Sparse Gaussian Process Audio Source Separation Using Spectrum Priors in the Time-Domain

Pablo A. Alvarado*, Mauricio A. Álvarez⋆, Dan Stowell*

*Queen Mary University of London, ⋆The University of Sheffield

Motivation

- Source separation (SS) aims to infer latent signals from a mixture [1].
- Time-frequency SS methods often discard phase. Thus, approximations are required, corrupting the reconstruction [4].
- Time-domain SS approaches based on Gaussian processes (GP) circumvent phase approximation [4]. GPs are distributions over functions.
- GPs are intractable for large audio signals, as the computational complexity of inference scales cubically with the data size. Also, GP predictions depend deeply on the kernel/prior.
- We analysed whether combining spectrum-inspired kernels and variational sparse GPs inference leads to more efficient and accurate SS models.

Source separation example using the proposed method:

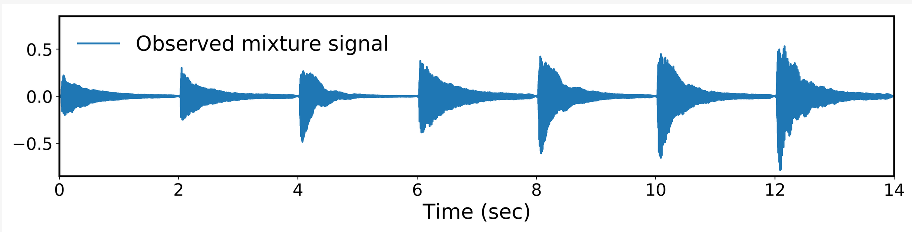


Fig. 1: Mixture signal.

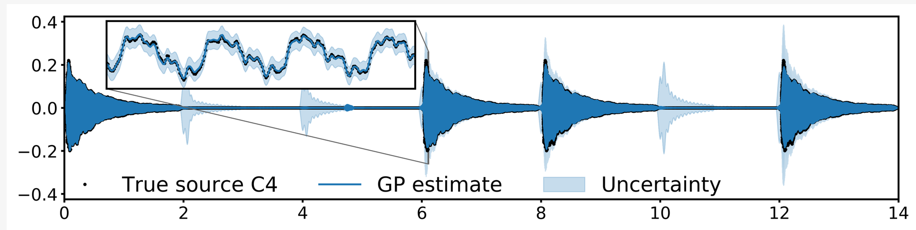


Fig. 2: Reconstructed source C4.

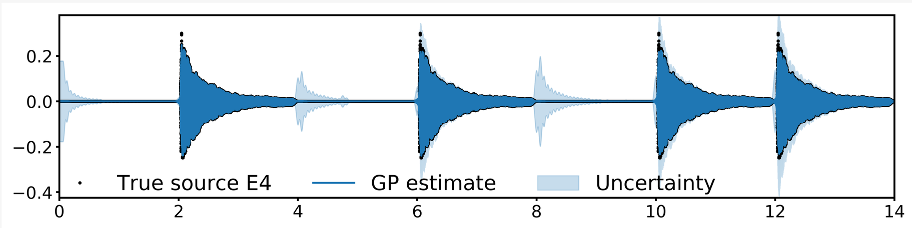


Fig. 3: Reconstructed source E4.

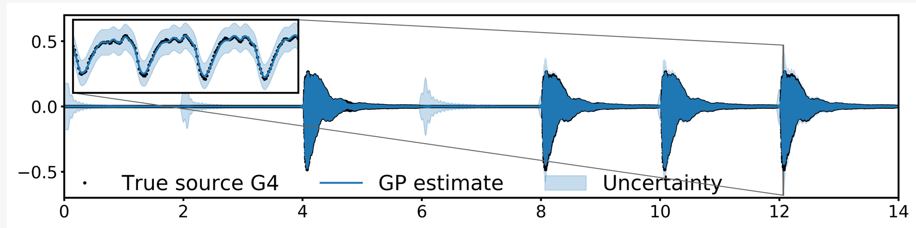


Fig. 4: Reconstructed source G4.

Method

- **Test data:** $\{y_i,\ t_i\}_{i=1}^n$, where $y_i \in \mathbb{R}$ is the i-th audio waveform sample of the mixture, at time $t_i \in \mathbb{R}$.
- **Train data:** isolated (single pitch) music notes, $\{\mathbf{g}^{(j)}\}_{j=1}^J$, where $\mathbf{g}^{(j)} \in \mathbb{R}^{\tilde{n}}$.
- **Regression model:** The mixture is modelled as a sum of GP sources, i.e. $y_i = f(t_i) + \epsilon_i$, where $f(t_i) = \sum_{j=1}^J s_j(t_i)$, and $\epsilon_i \sim \mathcal{N}(0, \nu^2)$.
- **Prior:** sources are GPs, $s_j(t) \sim \mathcal{GP}(0, k_j(t,t'))$. Thus, $f(t) \sim \mathcal{GP}(0, \sum_{j=1}^J k_j(t,t'))$, $\mathbf{s}_j \sim \mathcal{N}(\mathbf{0},\ \mathbf{K}_{s_j})$, and $\mathbf{f} \sim \mathcal{N}(\mathbf{0},\ \mathbf{K}_f)$, where $\mathbf{s}_j = [s_j(t_i)]_{i=1}^n$, $\mathbf{f} = [f(t_i)]_{i=1}^n$, $\mathbf{K}_{s_j}[i,j] = k_j(t_i, y_j)$, and $\mathbf{K}_f = \sum_{j=1}^J \mathbf{K}_{s_j}$.
- **Covariance:** we used spectral mixture (SM) kernels

$$k_j(\tau) = \sigma_j^2 \exp\left(-\frac{\tau}{\ell_j}\right) \times \sum_{d=1}^{D} \alpha_{jd}^2 \cos(\omega_{jd}\,\tau), \tag{1}$$

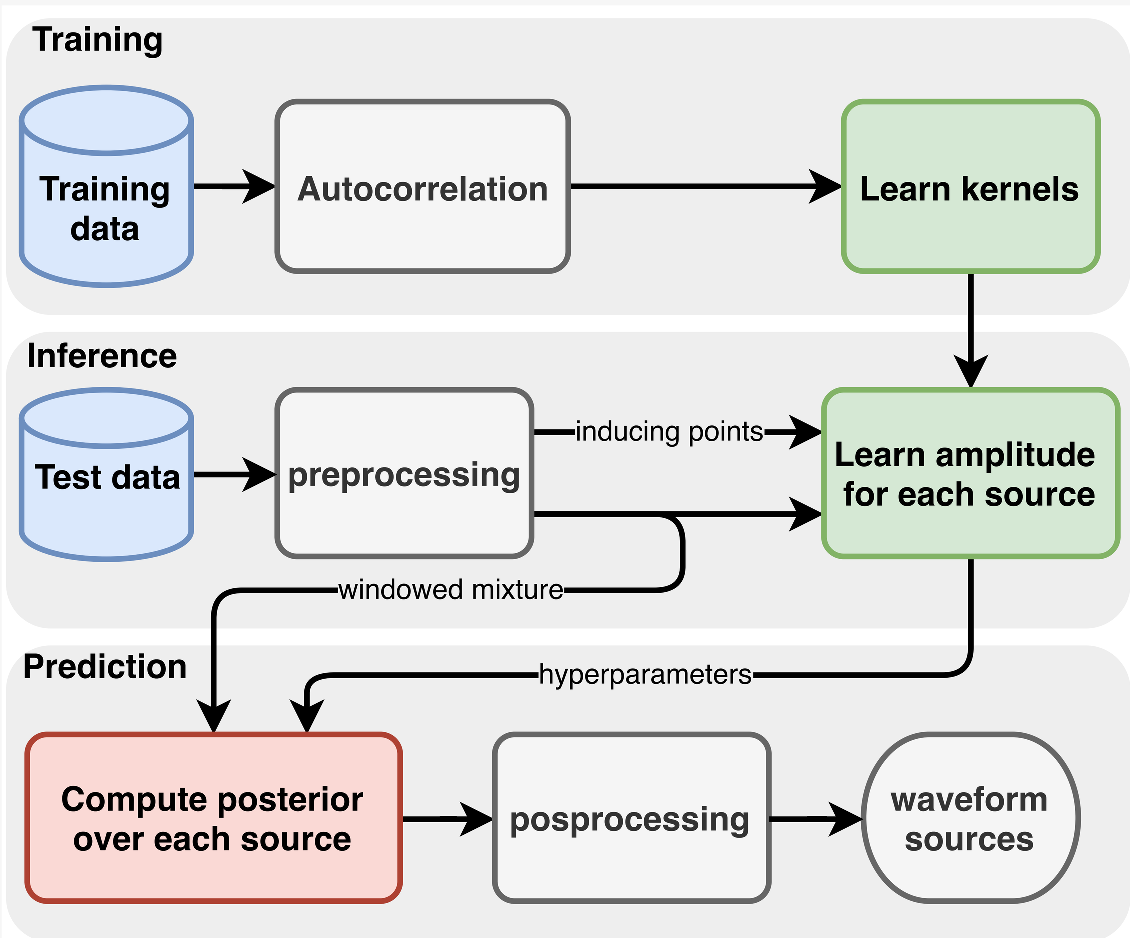
with $\boldsymbol{\theta}_j = \{\sigma_j^2,\ \ell_j,\ [\alpha_{jd}^2,\ \omega_{jd}]_{d=1}^D\}$, and $\tau = |t - t'|$ [3].

Likelihood: $\mathbf{y} \mid \mathbf{f} \sim \mathcal{N}(\mathbf{y} \mid \mathbf{f},\ \nu^2\mathbf{I})$, where $\mathbf{y} = [y_i]_{i=1}^n$.

Posterior:

$$\mathbf{s}_j \mid \mathbf{y} \sim \mathcal{N}\left(\mathbf{s}_i \mid \mathbf{K}_{s_j}^\top \mathbf{H}^{-1}\mathbf{y},\ \ \hat{\mathbf{K}}_{s_j}\right), \tag{2}$$

where $\mathbf{H} = \mathbf{K}_f + \nu^2\mathbf{I}$, and $\hat{\mathbf{K}}_{s_j} = \mathbf{K}_{s_j} - \mathbf{K}_{s_j}^\top \mathbf{H}^{-1}\mathbf{K}_{s_j}$.



Inference:
The kernels were initialized by minimizing

$$L(\boldsymbol{\theta}_j) = \frac{1}{N_c}\sum_{i=1}^{N_c}\left[k_j(\hat{\tau}_i) - C_j(\hat{\tau}_i)\right]^2, \tag{3}$$

$$C_j(\hat{\tau}) = \frac{1}{T}\int_0^T g^{(j)}(x+\hat{\tau})\ g^{(j)}(x)\ \mathrm{d}x, \tag{4}$$

where $C_j(\cdot)$ is the autocorrelation of the j-th training signal. To handle long signals, we windowed $\mathbf{y}$ into frames $\{\hat{\mathbf{t}}^{(w)},\ \hat{\mathbf{y}}^{(w)}\}_{w=1}^W$, and optimized (5) with respect to $\{\sigma_j^2\}_{j=1}^J$, using inducing variables $\mathbf{u} = [f(z_i)]_{i=1}^m$, at points $\mathbf{z} = [z_i]_{i=1}^m$.

$$\mathcal{L} \triangleq \log\mathcal{N}\left(\hat{\mathbf{y}}^{(w)} | \mathbf{0}, \mathbf{Q}_{\hat{n}\hat{n}} + \nu^2\mathbf{I}\right) - \frac{1}{2\nu^2}\mathrm{tr}\left(\mathbf{K}_{\hat{n}\hat{n}} - \mathbf{Q}_{\hat{n}\hat{n}}\right), \tag{5}$$

where $\mathbf{Q}_{\hat{n}\hat{n}} = \mathbf{K}_{\hat{n}m}\mathbf{K}_{mm}^{-1}\mathbf{K}_{m\hat{n}}$, $\mathbf{K}_{\hat{n}m}[i,j] = k_f(t_i^{(w)}, z_j)$, $\mathbf{K}_{mm}[i,j] = k_f(z_i, z_j)$, and $t_i^{(w)} = \mathbf{t}^{(w)}[i]$ ([2]). We computed (2) for each window, and merged the reconstructed sources.

Experiments:

- We used the dataset analysed in [4]: three mixture signals (piano, electric guitar, clarinet) sampled at 16kHz.
- Each mixture last 14 seconds, and has the sequence of events C4, E4, G4, C4+E4, C4+G4, E4+G4, C4+E4+G4.
- Compared methods: LD-PSDTF (positive semi-definite tensor factorization), KL-NMF (Kullback-Leibler NMF), and IS-NMF (Itakura-Saito NMF).
- The first three isolated events were used for training.

Results

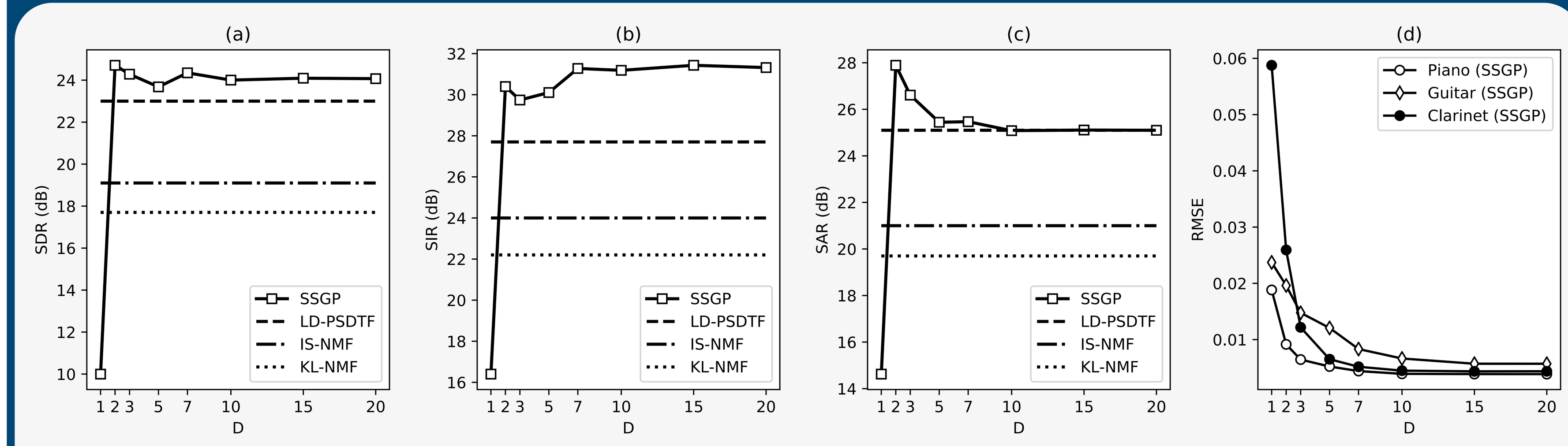


Fig. 6: Evaluation metrics versus kernel complexity. SDR (a), SIR (b), SAR (c), RMSE (d).

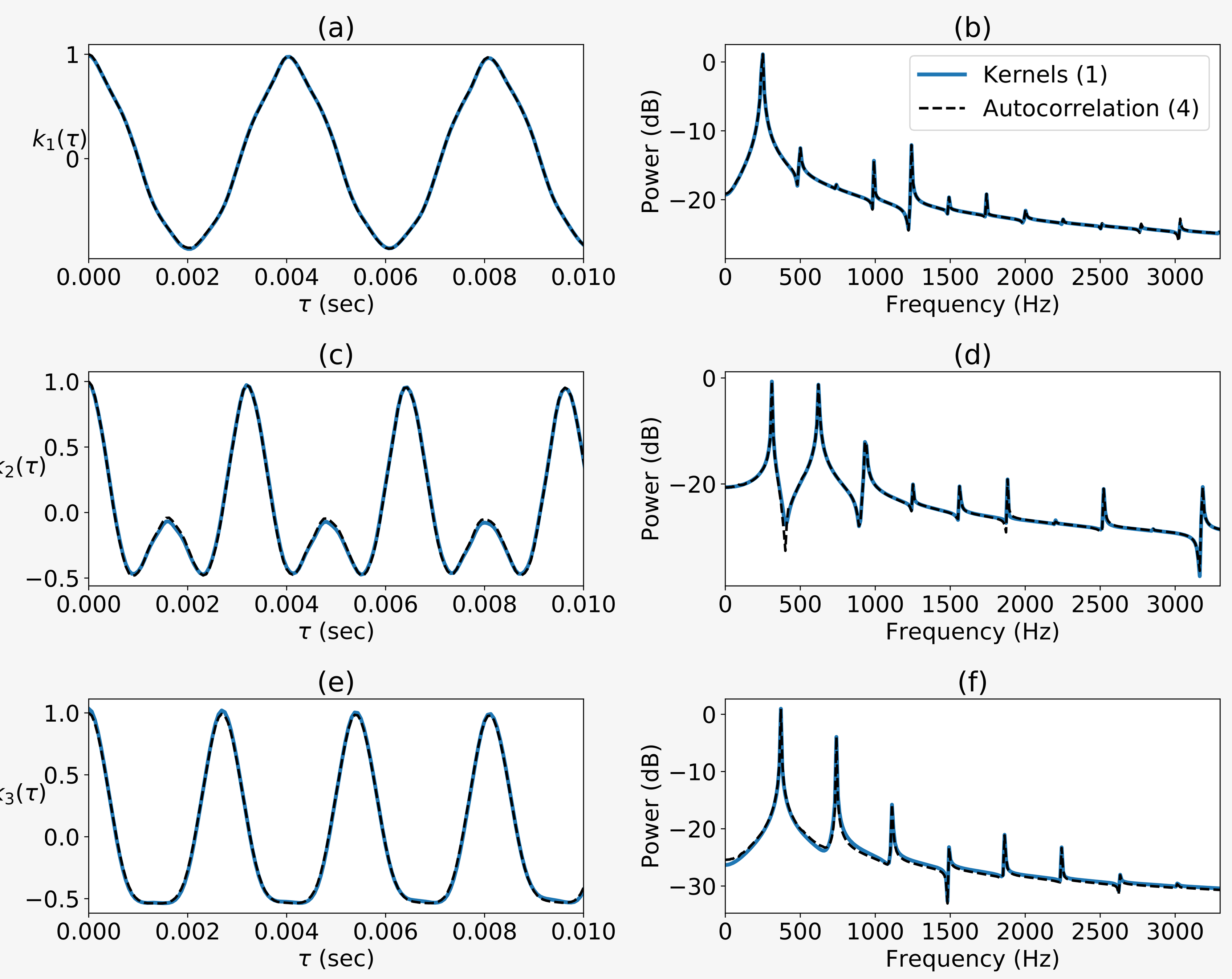


Fig. 7: Learned kernels for piano notes (left column). Corresponding log-spectral density (right column).

Results

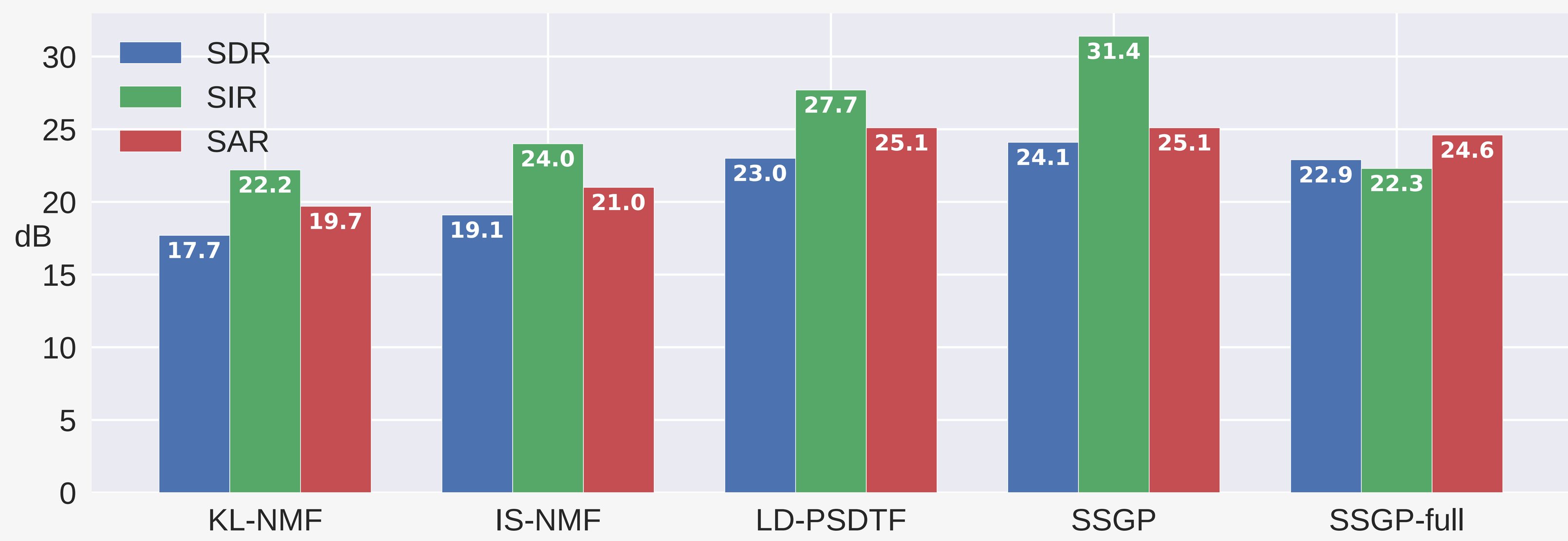


Fig. 8: Source separation metrics. Proposed method: SSGP.

- SSGP presented the highest SDR and SIR metrics (Fig. 8).
- SSGP reduced the optimization time by **98.12**% compared to the full GP model (Fig. 9).
- The learned kernels showed distinctive spectral patterns for each source (Fig. 7), suggesting SM kernels are suitable for learning intricate frequency content.
- SSGP is robust to kernel selection when the number of components in the source kernels is greater than three (Fig. 6).
- RMSE decreased exponentially with D, suggesting that increasing the number of components in the kernel leads to more accurate waveform reconstructions (Fig. 6(d)).

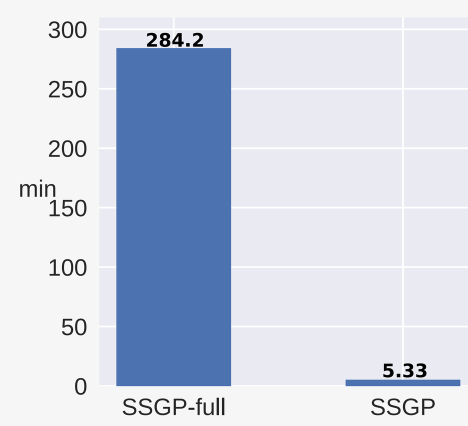


Fig. 9: Optimisation time.

Conclusions

- Combining variational sparse GPs and SM kernels enables time-domain source separation GP models to reconstruct audio sources in an efficient and informed manner, without compromising performance.
- Suitable spectrum priors over the sources are essential to improve source reconstruction.
- SSGP can be used for other applications such as multipitch-detection, where low interference between sources (SIR) is more relevant than reconstruction artifacts (SAR).
- Code available at `https://github.com/PabloAlvarado/ssgp`.

[1] Liutkus, A., Badeau, R., and Richard, G. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing 59*, 7 (July 2011), 3155–3167.

[2] Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2009), pp. 567–574.

[3] Wilson, A. G., and Adams, R. P. Gaussian process kernels for pattern discovery and extrapolation. *30th International Conference on Machine Learning (ICML)* (2013), 1067–1075.

[4] Yoshii, K., Tomioka, R., Mochihashi, D., and Goto, M. Beyond NMF: Time-domain audio source separation without phase reconstruction. In *14th International Society for Music Information Retrieval Conference (ISMIR)* (2013), pp. 369–374.

Acknowledgement
Pablo A. Alvarado is funded by Colciencias scholarship 679. Mauricio A. Álvarez is partially financed by the EPSRC Research Project EP/N014162/1. Dan Stowell is supported by EPSRC Fellowship EP/L020505/1.