A Model You Can Hear: Audio Identification with Playable Prototypes

Romain Loiseau[1, 2] Baptiste Bouvier[3] Yann Teytaut[3]
Elliot Vincent[1, 4] Mathieu Aubry[1] Loïc Landrieu[2]

[1]LIGM, ENPC [2]LASTIG, IGN [3]STMS, IRCAM [4]Inria and DIENS

romainloiseau.fr/a-model-you-can-hear/

ISMIR 2022 BENGALURU

Motivation

Context:

- Recent methods often rely on representations in high-dimensional abstract space. Those methods perform well but are difficult to interpret.

Contributions:

- We adapt the **transformation-invariant clustering** paradigm for audio in both supervised and unsupervised settings.
- We provide an audio identification model based on **prototypical sounds** that can be heard directly.
- Our model reaches **state-of-the-art** results for **audio classification and clustering** tasks while remaining easily **interpretable**.

Method

For each **input audio clip** x of **label** y characterized by its **log-spectrogram** with T time steps and F frequency bins, we define the following reconstruction model and losses for the k-th **prototype**.

Deep Transformation-Invariant Prototyping:

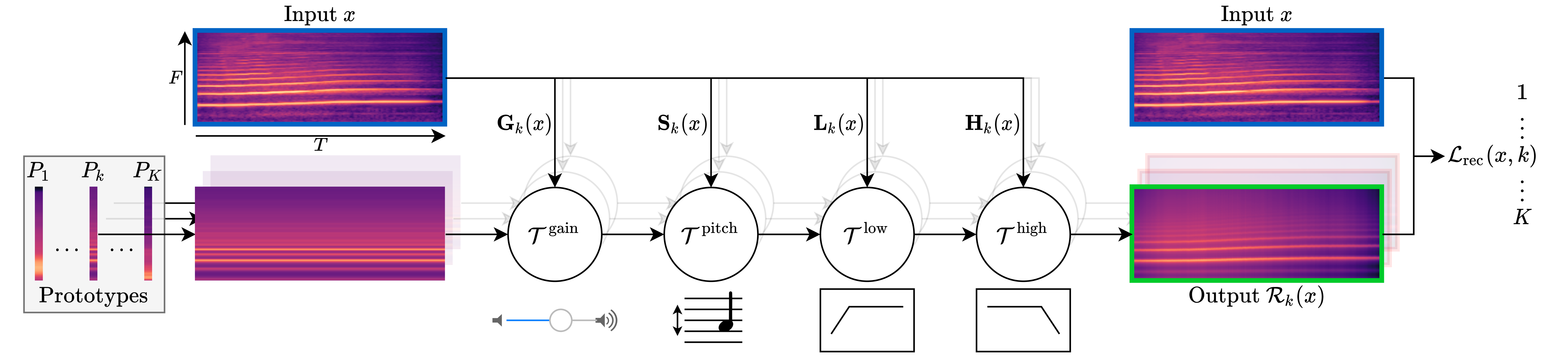
$$\mathcal{R}_k(x)[t] = \mathcal{T}^{\text{high}}_{\mathbf{H}_k(x)[t]} \circ \mathcal{T}^{\text{low}}_{\mathbf{L}_k(x)[t]} \circ \mathcal{T}^{\text{pitch}}_{\mathbf{S}_k(x)[t]} \circ \mathcal{T}^{\text{gain}}_{\mathbf{G}_k(x)[t]}(P_k)$$

Supervising Reconstruction and Cross-Entropy:

$$\mathcal{L}_{\text{rec}}(x, k) = \frac{1}{T} \sum_{t=1}^{T} \left\| x[t] - \mathcal{R}_k(x)[t] \right\|^2$$

$$\mathcal{L}_{\text{ce}}(x, y) = -\log \left(\exp\left(-\beta \mathcal{L}_{\text{rec}}(x, y)\right) / \sum_{k=1}^{K} \exp\left(-\beta \mathcal{L}_{\text{rec}}(x, k)\right) \right)$$

Method Overview



Given an **input sound**, we predict for each prototype a *gain*, a *pitch* shift, as well as *low* and *high frequency filters* at each timestamp to generate the **output**. Prototypes and transformations are learned jointly using a reconstruction loss in either a supervised or unsupervised setting.

Audio Identification Results

	OA	AA
SOL [1,2]		
Autoencoder + K-means	28.7	12.3
† APNet **[7]** + K-means	**37.3**	**18.2**
Ours w/o supervision	34.5	15.4
LibriSpeech [6]		
Autoencoder + K-means	11.0	11.1
† APNet **[7]** + K-means	36.3	36.4
Ours w/o supervision	**48.6**	**49.5**

Clustering Results. Clustering performances on the test sets.

†: Note that APNet requires labels at training time.

$$\mathcal{L}_{\text{clustering}}(x) = \min_{k=1}^{K} \mathcal{L}_{\text{rec}}(x, k)$$

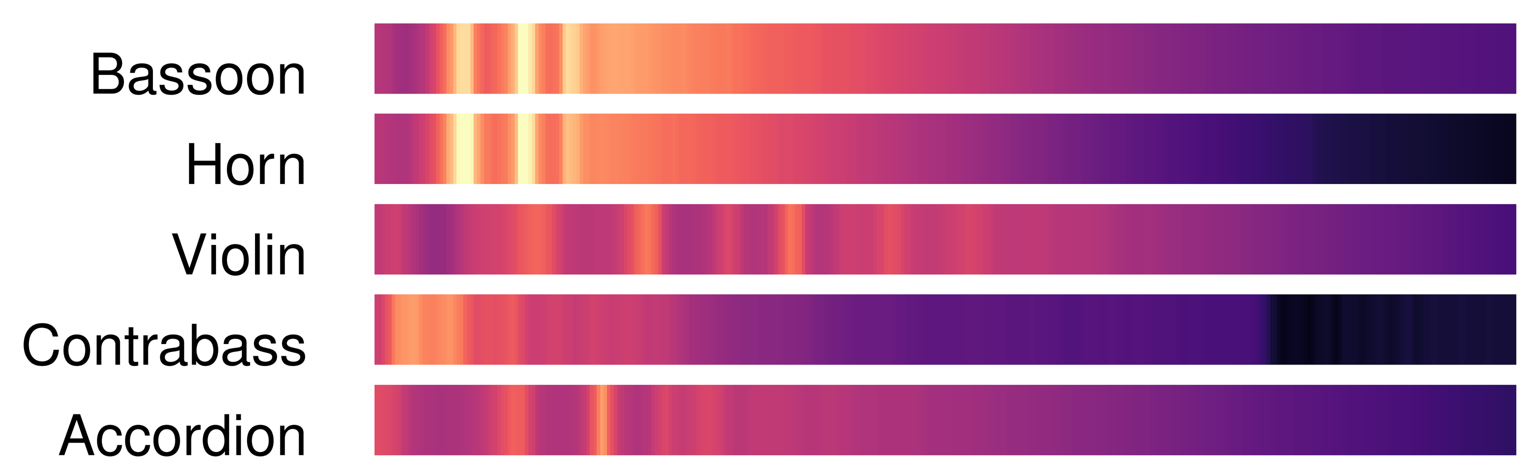
	OA	AA	$\mathcal{L}_{\text{rec}}$
SOL [1,2]			
Direct Classification	97.8	94.8	—
APNet **[7]**	95.3	91.3	**0.1**
Ours w supervision	**99.3**	**95.8**	2.6
LibriSpeech [6]			
Direct Classification	99.4	99.5	—
APNet **[7]**	97.8	97.8	**0.2**
Ours w supervision	**99.9**	**99.9**	2.6

Classification Results. Accuracy and reconstruction error.

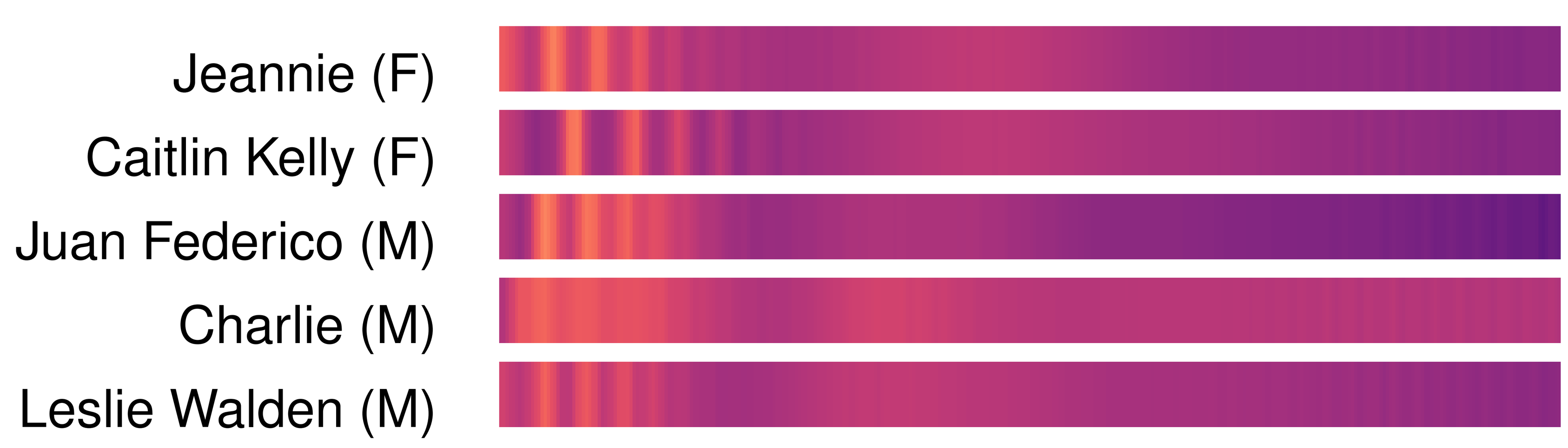
$$\mathcal{L}_{\text{classif}}(x, y) = \mathcal{L}_{\text{rec}}(x, y) + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}}(x, y)$$

Meaningful and Interpretable Prototypes

- Prototypes learn characteristics of their assigned class that go beyond simple harmonic components.
- The model captures elements of the **timbre**, such as their spectral envelopes under various conditions: notes, techniques, words, etc.
- Such results pave the way for **instrumental or vocal timber transfers** as natural applications.



Prototypes learned on the instrument dataset SOL **[1,2]**



Prototypes learned on the speech dataset LibriSpeech **[6]**

Bibliography

[1] Guillaume Ballet *et al*. Journées d'Informatique Musicale, 1999 **[2]** Carmine Emanuele Cella *et al*. ICMC, 2020 **[3]** Romain Loiseau *et al*. 3DV, 2021 **[4]** Vincent Lostanlen *et al*. DLfM, 2018 **[5]** Tom Monnier *et al*. NeurIPS, 2020 **[6]** Vassil Panayotov *et al*. ICASSP, 2015 **[7]** Pablo Zinemanas *et al*. Electronics, 2021