Unsupervised Harmonic Parameter Estimation Using Differentiable DSP and Spectral Optimal Transport

Bernardo Torres Geoffroy Peeters Gaël Richard

LTCI, Telecom Paris, Institut Polytechnique de Paris

Context

Differentiable Digital Signal Processing (DDSP): training neural networks to estimate parameters of signal models (e.g sinusoidal frequency, amplitudes) using synthesis and reconstruction.

Main takeaways

Spectral Optimal Transport (SOT) compares audio measuring frequency displacement of spectral frames

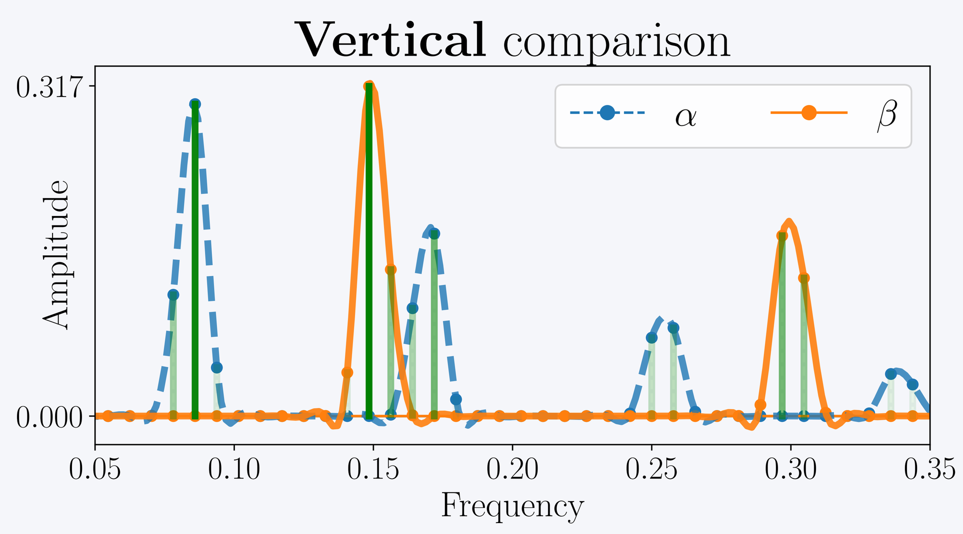
✓ Improves pitch accuracy and reconstruction error when estimating jointly the f_0 and amplitudes of a DDSP harmonic synthesizer (no supervision)

! Sensitive to spectrum normalization and leakage

Paper: Code:

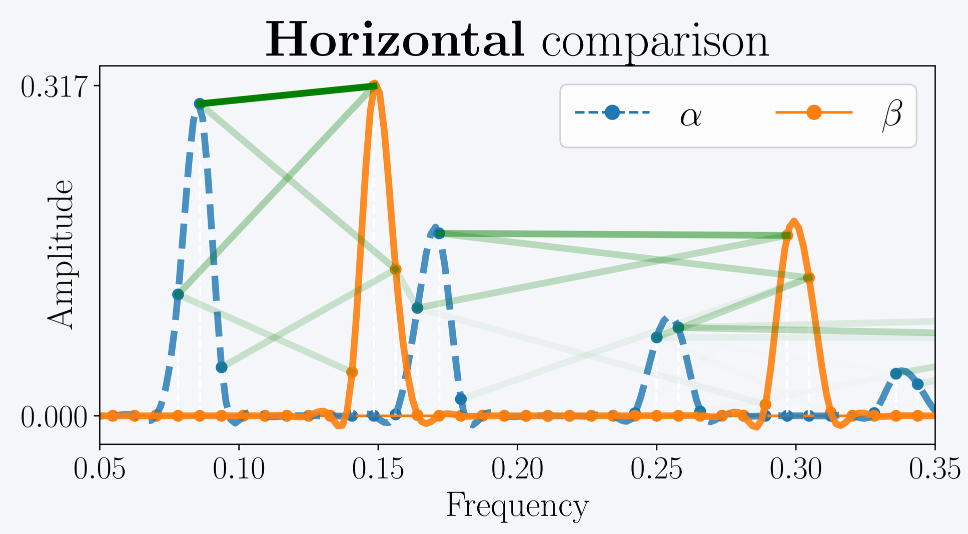
Overview

Usual reconstruction losses (e.g. Multi-Scale Spectral loss) → **Vertical** comparison



→ Bad gradient orientation w.r.t. frequency parameters, many local minima

This work: Spectral Optimal Transport (SOT) → **Horizontal** comparison



→ Better gradients for oscillator frequency estimation

How does Spectral Optimal Transport work ?

Time-frequency transforms $\mathbf{S}, \hat{\mathbf{S}}$ → (Normalize each frame j) → Consider spectra $(\mathbf{S}^{(j)}, \hat{\mathbf{S}}^{(j)})$ as discrete distributions (α, β) → Quantile functions (inverse CDFs) $\mathbf{F}^{-1}_\alpha$, $\mathbf{F}^{-1}_\beta$ → Wasserstein distance (closed form solution in 1D)

$$\mathcal{W}_p(\alpha,\beta)^p = \int_0^1 \left|F_\alpha^{-1}(r) - F_\beta^{-1}(r)\right|^p dr$$

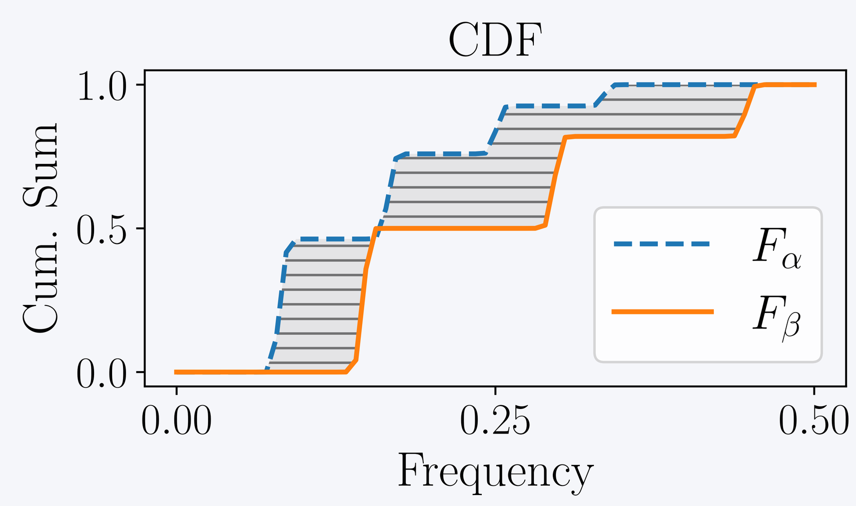
→ (Use as loss function (average over time)) → Improved frequency localization

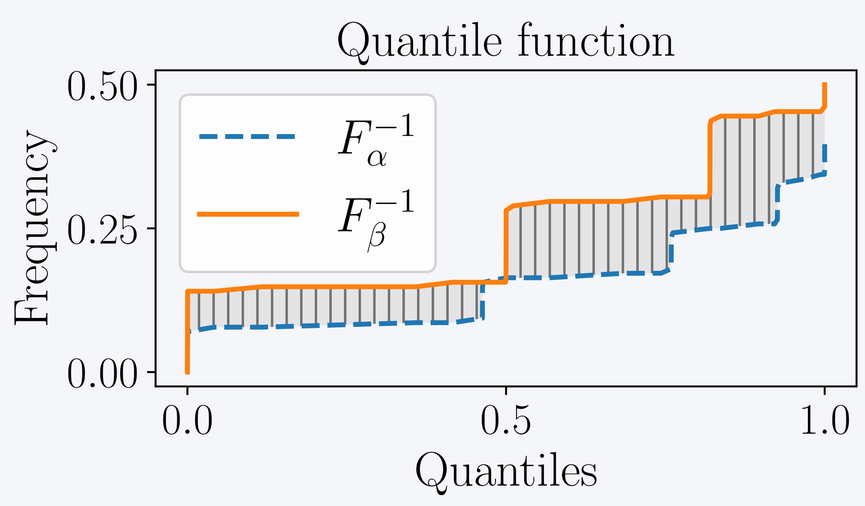
1D Wasserstein distance between normalized frame spectra

$$\mathcal{L}_{\text{SOT}}(s,\hat{s}) = \frac{1}{L}\sum_{j=0}^{L-1} \mathcal{W}_p(\mathbf{S}_\gamma^{2\,(j)}, \hat{\mathbf{S}}_\gamma^{2\,(j)})^p$$

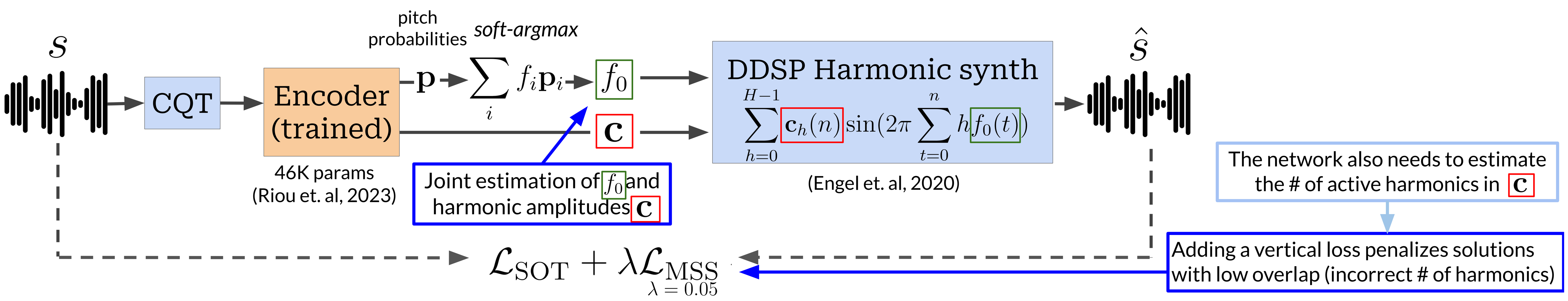
$\mathbf{S}_\gamma = |\text{STFT}_\gamma(s)|$ (magnitude STFT at scale/window size) γ $\quad \alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$

CDF: $F_\alpha(t) = \sum_{i=1}^n u(t - x_i)\mathbf{a}_i$ Quantile function: $F_\alpha^{-1}(r) = \inf\{x \in \mathbb{R} : F_\alpha(x) \geq r\}$





Unsupervised harmonic parameter estimation autoencoder



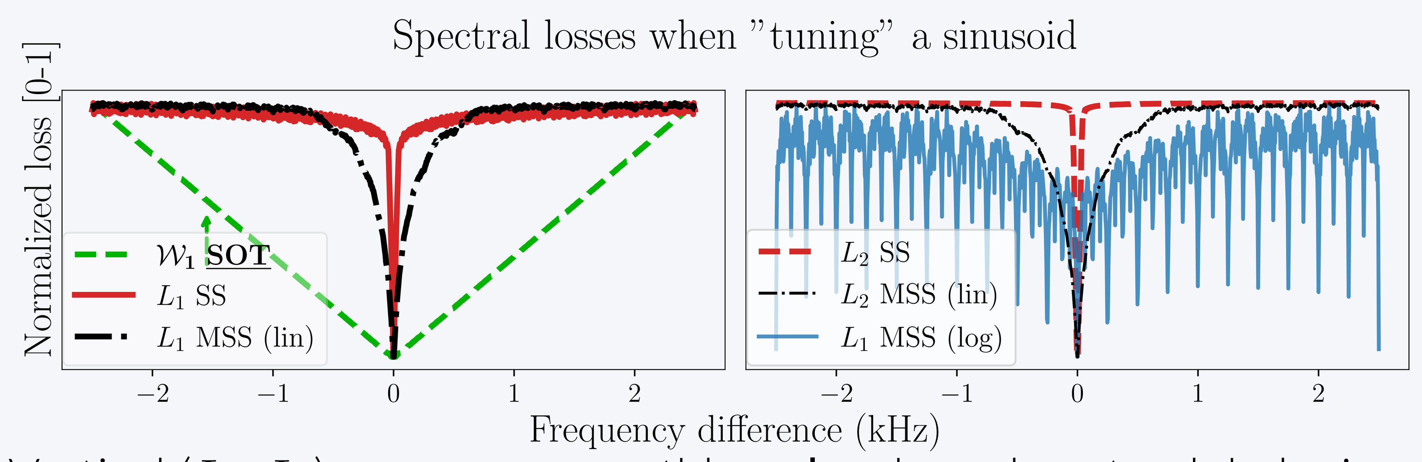
Results

Baseline: Multi-Scale spectral loss: $\mathcal{L}_{\text{MSS}}(s,\hat{s}) = \sum_{\gamma\in\Gamma} \underbrace{\left\|\mathbf{S}_\gamma - \hat{\mathbf{S}}_\gamma\right\|_1}_{\text{linear loss } (\mathcal{L}_{\text{lin}})} + \underbrace{\left\|\log(\mathbf{S}_\gamma) - \log(\hat{\mathbf{S}}_\gamma)\right\|_1}_{\text{log loss } (\mathcal{L}_{\text{log}})}$

Synthetic dataset varying f_0, harmonic amplitudes, and # of harmonics ([1-8])

Evaluation on pitch estimation and reconstruction metrics

Variations					Mean/Median (STD) test metrics (5 runs)		
$\mathcal{L}$	γ ($\mathcal{L}_{\text{SOT}}$)	Γ ($\mathcal{L}_{\text{lin}}$)	LogF	f_{cut}	LSD [dB] ↓	RPA [%] ↑	RCA [%] ↑
LIN	-	Γ_0	-	-	46.4 /58.0 (21.4)	20.2 /0.2 (44.6)	26.9 /3.9 (42.7)
MSS	-	Γ_0	-	-	80.5 /82.6 (15.1)	1.4 /0.1 (2.7)	4.0 /3.2 (4.5)
SOT	2048	Γ_0	×	✓	**23.5** /**24.5** (3.5)	**75.0** /**99.7** (43.2)	**99.2** /**99.8** (1.6)
SOT	512	Γ_0	×	✓	40.5 /26.6 (23.5)	42.9 /63.6 (39.4)	62.3 /75.2 (42.6)
SOT	512	Γ_0	✓	✓	25.9 /25.0 (2.5)	55.4 /63.7 (36.1)	86.8 /95.6 (16.2)
SOT	2048	Γ_0	×	×	70.6 /77.6 (31.8)	23.7 /20.0 (30.3)	46.0 /45.0 (36.4)
SOT	2048	{512}	×	✓	97.9 /101.1 (32.5)	14.1 /4.7 (25.5)	28.6 /11.6 (32.6)



Vertical (L_1, L_2) converge smoothly **only** when close to global min.

✓Horizontal SOT has good gradient orientation

✗ High sensitivity to initialization (specially MSS baseline)

✓SOT improves on MSS

✓Larger window size, logarithmic frequency scaling and frequency cutoff → improved metrics

! Uncertainty in the number of harmonics → tricky optimization