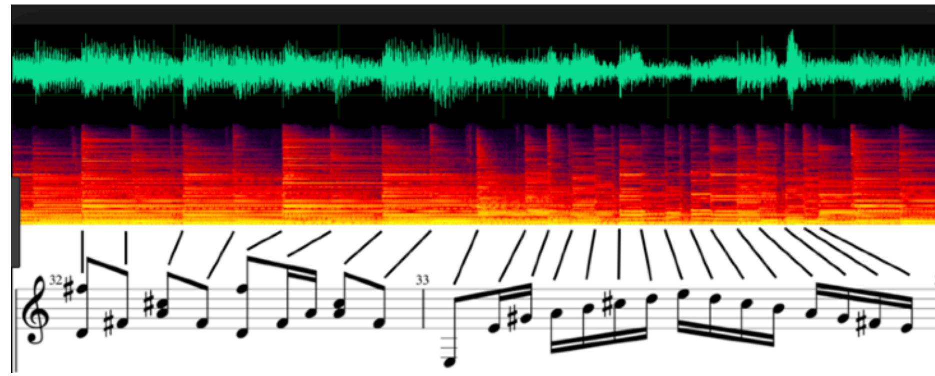
2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

Introduction

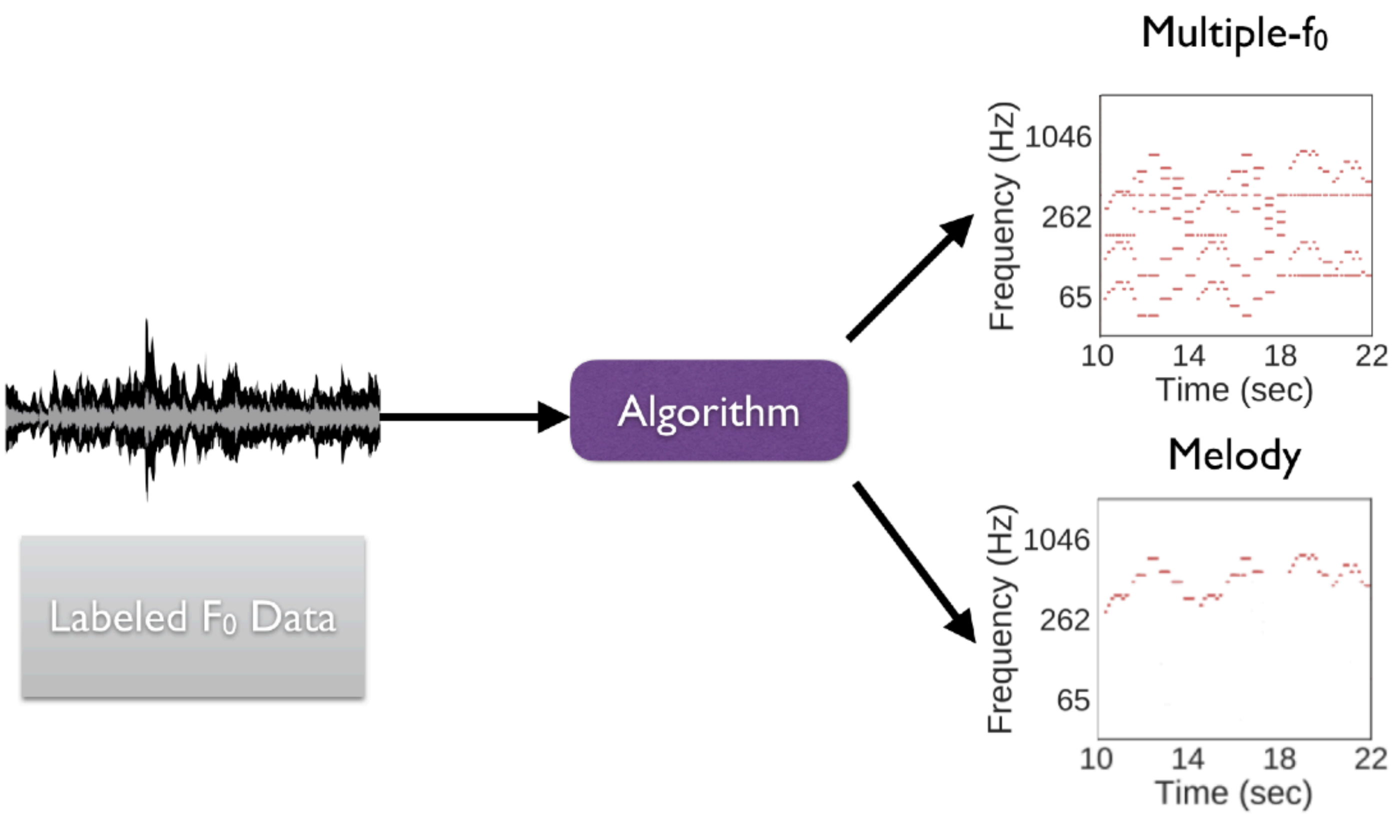
- **Goal** :
  - obtain a transcription of the content of an audio signal
    - which notes ? which chords ? which instruments ?
    - equivalent to Automatic Speech Recognition
  - difficulty :
    - all instruments are super-imposed in the audio signal
  - sub-tasks :
    - only transcribe the notes / f0 (in Hz)
    - only transcribe the dominant melody
- **How can we do that ?**
  - many methods proposed in the past : sinusoidal model, NMF, PLCA, ... ,
  - large improvement since 2017 using ... deep learning



[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

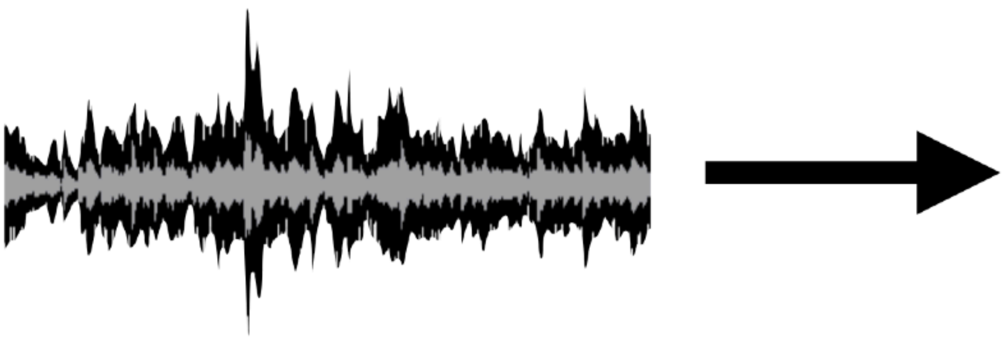
Introduction

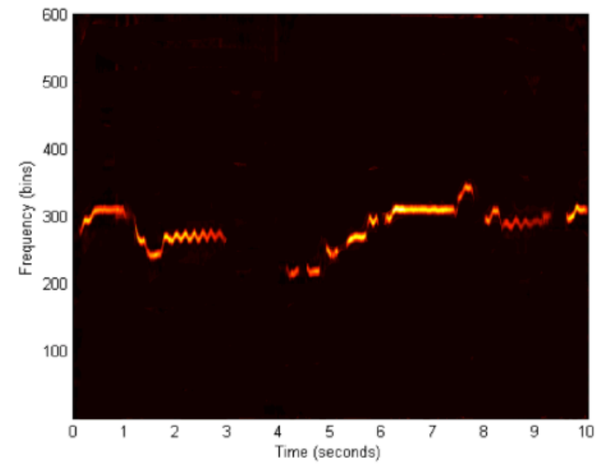


[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

Learning a "Saliency" Representation

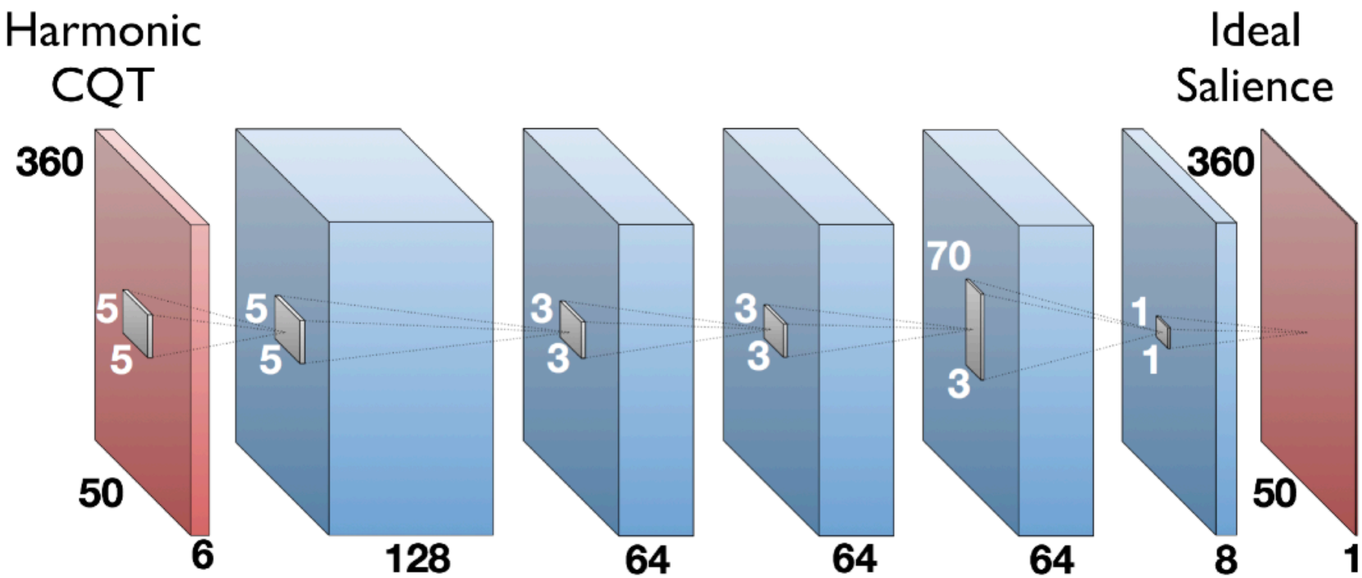




[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

Using Convolutional Neural Network

- **Inputs** :
  - magnitude spectrogram **with depth** (Harmonic CQT)
- **Outputs** :
  - image of the same size of input but with only saliency information

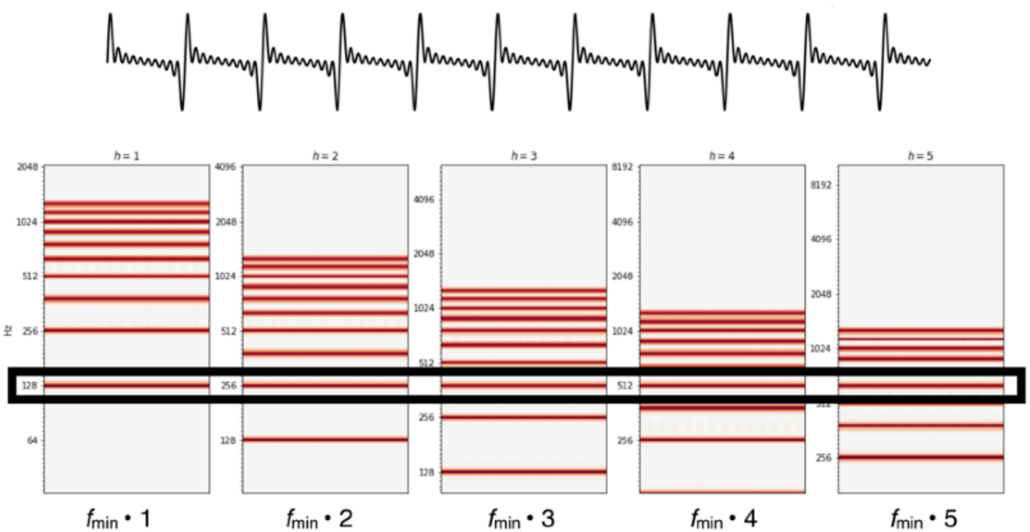


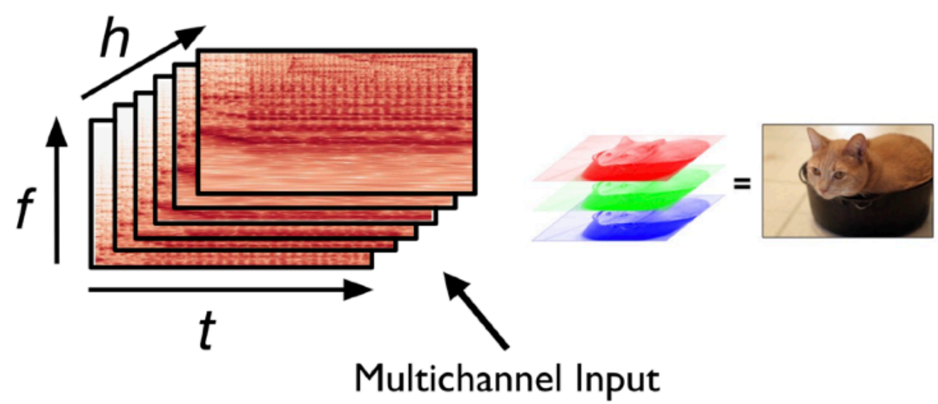
$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

Input Representation: a Harmonic CQT

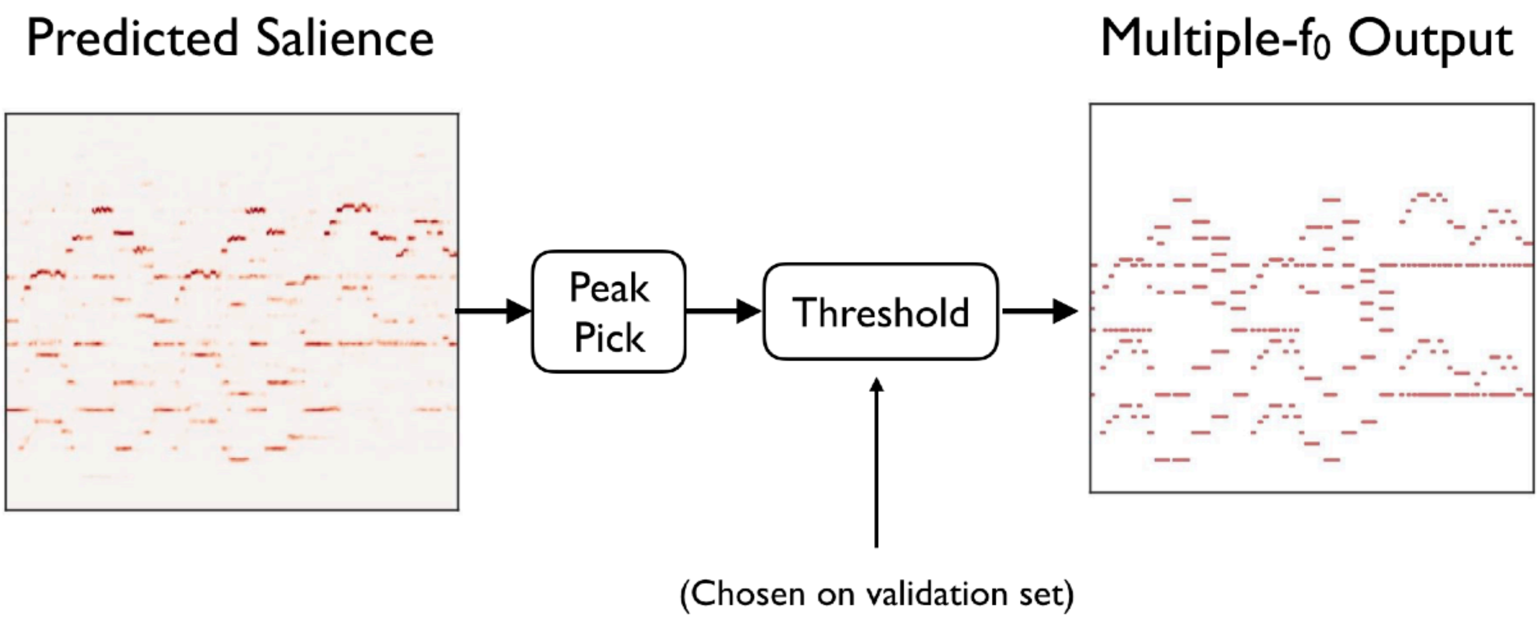




[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

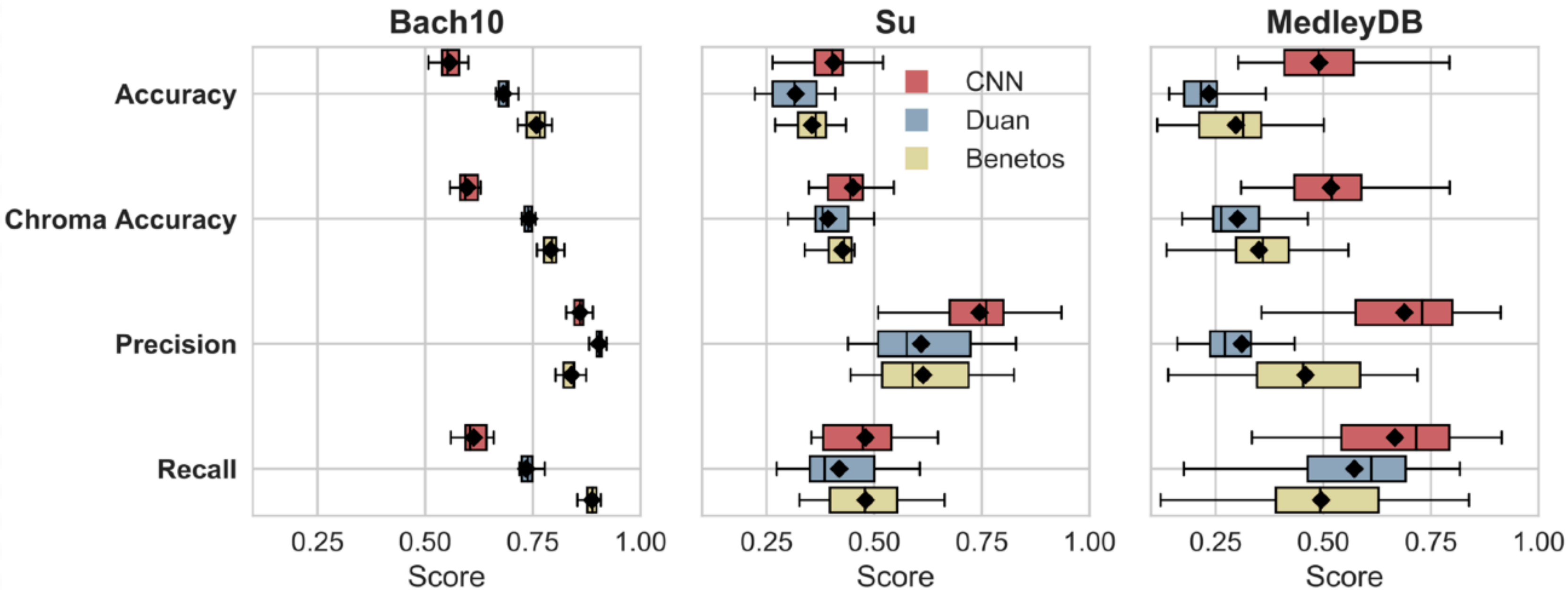
2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

Multiple-f_0 Estimation



[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

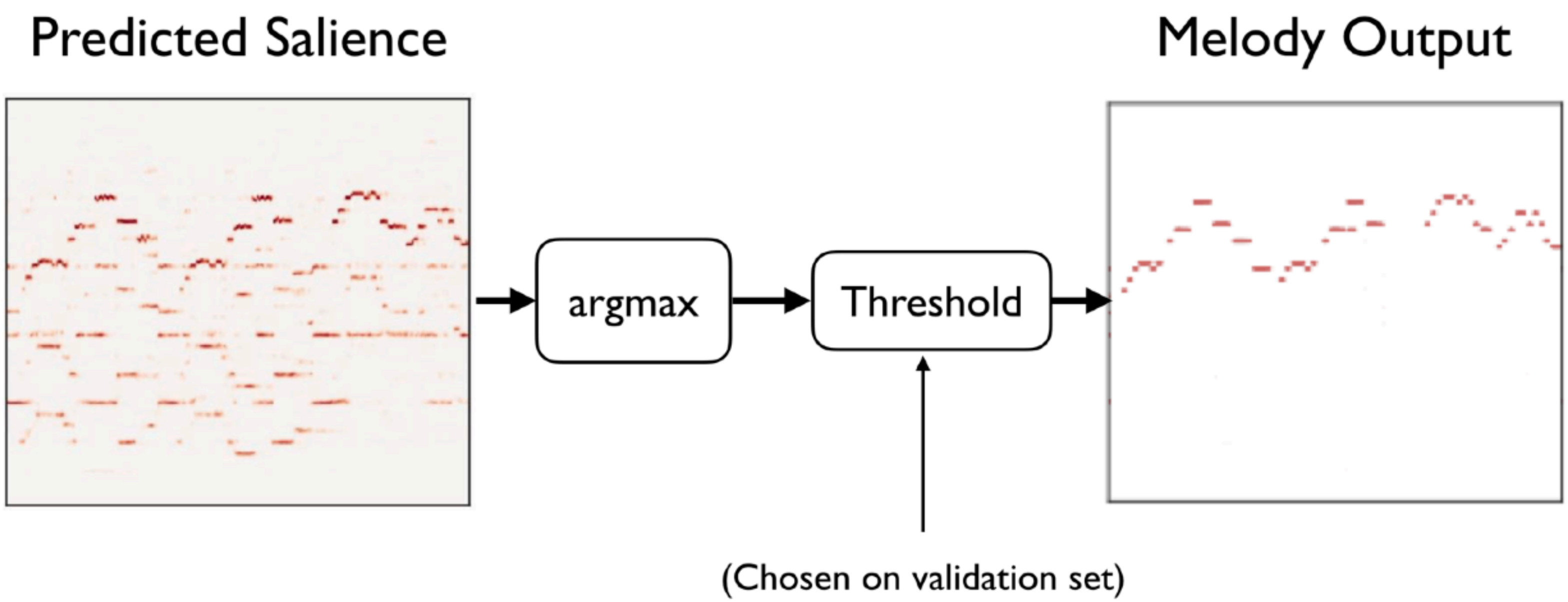
Multiple-f_0 Estimation: Results



[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

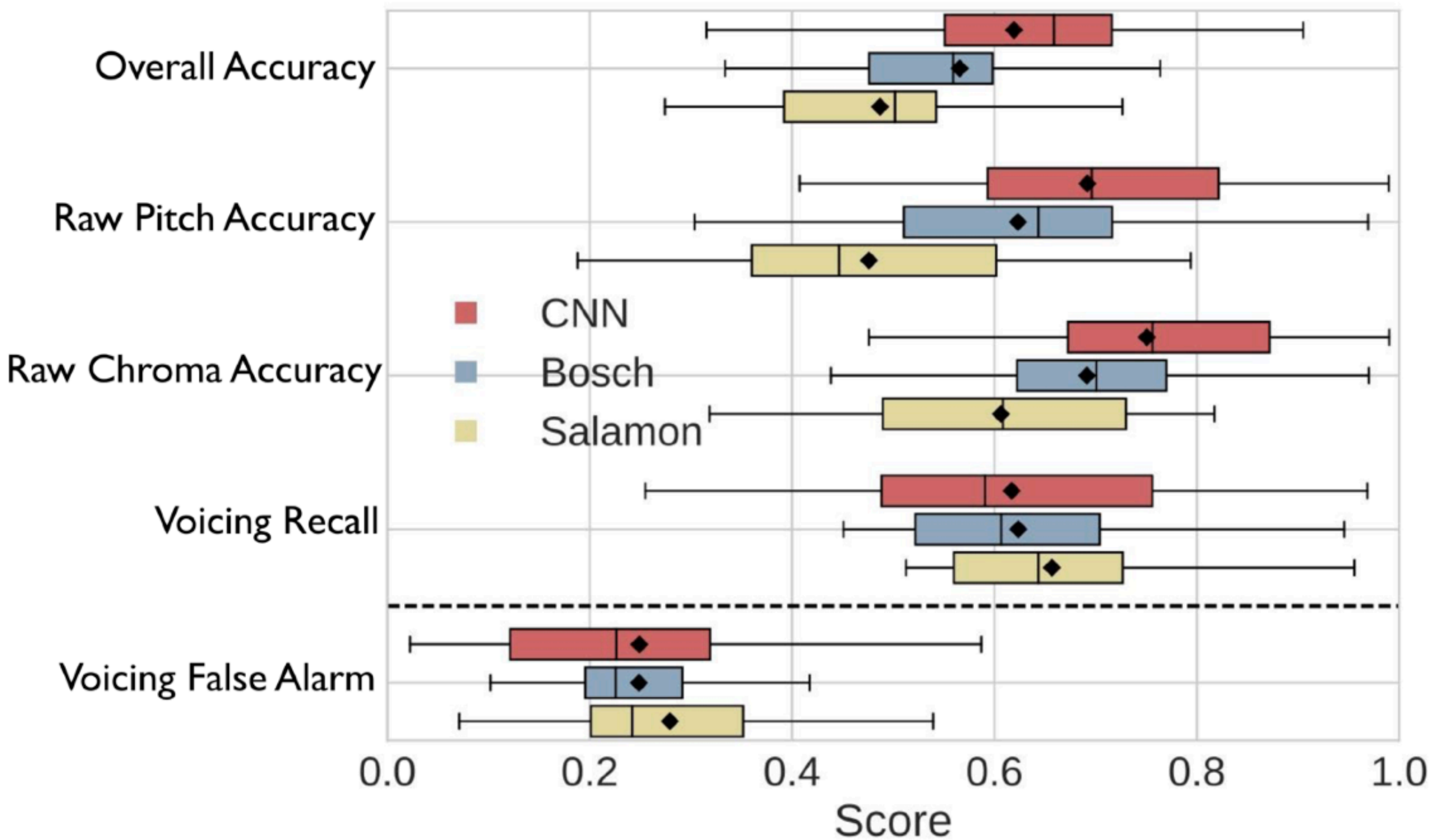
Dominant Melody Estimation



[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

2017 → Music Multiple-f_0 Estimation using Harmonic-CQT

Dominant Melody Estimation: Results

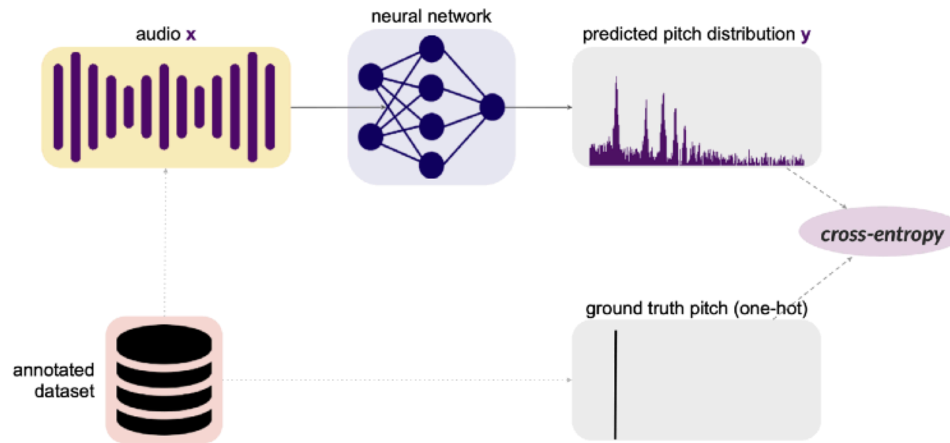


[R. Bittner et al. "**Deep salience representations for f0 estimation in polyphonic music**", in ISMIR, 2017] LINK

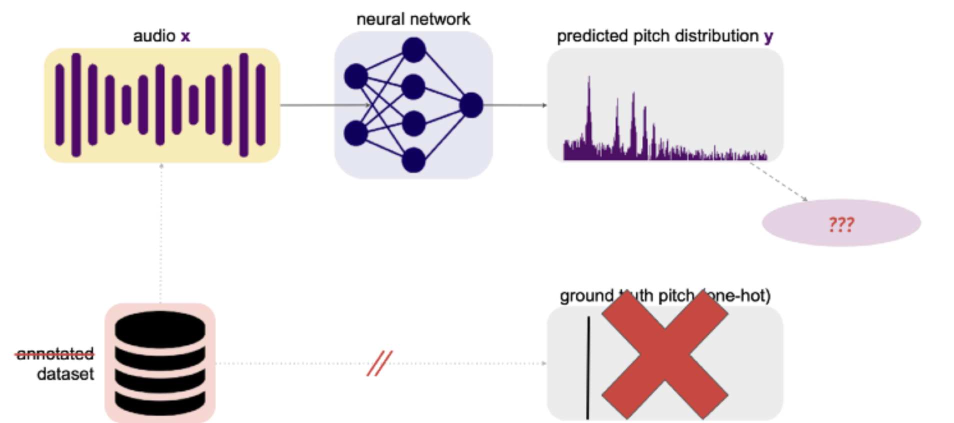
2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

Introduction

- **Traditional** way of training a neural network
    - **Supervised** training (compare prediction to ground-truth)



- **Our goal**:
    - train a neural network for pitch estimation **without** ground-truth annotations
    - = **Self-Supervised-Learning** (SSL)

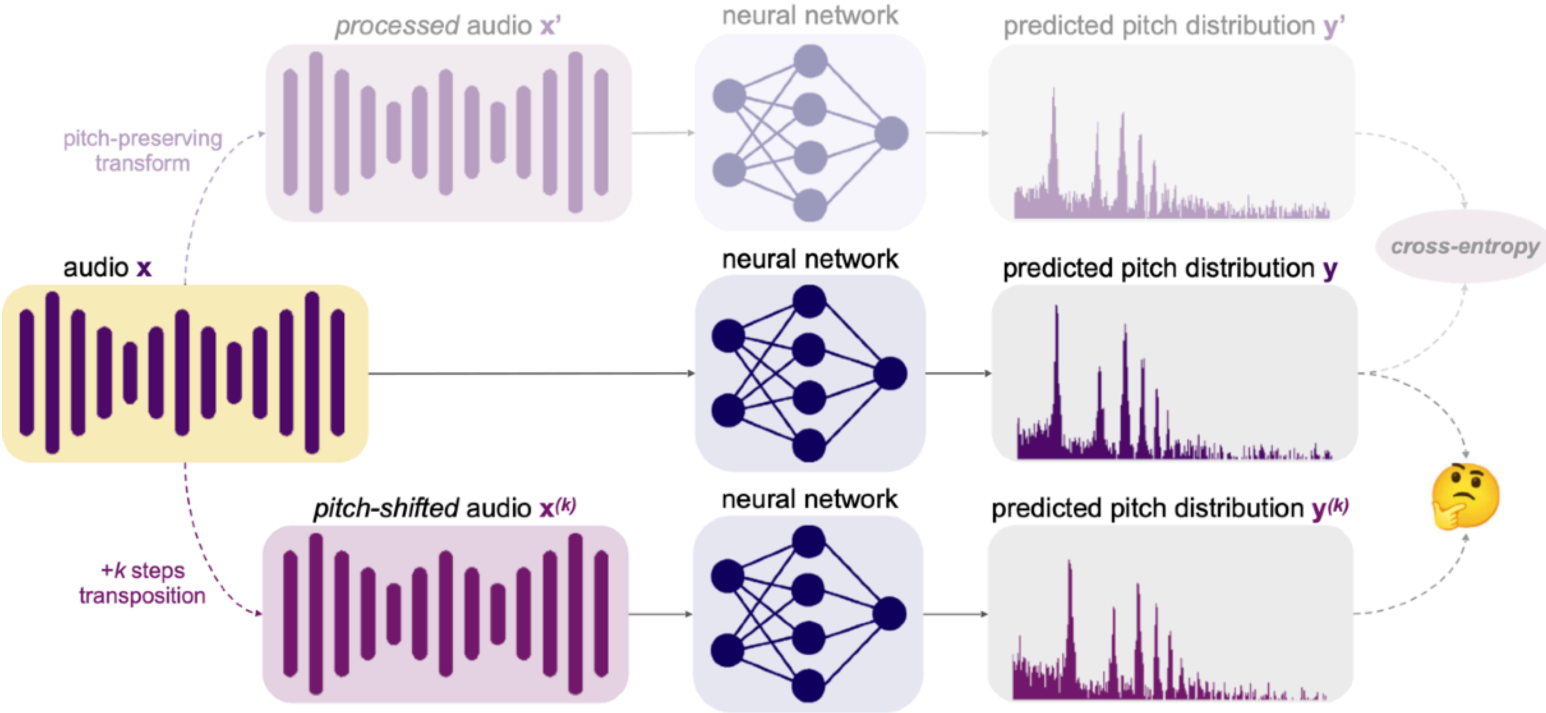


[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

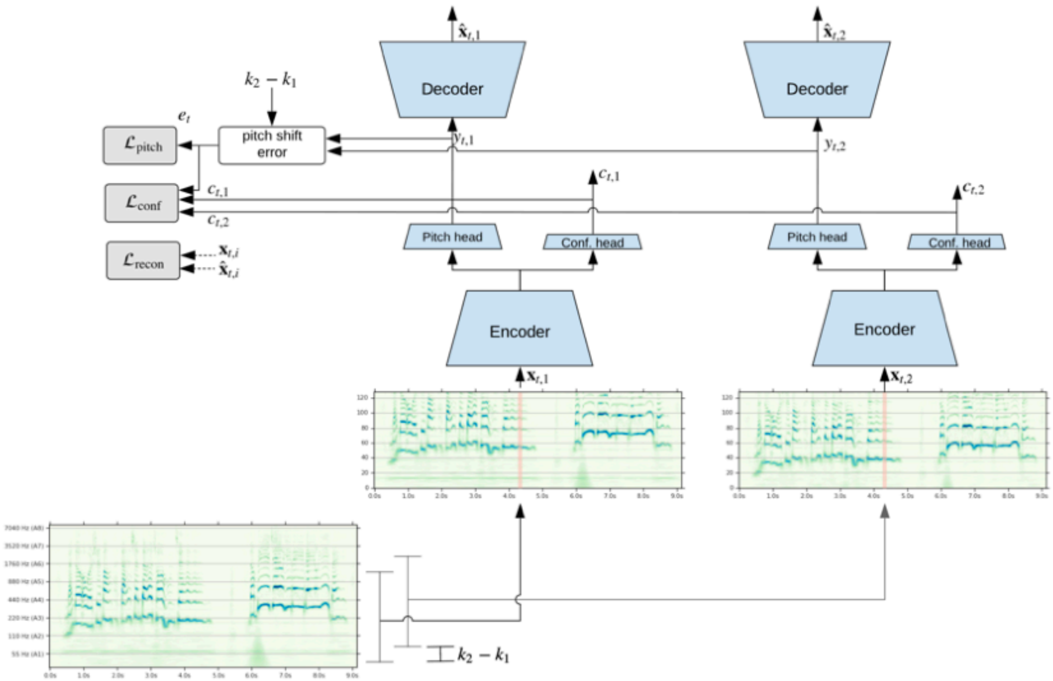
Self-Supervised-Learning

- **Siamese Network, SimCLR,**
  - given two views x, x' (augmented versions) of the same data x trains a neural network such that the corresponding outputs y, y' are similar ⇒ feature learning
  - can we make the output equivariant instead of invariant ?



[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK

2019 → SPICE: Self-supervised Pitch Estimation



equivariance loss

$$e_t = |(y_{t,1} - y_{t,2}) - \sigma(k_{t,1} - k_{t,2})|$$

with

$$\mathcal{L}_{pitch} = \frac{1}{T}\sum_t h(e_t)$$

$$h(x) = x^2/2 \text{ if } |x| \leq \tau$$

$$h(x) = \tau^2/2 + \tau(|x| - \tau) \text{ if } |x| \leq \tau$$

$$\sigma = \frac{1}{B[\log_2(f_{max}/f_{min})]}$$

reconstruction loss

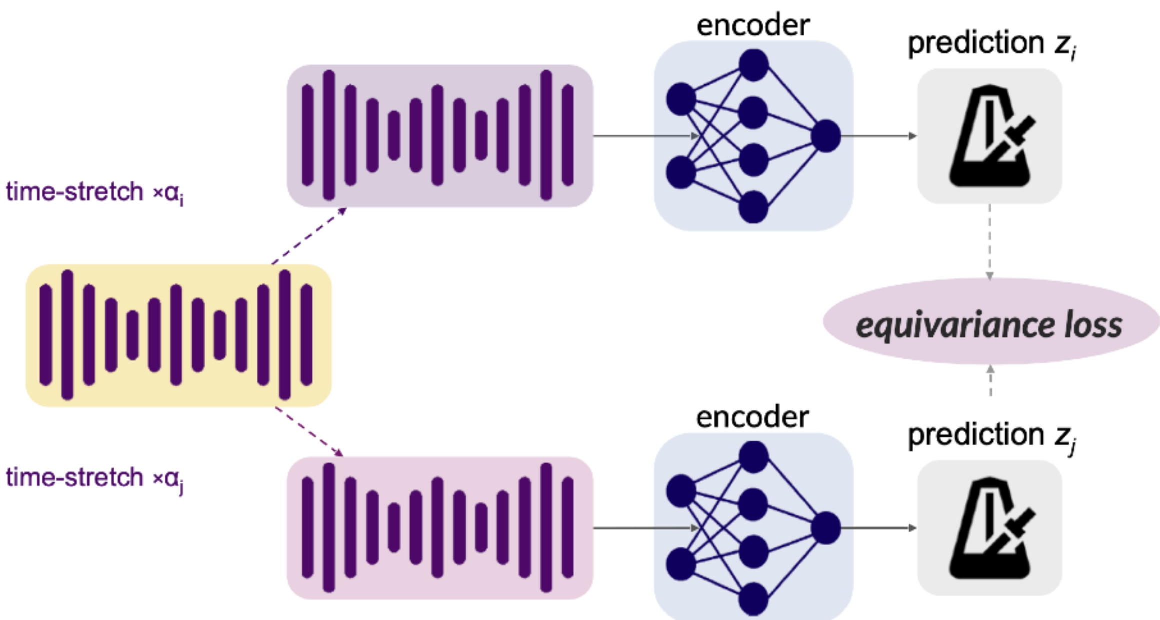
$$\mathcal{L}_{recon} = \frac{1}{T}\sum_t ||x_{t,1} - \hat{x}_{t,1}||_2^2 + ||x_{t,2} - \hat{x}_{t,2}||_2^2$$

total loss

$$\mathcal{L} = w_{pitch}\mathcal{L}_{pitch} + w_{recon}\mathcal{L}_{recon}$$

[B. Gfeller et al., "SPICE: Self-supervised pitch estimation" in . IEEE/ACM TASLP, 28:1118–1128, 2020.] LINK

2022 → Equivariant Self-Supervision for Tempo Estimation



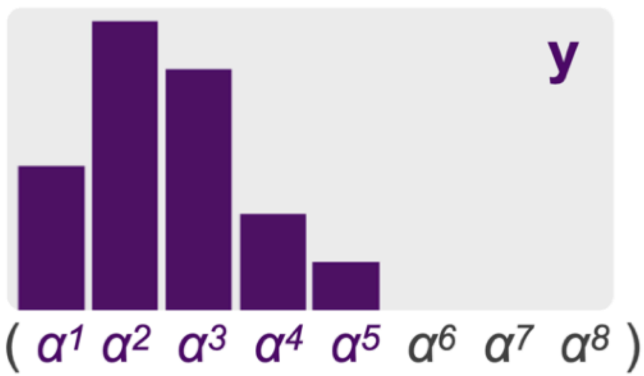
$$\mathcal{L} = \left| \frac{z_i}{z_j} - \frac{\alpha_i}{\alpha_j} \right|$$

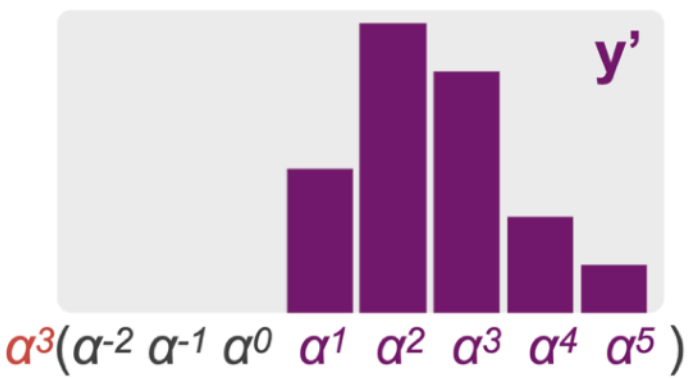
[E. Quinton, "**Equivariant Self-Supervision for Musical Tempo Estimation**", ISMIR, 2022] LINK

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

Self-Supervised-Learning

- **Our proposal: Class-based equivariance loss**
  - scalar product between the softmax output by a geometric series α





- Define **a** = $(\alpha, \alpha^2, \ldots, \alpha^d)^\top$, $\alpha > 0$
- Compute the **scalar** products $\mathbf{a}^\top\mathbf{y}$ and $\mathbf{a}^\top\mathbf{y}'$
- If **y** and **y'** are equal up to a shift of k, then

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

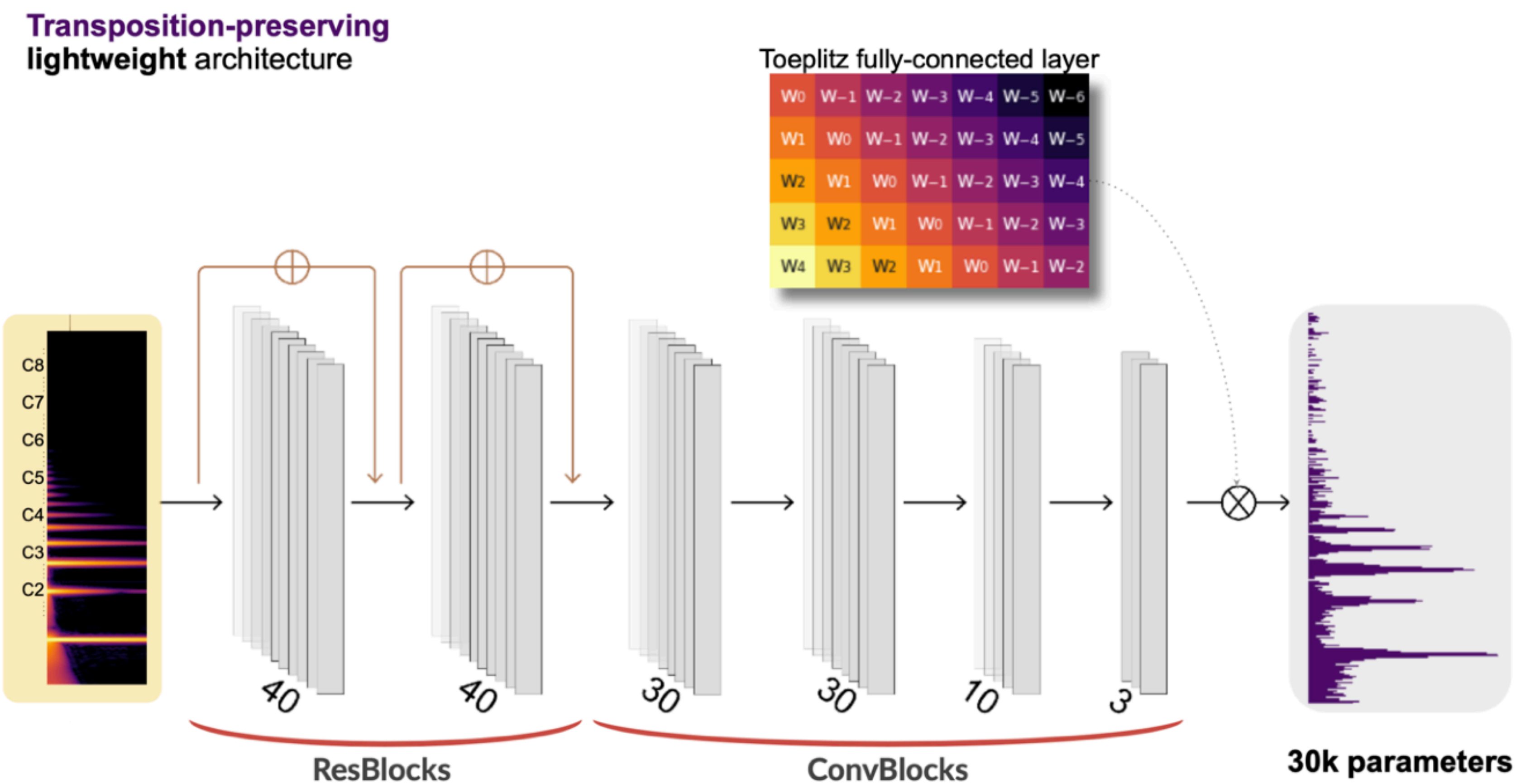
- **Equivariance** loss:

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

- translation of k between **y** and **y'** $\boxed{?}$ equivariance loss = 0

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

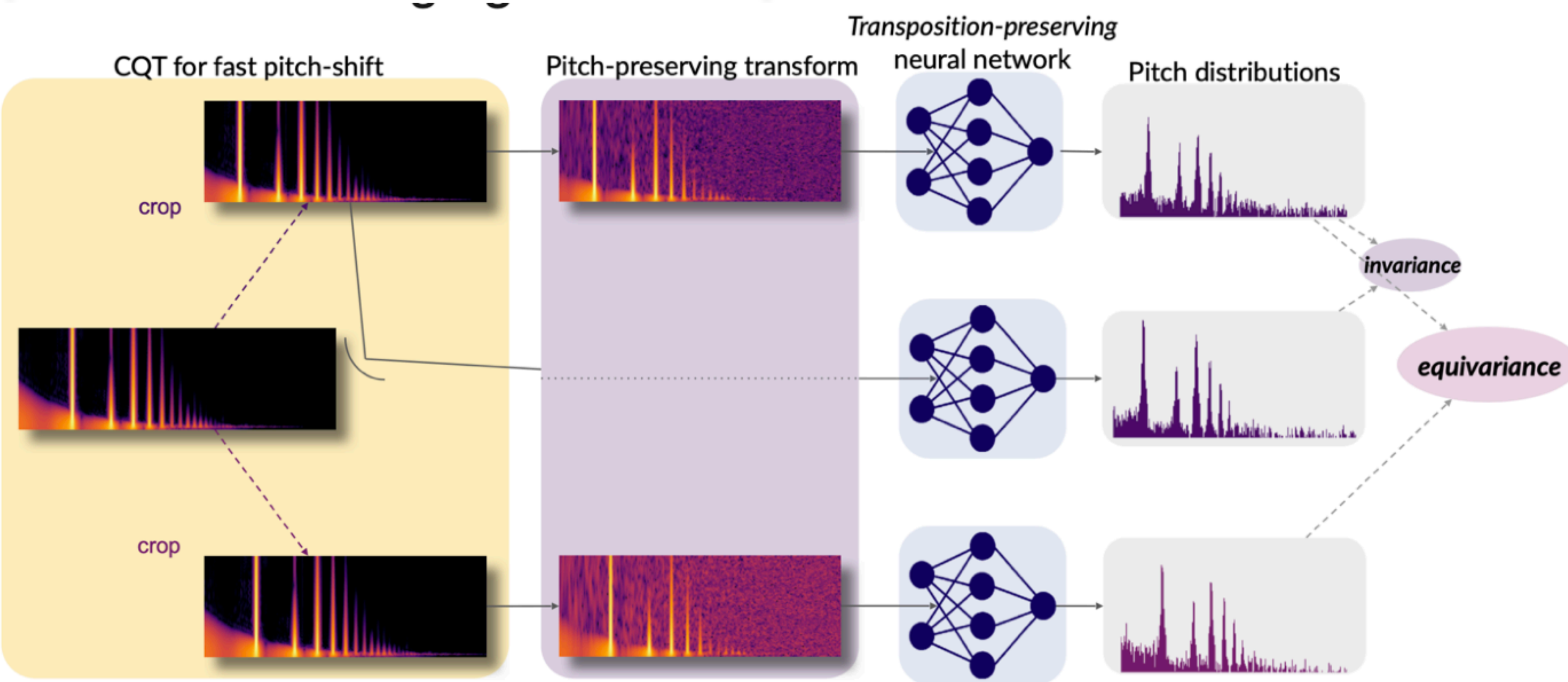
Lightweight architecture



[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

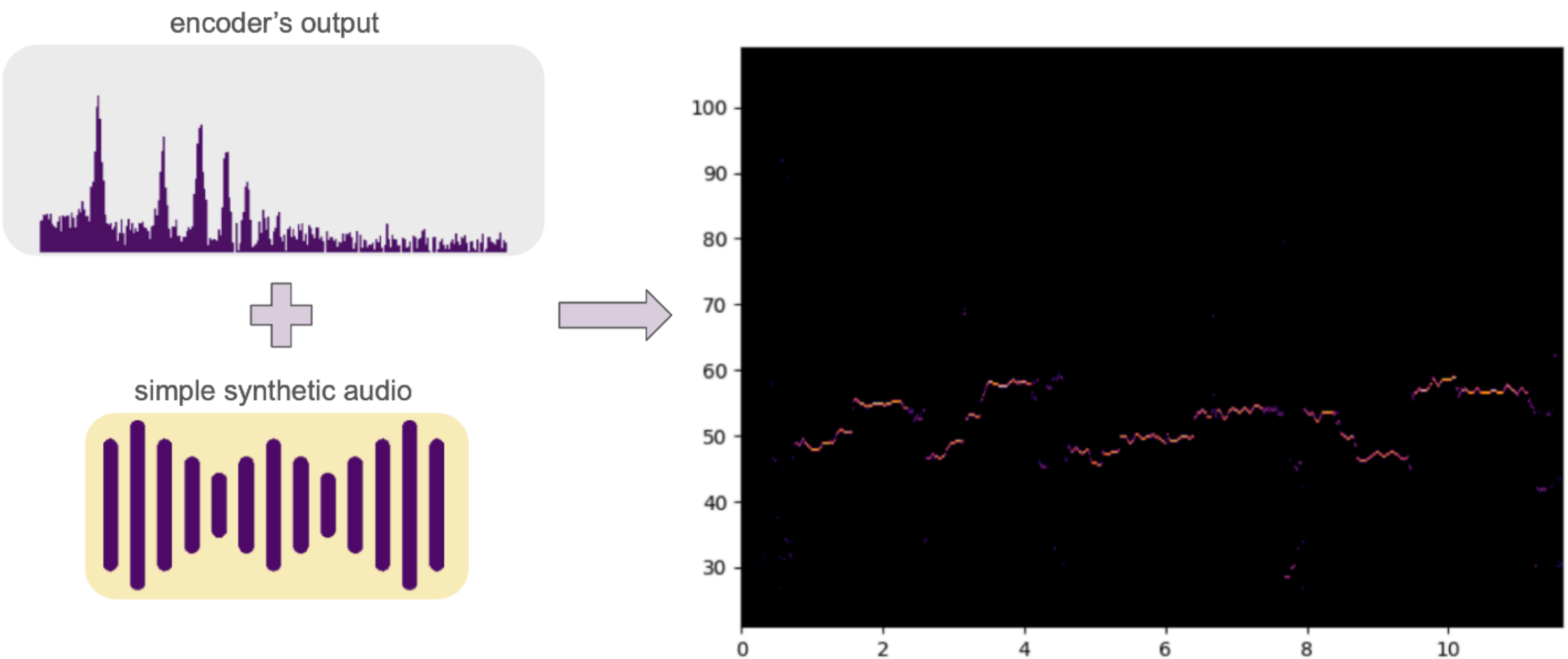
Training data



[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

From relative pitch to absolute pitch



[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK

2023 → PESTO: Self-Supervised Pitch Estimation: Equi-variance

Evaluation

Model	# params	Trained on	Raw Pitch Accuracy *MIR-1K*	Raw Pitch Accuracy *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%**

- ➔ SOTA in self-supervised pitch estimation
- ➔ Strong generalization performances
- ➔ Extremely lightweight model

[A. Riou, St. Lattner, G. Hadjeres, G. Peeters, "**PESTO, Pitch Estimation with Self-supervised Transposition-equivariant**", ISMIR, 2023] LINK