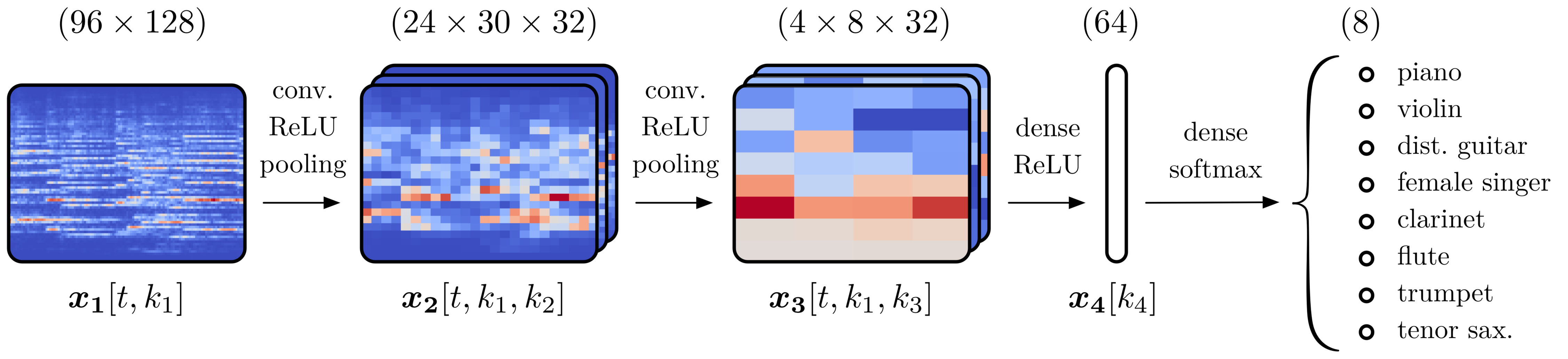
Deep Convolutional Networks on the Pitch Spiral for Musical Instrument Recognition

Vincent Lostanlen and Carmine-Emanuele Cella
École normale supérieure de Paris, PSL Research University, France

Deep convolutional networks (ConvNets) owe their success to two assumptions:
1. locality of correlations, and
2. stationarity of statistics.

Yet, the constant-Q transform (CQT) does not comply with them over the full hearing range.

What convolutional architectures for time-frequency representations ?

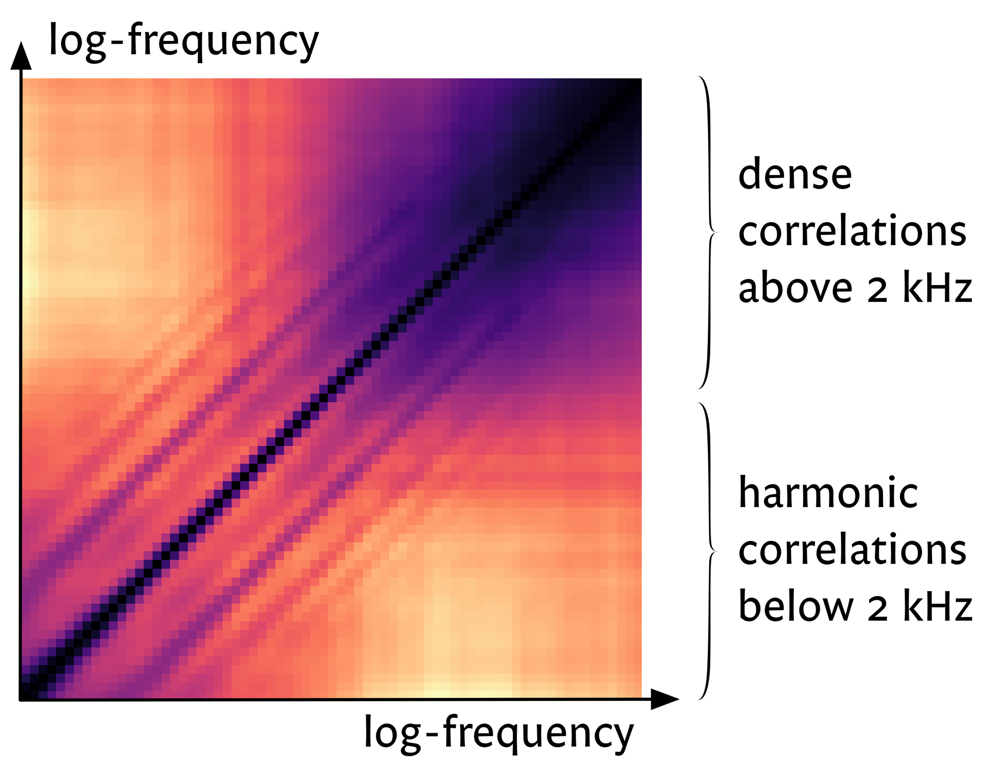


Two convolutional layers and two densely connected layers: a commonly used ConvNet architecture for MIR.
Trained with Adam optimizer, on stochastic cross-entropy, over normalized batches of size 32.
Dropout of 50% of the activations is applied at the last two layers.

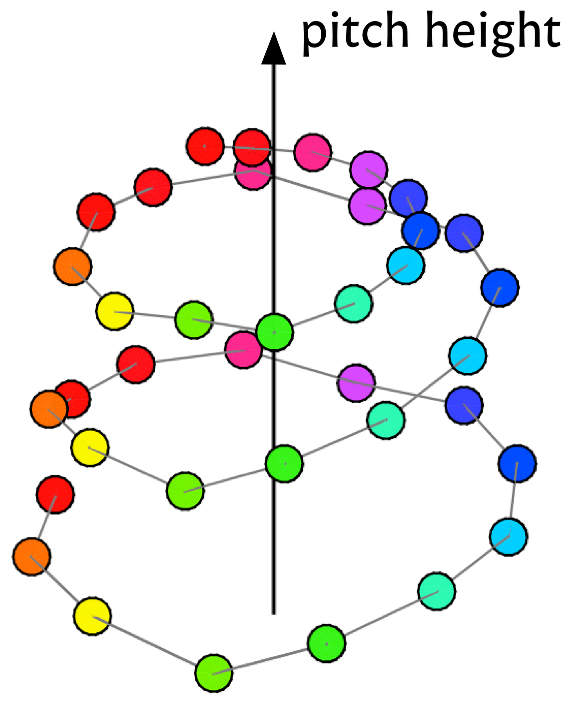
Problem 1: Locality of correlations ?

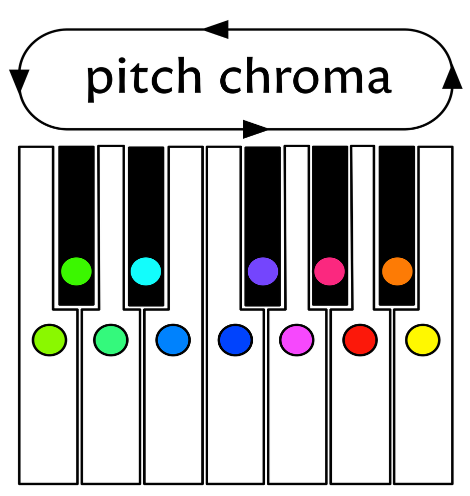
« Local neighborhoods in frequency do not share the same relationship » [Humphrey 2013].

We computed the covariance matrix between CQT coefficients in the RWC dataset of isolated notes.

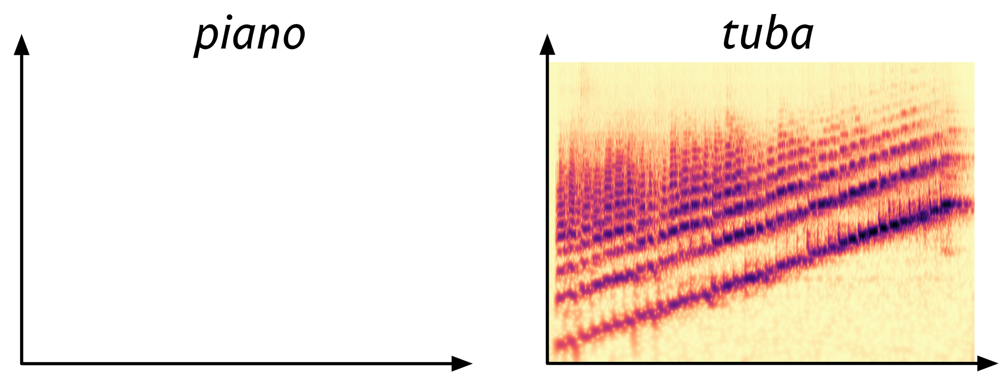


Isomap embedding [Le Roux 2007] of harmonic correlations reveals the pitch helix [Shepard 1965].





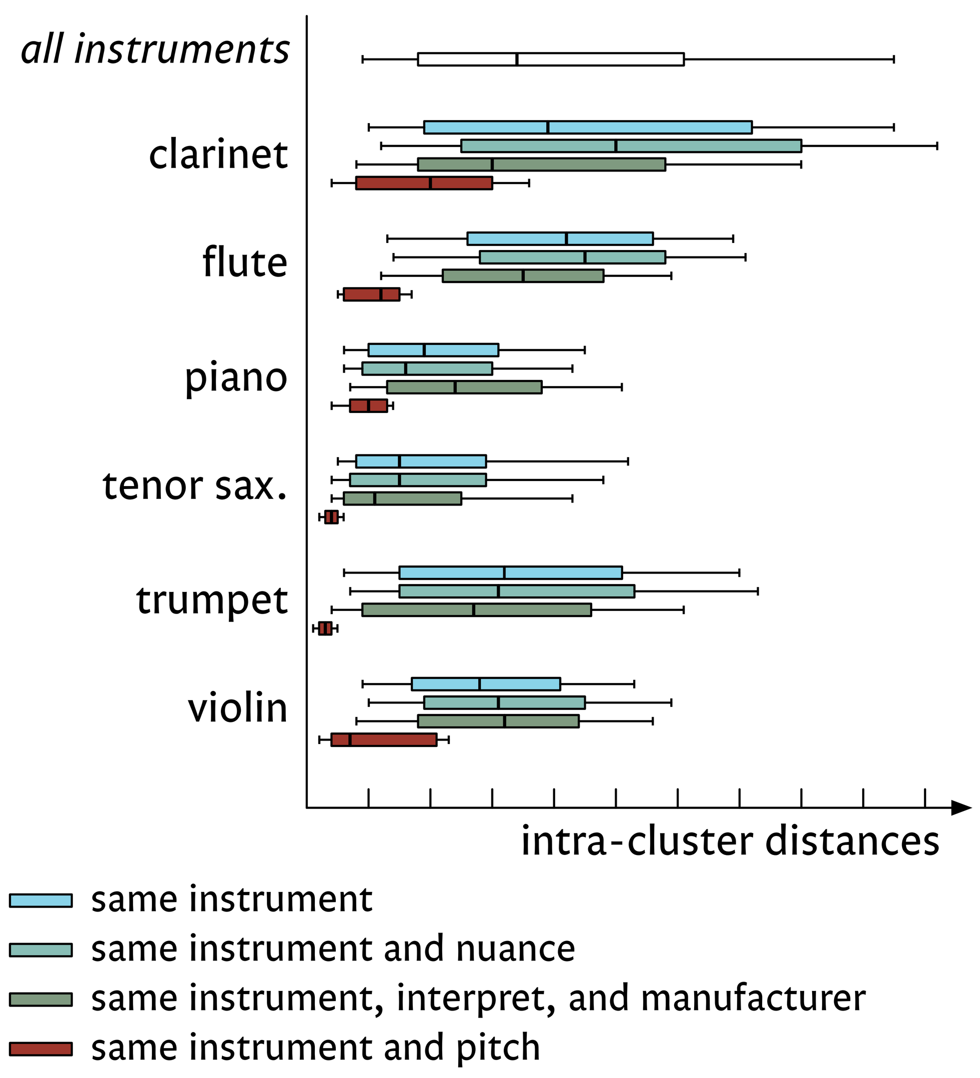
Problem 2: Stationarity of statistics ?



Source-filter interpretation:
the source is transposed by pitch shift while
the overall spectral envelope remains unchanged.

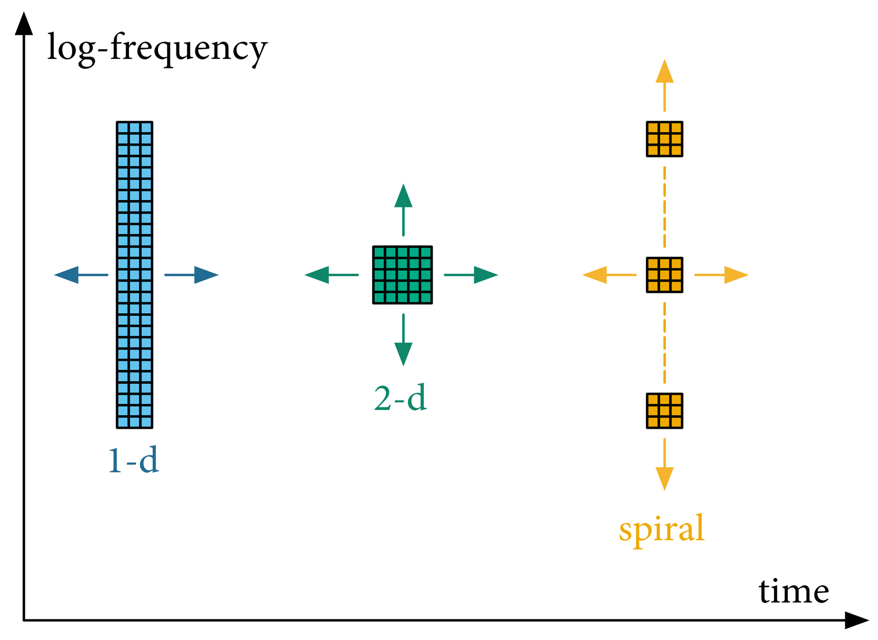
We computed pairwise distances in mel-frequency cepstral coefficients (MFCC) of isolated notes in the RWC dataset.

The DCT involved in MFCC yields the optimal basis under the assumption of stationarity.



Yet, MFCC are affected by realistic pitch shifts despite being designed to be invariant to frequency transposition of pure tones.

Solution: improved weight sharing



Distance ~1/n, unevenness ~1/n^2
At high frequencies, transposed pitches have similar spectra up to some additive bias.
We use 1-d convolutions above 2 kHz.

The probability of two randomly chosen partials between 1 and n to be in octave relationship is ~1/n.
We use spiral convolutios below 2 kHz.

Experiments in musical instrument classification with MedleyDB [] for training and solosDB [Joder] for testing.

MFCC, random forest classifier	38.6
2-d ConvNet	30.9
& spiral ConvNet	28.3
& 1-d ConvNet	26.0
1-d scattering [Andén 2014]	32.0
2-d scattering [Andén 2015]	22.0
spiral scattering [Lostanlen 2015]	19.9

Hybridyzing convolutional layers with multiple weight sharing strategies improves classification accuracy of ConvNets with respect to the traditional 2-d architecture.
However, the state of the art is obtained by a deep scattering network, in which learned convolutional kernels are replaced by wavelets.

References

Andén, Lostanlen, and Mallat. *Joint Time-frequency Scattering for Audio Classification*, MLSP 2015.
Goto, Hashigushi, Nikimura, and Oka. *RWC Music Database*, ISMIR 2003.
Warren, Uppenkamp, Patterson, and Griffiths. *Separating Pitch Chroma and Pitch Height in the Human Brain*, PNAS 2003.
Le Roux, Bengio, Lamblin, Joliveau, and Kégl. *Learning the 2-d topology of images*, NIPS 2007.
Shepard. *Circularity in Judgments of Relative Pitch*, JASA 1964.
Humphrey, Bello, and Le Cun. *Feature Learning and Deep Architectures: New Directions for Music Informatics*, JIIS 2013.

The source code to reproduce experiments is available at

`www.github.com/lostanlen/ismir2016`

This work is supported by the ERC InvariantClass 320959 grant.