– PESTO –

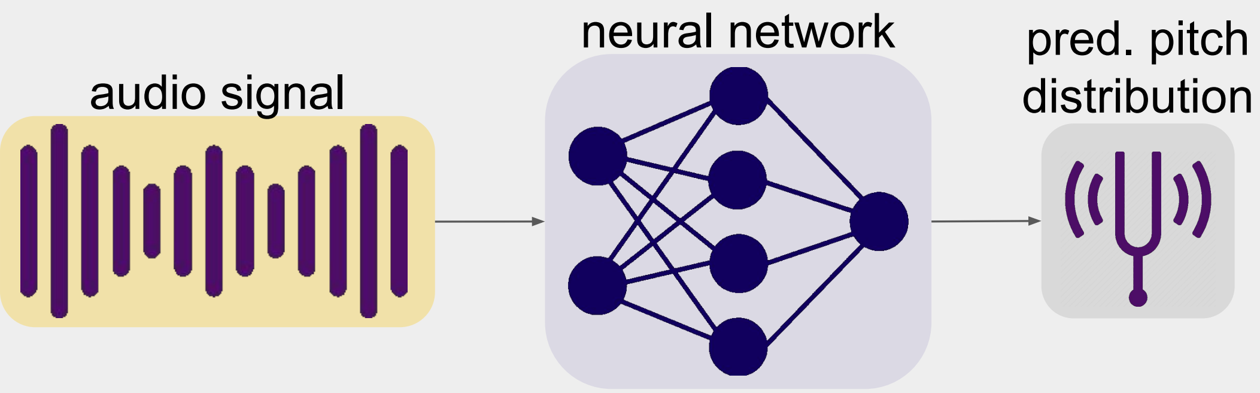
Pitch Estimation with Self-supervised Transposition-equivariant Objective

Alain Riou[1,2], Stefan Lattner[2], Gaëtan Hadjeres[3], Geoffroy Peeters[1]

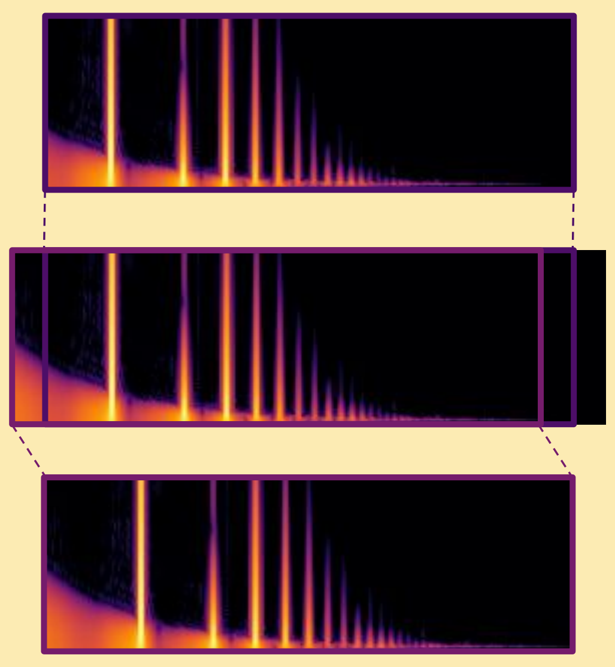
[1]Télécom Paris, [2]Sony CSL, [3]Sony AI

Pitch estimation **without annotations**

- Pitch estimation as a **classification** problem
- **SSL approach: no labels** required
- Compatible with music styles for which no **annotated** examples



CQT as a proxy for **pitch-shift**



- We compute the **CQT** of the input signal
- CQT's frequency scale is **logarithmic**
  ➔ translation = pitch-shift
- One pitch prediction per frame
  ➔ prediction resolution = CQT hop size
- Originally introduced in **SPICE**

Transposition-equivariant objective

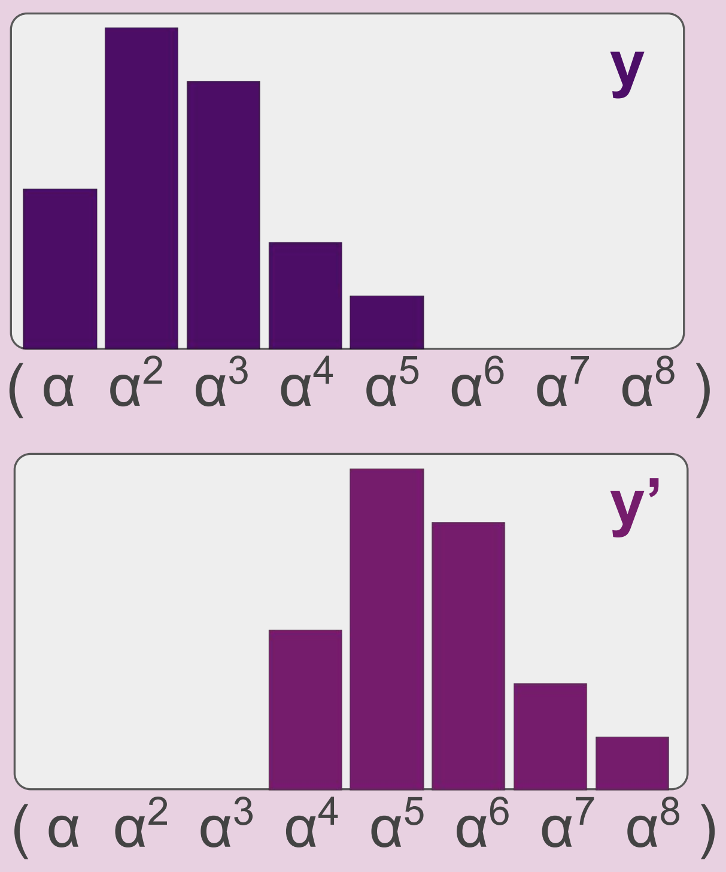
- Define **a** = (α, α², …, αᵈ)ᵀ, α > 0.
- Let **y**, **y'** ∈ [0,1]ᵈ be two distributions. If **y** and **y'** are equal up to a shift of *k*

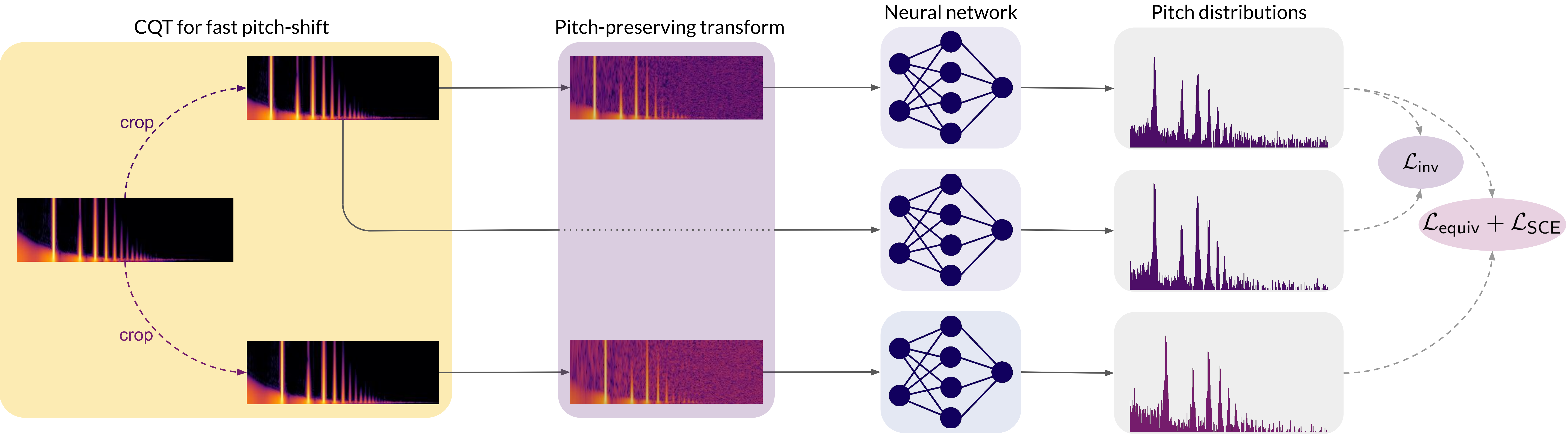
$$\mathbf{a}^\top \mathbf{y}' = \alpha^k \mathbf{a}^\top \mathbf{y}$$

- Hence our **equivariance** loss:

$$\mathcal{L}_{\text{equiv}}(\mathbf{y}, \mathbf{y}', k) = \left\| \frac{\mathbf{a}^\top \mathbf{y}'}{\mathbf{a}^\top \mathbf{y}} - \alpha^k \right\|$$

- This loss is **null** when **y'** is a **translation** of **y**
- As a regularization, we also minimize the **shifted cross-entropy** $\mathcal{L}_{\text{SCE}}$ between **y** and **y'** translated by *k* bins




CQT for fast pitch-shift
Pitch-preserving transform
Neural network
Pitch distributions
crop
crop
$\mathcal{L}_{\text{inv}}$
$\mathcal{L}_{\text{equiv}} + \mathcal{L}_{\text{SCE}}$

Pitch-preserving transforms for improving robustness

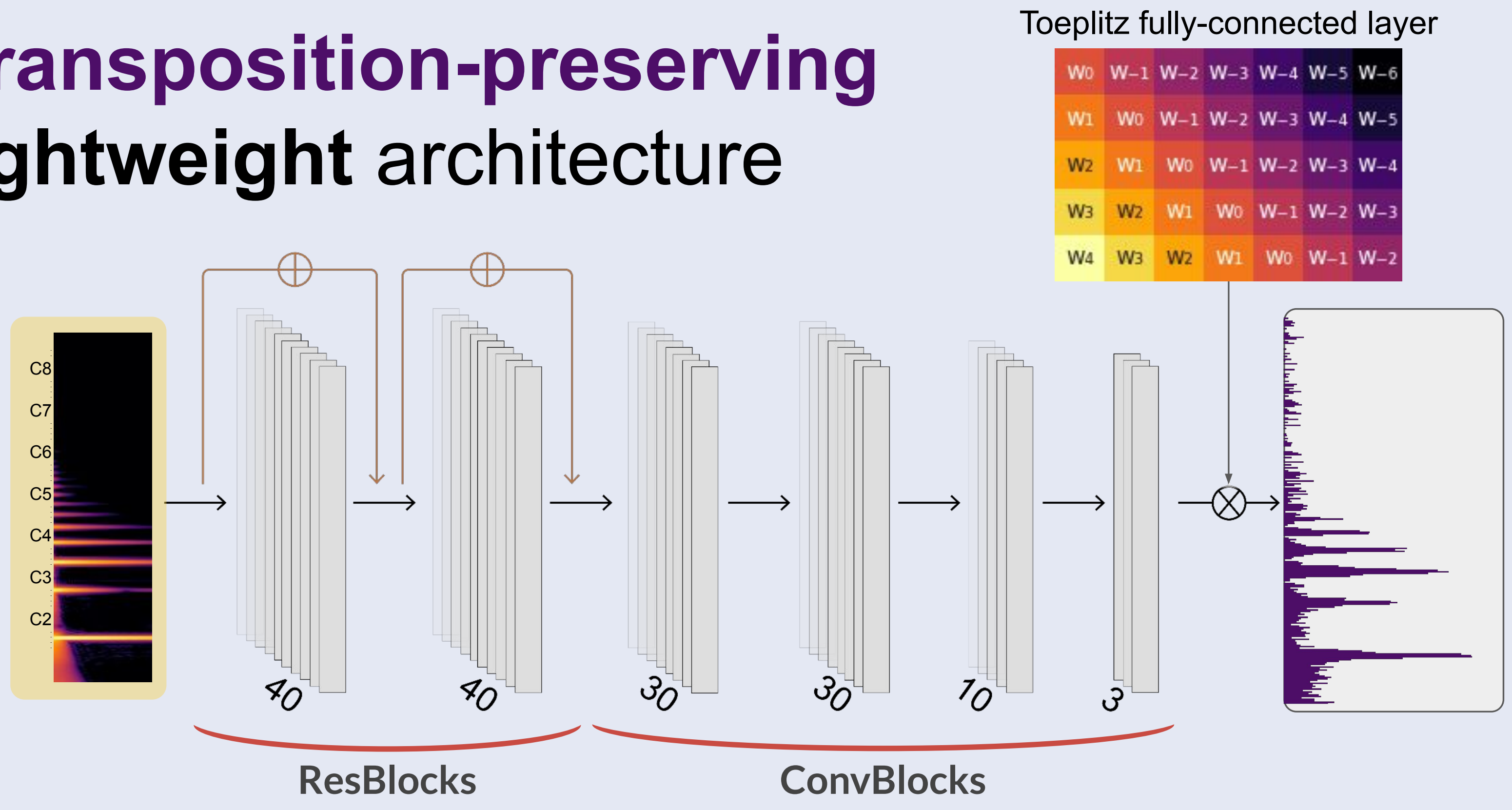
- **Pitch-preserving transforms** are applied to the signals for the model to see audios with **same pitch** but **different timbre**
- The model aims to minimize the **cross-entropy** between distributions of audios that share the same pitch
- When possible, mixing **background music** with different SNR makes the model more **robust**

Experimental results

			Raw Pitch Accuracy	
Model	# params	Trained on	*MIR-1K*	*MDB-stem-synth*
SPICE [19]	2.38M	private data	90.6%	89.1%
DDSP-inv [45]	-	*MIR-1K / MDB-stem-synth*	91.8%	88.5%
PESTO (ours)	28.9k	*MIR-1K*	**96.1%**	94.6%
PESTO (ours)	28.9k	*MDB-stem-synth*	93.5%	**95.5%**
CREPE [16]	22.2M	many (supervised)	**97.8%**	**96.7%**

- Trained on ***MIR-1K*** or ***MDB-stem-synth***
- Strong **generalization** performances
- **Outperforms** SSL baselines even in the cross-dataset scenario
- Much **more lightweight** and **faster** than **CREPE**
- Equivariance loss and Toeplitz fully-connected layer are **crucial**

Transposition-preserving **lightweight** architecture



W0	W−1	W−2	W−3	W−4	W−5	W−6
W1	W0	W−1	W−2	W−3	W−4	W−5
W2	W1	W0	W−1	W−2	W−3	W−4
W3	W2	W1	W0	W−1	W−2	W−3
W4	W3	W2	W1	W0	W−1	W−2

- The architecture is mostly **1d convolutions** and **elementwise operations**
- Thanks to the **Toeplitz** linear layer, translations are completely preserved
  ➔ If the CQT is shifted, then the probability density is shifted accordingly
- Overall architecture has less than **30k parameters!**

Conclusion

- **SOTA** in self-supervised pitch estimation
- Can be trained on **any audio**: suited for non-Western music
- 12x faster than **real-time** on CPU
- **Code** and **pretrained models** available online
- Pip-installable package: `pip install pesto-pitch`

Paper

Code

Sony CSL

SONY

ADASP
audio data analysis signal processing

TELECOM Paris

IP PARIS