Translation-Equivariant SSL For Pitch Estimation with Optimal Transport

Bernardo Torres* Alain Riou* Gaël Richard Geoffroy Peeters

LTCI, Télécom-Paris, Institut Polytechnique de Paris

tl;dr: We replace the equivariance losses of PESTO by an Optimal Transport-based loss

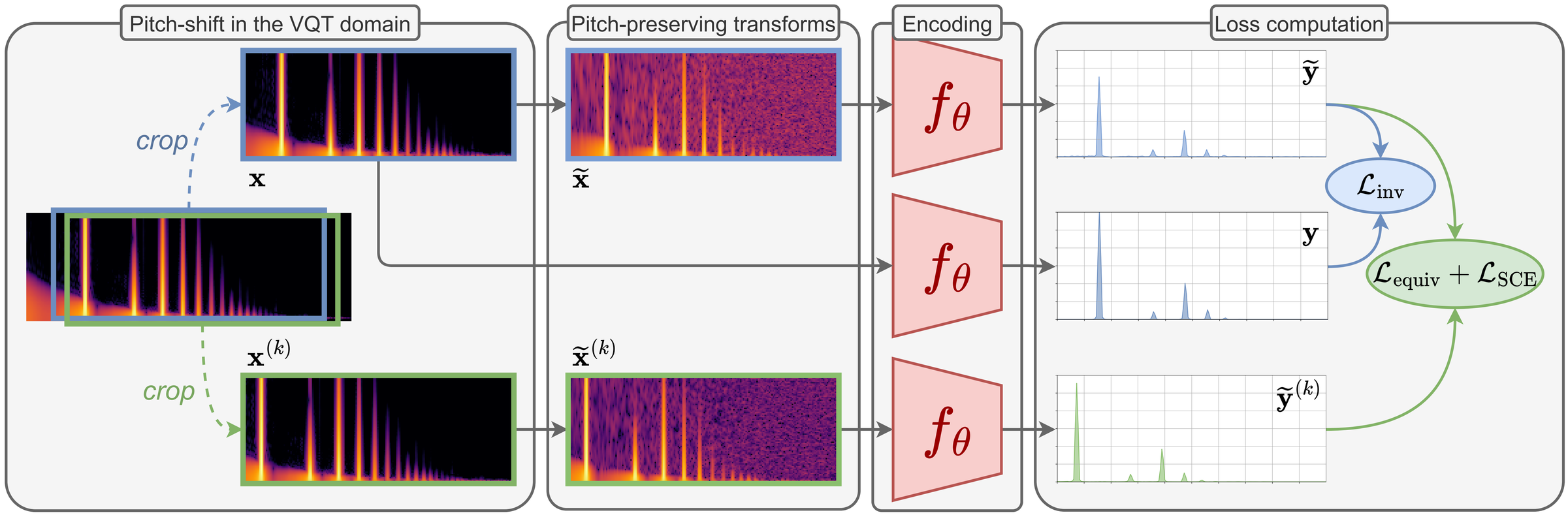

Figure 1. Overview of the PESTO model. We apply both pitch-shifting and pitch-preserving transfoms to individual VQT frames. Then we jointly optimize an **invariance** criterion between frames that have the same pitch, and an **equivariance** criterion between frames that are pitch-shifted.

Background: PESTO (ISMIR '23, TISMIR '25)

Three losses to promote invariance and equivariance:

$$\mathcal{L}_{\text{PESTO}} = \underbrace{\lambda_{\text{inv}}\mathcal{L}_{\text{inv}}}_{\text{invariance}} + \underbrace{\lambda_{\text{equiv}}\mathcal{L}_{\text{equiv}} + \lambda_{\text{SCE}}\mathcal{L}_{\text{SCE}}}_{\text{equivariance}} \tag{1}$$

- **Invariance loss:** Cross-entropy between pitch distributions

$$\mathcal{L}_{\text{inv}}(\mathbf{y}, \widetilde{\mathbf{y}}) = \sum_{i=1}^{K} \tilde{y}_i \log y_i \tag{2}$$

- **Equivariance loss:** Map distributions to a scalar proportional to the fundamental frequency

$$\mathcal{L}_{\text{equiv}}\left(\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}}^{(k)}, k\right) = \left\| \frac{(\alpha, \alpha^2, \ldots, \alpha^K) \cdot \widetilde{\mathbf{y}}^{(k)}}{(\alpha, \alpha^2, \ldots, \alpha^K) \cdot \widetilde{\mathbf{y}}} - \alpha^k \right\| \tag{3}$$

- **Regularization loss:** Shifted cross-entropy between pitch distributions

$$\mathcal{L}_{\text{SCE}}(\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}}^{(k)}, k) = \sum_{i=1}^{K} \tilde{y}_{i+k}^{(k)} \log \tilde{y}_i \tag{4}$$

Experimental results

		MIR-1K		MDB		PTDB	
		RPA	RCA	RPA	RCA	RPA	RCA
PESTO	MIR-1K	97.7	98.0	94.8	95.9	87.7	90.3
	MDB	94.6	96.1	97.0	97.1	88.3	89.9
	PTDB	95.6	96.9	96.3	96.6	89.7	91.2
PESTO-OT	MIR-1K	97.8	98.1	86.6	95.1	88.0	90.1
	MDB	91.6	94.0	95.3	95.6	88.3	89.8
	PTDB	86.4	92.9	93.7	94.5	89.0	90.8

Table 1. Comparison of our model with the PESTO model on different datasets. Rows and columns correspond to training and evaluation sets, respectively.

Our new loss based on Optimal Transport

Only one loss for invariance and one for equivariance:

$$\mathcal{L}_{\text{PESTO-OT}} = \underbrace{\lambda_{\text{inv}}\mathcal{L}_{\text{inv}}}_{\text{invariance}} + \underbrace{\lambda_{\text{OT}}\mathcal{L}_{\text{OT}}}_{\text{equivariance}} \tag{5}$$

- **Invariance loss:** same as before
- **Optimal Transport loss:** Wasserstein distance between pitch distributions

$$\mathcal{L}_{\text{OT}}\left(\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}}^{(k)}, k\right) = \mathcal{W}_2\left(\widetilde{\mathbf{y}}, \tau_{-k}(\widetilde{\mathbf{y}}^{(k)})\right). \tag{6}$$

where
- $\mathcal{W}_2 : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ is the 2-Wasserstein distance
- $\tau_b : \mathbb{R}^K \to \mathbb{R}^K$ shifts the distribution by b bins

Optimal Transport has nice theoretical properties

- **Symmetry:** $\mathcal{L}_{\text{OT}}(y_1, y_2, k) = \mathcal{L}_{\text{OT}}(y_2, y_1, -k)$.
- **Invariance:** $\mathcal{W}_p(y_1, y_2) = \mathcal{W}_p(\tau_k(y_1), \tau_k(y_2))$.
- **Linear scaling under τ_k:** $\mathcal{W}_2(y, \tau_k(y)) \propto |k|$.
- **Stability:** $\mathcal{L}_{\text{OT}}$ avoids the large floating-point powers α^i that can cause numerical instability in $\mathcal{L}_{\text{equiv}}$.

Take-aways

- Not as delicious as the original PESTO, but promising proof-of-concept
- Wasserstein distance is a good alternative to cross-entropy when "distance" between classes matters
- Opens up new possibilities: fine-tuning, circular optimal transport...

This LBD

Our new TISMIR!