernel:

$$k^{(d)}(t,t') \;=\; \sigma_d^2 \exp(-\frac{(t-t')^2}{2l_d^2})$$

(Wilson and Adams, 2013) showed that any stationary kernel can be approximated by such kernels at arbitrary precision. The strength of these kernels is that, while they allow a large expressiveness/modelling flexibility, they have a simple closed form that leads to straightforward analytic inference. Later, (Alvarado and Stowell, 2017) adapted SE spectral mixture kernels by using Lorentzian mixtures instead of Gaussian, resulting in the Matérn-$\frac{1}{2}$ spectral mixture kernel where each canonical component is:

$$k^{(d)}(t,t') \;=\; \sigma_d^2 \exp(-\frac{|t-t'|}{l_d})$$

where σ_d^2 and l_d refers to the variance and length-scale parameters, respectively. This adaptation was then used in (Alvarado et al., 2019) for source separation.

While these models are appealing, inference remains costly because of the need to invert covariance matrices. On the other hand, (Solin and Särkkä, 2014) showed that quasi-periodic covariances can be formulated as state space models, which can be solved using classical Kalman filtering. This has the strong advantage of reducing the inference complexity from cubic to linear time. Using the link between quasi-periodic covariances and state space models, (Wilkinson et al., 2019; Wilkinson, 2019) showed that this Matérn-$\frac{1}{2}$ SM-GP model is equivalent to the probabilistic phase vocoder (PPV) which was defined in (Cemgil and Godsill, 2005) as a complex time-frequency first-order auto-regressive process:

$$\begin{aligned} s_{d,t_i} &= \psi_d e^{i\Omega_d} s_{d,t_{i-1}} + \rho_d \xi_{d,t_i} \\ y_i &= \sum_{d=1}^{D} \mathrm{Re}[s_{d,t_i}] + \epsilon_i \end{aligned}$$

where $s_{d,t_i} \in \mathbb{C}$ is a complex phasor, $\xi_{d,t_i} \sim \mathcal{CN}(0,1)$ is a complex Gaussian noise and $\epsilon_i \sim \mathcal{N}(0, \sigma_{y_i}^2)$. The parameters ψ_d and ρ_d^2 represent the process and noise variances, respectively, whereas Ω_d is the instantaneous angular frequency. If Δt denotes the time step size: $\Delta_t = t_i - t_{i-1}$, the parameters are identified as: $\psi_d = \exp(-\frac{\Delta t}{l_d})$, $\rho_d = \sigma_d^2(1 - \exp(-\frac{2\Delta t}{l_d}))$ and $\Omega_d = w_d$. This shows that w_d determines the center frequency, the length-scale l_d characterizes the filter bandwidth and the variance σ_d^2 represents the scale.

Recognizing that existing state-of-the-art methods for probabilistic TF analysis can be viewed as modifications of the PPV, this implies that SM-GP can be interpreted as probabilistic filter banks in which the length-scales determine the bandwidth of the filters. In particular, the establishment of this link allows the consideration of higher-order Markov extensions of PPV by simply using SM-GP models with smoother Matérn kernels than the exponential kernel:

$$k^{(d)}(t,t') = \sigma_d^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \big(\sqrt{2\nu}\frac{|t-t'|}{l_d}\big)^{\nu} K_\nu\big(\sqrt{2\nu}\frac{|t-t'|}{l_d}\big)$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind and σ_d^2, l_d and ν_d refer to the variance, the length-scale and smoothness hyperparameters, respectively. For instance, considering Matérn-$\frac{3}{2}$ ($\nu = 3/2$) and Matérn-$\frac{5}{2}$ ($\nu = 5/2$) correspond to a second-order AR(2) and third order AR(3) PPV, respectively. Such higher-order models are suitable to model signals with slowly varying instantaneous frequencies, a characteristic of many real-world signals such as speech.

Inference in these models can be performed via a Kalman smoothing whose computational complexity scales linearly in the sample size N but cubicly increases with the state space dimension: $O(N\eta^3)$. Given a Matérn-ν covariance, the state space dimension is $\eta = (2\nu + 1)D$. Thus, inference can still be a burden in high-order models. We also mention that the SE spectral mixture ($\nu \to \infty$) cannot be (exactly) considered in this formulation.

3.2. Hyperparameter learning

The parameters of the SM time-frequency representation are $\theta_{\text{sm}} = (\theta_k^d, w_d)_{d=1}^D$, where θ_k^d are the parameters of the canonical covariance $k^{(d)}(t,t')$ and w_d are the filter bank center frequencies. One way to learn the SM-GP hyperparameters $\theta = (\theta_{\text{sm}}, \gamma^2)$, including the noise variance, is to minimize the energy function which is equivalent

to maximize the marginal likelihood:

$$p(\mathbf{y}|\theta) = \mathcal{N}(0, \sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2 \mathbf{I}_N) \tag{1}$$

where $\mathbf{K}_{i,j}^{(d)} = k^{(d)}(t_i, t_j)$, $\mathbf{O}_{i,j}^{(d)} = cos(w_d(t_i - t_j))$ for $i, j = 1, \ldots, N$ and $\odot$ denotes the component-wise multiplication. Optimization in the time domain is cumbersome because it requires inversion of large covariance matrices. To overcome this issue, (Turner and Sahani, 2014) proposed a frequency-domain method by considering the Fourier transform of the marginal likelihood. Let $\tilde{\mathbf{y}} = (\tilde{y}_j)_{j=1}^N$ be the signal spectrum and $S^{(d)}(.)$ the d-th spectral density associated to the covariance $k^{(d)}(.)$. Therefore, the d-th spectral density associated to $c^{(d)}(.)$ is $\frac{1}{2}[S^{(d)}(w - w_d) + S^{(d)}(w + w_d)]$. Thus, the model' spectral density is $\gamma_{y,i}(\theta) = \frac{1}{2} \sum_{d=1}^{D} [S^{(d)}(w_i - w_d) + S^{(d)}(w_i + w_d)] + N\gamma^2$ where $w_i = 2\pi \frac{t_i}{N}$ is the centre frequency of bin i. Finally, the log marginal likelihood in the frequency domain, also called the "*whittle*" log likelihood, is given by:

$$\log p(\mathbf{y}|\theta) = -\frac{1}{2} \sum_{i=1}^{N} \Big(\log(\gamma_{y,i}(\theta)) + \frac{|\tilde{y}_i|^2}{\gamma_{y,i}(\theta)} \Big) - \frac{N}{2} \log(2\pi) \tag{2}$$

The details of calculations leading to this expression are given in Appendix A. The model hyperparameters can then be fitted to minimize the dissimilarity between the model spectrum and the signal power spectrum by, for instance, a gradient descent algorithm. A significant advantage of this method is that the log marginal likelihood evaluation costs only $O(N+D)$, if the spectral density is known and analytic (such as for SE and Matérn kernels). This is because the computation of $\log p(\mathbf{y}|\theta)$ is primarily dominated by the necessity to invert the vector $(\gamma_{y,1}(\theta), \ldots, \gamma_{y,N}(\theta))^T$, where each element represents the sum of D terms.

However, a practical issue with the learning algorithm in the frequency domain is that the signal power spectrum $|\tilde{y}_i|^2$ may have significant variations which can yield numerous local optima during the optimization process. This issue can be handled, for instance, by replacing the spectrum by the Welch's periodogram (Welch, 1967) computed from multiple data segments. This yields a smoothed spectrum and helps reducing the number of local optima during optimization, with its effect gradually diminishing as the number of data segments used is reduced, ultimately allowing the process to better fit the original spectrum.

As we will see in the experiments, this approach is not only significantly more

efficient than the time-domain approach but also significantly more accurate.

4. Laplace operator for covariance approximation

The Mercer's theorem states that, for a zero mean GP with covariance $k(t, t')$, there exists an orthogonal basis (ϕ_l) in $\mathbb{L}^2$ and eigenvalues λ_l such that:

$$k(t, t') = \sum_{l=1}^{\infty} \lambda_l \phi_l(t) \phi_l(t'); \quad t, t' \in \mathbb{R}$$

The eigenvalues λ_l and eigenfunctions ϕ_l can be obtained from the integral operator and the solution is provided by the Fredholm integral equation Ghanem and Spanos (1991). For stationary covariance functions (kernels), (Solin and Särkkä, 2020) proposed an approximation of the covariance based on expansion of the Laplace operator in a compact subset. In the univariate case, given a compact interval $[-T, T] \subset \mathbb{R}$, the approximation is given as:

$$k(t, t') \approx \sum_{l=1}^{\infty} S(\sqrt{\gamma_l}) \psi_l(t) \psi_l(t'); \quad t, t' \in [-T, T]$$

where $S(.)$ is the spectral density of the kernel $k(.)$ and the eigenfunctions and eigenvalues are:

$$\psi_l(t) = \frac{1}{\sqrt{T}} \sin(l\pi \frac{(t+T)}{2T}) \text{ and } \gamma_l = (\frac{l\pi}{2T})^2$$

An advantage of this approach is that the kernel can be approximated by point-wise evaluations of the spectral density and sinusoidal eigenfunctions which are independent of the hyperparameters of the kernel. This allows fast inference and learning when the spectral density is known analytically such as for the SE:

$$S^{\text{SE}}(w) = \sigma^2 \sqrt{2\pi l^2} \exp(-2\pi^2 l^2 w^2)$$

and the Matérn-ν kernels:

$$S^{\text{Matérn}}(w) = 2\sigma^2 \pi^{1/2} \frac{\Gamma(2\nu)}{\Gamma(\nu + 1/2)} \left(\frac{\sqrt{2\nu}}{l} \right)^{2\nu} \left((\frac{\sqrt{2\nu}}{l})^2 + w^2 \right)^{-\nu - 1/2}$$

Moreover, (Solin and Särkkä, 2020) showed that the M-order truncation approximation:

$$k_M(t,t') = \sum_{l=1}^{M} S(\sqrt{\gamma_l})\psi_l(t)\psi_l(t') \tag{3}$$

converges uniformly to $k(.,.)$:

$$\lim_{T\to\infty}\Big(\lim_{M\to\infty} k_M(t,t')\Big) = k(t,t')$$

The convergence behavior is notably influenced by both the eigenvalues $S(\sqrt{\gamma_l})$ and the differentiability characteristics of the covariance function $k(.,.)$. In the work of (Takhanov, 2023) it was demonstrated that the rate of uniform convergence varies based on the eigenvalue decay rate. Specifically, for a covariance function $k(.,.)$ that is twice as differentiable (2β times), the truncated covariance $k_M(.,.)$ approximates $k(.,.)$ with an error of $O\big((\sum_{l=M+1}^{\infty} S(\sqrt{\gamma_l}))^{\frac{\beta}{\beta+1}}\big)$. For covariance functions that are infinitely differentiable, the error rate is even more favorable, falling within $O\big((\sum_{l=M+1}^{\infty} S(\sqrt{\gamma_l}))^{1-\epsilon}\big)$ for any $\epsilon > 0$. In summary, smoother covariance functions tend to exhibit faster convergence rates, whereas less smooth or non-differentiable covariance functions may manifest slower or no convergence.

If $\mathbf{\Psi} = \begin{pmatrix} \psi_1(t_1) & \dots & \psi_M(t_1) \\ \vdots & \dots & \vdots \\ \psi_1(t_N) & \dots & \psi_M(t_N) \end{pmatrix}$ and $\mathbf{S} = \mathrm{diag}(S(\sqrt{\gamma_1}), \dots, S(\sqrt{\gamma_M}))$ then the approximate log marginal likelihood adapted to the truncated covariance function is:

$$\log p(\mathbf{y}|\theta) \approx -\frac{1}{2}\log|\mathbf{\Psi S \Psi}^T + \gamma^2\mathbf{I}_N| - \frac{1}{2}\mathbf{y}^T(\mathbf{\Psi S \Psi}^T + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi)$$

Figure 1 shows an example of a mixture of five Matérn covariance functions ($D = 5$) and their approximations for different truncation order M (we set $T = 2\pi$). It can be observed that, for the SE mixture, an order as low as $M = 12$ yields a very good approximation, while for Matérn mixtures a higher M is required. In the following, we will use this M-order approximation to perform inference in SM-GP.

5. Inference in SM-GP using the reduced-rank covariance approximation

Consider the probabilistic time-frequency representation given by the SM-GP model:

$$\begin{aligned} f^{(d)}(t) &\sim \mathcal{GP}(0, c^{(d)}(t,t')) \\ y_i &= \sum_{d=1}^{D} f^{(d)}(t_i) + \epsilon_i; \quad i = 1, ..., N \end{aligned}$$

where the covariance function of d-th component is:

$$c^{(d)}(t,t') = cos(w_d(t-t')) \times k^{(d)}(t,t'); \quad t, t' \in \mathbb{R} \tag{4}$$

and where $k^{(d)}$ is a canonical kernel with a spectral density $S^{(d)}$. Under the assumption that all the subband components $f^{(d)}$ are independent, the covariance $c(.)$ of the above SM-GP model is:

$$c(t,t') = \sum_{d=1}^{D} c^{(d)}(t,t')$$

Let $\mathbf{t} = (t_i)_{i=1}^N$, $\mathbf{y} = (y_i)_{i=1}^N$, $\mathbf{f} = (f^{(1)}(\mathbf{t}), \ldots, f^{(D)}(\mathbf{t}))^T \in \mathbb{R}^{DN \times 1}$ the vector of all stacked realizations and $\mathbf{C}^{(d)} = c^{(d)}(\mathbf{t}, \mathbf{t})$ the d-th prior covariance matrix, then the prior and the likelihood terms are:

$$\begin{aligned} p(\mathbf{f}) &= \mathcal{N}(0, \mathbf{\Gamma}) \\ p(\mathbf{y}|\mathbf{f}) &= \mathcal{N}(\mathbf{P}\mathbf{f}, \gamma^2 \mathbf{I}_N) \end{aligned}$$

where $\mathbf{\Gamma}$ is a block diagonal matrix that collects all prior covariance matrices:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{C}^{(1)} & 0 & \ldots & 0 \\ 0 & \mathbf{C}^{(2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{C}^{(D)} \end{pmatrix}$$

and the matrix $\mathbf{P}$ select the contributing entries of $\mathbf{f}$ at each time-point:

$$\mathbf{P} = (\mathbf{I}_N \ldots \mathbf{I}_N)$$

From Bayes' rule, the posterior distribution over $\mathbf{f}$ is also a Gaussian: $p(\mathbf{f}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Sigma}_{\text{post}})$ with a mean and a covariance matrix expressed as:

$$\boldsymbol{\mu}_{\text{post}} = (\boldsymbol{\mu}_{\text{post}}^{(1)}, \ldots, \boldsymbol{\mu}_{\text{post}}^{(D)})^T = \boldsymbol{\Gamma}\mathbf{P}^T(\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y} \quad (5)$$

$$\boldsymbol{\Sigma}_{\text{post}} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{P}^T(\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T + \gamma^2\mathbf{I}_N)^{-1}\mathbf{P}\boldsymbol{\Gamma} \quad (6)$$

The d-th component in (5) is given by:

$$\boldsymbol{\mu}_{\text{post}}^{(d)} = \mathbf{C}^{(d)}(\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y}$$

Applying (4), the d-th prior covariance matrix is updated according to:

$$\mathbf{C}^{(d)} = \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)}, \text{where } \mathbf{O}^{(d)} \text{and } \mathbf{K}^{(d)} \text{are defined in (1)}$$

Using the fact that $\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T = \sum_{d=1}^{D} \mathbf{C}^{(d)}$, the d-th posterior mean can be rewritten as:

$$\boldsymbol{\mu}_{\text{post}}^{(d)} = \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)}\big(\sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N\big)^{-1}\mathbf{y}$$

Noting that $\boldsymbol{\Sigma}_y = \sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N$ corresponds to the signal covariance matrix, the d-th posterior is given by:

$$\boldsymbol{\mu}_{\text{post}}^{(d)} = \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)}\boldsymbol{\Sigma}_y^{-1}\mathbf{y}$$

We now use the approximation (3) for each kernel $k^{(d)}$, $c(.)$ can be thus approximated by:

$$c(t, t') \approx \sum_{d=1}^{D} cos(w_d(t - t')) \times \sum_{l=1}^{M} S^{(d)}(\sqrt{\gamma_l})\psi_l(t)\psi_l(t')$$

Let $\boldsymbol{\Psi} = \begin{pmatrix} \psi_1(t_1) & \ldots & \psi_M(t_1) \\ \vdots & \ldots & \vdots \\ \psi_1(t_N) & \ldots & \psi_M(t_N) \end{pmatrix}$ and $\mathbf{S}^{(d)} = \text{diag}(S^{(d)}(\sqrt{\gamma_1}), \ldots, S^{(d)}(\sqrt{\gamma_M}))$, then $\mathbf{K}^{(d)}$ and $\boldsymbol{\Sigma}_y$ can be approximated, respectively, by:

$$\hat{\mathbf{K}}^{(d)} = \boldsymbol{\Psi}\mathbf{S}^{(d)}\boldsymbol{\Psi}^T$$

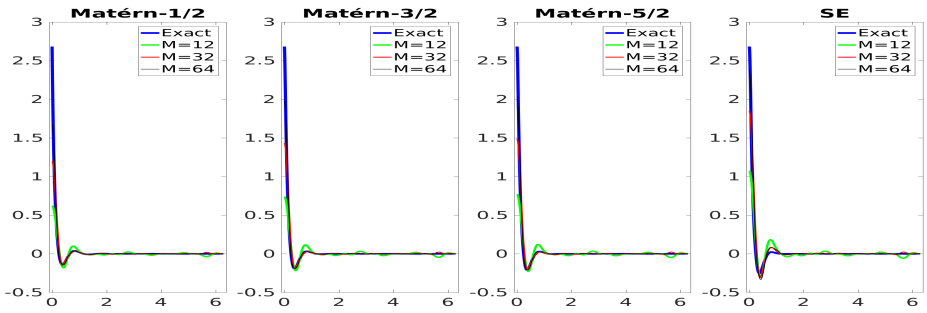

Figure 1: Approximations of the mixture of 5 Matérn and SE covariance functions on $[0, 2\pi]$ with $M = 12$, $M = 32$ and $M = 64$ eigenfunctions.

$$\hat{\boldsymbol{\Sigma}}_y = \sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \boldsymbol{\Psi}\mathbf{S}^{(d)}\boldsymbol{\Psi}^T + \gamma^2 \mathbf{I}_N$$

Thus, the d-th posterior mean can be approximated by:

$$\hat{\boldsymbol{\mu}}^{(d)}_{\text{post}} = \mathbf{O}^{(d)} \odot \boldsymbol{\Psi}\mathbf{S}^{(d)}\boldsymbol{\Psi}^T \hat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{y} \tag{7}$$

We now show that this approximation leads to a significant complexity reduction for the posterior mean computation. Indeed, applying the trigonometric formula, $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, then:

$$\mathbf{O}^{(d)} \odot \boldsymbol{\Psi}\mathbf{S}^{(d)}\boldsymbol{\Psi}^T = \mathbf{X}^{(d,1)}{\mathbf{X}^{(d,1)}}^T + \mathbf{X}^{(d,2)}{\mathbf{X}^{(d,2)}}^T$$

where $\mathbf{X}^{(d,1)}$ and $\mathbf{X}^{(d,2)}$ are the $N \times M$ matrices with entries:

$$\mathbf{X}^{(d,1)}_{i,l} = \sqrt{S^{(d)}(\sqrt{\gamma_l})}\psi_l(t_i)cos(w_d t_i)$$
$$\mathbf{X}^{(d,2)}_{i,l} = \sqrt{S^{(d)}(\sqrt{\gamma_l})}\psi_l(t_i)sin(w_d t_i)$$

for $i = 1, \ldots, N$ and $l = 1, \ldots, M$. Applying the matrix inversion lemma 1 (see

Appendix B), we get:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_y^{-1} &= \big(\sum_{d=1}^{D} \mathbf{X}^{(d,1)}\mathbf{X}^{(d,1)^T} + \mathbf{X}^{(d,2)}\mathbf{X}^{(d,2)^T} + \gamma^2\mathbf{I}_N\big)^{-1} \\ &= \gamma^{-2}\mathbf{I}_N - \gamma^{-4}[\mathbf{X}^{(1,1)}, \mathbf{X}^{(1,2)}, \ldots, \mathbf{X}^{(D,1)}, \mathbf{X}^{(D,2)}]\mathbf{H}^{-1}[\mathbf{X}^{(1,1)}, \mathbf{X}^{(1,2)}, \ldots, \mathbf{X}^{(D,1)}, \mathbf{X}^{(D,2)}]^T\end{aligned}$$

where $\mathbf{H}$ is the $2MD \times 2MD$ matrix:

$$\mathbf{H} = \mathbf{I}_{2MD} + \gamma^{-2} \underbrace{\begin{pmatrix} \mathbf{X}^{(1,1)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(1,1)^T}\mathbf{X}^{(1,2)} & \ldots & \mathbf{X}^{(1,1)^T}\mathbf{X}^{(D,1)} & \mathbf{X}^{(1,1)^T}\mathbf{X}^{(D,2)} \\ \mathbf{X}^{(1,2)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(1,2)^T}\mathbf{X}^{(1,2)} & \ldots & \mathbf{X}^{(1,2)^T}\mathbf{X}^{(D,1)} & \mathbf{X}^{(1,2)^T}\mathbf{X}^{(D,2)} \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ \mathbf{X}^{(D,1)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(D,1)^T}\mathbf{X}^{(1,2)} & \ldots & \mathbf{X}^{(D,1)^T}\mathbf{X}^{(D,1)} & \mathbf{X}^{(D,1)^T}\mathbf{X}^{(D,2)} \\ \mathbf{X}^{(D,2)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(D,2)^T}\mathbf{X}^{(1,2)} & \ldots & \mathbf{X}^{(D,2)^T}\mathbf{X}^{(D,1)} & \mathbf{X}^{(D,2)^T}\mathbf{X}^{(D,2)} \end{pmatrix}}_{\mathbf{F}}$$

In the simple case when $D = 1$, it is easy to see that the evaluation of the approximate posterior mean (7) is dominated by the cost of constructing the $2M \times 2M$ matrix $\mathbf{F} = \begin{pmatrix} \mathbf{X}^{(1,1)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(1,1)^T}\mathbf{X}^{(1,2)} \\ \mathbf{X}^{(1,2)^T}\mathbf{X}^{(1,1)} & \mathbf{X}^{(1,2)^T}\mathbf{X}^{(1,2)} \end{pmatrix}$ which requires $O(M^2N)$ operations (each of the four sub-matrices is the product of $M \times N$ and $N \times M$ matrices). In the general case when $D > 1$, $\mathbf{F}$ is a $2MD \times 2MD$ matrix containing $4D^2$ sub-matrices each of which requires $O(M^2N)$ operations to be constructed. The cost to invert the matrix $\mathbf{H}$, $O(M^3D^3)$, is dominated by $O(M^2D^2N)$ when the sample size N is high enough i.e. $N > MD$, which is generally the case for audio signals. In total, the computational cost is linear in time and quadratic in both the number of mixtures the and truncation order: $O(M^2D^2N)$. Recall that the cost of inference with SM-GP based on Kalman filtering, $O(N\eta^3)$, cubicly increases in the state space dimension $\eta = (2\nu + 1)D$. Therefore, the proposed method achieves a lower computational complexity compared to the SSM approach under specific conditions, specifically when $M^2 < (2\nu + 1)^3 D$. Moreover, it has additional advantage of being able to handle any stationary kernel, such as SE, not only limited-orders Matérn ones as the SSM approach.

6. Experimental results

In this section, we carry out experiments to illustrate the practical benefits of the approach. Using voiced speech, we start by showing that the the frequency-domain

hyperparameter learning is effective as it can lead to the modeling and identification of key speech features. We then perform denoising experiments using the same data, another dataset of real vowels as well as synthetic data. We compare our approach to the GTF method (Turner and Sahani, 2014) (cubic in time) and the SSM method (Wilkinson et al., 2019) based on Kalman filtering (cubic in state dimension). Matlab codes implementing these methods have been made available in Unifying-Prob-Time-Freq. A matlab code of our method, which is an adaptation of the Turner and Wilkinson codes, is provided in Reduced-rank SM-GP. An important improvement in our code is the integration of the estimation of noise variance, which is not implemented in the other codes. We also report observations from experiments using GPyTorch which implements hyperparameter learning using the classical time-domain approach and performs canonical SM-GP inference, with full inversion of the covariance matrix (cubic in time as GTF). In all these experiments, we use Matérn-$\frac{5}{2}$ as the canonical kernel $k^{(d)}$ because $\nu = \frac{5}{2}$ is the highest order implemented by Wilkinson in his code. We set the truncation order to $M = 12$ based on the observation mentioned in Section 4.

6.1. Fundamental and formant frequency estimation

In this experiment, we illustrate the ability of the frequency-domain hyperparameter learning method described in section 3.2 to adaptively learn adequate TF representations. To do so, we use short and steady vowels for which the most important spectral cues are known, the trajectories of the fundamental frequency F_0 and of the first 3 formants, F_1, F_2 and F_3. The shortness and steadiness of the vowels make the stationarity assumption viable. The vowel are extracted from Vowel Database, where these frequencies were manually annotated frame-by-frame. The left panel of Figure 2 shows the spectrograms of 3 adult-male vowels taken from the words HAD, HEAD and HEED, along with the manual annotation of $F_i, i = 0, ..., 3$.

If the learning method is effective, the set of learned centre-frequencies w_d should contain frequencies close to at least some of the F_i, in average. Interestingly, we found that this not only occurs but also that these frequencies correspond to high variances of subband models. More precisely, the highest σ_d^2 are associated with w_d which are close to the average of the F_i values. The right panel of Figure 2 shows the spectrograms of the posterior mean signals, with $D = 20$, where the green lines indicate the highest four w_d. It can be seen that each w_d is close to the average of a F_i (for HEED, 2 of the 4 highest w_d have almost the same value which is close to the average of F_2). This interesting behaviour is coherent with the expectation that subbands where most of the energy is concentrated should be reflected in the GP model by a high model variance of these subbands.

$\lvert median(F_j) - \hat{F}_j \rvert (Hz)$	HAD	HEAD	HEED
$j = 0$ (fundamental frequency)	3.2	0.2	0.6
$j = 1$ (first formant)	19.7	17.8	16.6
$j = 2$ (second formant)	28.1	34.6	26.7
$j = 3$ (third formant)	31.9	50.2	623.1

Table 1: Absolute difference between each F_i median and its estimation.

Table 1 presents the absolute value of the difference between each "true" F_j and its estimation $\hat{F}_j$ using this automatic detection. By "true", we mean the median of the F_j trajectory computed by Praat (Boersma and Van Heuven, 2001), which we visually checked that they are close to the manual values in Figure 2-left. We did so because the numerical values of the manual annotation are not provided in the database. We chose the median, instead of the average, in order to discard some the F_j values at the end of the vowels which increase the variance of the F_j. As can be expected, the learning does not yield monotonic w_d, that is, the highest σ_d^2 can correspond to any F_j, we thus order them to obtain $\hat{F}_j$. These results show that the estimation is very accurate, except for F_3 of HEED where the model failed to capture the subband around the 3rd formant, and confirms the effectiveness of the frequency-domain learning approach. This behaviour also opens a promising new perspective for automatic probabilistic estimation of the fundamental and formant frequencies, but this is beyond the scope of this paper.

6.2. Denoising

Simulated data: We constructed a signal of a sample size $N = 1000$ as the sum of $D = 5$ subbands simulated from a Matérn-$\frac{5}{2}$ SM-GP. The hyperparameters were fixed and chosen so that the filter banks center frequencies w_d are uniformly distributed between 50Hz and 8000Hz, the bandwidths l_d are uniformly distributed between 100Hz and 500Hz and with equal variance $\sigma_d^2 = 0.01$ for $d = 1, \ldots, 5$. A Gaussian noise was added to this (clean) signal at different signal-to-noise-ratio (SNR) between -5dB and 5dB. Then, the posterior mean was computed using the noisy observations and considered as an estimation of the clean signal. Figure 3 shows the SNR improvement of the estimated signal using our method (proposal), the GTF method (GTF) and the SSM method based on Kalman filetring (Kalman). In this example, GTF yields the best performance which is natural as it is, theoretically, the optimal method involving the inversion of the full covariance matrix. However, our method not only yields a very close performance to GTF but also outperforms the SSM method. We found that the latter is less accurate than the other methods in

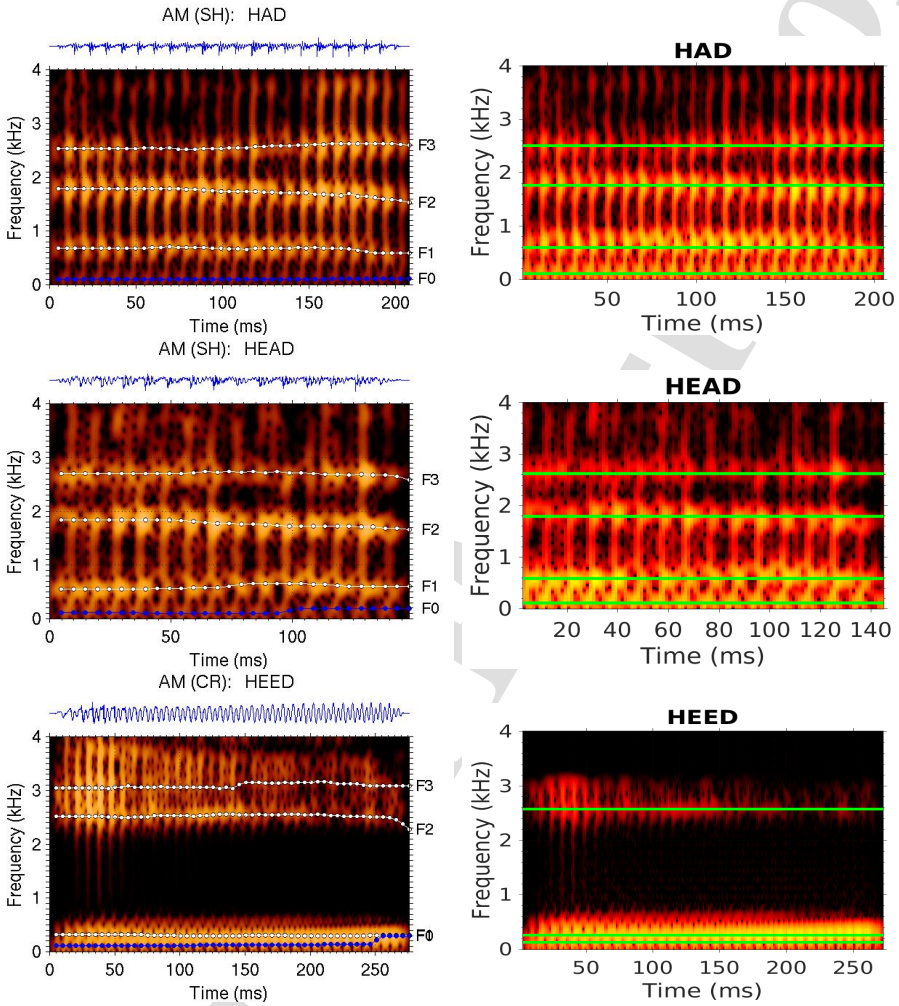

Figure 2: Spectrograms of the original vowels (left panel) and the estimated posterior mean signal (right panel) using a SM-GP with a mixture of $D = 20$ Matérn-$\frac{5}{2}$ kernels. The green lines indicate the location of the center frequencies w_d associated with the 4 highest variances σ_d^2.

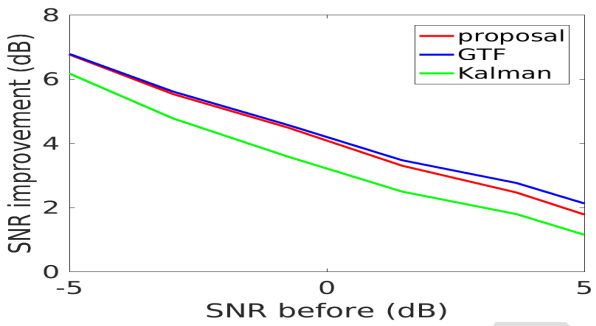

Figure 3: Denoising of simulated data: SNR improvement using the Matérn-$\frac{5}{2}$ SM-GP.

Method	Time (s)
Proposal	**0.48**
Kalman	0.54
GTF	0.64

Table 2: The elapsed time in seconds (s).

recovering the subband components. This observation stands also when considering Matérn-$\frac{1}{2}$ and Matérn-$\frac{3}{2}$ kernels. In term of computational complexity, as shown in Table 2, the practical computation time confirms the theoretical expectation of a better computational efficiency of our method.

Real data: In this experiment, a Gaussian noise was added to each of the 3 vowels (used in the previous section) at different SNRs between -5dB and 5dB. Then, the posterior mean was computed using the noisy observations and considered as an estimation of the clean signal. Figure 4 shows the SNR improvement obtained by using our method, the GTF and the SSM method. We recall that (Turner and Sahani, 2014) showed that GTF was competitive with several standard denoising techniques. The results of Figure 4 show that, globally, our method outperforms the other methods, particularly in low SNR. They also indicate that, besides the computational load, inverting the full covariance (as GTF does) is not a guaranty of a better performance when processing real-world data. We also tested GPyTorch but it yielded poor noise variance estimates, which in turn yielded poor estimates of the clean signal. This indicates that the hyperparameter temporal-domain estimation (of GPyTorch) is very to noise and confirms that the superiority of the frequency-domain learning approach. Overall, in addition to the computational gain, these

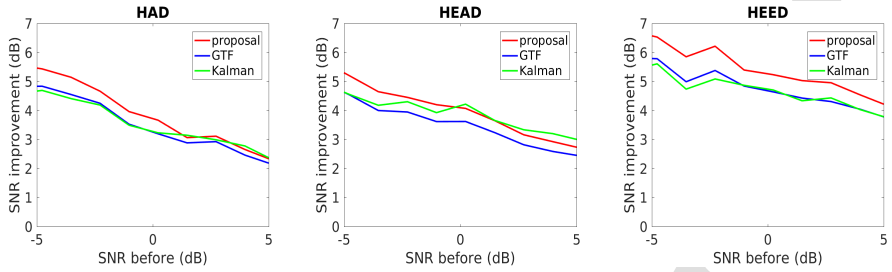

Figure 4: Denoising of the 3 vowels: SNR improvement using the Matérn-$\frac{5}{2}$ SM-GP.

results indicate that our approach to posterior mean estimation outperforms the two other techniques in the task of denoising.

In order to illustrate the generalization ability of our method, we carried out a second experiment using the PC-GITA database (Orozco-Arroyave et al., 2014). The latter consist in recordings of different speech tasks form 50 healthy and 50 non-healthy speakers. We considered only the sustained vowels subset from the healthy speakers, from which we extracted 10 randomly chosen speech files. The average duration of the vowels is about 1 second. The signals, which sampling frequency is 44.1Kz, were downsampled to 8Kz. Figure 5 (a) shows the averaged SNR improvement obtained by using our method, the GTF and the SSM method. It shows that our method considerably outperforms the others for large noise variances (low SNRs). Figure 5 (b) shows an example of the power spectrum of a vowel, in pink, its noisy version in black, and the estimated power spectrum using the three methods. It shows that our method is slightly better over the first third of frequencies ($<$ 1200Kz), all the methods do not perform well on the second third (between 1200Kz and 3300Hz), and our method significantly outperforms the others over high frequencies ($>$ 3300Hz). Overall, these results indicate that our method has a good generalization ability.

7. Conclusion

We proposed a new approach to design probabilistic time-frequency representations of signals. This approach capitalizes on recent findings showing the equivalence between spectral mixture Gaussian processes and a large class of probabilistic time-frequency models. Based on a novel reduced-rank approximation of covariance functions, we developed an algorithm which allows faster inference than a recently proposed method. This was achieved because the algorithm bi-passes the formulation of

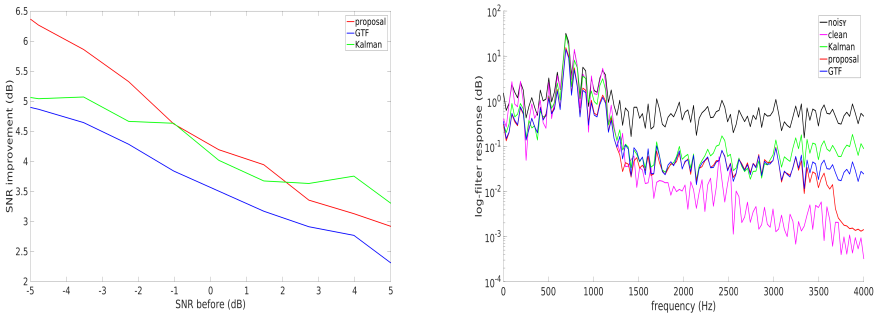

Figure 5: (a) Averaged SNR improvement over the vowels subset. (b) Example of a clean, noisy and estimated spectra of a vowel.

SM-GP in terms of state space models, in which inference is performed by Kalman filtering which is cubic in state dimension. Moreover we experimentally showed, using speech data, that the time-frequency method for hyperparameter learning is significantly more accurate than the time-domain method, as it can reveal important spectral cues in the signal. Finally, denoising experiments using speech data as well as simulated data show that our method yields better performances. Though our algorithm reduces the computation complexity, it is still quadratic in the covariance approximation order and the number of mixture components. This may constitute a limitation in some applications. Our future objective is to make the complexity linear in these parameters, this is purpose of our ongoing research.

Appendix A

The log marginal likelihood of the SM-GP model in the temporal domain is given by:

$$\log p(\mathbf{y}|\theta) = -\frac{1}{2}\mathbf{y}^T(\sum_{d=1}^{D}\mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y} - \frac{1}{2}\log|\sum_{d=1}^{D}\mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N| - \frac{N}{2}\log(2\pi)$$

In order to reduce the complexity of evaluating $\log p(\mathbf{y}|\theta)$ the spectral approach consists of approximating the covariance matrix $c^{(d)}(t,t') = cos(w_d(t-t')) \times k^{(d)}(t,t')$

by:

$$c^{(d)}(t,t') \approx \sum_{j=1}^{N} FT_{t,j} \times \gamma_j^{(d)} \times FT_{t',j}^{*}$$

where $\gamma_j^{(d)} = \frac{1}{2}[S^{(d)}(w_j - w_d) + S^{(d)}(w_j + w_d)]$ refers to d-th spectral density associated to c^d and $FT_{t,j} = e^{iw_j}$ is the complex exponential basis evaluated at $w_j = 2\pi\frac{t_j}{N}$. Consequently, the covariance matrix $\mathbf{C}^{(d)} = \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)}$ can be approximately decomposed as:

$$\mathbf{C}^{(d)} \approx \mathbf{\Phi}\mathbf{R}^{(d)}\mathbf{\Phi}^{*}$$

where $\mathbf{\Phi} = \begin{pmatrix} FT_{t_1,t_1} & \dots & FT_{t_1,t_N} \\ \vdots & \dots & \vdots \\ FT_{t_N,t_1} & \dots & FT_{t_N,t_N} \end{pmatrix}$ and $\mathbf{R}^{(d)} = \begin{pmatrix} \gamma_1^{(d)} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \gamma_N^{(d)} \end{pmatrix}$. Based on the last decomposition we get:

$$\begin{aligned} \mathbf{y}^T(\sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N)^{-1}\mathbf{y} &\approx \sum_{j=1}^{N} \frac{|\tilde{y}_j|^2}{\gamma_{y,j}} \\ \log|\sum_{d=1}^{D} \mathbf{O}^{(d)} \odot \mathbf{K}^{(d)} + \gamma^2\mathbf{I}_N| &\approx \sum_{j=1}^{N} \log(\gamma_{y,j}) \end{aligned}$$

where $\tilde{\mathbf{y}} = (\tilde{y}_j)_{j=1}^{N}$ is the signal spectrum and $\gamma_{y,j} = \sum_{d=1}^{D} \gamma_j^{(d)} + N\gamma^2$ is the model' spectral density. Therefore, the marginal likelihood in the frequency domain called the "*whittle*" likelihood is:

$$\begin{aligned} p(\mathbf{y}|\theta) &= (2\pi)^{-\frac{N}{2}} \times \prod_{j=1}^{N} \gamma_{y,j}^{-\frac{1}{2}} \times \exp\big(-\frac{1}{2}\frac{|\tilde{y}_j|^2}{\gamma_{y,j}}\big) \\ &= \mathcal{N}\big(\tilde{\mathbf{y}}|0, \mathrm{diag}(\gamma_{y,1}, \dots, \gamma_{y,N})\big) \end{aligned}$$

This completes the expression of the "*whittle*" log likelihood in (2).

The "*whittle*" log likelihood captures the trade-off between model fit and model complexity, and optimizing it provides a principled way to determine the hyperparameters that best explain the observed data in the context of the SM-GP. One way to estimate the hyperparameters of the SM-GP model is to maximize the "*whittle*"

log likelihood by evaluating the partial derivatives w.r.t $\theta = (\theta_{\text{sm}}, \gamma^2)$:

$$\frac{\partial \log p(\mathbf{y}|\theta)}{\partial \theta} = -\frac{1}{2}\sum_{j=1}^{N}\Big(\frac{1}{\gamma_{y,j}} - \frac{|\tilde{y}_j|^2}{\gamma_{y,j}^2}\Big) \times \frac{\partial \gamma_{y,j}}{\partial \theta}$$

Appendix B

Lemma 1 (Matrix inversion lemma (Golub and Van Loan, 1996))**.** *If* $\mathbf{A}$ *is invertible square matrix and* $\mathbf{U}_j$ $(j = 1, \ldots, J)$ *are* $N \times M$ *rectangular matrices then:*

$$(\mathbf{A} + \sum_{j=1}^{J} \mathbf{U}_j \mathbf{U}_j^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}[\mathbf{U}_1, \ldots, \mathbf{U}_J]\mathbf{H}^{-1}[\mathbf{U}_1, \ldots, \mathbf{U}_J]^T\mathbf{A}^{-1}$$

where $\mathbf{H}$ *is* $MJ \times MJ$ *matrix satisfying*

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_M + \mathbf{U}_1^T\mathbf{A}^{-1}\mathbf{U}_1 & \ldots & \mathbf{U}_1^T\mathbf{A}^{-1}\mathbf{U}_J \\ \vdots & \vdots & \vdots \\ \mathbf{U}_J^T\mathbf{A}^{-1}\mathbf{U}_1 & \ldots & \mathbf{I}_M + \mathbf{U}_J^T\mathbf{A}^{-1}\mathbf{U}_J \end{pmatrix}$$

□

References

Akhiezer, N.I., Glazman, I.M., 2013. Theory of linear operators in Hilbert space. Dover Books on Mathematics, Dover Publications.

Alvarado, P.A., Alvarez, M.A., Stowell, D., 2019. Sparse Gaussian process audio source separation using spectrum priors in the time-domain, in: 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 995–999.

Alvarado, P.A., Stowell, D., 2017. Efficient learning of harmonic priors for pitch detection in polyphonic music. ArXiv abs/1705.07104.

Boersma, P., Van Heuven, V., 2001. Speak and unspeak with PRAAT. Glot Int 5, 341–347.

Cemgil, A.T., Godsill, S.J., 2005. Probabilistic phase vocoder and its application to interpolation of missing values in audio signals, in: 2005 13th European Signal Processing Conference, pp. 1–4.

Cohen, L., 1995. Time-frequency analysis: theory and applications. Prentice-Hall, Inc., USA.

Ghanem, R.G., Spanos, P.D., 1991. Stochastic finite elements: a spectral approach. Springer-Verlag, Berlin, Heidelberg.

Golub, G.H., Van Loan, C.F., 1996. Matrix computations. Third ed., The Johns Hopkins University Press.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82, 35.

Liutkus, A., Badeau, R., Richard, G., 2011. Gaussian processes for underdetermined source separation. IEEE Transactions on Signal Processing 59, 3155–3167.

Mallat, S., 2008. A Wavelet tour of signal processing, The sparse way. 3rd ed., Academic Press, Inc., USA.

Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C., Nöth, E., 2014. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland. pp. 342–347.

Plataniotis, K., 2000. Gaussian mixtures and their applications to signal processing. Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems. pp. 1–33.

Qi, Y., Minka, T.P., Picara, R.W., 2002. Bayesian spectrum estimation of unevenly sampled nonstationary data, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II–1473–II–1476.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. Adaptive computation and machine learning, MIT Press.

Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs, in: Advances in Neural Information Processing Systems, MIT Press. p. 1257–1264.

Solin, A., Särkkä, S., 2014. Explicit link between periodic covariance functions and state space models, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (PMLR), Reykjavik, Iceland. pp. 904–912.

Solin, A., Särkkä, S., 2020. Hilbert space methods for reduced-rank Gaussian process regression. Statistics and Computing 30, 419–446.

Suwandi, R.C., Lin, Z., Sun, Y., Wang, Z., Cheng, L., Yin, F., 2022. Gaussian process regression with grid spectral mixture kernel: distributed learning for multidimensional data, in: 2022 25th International Conference on Information Fusion (FUSION), pp. 1–8.

Takhanov, R., 2023. On the speed of uniform convergence in Mercer's theorem. Journal of Mathematical Analysis and Applications 518, 126718.

Turner, R.E., Sahani, M., 2014. Time-frequency analysis as probabilistic inference. IEEE Transactions on Signal Processing 62, 6171–6183.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics 15, 70–73.

Wilkinson, W., 2019. Gaussian process modelling for audio signals. Ph.D. thesis. London.

Wilkinson, W.J., Riis Andersen, M., Reiss, J.D., Stowell, D., Solin, A., 2019. Unifying probabilistic models for time-frequency analysis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, United Kingdom. pp. 3352–3356.

Williams, C.K.I., Barber, D., 1998. Bayesian classification with Gaussian processes. IEEE Trans. Pattern Anal. Mach. Intell. 20, 1342–1351.

Wilson, A.G., Adams, R.P., 2013. Gaussian process kernels for pattern discovery and extrapolation, in: ICML (3), Proceedings of Machine Learning Research (PMLR), Atlanta, Georgia, USA. pp. 1067–1075.

Yin, F., Pan, L., Chen, T., Theodoridis, S., Luo, Z.Q.T., Zoubir, A.M., 2020. Linear multiple low-rank kernel based stationary Gaussian processes regression for time series. IEEE Transactions on Signal Processing 68, 5260–5275.

Highlights (for review)

We express our gratitude to the Editor-in-Chief and the reviewers for their supportive comments and feedback. We have addressed all the comments and updated the paper accordingly.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: